

# **Functional Studies of Genome Architectural Proteins CTCF and ZNF143**

Rosie Rong Shen

Division of Structural Biology  
The Institute of Cancer Research  
University of London

This thesis is submitted for the degree of Doctor of Philosophy.

July 2021

# **Declaration**

The work described in this thesis was carried out at the Institute of Cancer Research, London, under the supervision of Professor Alessandro Vannini.

I, Rosie Rong Shen, declare that the work presented in this thesis is my own and, where information has been derived from other sources, I confirm that this had been indicated in the thesis.

Signature



Date

28/7/2021

# Acknowledgements

First of all, I am very grateful for having Professor Alessandro Vannini as my supervisor for the past four years. His immense knowledge and continuous support have made my PhD an enjoyable experience and I have learnt a lot more than I thought possible.

Next, I would like to thank Ewan, Guiller and Thanga who were always here to help, and I am thankful for their tremendous support and invaluable advices. I would also like to thank Erin for the inspirational and motivational chats along the way. I am grateful to have amazing team members, past and present, for creating a friendly and cooperative atmosphere: Jerome, Helen, Andri, Laura, Yibo and Yehuda.

My immense gratitude goes to the whole Structural Biology department for the inspirational advices and chats we had during lunchtime and tea breaks. In particular I owe a huge thank you to Fabienne for teaching me electron microscopy with great patience, Steve for helping me with crystallography and answering all the questions I had with different biophysical equipment, Chris for solving all the IT problems I encountered and Jane for managing the lab organisation.

I would like to thank my amazing family for everything they have given me; thank you Mum, Dad and George. A huge thank you to my incredible friends who were always here for me. Lastly, I would like to thank Anfernee for his understanding and encouragement that has kept me going over the last few years.

# Abstract

The genome architecture and gene transcription are interdependent. The way chromatin is packed within the nucleus controls when and how genes are expressed. A lot of chromatin interacting proteins are involved to ensure regulated gene transcription. The highly conserved zinc finger protein CCCTC-binding factor (CTCF) is one of the core genome architecture proteins that acts as a transcription insulator and activator where it regulates long-range chromatin loops and contributes to the establishment of topological-associating domains (TAD). Genome-wide analyses suggested that zinc finger 143 (ZNF143), another ubiquitously expressed zinc finger transcription factor is also a key regulator for CTCF-mediated chromatin interactions to establish conserved chromatin loops. In addition, chromatin remodelling protein chromodomain helicase DNA binding protein 8 (CHD8) has also been shown to colocalise extensively at CTCF and ZNF143 binding sites to modulate nucleosome positions near chromatin loop anchors. Recent studies are increasingly describing the importance of the crosstalk between these chromatin proteins and how they play a significant role in the three-dimensional genome architecture, however a lot of the underlying mechanisms of chromatin organisation remain unclear. This thesis focuses on the functional roles of three chromatin proteins, CTCF, ZNF143 and CHD8, to gain an insight into how these proteins might interact with each other and probe their role in chromatin organisation.

Purification protocols of human CTCF, ZNF143 and CHD8 were established. Detailed functional assays were carried out on the individual proteins to provide insights on how they might act on the genome. Their DNA- and nucleosome-binding properties were investigated with biochemical and biophysical techniques. Direct interactions between these proteins were further explored with pull-down assays. Structural analysis was performed on the CTCF-CHD8 complex using negative stain and cryo-electron microscopy. While CHD8 was able to interact with both CTCF and ZNF143, direct

interaction between CTCF and ZNF143 was not observed, indicating that CHD8 might be a cofactor that mediates the association of CTCF and ZNF143 at chromatin loop anchors. Together, the biochemical and biophysical studies presented in this thesis provide a fitting contribution to our current knowledge about the molecular mechanisms in genome organisation.

# Statement of COVID Impact

The COVID-19 pandemic has led to disruptions that prevented me from completing the amount of work that I initially planned. Due to lockdowns and travel restrictions, I was not able to carry out experimental work for four months between March 2020 to July 2020. Whilst my work in this thesis revealed interesting properties of CTCF, ZNF143 and CHD8 alone and together, there are a few areas I would have worked on if there were no COVID disruptions, specifically the structural determination of the CTCF-CHD8 complex. Although I have generated a 3D model based on the negative stain data and carried out preliminary cryo-EM data collection, the quality of the cryo-EM data set could have been greatly improved if given more time to optimise. Understanding the structure of the CTCF-CHD8 complex would reveal the mechanism of the CTCF-CHD8 interaction and its prominent role in chromatin regulation which would add a lot to the current knowledge on the molecular mechanisms in genome organisation. In particular, this would also shed light into how CTCF is not able to interact with CHD8 and consensus DNA simultaneously which might explain how these proteins are recruited to the CTCF-binding sites across the genome. Optimisations that I would have made include collecting more data to increase the number of particles for 2D classification, optimising buffer conditions with differential scanning fluorimetry (DSF), introducing an extra purification step and moderating the cryogenic conditions and grid to improve imaging. In addition, *in vitro* experiments including pull-down assays with purified protein constructs could aid and corroborate our understanding on the interaction interface between CTCF and CHD8 based on the cross-link mass spectrometry experiment. Whilst it is important to probe the CTCF-CHD8 complex, it would also be intriguing to unravel how this complex might modulate nucleosome remodelling activities. I would have carried out *in vitro* EMSA experiments to understand if CTCF, CHD8 and nucleosome form a complex and nucleosome remodelling assays would also have been performed to explore the remodelling activities of CHD8 in the presence of CTCF. Together, these experiments would elucidate the mechanism behind

CTCF-CHD8 nucleosome repositioning and reveal their significant role in genome organisation. As a result of the pandemic, I have planned these as future experiments. Further details can be found in the discussion chapter section 7.4.1.

Signature (Student: Rosie Rong Shen)



Date

28/7/2021

Signature (Supervisor: Professor Alessandro Vannini)



Date

28/7/2021

# List of Abbreviations

<b>Å</b>	Angstrom
<b>2D</b>	Two dimensional
<b>3C</b>	Chromosome conformation capture
<b>3D</b>	Three dimensional
<b>A<sub>260</sub></b>	Absorbance at 260 nm
<b>A<sub>280</sub></b>	Absorbance at 280 nm
<b>ADP</b>	Adenosine diphosphate
<b>AMP-PNP</b>	adenosine 5'-(β,γ-imido)triphosphate
<b>ATP</b>	Adenosine triphosphate
<b>ATPyS</b>	Adenosine 5'-(γ-thio)triphosphate
<b>bp</b>	Base-pair
<b>CHD</b>	Chromodomain helicase DNA-binding protein
<b>CHD8</b>	Chromodomain helicase DNA-binding protein 8
<b>ChIP-seq</b>	Chromatin immunoprecipitation followed by high throughput sequencing
<b>co-IP</b>	Co-immunoprecipitation
<b>coA</b>	Co-enzyme A
<b>CpG</b>	Cytosine-guanine dinucleotide
<b>CTCF</b>	CCCTC-binding factor
<b>CTD</b>	C-terminal domain
<b>CV</b>	Column volume
<b>Cy5</b>	Cyanine-5
<b>DLS</b>	Dynamic light scattering
<b>DNA</b>	Deoxyribonucleic acid
<b>dsDNA</b>	Double-stranded DNA
<b>DSE</b>	Distal sequence element
<b>DTT</b>	Dithiothreitol
<b>e<sup>-</sup></b>	Electron
<b><i>E. coli</i></b>	<i>Escherichia coli</i>
<b>EDTA</b>	Ethylenediaminetetraacetic acid
<b>EM</b>	Electron microscopy
<b>EMSA</b>	Electrophoretic mobility shift assay

<b>EtBr</b>	Ethidium bromide
<b>FP</b>	Fluorescence polarisation
<b>FT</b>	Flowthrough
<b>GF</b>	Gel filtration
<b>GST</b>	Glutathione S-transferase
<b>HA</b>	Human influenza hemagglutinin
<b>HEK293</b>	Human embryonic kidney 293 cells
<b>HEP</b>	Heparin
<b>HEPES</b>	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
<b>HP</b>	High Performance
<b>IPTG</b>	Isopropyl b-D-1-thiogalactopyranoside
<b>kb</b>	Kilobase
<b>K<sub>d</sub></b>	Rate of dissociation
<b>kDa</b>	Kilodalton
<b>LCMS</b>	Liquid chromatography mass spectrometry
<b>M</b>	Molar (moles per litre)
<b>MALS</b>	Multi-angle light scattering
<b>mg/ml</b>	Milligrams per millilitre
<b>ml</b>	Millilitre
<b>mM</b>	Millimolar
<b>MS</b>	Mass spectrometry
<b>MW</b>	Molecular weight
<b>ng</b>	Nanogram
<b>nm</b>	Nanometre
<b>NRL</b>	Nucleosome repeat length
<b>NTD</b>	N-terminal domain
<b>°C</b>	Degree Celsius
<b>OD<sub>600</sub></b>	Optical density, measured at 600 nm
<b>PAGE</b>	Polyacrylamide gel electrophoresis
<b>PBS</b>	Phosphate buffered saline
<b>PCR</b>	Polymerase Chain Reaction
<b>pI</b>	Isoelectric point
<b>RELION</b>	Regularised likelihood optimisation
<b>R<sub>g</sub></b>	Radius of gyration
<b>R<sub>h</sub></b>	Hydrodynamic radius
<b>RNA</b>	Ribonucleic acid

<b>SDS-PAGE</b>	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
<b>SEC</b>	Size exclusion chromatography
<b>snRNA</b>	Small nuclear RNA
<b>Strep</b>	Streptactin
<b>TAD</b>	Topologically associating domain
<b>TBE</b>	Tris/Borate/EDTA
<b>TBST</b>	Tris buffered saline tween 20
<b>UV</b>	Ultraviolet
<b>v/v</b>	Volume/volume
<b>ZF</b>	Zinc finger
<b>ZNF143</b>	Zinc finger protein 143
<b>µg/ml</b>	Micrograms per millilitre
<b>µl</b>	Microlitre
<b>µM</b>	Micromolar

# Table of Contents

<b><i>Declaration</i></b> .....	<b>2</b>
<b><i>Acknowledgements</i></b> .....	<b>3</b>
<b><i>Abstract</i></b> .....	<b>4</b>
<b><i>Statement of COVID Impact</i></b> .....	<b>6</b>
<b><i>List of Abbreviations</i></b> .....	<b>8</b>
<b><i>Table of Contents</i></b> .....	<b>11</b>
<b><i>List of Figures</i></b> .....	<b>17</b>
<b><i>List of Tables</i></b> .....	<b>20</b>
<b>1      <i>Introduction</i></b> .....	<b>21</b>
1.1 <b>Overview</b> .....	<b>21</b>
1.2 <b>Chromatin Organisation</b> .....	<b>22</b>
1.2.1    Hierarchical Folding of Chromatin .....	22
1.2.2    Chromosome Territories .....	23
1.2.3    Active (A) and Inactive (B) Compartments.....	24
1.2.4    Topologically Associated Domains (TADs) .....	24
1.2.5    Sub-TADs and Chromatin Loops.....	25
1.2.6    Nucleosome .....	26
1.2.7    Techniques Used to Study the Genome Architecture.....	27
1.3 <b>The Role of Chromatin in Gene Transcription Regulation</b> .....	<b>28</b>
1.3.1    Eukaryotic Transcription .....	28
1.3.2    Long-Range Enhancer-Promoter Contacts.....	29
1.3.3    Chromatin Remodelling .....	33
1.4 <b>Chromatin Architecture Protein CTCF</b> .....	<b>43</b>
1.4.1    Domain Architecture.....	43
1.4.2    Mechanisms of CTCF Binding to DNA.....	44
1.4.3    CTCF-Mediated Enhancer-Promoter Looping Within TADs .....	46
1.4.4    CTCF-Mediated Insulation at TAD Boundaries .....	48
1.4.5    Nucleosome Positioning.....	49
1.4.6    Influences of CTCF Depletion .....	50
1.5 <b>Transcription Factor ZNF143</b> .....	<b>51</b>
1.5.1    Domain Architecture.....	51

1.5.2	Mechanisms of ZNF143 Binding to DNA .....	52
1.5.3	Regulation of ZNF143 at Gene Promoters .....	54
1.5.4	Potential Role as a Chromatin-Looping Factor.....	56
<b>1.6</b>	<b>Chromatin Remodelling Protein CHD8.....</b>	<b>57</b>
1.6.1	Domain Architecture.....	57
1.6.2	Role in Chromatin Remodelling .....	59
1.6.3	The Interplay with 3D Architectural Proteins.....	59
<b>1.7</b>	<b>Perspectives and Aims.....</b>	<b>60</b>
<b>2</b>	<b>Materials and Methods.....</b>	<b>62</b>
<b>2.1</b>	<b>Molecular Cloning .....</b>	<b>62</b>
2.1.1	Overview .....	62
2.1.2	Plasmids constructs.....	62
2.1.3	Polymerase Chain Reaction (PCR).....	65
2.1.4	Cloning using the Gibson Assembly cloning system.....	66
2.1.5	Transformation into Stellar Competent Cells.....	66
2.1.6	DNA Purification from Agarose Gel.....	67
2.1.7	QuikChange Mutagenesis .....	67
<b>2.2</b>	<b>Protein Expression .....</b>	<b>68</b>
2.2.1	Protein expression in <i>E. coli</i> .....	68
2.2.2	Protein Expression in Insect Cells.....	68
<b>2.3</b>	<b>Protein Purification .....</b>	<b>70</b>
2.3.1	Overview .....	70
2.3.2	Purification Protocol of Human CTCF.....	70
2.3.3	Purification Protocol of Human ZNF143 constructs.....	72
2.3.4	Purification Protocol of Human CHD8.....	73
2.3.5	Nucleosome Reconstitution .....	74
<b>2.4</b>	<b>Protein Analysis and Quantification .....</b>	<b>77</b>
2.4.1	Theoretical Molecular Weights and Extinction Coefficients .....	77
2.4.2	Protein Concentration Quantification.....	78
2.4.3	SDS-PAGE Analysis .....	79
2.4.4	Western Blot Analysis .....	79
2.4.5	Silver Stain Analysis.....	80
<b>2.5</b>	<b>Biochemical and Functional Analysis .....</b>	<b>80</b>
2.5.1	Limited Proteolysis.....	80

2.5.2	Electrophoretic Mobility Shift Assay (EMSA) .....	81
2.5.3	Insect Cell Co-expression Pulldown Assay .....	82
2.5.4	Recombinant Protein Pulldown Assay .....	83
2.5.5	Analytical Gel Filtration Chromatography.....	83
2.5.6	Nucleosome Remodelling Assay .....	84
2.5.7	Co-immunoprecipitation Pulldown Assay.....	84
<b>2.6</b>	<b>Biophysical Analysis .....</b>	<b>87</b>
2.6.1	Size Exclusion Chromatography Coupled to Multi-Angle Light Scattering (SEC-MALS)87	
2.6.2	Mass Photometry.....	88
2.6.3	Fluorescence Polarisation .....	90
2.6.4	ATP Hydrolysis Assay.....	92
2.6.5	Dynamic Light Scattering.....	94
2.6.6	Cross-Link Mass Spectrometry.....	94
<b>2.7</b>	<b>X-ray Crystallography .....</b>	<b>95</b>
2.7.1	Overview .....	95
2.7.2	Protein Crystal Formation .....	96
2.7.3	Sample Preparation.....	99
2.7.4	Data Collection .....	100
<b>2.8</b>	<b>Electron Microscopy.....</b>	<b>100</b>
2.8.1	Overview .....	100
2.8.2	Sample preparation.....	101
2.8.3	The Transmission Electron Microscope .....	101
2.8.4	Electron Scattering and Image Formation .....	105
2.8.5	Image Processing and Analysis.....	108
2.8.6	Negative-Stain Electron Microscopy Data Collection.....	111
2.8.7	Cryo-Electron Microscopy Data Collection .....	113
<b>3</b>	<b>Purification and Characterisation of Human CTCF .....</b>	<b>115</b>
<b>3.1</b>	<b>Purification of Recombinant Human CTCF .....</b>	<b>115</b>
3.1.1	Purification Optimisation Trials of CTCF.....	115
3.1.2	Optimised Purification Protocol of Human CTCF .....	119
3.1.3	Purification of ybbR-tagged CTCF.....	121
3.1.4	Purification of N- and C-terminal Domains .....	122
<b>3.2</b>	<b>CTCF Shows Concentration-Dependent Oligomerisation.....</b>	<b>124</b>
3.2.1	Mass Photometry Analysis.....	124
3.2.2	Dynamic Light Scattering Analysis.....	125

<b>3.3 DNA Binding Studies of CTCF .....</b>	<b>126</b>
3.3.1 Purified CTCF binds to DNA.....	126
3.3.2 CTCF-DNA Forms a Stable Complex Over Gel Filtration .....	129
3.3.3 CTCF-DNA Negative Stain Studies .....	130
<b>3.4 Nucleosome Binding Studies .....</b>	<b>132</b>
3.4.1 CTCF Interacts with Nucleosomes.....	132
<b>3.5 Summary.....</b>	<b>133</b>
<b>4 Purification and Characterisation of Human ZNF143 .....</b>	<b>135</b>
<b>4.1 Purification of Recombinant Human ZNF143 .....</b>	<b>135</b>
4.1.1 Purification of <i>E. coli</i> Expressed ZNF143 .....	135
4.1.2 Contaminant Observed During Purification .....	137
4.1.3 Optimised Full-length ZNF143 Purification .....	138
4.1.4 Full-length ZNF143 Exists as Monomer in Solution.....	140
4.1.5 Design of a More Stable Construct ZNF143 <sub>103-466</sub> .....	140
4.1.6 Purification of ZNF143 <sub>103-466</sub> .....	142
<b>4.2 DNA Binding Studies of ZNF143 .....</b>	<b>144</b>
4.2.1 ZNF143-DNA Forms a Stable Complex Over Gel Filtration.....	144
4.2.2 N- and C-terminal region of ZNF143 do not bind DNA.....	145
4.2.3 ZNF143 Binds to A Range of DNA Constructs.....	148
4.2.4 ZNF143 Binds to Constructs with Non-Consensus Sequence.....	151
4.2.5 ZNF143 Prefers Z-motif Over Non-Consensus DNA .....	152
4.2.6 Binding Affinities Between ZNF143 and Different DNA Constructs .....	155
4.2.7 ZNF143 Interacts with Nucleosomes .....	157
<b>4.3 ZNF143 Crystallography Trials .....</b>	<b>158</b>
4.3.1 Crystallisation Trial with Apo ZNF143 .....	159
4.3.2 Crystallisation Trial with ZNF143-DNA .....	160
4.3.3 Crystallisation Trial with Apo ZNF143 <sub>103-466</sub> .....	162
4.3.4 Crystallisation Trial with ZNF143 <sub>103-466</sub> -DNA .....	163
<b>4.4 ZNF143 and Its Potential Interacting Partners .....</b>	<b>165</b>
<b>4.5 Summary.....</b>	<b>169</b>
<b>5 Purification and Characterisation of Human CHD8 .....</b>	<b>171</b>
<b>5.1 Purification of Recombinant Human CHD8 .....</b>	<b>171</b>
5.1.1 Purification Protocol of CHD8 .....	171

5.1.2	Purified CHD8 is Phosphorylated .....	174
5.1.3	Purification of the ATPase-Deficient CHD8 K842R Mutant .....	176
5.1.4	Purified Wild Type CHD8 is Functional and Active.....	177
<b>5.2</b>	<b>Biophysical Analysis of Purified CHD8.....</b>	<b>178</b>
5.2.1	CHD8 Is Monomeric in Solution .....	178
5.2.2	CHD8 Negative Stain Analysis .....	179
<b>5.3</b>	<b>Nucleosome Binding Studies of CHD8.....</b>	<b>181</b>
5.3.1	CHD8 Binds to A Range of Different DNA Lengths .....	181
5.3.2	Nucleosome Sequence Design .....	182
5.3.3	CHD8 Binds to Nucleosomes.....	184
5.3.4	Binding of CHD8 to Nucleosome is not ATP-Dependent.....	184
5.3.5	CHD8 Remodels Nucleosomes .....	185
5.3.6	Negative Stain Trials and Analysis.....	188
5.3.7	Cryo-EM Trial.....	194
<b>5.4</b>	<b>Summary.....</b>	<b>196</b>
<b>6</b>	<b><i>Interaction Studies of Genome Architecture Proteins .....</i></b>	<b>197</b>
<b>6.1</b>	<b>Interaction Studies Between CTCF and CHD8.....</b>	<b>197</b>
6.1.1	CHD8 Interacts Directly with CTCF .....	197
6.1.2	CTCF and CHD8 Form a Stable Complex Over Gel Filtration .....	198
6.1.3	Negative Stain EM Trials and Analysis.....	199
6.1.4	Cryo-EM Trial and Analysis.....	205
6.1.5	Domains of CTCF and CHD8 Involved in Their Interaction .....	206
6.1.6	CTCF Cannot Interact with CHD8 and DNA Simultaneously.....	212
<b>6.2</b>	<b>Interactions Studies of ZNF143 .....</b>	<b>219</b>
6.2.1	ZNF143 and CTCF .....	219
6.2.2	ZNF143 and CHD8 .....	220
<b>6.3</b>	<b>Interaction Studies Between ZNF143, CTCF and CHD8 .....</b>	<b>225</b>
<b>6.4</b>	<b>Interaction Studies with Other Chromatin Proteins .....</b>	<b>228</b>
6.4.1	CTCF Does Not Interact with CHD4 .....	228
<b>6.5</b>	<b>Summary.....</b>	<b>229</b>
<b>7</b>	<b><i>Discussion .....</i></b>	<b>231</b>
<b>7.1</b>	<b>Characterisation of Human CTCF .....</b>	<b>231</b>
7.1.1	Generation of Functional Human Full-Length CTCF Protein .....	231

7.1.2	Concentration Dependent Oligomerisation of CTCF.....	232
7.1.3	CTCF Binds to DNA Lacking its Consensus Site.....	234
7.1.4	Nucleosome Binding Studies of CTCF.....	235
<b>7.2</b>	<b>Characterisation of Human ZNF143 .....</b>	<b>236</b>
7.2.1	Generation of Functional Human Full-Length ZNF143 Protein.....	236
7.2.2	ZNF143 Binds to A Variety of DNA Sequences .....	237
7.2.3	ZNF143 Shows Strong Binding to Nucleosomes.....	239
7.2.4	Crystallisation Challenges of ZNF143 .....	239
7.2.5	Potential Interacting Partners of ZNF143.....	241
<b>7.3</b>	<b>Characterisation of Human CHD8 .....</b>	<b>242</b>
7.3.1	Generation of Functional Human Full-Length CHD8 Protein .....	242
7.3.2	CHD8 Performs Nucleosome Sliding through ATP Hydrolysis.....	243
<b>7.4</b>	<b>The Interaction between Chromatin Proteins CTCF, ZNF143 and CHD8 .....</b>	<b>246</b>
7.4.1	The Interplay Between CTCF and CHD8 .....	246
7.4.2	The Interplay Between CTCF and ZNF143 .....	249
7.4.3	The Interplay Between CHD8 and ZNF143 .....	250
7.4.4	The Interplay Between CTCF, ZNF143 and CHD8 .....	252
<b>7.5</b>	<b>Future Directions .....</b>	<b>254</b>
<b>7.6</b>	<b>Conclusions .....</b>	<b>255</b>
<b>Appendix.....</b>		<b>257</b>
<b>Bibliography.....</b>		<b>258</b>

# List of Figures

Figure 1.1 Overview 3D Genome Organisation .....	23
Figure 1.2 Determinants for Enhancer-Promoter Specificity.....	31
Figure 1.3 Dynamic Properties of Nucleosomes.....	35
Figure 1.4 Domain Organisation of Chromatin Remodellers .....	36
Figure 1.5 Domain Architecture of the Three CHD Sub-Families .....	39
Figure 1.6 Structural Changes of Chd1 During ATPase Activation .....	41
Figure 1.7 Comparison of Structures of Known Nucleosome-CHD Remodellers....	42
Figure 1.8 Schematic Representation of Human CTCF Domains .....	44
Figure 1.9 CTCF ZF3-7 Form Base-Specific Contacts .....	45
Figure 1.10 The Loop Extrusion Model.....	48
Figure 1.11 Schematic Representation of Human ZNF143 Domains.....	52
Figure 1.12 ZNF143 Consensus Sequence.....	53
Figure 1.13 Transcription Pre-Initiation Complex at snRNA Promoters.....	55
Figure 1.14 Schematic Representation of Human CHD8 .....	58
Figure 2.1 SEC-MALS Setup .....	88
Figure 2.2 Principle of Mass Spectrometry .....	89
Figure 2.3 Schematic of Fluorescence Polarisation Assay Principle .....	91
Figure 2.4 ATP/NADH Coupled Assay .....	93
Figure 2.5 Vapour Diffusion Crystallisation Experiment Setup .....	97
Figure 2.6 Phase Diagram for Crystallisation of Macromolecules .....	98
Figure 2.7 Electron Scattering and Image Contrast.....	106
Figure 3.1 First CTCF Purification Trial .....	116
Figure 3.2 Four Methods Used to Remove DNA Contamination.....	117
Figure 3.3 CTCF Purification Optimisation .....	119
Figure 3.4 Optimised Purification Protocol of CTCF .....	121
Figure 3.5 Gel Filtration Elution Profile of CTCF <sub>Alexa488</sub> .....	122
Figure 3.6 Purification of CTCF <sub>1-240</sub> and CTCF <sub>591-727</sub> .....	124
Figure 3.7 Mass Photometry Analysis .....	125
Figure 3.8 Dynamic Light Scattering Analysis .....	126
Figure 3.9 EMSA of CTCF-DNA Complex .....	127
Figure 3.10 CTCF Binds to Non-Consensus DNA.....	128
Figure 3.11 CTCF-DNA Formed a Stable Complex over Gel Filtration .....	130
Figure 3.12 Negative Stain EM Micrographs of CTCF-DNA Complex.....	131
Figure 3.13 Interaction Between CTCF and Nucleosome .....	133

Figure 4.1 Purification of <i>E. coli</i> Expressed Full-length ZNF143 .....	137
Figure 4.2 Purification of Insect Cell Expressed Full-length ZNF143 .....	139
Figure 4.3 ZNF143 Behaves as a Monomer in Solution .....	140
Figure 4.4 Trypsin Digestion of ZNF143 .....	142
Figure 4.5 Purification of ZNF143 <sub>103-466</sub> .....	143
Figure 4.6 ZNF143 Forms a Stable Complex with DNA .....	145
Figure 4.7 N- and C-terminal Domains Do Not Affect ZNF143 DNA-Binding .....	147
Figure 4.8 ZNF143 Binding to its Consensus Motifs.....	149
Figure 4.9 Additional ZNF143 Binding Sites Present.....	150
Figure 4.10 ZNF143 Binds to Non-Consensus DNA Sequences .....	152
Figure 4.11 ZNF143 Prefers DNA Sequences with Z-motif .....	154
Figure 4.12 Fluorescence Polarisation of ZNF143 and Different DNA Sequences	157
Figure 4.13 ZNF143 Binds to Nucleosomes .....	158
Figure 4.14 Full-length ZNF143-DNA Crystal Leads .....	162
Figure 4.15 Crystallisation Trials with ZNF143 <sub>103-466</sub> -DNA Complex.....	164
Figure 4.16 Co-Immunoprecipitation of Endogenous ZNF143 .....	167
Figure 5.1 Purification of StrepII-CHD8 .....	173
Figure 5.2 Purification of StrepII-CHD8-FLAG .....	174
Figure 5.3 CHD8 Treated with Pro-Q Diamond Stain.....	175
Figure 5.4 Purification of CHD8 <sub>K842R</sub> .....	177
Figure 5.5 CHD8 ATPase Activity.....	178
Figure 5.6 CHD8 Mass Photometry Analysis.....	179
Figure 5.7 CHD8 Negative Stain Analysis .....	180
Figure 5.8 CHD8 DNA-Binding Study .....	182
Figure 5.9 CHD8-Nucleosome Binding Assay .....	184
Figure 5.10 Nucleosome Binding is not ATP-dependent.....	185
Figure 5.11 CHD8 Nucleosome Remodelling Assays .....	187
Figure 5.12 CHD8-Nucleosome Gel Filtration Negative Stain Trial .....	189
Figure 5.13 CHD8-Nucleosome Glycerol Gradient Negative Stain Trial .....	190
Figure 5.14 CHD8 Negative Stain Analysis .....	193
Figure 5.15 Refined CHD8-Nucleosome 3D Map.....	193
Figure 5.16 CHD8-Nucleosome Cryo-EM Trial.....	195
Figure 6.1 Direct Interaction Observed Between CTCF and CHD8.....	198
Figure 6.2 Analytical Gel Filtration of CTCF and CHD8.....	199
Figure 6.3 First CTCF-CHD8 Complex Negative Strain Trial .....	201
Figure 6.4 Cross-Linked CTCF-CHD8 Complex Negative Stain Trial .....	202

Figure 6.5 Negative Stain Analysis of Cross-Linked CTCF-CHD8 Complex .....	205
Figure 6.6 Cryo-EM Trial.....	206
Figure 6.7 Cross-Link Mass Spectrometry.....	209
Figure 6.8 CTCF Domain Involved the CTCF-CHD8 Interaction .....	210
Figure 6.9 CHD8 BRK Domain is not Involved in CTCF-CHD8 Interaction .....	211
Figure 6.10 EMSA Between CHD8 and CTCF-DNA Complex .....	213
Figure 6.11 CHD8 Does Not Interact with Pre-formed CTCF-DNA Complex .....	214
Figure 6.12 EMSA Between DNA and CTCF-CHD8 Complex .....	216
Figure 6.13 Consensus DNA Disrupts the Interaction between CTCF and CHD8	217
Figure 6.14 Interaction Between CTCF, CHD8 and Non-consensus DNA.....	218
Figure 6.15 ZNF143 Does Not Interact with CTCF.....	220
Figure 6.16 Direct Interaction Between ZNF143 and CHD8.....	221
Figure 6.17 Analytical Gel Filtration of ZNF143 and CHD8 .....	222
Figure 6.18 ZNF143 Domain Involved in the ZNF143-CHD8 Interaction .....	223
Figure 6.19 CHD8 Domains Involved in the ZNF143-CHD8 Interaction.....	225
Figure 6.20 Interaction Between CTCF, ZNF143 and CHD8.....	227
Figure 6.21 No Direct Interaction Between CHD4 and CTCF.....	229

# List of Tables

Table 2.1 List of Vectors Used for Cloning Different Protein Constructs .....	63
Table 2.2 List of CTCF Constructs.....	63
Table 2.3 List of ZNF143 Constructs .....	63
Table 2.4 List of CHD8 Constructs .....	64
Table 2.5 List of Other Protein Constructs.....	64
Table 2.6 Sequences of Protein Tags and Cleavage Sites Used .....	65
Table 2.7 Primers Used for CHD8 Mutagenesis.....	67
Table 2.8 Molecular Weights, Isoelectric Point and Extinction Coefficients of All Constructs.....	78
Table 2.9 Antibodies Used for Western Blots .....	80
Table 2.10 Different Typhoon Imaging Filters Used for Dye Visualisation.....	82
Table 2.11 List of Commercial Screens Used for Crystallisation Trials .....	100
Table 2.12 Glycerol Percentage Used for the Glycerol Gradient Experiment.....	112
Table 3.1 Hydrodynamic Radius and Polydispersity of Different CTCF Species...	126
Table 3.2 DNA Construct with CTCF-Consensus Binding Site.....	127
Table 3.3 DNA Construct Without CTCF-Consensus Sequence.....	128
Table 3.4 DNA Sequence Used for Nucleosome Assembly .....	132
Table 4.1 DNA Constructs Used for ZNF143 Binding Studies.....	149
Table 4.2 84bp DNA Used for ZNF143-DNA Binding Study .....	150
Table 4.3 DNA Constructs without ZNF143 Consensus Sites.....	152
Table 4.4 Artificial DNA Constructs for ZNF143 DNA Binding Studies.....	157
Table 4.5 Previous Zinc Finger Domain Crystal Formation Conditions .....	159
Table 4.6 Screening Kits Used for Full-length ZNF143 Crystallisation .....	160
Table 4.7 Screening Kits Used for Full-length ZNF143-DNA Crystallisation .....	161
Table 4.8 Screening Kits Used for ZNF143 <sub>103-466</sub> Crystallisation Trials .....	163
Table 4.9 Screening Kits Used for ZNF143 <sub>103-466</sub> -DNA Crystallisation Trials .....	163
Table 4.10 Crystallisation Conditions with Potential Crystal Leads .....	165
Table 4.11 Immunoprecipitation Results of ZNF143.....	168
Table 5.1 ATP Hydrolysis Rate of CHD8 Under Different Conditions.....	178
Table 5.2 DNA Sequences Used for CHD8 Binding Studies .....	181
Table 5.3 DNA Sequence Used for Nucleosome Assembly .....	183

# 1 Introduction

## 1.1 Overview

Deoxyribonucleic acid (DNA) is the fundamental basis of all life forms. It carries the genetic material for development, function and growth in humans. DNA folds into a more compact form to fit within the nucleus of a cell with the help of histones, transcription factors and noncoding RNAs. This protein-DNA complex is called the chromatin. There are multiple levels of chromatin packing in human cells. From chromatin fibres to topologically associated domains (TADs) and to chromosome territories, the way chromatin is packaged within the nucleus controls when and how genes are expressed (Bonev and Cavalli, 2016; Rowley and Corces, 2018). The genome architecture and transcription are interdependent. The detail of how DNA fits within a nucleus while it can be readily reorganised to control gene expression remains elusive. A lot of chromatin proteins are involved to ensure regulated gene transcription, such as genome architecture proteins CTCF (CCCTC-binding factor) and ZNF143 (zinc-finger protein 143), which are involved in long-range interactions through forming chromatin loops; and chromatin remodelling protein CHD8 (chromodomain helicase DNA binding protein 8) where it utilises ATP to modulate nucleosome position on the DNA (Ong and Corces, 2014; Alpsoy, Sood and Dykhuizen, 2021; Huning and Kunkel, 2021). Recent studies are increasingly describing the importance of the crosstalk between these chromatin proteins and how they play a significant role in the three-dimensional (3D) genome architecture, however a lot of the underlying mechanism of chromatin organisation remains unclear. This leads to the motivation for this thesis, which focuses on the functional roles of three chromatin proteins, CTCF, ZNF143 and CHD8, to gain an insight into how these proteins might interact with each other and probe their role in chromatin organisation. An introduction to the role of chromatin in transcription gene regulation, and the background of CTCF, ZNF143 and CHD8 will be described in this chapter.

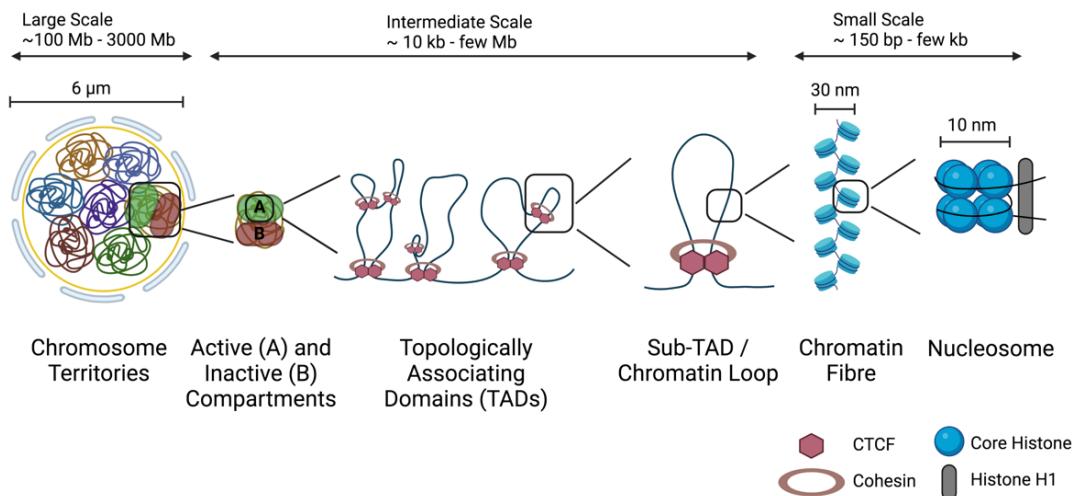
## 1.2 Chromatin Organisation

### 1.2.1 Hierarchical Folding of Chromatin

The way DNA is packaged within the nucleus is important for when and how genes are expressed. DNA in each human cell can be up to approximately 2 metres long in length. Several levels of chromatin organisation are required to compact DNA into chromosomes to fit within the nucleus that is approximately 6 µm in diameter. An efficient and organised genome packing allow the compacted DNA to be functionally accessible for DNA replication and repair, transcriptional control and chromosome translocation (Bonev and Cavalli, 2016). Chromatin is tightly regulated at each level to maintain a robust genome organisation yet at the same time flexible enough to adapt to changes during development and disease. Furthermore, the chromatin is a dynamic structure and the crosstalk between internal and external cues coordinates gene expression, these cues include post-translational modifications of histones, histone variants, chromatin remodelling, non-coding RNAs, transcription factors and architectural proteins (Alpsoy, Sood and Dykhuizen, 2021).

The 3D architecture of DNA is hierarchical (Figure 1.1). Within the nucleus, chromosomes and lamina-associated domains are segregated into different territories. These chromosomes tend to cluster into two types of compartments based on their transcription activities, termed as active (A) and inactive (B) compartments. Within the active (A) compartment, topologically associated domains (TADs) are present, these domains are mega bases long and are conserved across cell types. TAD boundaries are a crucial regulator of transcription as it restricts gene expression to within the boundary. Sub-TADs and chromatin loops form within TADs, these loops are critical for linking distal enhancer and promoter regions into close proximity for efficient gene transcription. Enhancer-promoter chromatin loops vary significantly across different cell types. Lastly in the most basic level, chromatin loops are formed by the fundamental units called nucleosomes (Bonev and Cavalli, 2016; Mishra and Hawkins, 2017; Alpsoy, Sood and Dykhuizen, 2021). With the recent advancement of next generation sequencing techniques, the resolution of the genome-wide chromatin interaction maps has improved substantially. It

has also become clear that the chromatin architecture is more complicated than previously anticipated (Bonev and Cavalli, 2016).



**Figure 1.1 Overview 3D Genome Organisation**

The 3D genome is hierarchical to ensure the compacted DNA is functionally accessible for transcriptional control. Within the nucleus, there are discrete chromosomal territories that are clustered into two types of compartments (active and inactive). Chromatins in the active compartment are further segregated into different regions called topologically associating domains (TADs). TADs are comprised of chromatin fibres forming self-associating loops to define TAD boundaries. Chromatin fibres are formed by nucleosomes where DNA is packaged around histone proteins.

### 1.2.2 Chromosome Territories

Chromosome territories were first discovered in the 1990s with the onset of the fluorescence *in situ* hybridisation (FISH) technique (Lichter *et al.*, 1988; Jauch *et al.*, 1990). Compartmentalisation of chromosomes was observed where individual entities tend to occupy distinct positions within the nucleus. In addition, the gene density of the chromosome is correlated to its nuclear position, where gene-poor chromosomes are usually found at the nuclear periphery and gene-rich chromosomes exhibiting high transcriptional activity are located at the interior of the nucleus (Foster and Bridger, 2005). Specific genomic regions with low levels of gene expression that preferentially locate at the nuclear periphery are termed lamina-associated domains (LADs). These domains exhibit low overall gene density and are tethered to the nuclear

lamina which are restricted from the transcription factories. Nonetheless, more than 30% of the genome forms LADs, suggesting its prominent role in the overall genome spatial organisation and gene repression (van Steensel and Belmont, 2017).

### 1.2.3 Active (A) and Inactive (B) Compartments

The genome-wide view of the interactions of the mammalian genome was revealed in 2009 (Lieberman-aiden *et al.*, 2009). These results showed that within each chromosome territory, chromatin interactions are further clustered into two types of compartments, active (A) and inactive (B) compartments based on the level of transcription activities. Most of the region in the active compartment A contain actively transcribed genes and histone modifications with a small portion of genes that are repressed. On the other hand, inactive compartment B is comprised of mostly transcriptionally repressed genes and gene-poor regions within the lamina-associated domains (Lieberman-aiden *et al.*, 2009). These two compartments are present in all cell types, but the particular genomic regions associated to each compartment can be redistributed during differentiation (Dixon *et al.*, 2015).

### 1.2.4 Topologically Associated Domains (TADs)

In the next level of genome organisation, megabase-sized segments termed topological associating domains (TADs) were discovered by chromosome conformation capture (3C) and they are the most prominent structural feature of genome organisation. The initial analysis of the human genome suggested a median TAD size of ~880 kb (Dixon *et al.*, 2012). They are highly conserved across species and the TAD regions are shared across different cell types (Dixon *et al.*, 2012; Nora *et al.*, 2012; Vietri Rudan *et al.*, 2015). Elements within a TAD region tend to self-associate and interact with each other at higher frequencies than with regions located in the adjacent domains.

TAD boundaries display insulator and barrier element activities which are crucial to restricting transcription events to within the regions (Dixon *et al.*, 2012). The weakening of TAD boundaries has been associated with a range

of diseases including cancer and developmental disorders (Lupiáñez, Spielmann and Mundlos, 2016; Akdemir *et al.*, 2020). The most notable feature in mammalian TAD boundaries is the strong association of CCCTC-binding factor (CTCF) and cohesin with TAD borders (Rao *et al.*, 2014; Akdemir *et al.*, 2020), where CTCF binding sites are detected at ~ 76% of all boundaries (Dixon *et al.*, 2012; Bonev and Cavalli, 2016). There are a few other features that establish the TAD boundaries, including histone modifications, housekeeping genes and short interspersed element (SINE) retrotransposons (Dixon *et al.*, 2012).

### 1.2.5 Sub-TADs and Chromatin Loops

With recent higher resolution studies, smaller domains (a few hundred kb) called sub-TADs were discovered, bringing the TAD median size down to ~185 kb with domain sizes ranging between 40 kb to 3 Mb (Rao *et al.*, 2014). Sub-TADs and chromatin loops form the basis of topologically associated domains and are highly correlated to gene expression. Although the positioning of TADs remain stable across cell types, changes in the chromatin structure are observed within the sub-domains during differentiation (Dixon *et al.*, 2015). 39% of TAD domain boundaries are connected to each other through loop domains defined by CTCF binding sites (Rao *et al.*, 2014). Chromatin loops vary in strength and length. They do not only define TAD and sub-TAD boundaries, but also connect distal enhancer and promoter elements, and recruit large protein complexes such as the preinitiation complex (PIC) and transcription factors to transcription start sites to facilitate and regulate transcription (Liu *et al.*, 2014).

Approximately 30% of chromatin loops are enhancer-promoter loops (Rao *et al.*, 2014), and these loops are confined to interactions within TAD boundaries, where they alter substantially between cell types (Dixon *et al.*, 2015). Studies suggested that 85% of CTCF binding sites are present within topological domains to mediate transcription interactions and CTCF motifs are present at the anchor of most long-range chromatin loops (Dixon *et al.*, 2012). Chromatin looping is coordinated together with another genome architecture protein

cohesin. It has been speculated that this ring-shaped protein complex physically holds the chromatin loops at CTCF binding sites (Parelho *et al.*, 2008). The specificity of enhancer-promoter loop formations will be discussed in section 1.3.2.

### 1.2.6 Nucleosome

Histones together with DNA are packed into a protein-dense structure called chromatin and each of the individual units, termed as nucleosome, form the most basic level of chromatin organisation (Alberts *et al.*, 2003). This 30 nm chromatin fibre is also described as ‘beads on a string’ where the string is the DNA, and each bead is a nucleosome core particle. Each nucleosome consists of a double-stranded DNA of ~146 bp wrapped around an octameric protein core with two molecules of each histones H2A, H2B, H3 and H4. Two histone H3-H4 dimers form the central tetramer and are capped by one H2A-H2B dimer on each end. The DNA wraps the histone octamer in 1.65 turns and there are 14 contact points between the histones and DNA (Luger *et al.*, 1997). These multiple interaction points make nucleosomes the most stable protein-DNA complexes under physiological conditions, thereby well-suited for its packaging function. The majority of nucleosome assembly occurs during DNA replication (Clapier *et al.*, 2017) and another histone protein, H1 is actively involved in connecting nucleosomes and stabilising the chromatin fibre (Hergeth and Schneider, 2015).

Nucleosome position is not static, they can be placed, moved and removed by chromatin remodelers. The length of linker DNA between two nucleosomes vary from a few nucleotides to 80 bp, and this variation is important for the diversity of gene regulation (Kornberg and Lorch, 1999). The genome wide nucleosome repeat length (NRL - average distance between centres of adjacent nucleosomes) changes during cell differentiation where higher NRL is correlated to regions with active transcription (Clarkson *et al.*, 2019).

### 1.2.7 Techniques Used to Study the Genome Architecture

The current understanding of the genome structure in 3D is largely dependent and limited by two types of approach, namely imaging and molecular cross-linking (Fraser *et al.*, 2015). A variety of microscopy imaging techniques, including fluorescence *in situ* hybridisation (FISH) can be used to visualise the distances between DNA segments in individual cells. The resolution and sensitivity of these assays are highly dependent on the size of the fluorescent probes and the light-capturing capability of the microscope. Hence these two factors limit the resolution to genomic distances at the sub 100 kb range (Fraser *et al.*, 2015).

The second technique, molecular crosslinking-based assays have been advanced substantially over the past two decades and nowadays the resolution of the genome-wide contact detection can be down to 1 kb (Rao *et al.*, 2014). These high-throughput assays capture DNA-DNA interactions in 3D space through cross-linking, followed by DNA shearing, and intramolecular ligation of DNA in close proximity. Following capture, the crosslinks are reversed and the interacting regions are sequenced by quantitative PCR (Dekker, 2002). This technique is termed chromosome conformation capture (3C), and the first generation was used to explore interactions between a pair of specified genomic loci. Later generations of the assays were developed, where chromosome conformation capture-on-chip (4C) assess contacts between one particular locus and the rest of the genome, and chromosome conformation capture carbon copy (5C), which studies contacts between all fragments within a given megabase genomic region (Fraser *et al.*, 2015). Capture Hi-C is the most comprehensive high-throughput assay that detects all genome-wide contacts and generates high resolution information about the genome (Fraser *et al.*, 2015). Lastly, chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) combines Hi-C and ChIP-seq (chromatin immunoprecipitation sequencing, a method used to analyse protein interactions with DNA) to give an overall view of interactions mediated by proteins of interests, such as CTCF and ZNF143 (Handoko *et al.*, 2011; Tang *et al.*, 2015; Ye *et al.*, 2020).

In summary, based on these techniques, the overall landscape of chromatin organisation has been developed and probed, but a lot of the molecular details of how chromatin architecture proteins act on the chromatin remain elusive. Therefore, further biochemical and structural studies would shed light into the molecular mechanisms of protein interactions.

## 1.3 The Role of Chromatin in Gene Transcription Regulation

### 1.3.1 Eukaryotic Transcription

Eukaryotic transcription is a complex system that is regulated at many levels to ensure specific genes are expressed in distinct cells to establish cell identity and function (Venters and Pugh, 2009). Transcription is performed by RNA polymerases (RNA Pol) to synthesize RNA using DNA as a template. There are three main types of RNA polymerases, including RNA polymerase I where it produces large ribosomal RNA precursor; RNA Pol II that synthesizes messenger RNA (mRNA) and non-coding RNA, and RNA Pol III that generates transfer RNA (tRNA) and small nuclear RNA (snRNA) (Venters and Pugh, 2009). The associated factors and mechanisms of regulation differ between the three types of polymerases (Cramer, 2019).

To allow transcription initiation, the RNA polymerase and its respective transcription factors must gain access to the core promoter region. Chromatin access is inhibited by nucleosomes. The nucleosomal DNA is hardly accessible by the DNA-binding factors for active transcription to take place, thereby nucleosomes must be removed or shifted (Cramer, 2019). Following chromatin remodelling, transcription is activated through the binding of site-specific chromatin proteins onto their cognate DNA motifs in response to cellular signals. These DNA motifs alter the local chromatin environment and facilitates the assembly of the pre-initiation complex (PIC), which is composed of general transcription factors (GTFs) and RNA polymerase (Venters and Pugh, 2009; Andersson and Sandelin, 2020). Each RNA polymerase recognises its own set of GTFs. These GTFs bind to the core promoter element that has a specific DNA sequence located near the transcription start

site (TSS) and recruit RNA polymerase to initiate transcription. The rate of RNA polymerase recruitment to the TSS can be modulated by the integration of DNA-binding transcription factors (TFs) (Venters and Pugh, 2009; Andersson and Sandelin, 2020). These transcription factors can bind near the promoter or enhancer regions, which are distal genomic regions with respect to the core promoter element. Multiple TFs can cooperate through *cis*-regulatory elements, and the communication between enhancer and promoter rely on the dynamic chromatin architecture, where chromatin loops are established to modulate transcription (Venters and Pugh, 2009; Cramer, 2019; Andersson and Sandelin, 2020).

Recent genome-wide expression profiling has improved our understanding on the organisation of the transcription machinery, in particular the mapping of protein interactions along genomic DNA and the positioning of nucleosomes throughout the genome (Venters and Pugh, 2009). Although chromatin packaging within the nucleus is an important factor that determines the transcriptional state of specific gene loci, it is becoming clearer that other events, including enhancer-promoter contacts and nucleosome remodelling significantly alter the transcriptional functional outcome (Li, Carey and Workman, 2007; Matharu and Ahituv, 2015; Schoenfelder and Fraser, 2019).

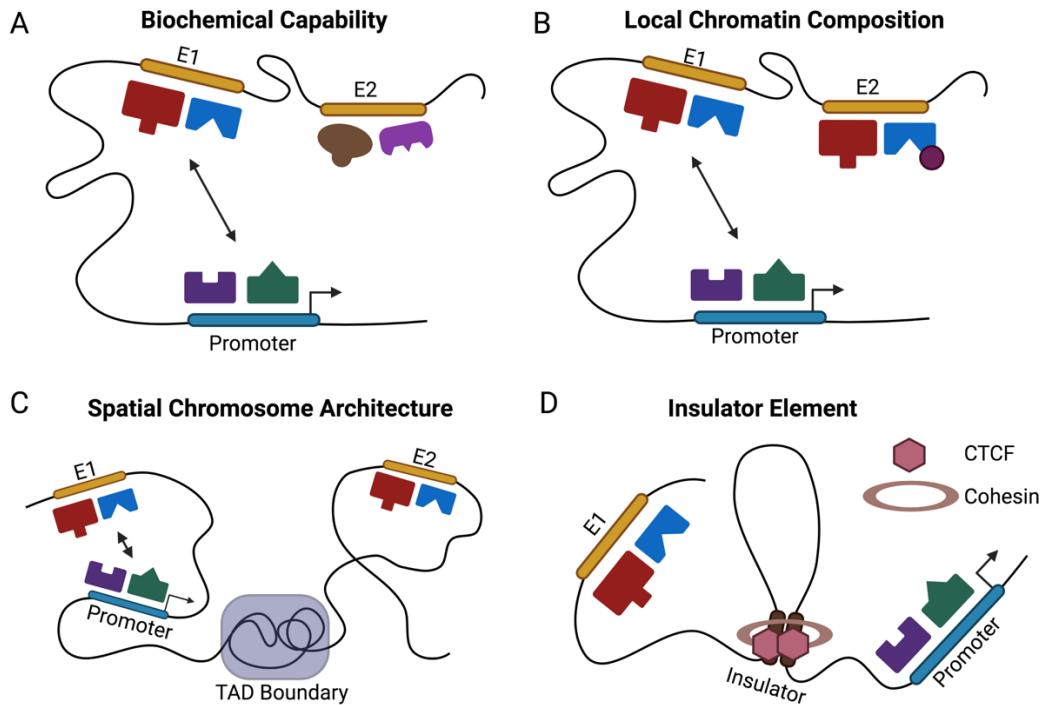
### 1.3.2 Long-Range Enhancer-Promoter Contacts

#### 1.3.2.1 Determinants for Enhancer-Promoter Specificity

*Cis*-gene regulatory elements mediate a variety of regulatory processes, from DNA looping to transcriptional and translational processes (Schoenfelder and Fraser, 2019). Together with site-specific transcription factors and chromatin regulators, they establish the molecular regulatory networks and control gene expression in a spatiotemporal manner (Matharu and Ahituv, 2015). There are a range of gene regulatory elements that have been characterised, including: “promoters”, which are essential for gene expression; “enhancers”, which target gene promoters to activate or increase transcription activity; “repressors”, which inhibit gene promoters; and “insulators”, which prevent non-cognate crosstalk between enhancers and promoters when present

between them (Matharu and Ahituv, 2015). While many promoters appear to be interacting with multiple enhancers and vice versa, a remarkable degree of specificity has been observed.

The total number of enhancers in the human genome is estimated to be around 1 million (Schoenfelder and Fraser, 2019). There are a few factors that can provide specificity to bring enhancers to their cognate promoters, including biochemical compatibility, local chromatin composition, spatial chromosome architecture, and insulator elements (van Arensbergen, van Steensel and Bussemaker, 2014) (Figure 1.2). Biochemical compatibility is encoded in the gene regulatory elements where specific DNA sequences bring transcription factors, co-factors and chromatin remodelling complexes to the same transcriptional loci. The intrinsic specificity of these protein combinations determines the enhancer-promoter selectivity (van Arensbergen, van Steensel and Bussemaker, 2014). Insulator elements can control the local chromatin structure and are bound by specific DNA-binding factors. One of the predominant models for how insulators work is the insulator-mediated looping model, where chromatin loops are mediated between the two insulator sites to prevent the physical interaction between enhancer and promoter, providing additional specificity to the enhancer-promoter contact (Herold, Bartkuhn and Renkawitz, 2012).



**Figure 1.2 Determinants for Enhancer-Promoter Specificity**

A: Biochemical specificity is encoded in the intrinsic properties of the transcription factors. Enhancer E1, but not E2 is compatible with the target promoter. B: The specificity of the transcription factors bound onto enhancer E2 has been modulated by the chromatin environment and hence the promoter selectively interacts with E1. C: Although both enhancers E1 and E2 are biochemically compatible with the promoter, the spatial chromosome architecture prevents E2 from contacting the promoter. D: The presence of insulator elements positioned in between the enhancer and promoter regions prevents the crosstalk through loop formation. Insulator-mediated loops are usually facilitated by architecture proteins CTCF and cohesin.

### 1.3.2.2 Direct Contact Between Enhancer and Promoter Elements

It has been postulated that the regulatory information to direct transcription is revealed through the direct contact between enhancers and promoters (Matharu and Ahituv, 2015). However, the chances where a physical contact between a distal enhancer and a promoter occur naturally are low unless the overall folding of the chromatin fibre brings the two elements into close proximity. Moreover, the relative contact frequency decreases drastically as the linear distance between the two elements increase. For instance, the contact frequencies of an enhancer-promoter pair at distance of 100 kb are

only ~ 2% of those at 2 kb (van Arensbergen, van Steensel and Bussemaker, 2014). Although there are enhancers that target the first promoter downstream or upstream, 30% functional enhancer-promoter interactions bypass at least one promoter and act over long linear distances (Wang *et al.*, 2018). The looping model is the most well-accepted and supported model on how the specificity between the enhancer-promoter pair can be achieved (Bulger and Groudine, 2011; Ong and Corces, 2011; Matharu and Ahituv, 2015). The looping model proposes that the contact between the enhancer and promoter element is facilitated by the enhancer-bound activating proteins and transcription factors at the promoter region, and these two elements come into close proximity to allow direct communication through the looping out of the intervening DNA region (Ptashne, 1986).

### 1.3.2.3 Looping Factors

Looping factors including architectural proteins, CTCF and cohesin, are found to be enriched at the anchors of strong chromatin interactions and are required to enable chromatin looping that act over long linear distances. Deletion of CTCF sites compromises the enhancer-promoter interactions and reduces transcription activities (Ren *et al.*, 2017; Kubo *et al.*, 2021). A recent study suggested that in addition to CTCF and cohesin, zinc finger protein 143 (ZNF143) is required at promoter regions to stimulate the formation of chromatin interaction with distal regulatory elements (Bailey *et al.*, 2015). ZNF143 was first characterised as a transcription factor, but recently it has been shown that ZNF143 acts as a positive driver of chromatin loops and has insulator properties similar to other architectural proteins (Bailey *et al.*, 2015). Increasing evidence is suggesting that ZNF143 is a key regulator for CTCF-bound promoter-enhancer loops as a large percentage of CTCF and ZNF143 DNA-binding motifs are distributed in converging orientations (Zhou *et al.*, 2021). In addition, enrichment of ZNF143 at CTCF binding sites has been reported, further hinting at a possible role in chromatin loop formation (Bailey *et al.*, 2015). Further details on how ZNF143 might facilitate chromatin looping will be discussed in section 1.5.4.

In addition to CTCF, cohesin and ZNF143, the transcriptional mediator complex and non-coding RNAs have also been shown to be involved in loop formations, acting as determinants for specific enhancer-promoter interactions (Matharu and Ahituv, 2015; Schoenfelder and Fraser, 2019).

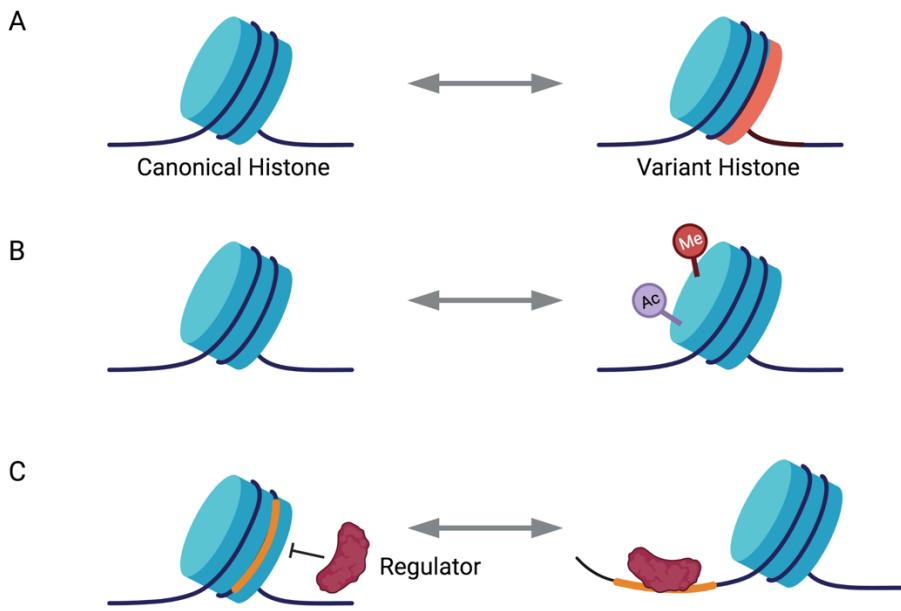
### 1.3.3 Chromatin Remodelling

#### 1.3.3.1 Dynamic Properties of Nucleosomes

Since the discovery of nucleosomes, it has been speculated that these repeating units on chromatin function beyond DNA compaction. The discovery of nucleosomes altering transcription *in vitro* in 1986 (Knezetic and Luse, 1986) has led to increasing chromatin research efforts. Further biochemical isolation and characterisation showed that the alteration of chromatin structure imposes a key regulatory step for all processes that act upon DNA (Kornberg and Lorch, 1999). Since nucleosomes are fundamentally repressive elements, active promoters are found in nucleosome-depleted regions (Cramer, 2019). With the genome wrapped around the histone cores, the nucleosomal DNA is hardly accessible by the DNA-binding factors for active gene expression and transcription to take place. The two main modes by which the chromatin is made more accessible for gene transcription to initiate are either covalent modification of histones (Fischle, Wang and Allis, 2003) or nucleosomes displaced by chromatin remodelling complexes to expose the underlying DNA with the use of energy from ATP hydrolysis (Yodh, 2013). Chromatin nucleosome remodelling and histone modification are two essential processes for eukaryotic transcription regulation. These two actions work in concert to restructure, mobilise and eject nucleosomes to regulate access of transcription machinery to the DNA (Kornberg and Lorch, 1999).

Nucleosomes have been shown to exhibit at least three dynamic properties *in vivo*, including compositional alteration, covalent modification and translational repositioning (Saha, Wittmeyer and Cairns, 2006) (Figure 1.3). Nucleosomes are constructed with the canonical histones (H2A, H2B, H3 and H4), but during transcription, non-canonical histone variants are deposited. For instance nucleosomes with histone H2A variant H2A.Z are highly enriched at gene

promoters to facilitate gene activation (Zhang, Roberts and Cairns, 2005). Histone covalent modification refers to the post-translational modifications of lysine residues on the histone amino-terminal tails that extend from the octamer core (Strahl and David Allis, 2000). Euchromatin has loose chromatin structure and euchromatin modifications are associated with active transcription, such as acetylation of histone 3 and histone 4 (H3 and H4) or methylation of lysine 4 on H3 (H3 K4me). Whereas in heterochromatin, chromatin are condensed and inactive for transcription, histones modifications are associated with the methylation of lysine 9 (H3 K9me) and lysine 27 (H3 K27me) on H3 (Strahl and David Allis, 2000). These modifications are distributed in distinct localised patterns and their locations are tightly regulated. Histone modifications have been postulated to make the chromatin more accessible for gene transcription factors through a number of methods, such as loosening the histone-DNA interactions through the change in the net charge of nucleosomes, and that modifications are recognised by specific chromatin proteins to promote transcription activities (Li, Carey and Workman, 2007). Lastly, nucleosomes are translocated or ejected to alternative positions along the DNA by chromatin remodelers to provide regulated access to DNA. The repositioning and ejection of nucleosomes allow the nucleosomes to be spaced properly during chromatin assembly and facilitate the orderly recruitment of transcription factors to specific genes during transcription (Saha, Wittmeyer and Cairns, 2006; Tyagi *et al.*, 2016).



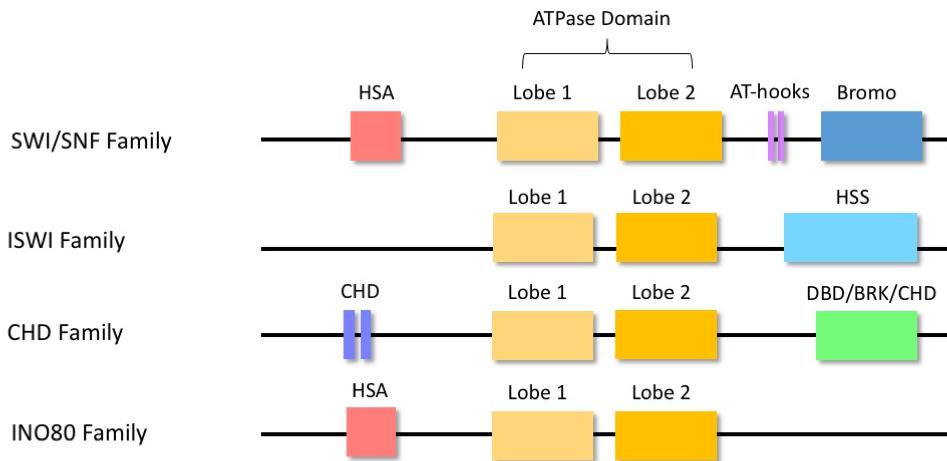
**Figure 1.3 Dynamic Properties of Nucleosomes**

A: The compositions of the histone core can be altered by chromatin remodelers. The most common alteration is from canonical histone H2A to variant histone H2A.Z. B: Methylation and acetylation modifications by histone-modifying enzymes have been postulated to enhance the binding of specific gene transcription factors. C: Nucleosome repositioning by chromatin remodelers allows the underlying nucleosomal DNA sequences (orange segment) to be more accessible to regulatory factors to facilitate transcription regulation.

### 1.3.3.2 Four Families of Chromatin Remodelers

A major class of chromatin regulators, termed chromatin-remodelling complexes utilise ATP hydrolysis to overcome the energy barrier created by the 14 histone-DNA contact points to catalyse a broad range of chromatin interactions including nucleosome sliding, nucleosome eviction and exchanging histone variants of the octamers (Tyagi *et al.*, 2016). ATP-dependent chromatin remodelers are classified into four distinct families: SWI/SNF (switch/sucrose-non-fermenting), ISW1 (imitation switch), CHD (chromodomain-helicase-DNA binding) and INO80 (inositol requiring 80) (Figure 1.4). All four families possess a helicase-like subunit of the SWI/SNF family of SF2 helicases. These families were assigned primarily based on their sequence homology within the catalytic ATPase subunit as well as their associated subunits. In general, ATP-dependent chromatin remodelers are multi-domain proteins that are comprised of the bromodomains, plant

homeodomains (PHD), chromodomains, SANT domains, and AT hook regions. All these domains work in conjunction to stabilise its interaction with histones and/or nucleosomal DNA (Saha, Wittmeyer and Cairns, 2006; Tyagi *et al.*, 2016; Clapier *et al.*, 2017).



**Figure 1.4 Domain Organisation of Chromatin Remodellers**

Chromatin remodellers are classified into four main families based on their domain architectures. The conserved domains of each family are highlighted. Lobe 1 and Lobe 2 form the basis of SF2-helicases and this ATPase domain is conserved throughout the four families. HSA: Helicase/SANT-associated, HSS: HAND-SANT-SLIDE, CHD: Chromodomain, DBD: DNA-binding domain, BRK: Brahma and Kismet.

In general, chromatin remodellers in the SWI/SNF family slide and unwrap nucleosomes along DNA to generate nucleosome-free DNA to allow chromatin access by sequence-specific transcription factors. The bromodomain at the C-terminal region targets acetylated histone tails to facilitate nucleosome remodelling. The additional HSA domain interacts with actin, where the presence of actin has been postulated to modulate the binding of SWI/SNF remodellers to the chromatin and stimulate ATPase activities (Tyagi *et al.*, 2016). ISWI ATPases are highly conserved in eukaryotes and are critical for the assembly and spacing of nucleosomes. Remodellers in the ISWI family demonstrate DNA-dependent ATPase activity and the binding to nucleosomal DNA is performed through its characteristic HSS domain. There are two main ISWI ATPases in human, SNF2H and SNF2L. In particular SNF2H has been shown to interact with cohesin and regulate chromatin

folding in loop domains (Tyagi *et al.*, 2016). The INO80 family influences nucleosome eviction and replaces canonical histones with histone variants. INO80 complexes contain a YY1 subunit, which is a zinc finger transcription factor that is crucial for growth and development regulatory genes. Similar to the SWI/SNF family, the presence of the HSA domain allows the interaction with actin components to facilitate chromatin remodelling (Tyagi *et al.*, 2016). Lastly, the characteristic feature in the CHD family is the presence of its tandem chromodomains that binds to methylated histones. This family can be further divided into three subfamilies (details in section 1.3.3.3). The main role of CHD proteins is to facilitate nucleosome assembly and ensure regular spacing between nucleosomes (Tyagi *et al.*, 2016).

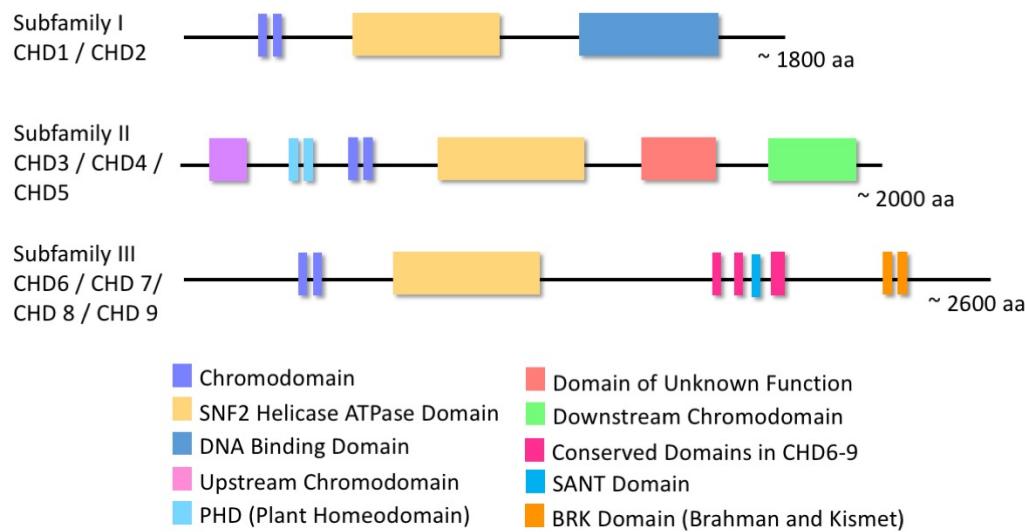
#### 1.3.3.3 Chromodomain-Helicase DNA Binding Proteins

Proteins in the chromodomain helicase DNA-binding (CHD) family are important regulators of the chromatin structure and CHD mutations have been implicated with numerous types of cancer (Li and Mills, 2014; Mills, 2017). CHD remodelers have been shown to be involved throughout the transcription process (initiation, elongation and termination), and are responsible for all three remodelling processes: assembly (spacing nucleosomes), access (exposing promoters) and editing (incorporating histone variants) (Tyagi *et al.*, 2016). They are highly conserved and are characterised by the N-terminal chromodomains next to a catalytic SF2 (superfamily2) helicase domain (Tyagi *et al.*, 2016). The chromodomain (chromatin organisation modifier) is an evolutionarily conserved sequence motif with three beta strands packed against a C-terminal alpha helix (Eissenberg, 2012). It acts as an interaction surface for a variety of chromatin factors (DNA, RNA and histones) associated with transcriptional regulation of genes. This motif is a 50 amino acid region of shared homology with epigenetic repressors heterochromatin protein 1 (HP1) and Polycomb (Pc) found in *Drosophila melanogaster* (Marfella and Imbalzano, 2007). The conserved set of amino acid motifs in the bi-lobed SF2-like ATPase domain consists of two tandem RecA-like folds that form an active-site cleft, which has also been found in proteins involved in chromatin assembly, transcription regulation and differentiation (Marfella and Imbalzano, 2007).

In mammalian cells, there are nine CHD proteins identified and they can be further divided into three sub-families based on the presence or absence of additional domains: group I CHD1-2, group II CHD3-5 and group III CHD6-9 (Figure 1.5). The CHD1 and CHD2 proteins in CHD group I contain a DNA-binding domain located in the C-terminal region. Recent evidence suggested that CHD1 can shift nucleosomes and induce regular nucleosome spacing (Lieleg *et al.*, 2015). In addition, CHD1 engages in nucleosome assembly over elongating RNA polymerase II and re-establishes the repressive chromatin structure (Tyagi *et al.*, 2016). Group II remodelers (CHD3, CHD4 and CHD5) lack the DNA-binding domain but contain a pair of N-terminal PHD zinc-finger like domains which recognise methylated histones. CHD4 in particular shows nucleosome spacing activities (Silva *et al.*, 2016) and is required for the maintenance of chromatin structures at sites of active transcription (Bornelöv *et al.*, 2018).

CHD6, CHD7, CHD8 and CHD9 belong to the third sub-family where additional function motifs were observed in the C-terminal region, including two BRK (Brahma and Kismet) domains and a SANT (switching-defective protein 3, adaptor 2, nuclear receptor co-repressor, transcription factor IIIB) domain. The BRK domain is specific to higher eukaryotes and act as transcriptional coactivator and corepressor through the interaction with chromatin substrates unique to higher eukaryotes. SANT domains have been shown to bind DNA and interact primarily with unmodified histone tails (Marfella and Imbalzano, 2007). Although these four proteins exhibit high degree of sequence identity (50-54%), their role within the genome are non-redundant because they recruit different subsets of sequence-specific transcription factors (Manning and Yusufzai, 2017). Furthermore, group III CHD enzymes are seen at the enhancer and promoter regions of genes by ChIP analysis, where they might regulate chromatin structure together with architecture proteins using their chromatin remodelling activities (de Dieuleveult *et al.*, 2016). Although sub-families I and III showed preference for binding to AT-rich substrates (Marfella and Imbalzano, 2007), CHD remodelers lack the intrinsic ability to target specific DNA sequences or genes.

Thereby how chromatin remodelling complexes are targeted to specific nucleosomes is an area that requires further investigation.



**Figure 1.5 Domain Architecture of the Three CHD Sub-Families**

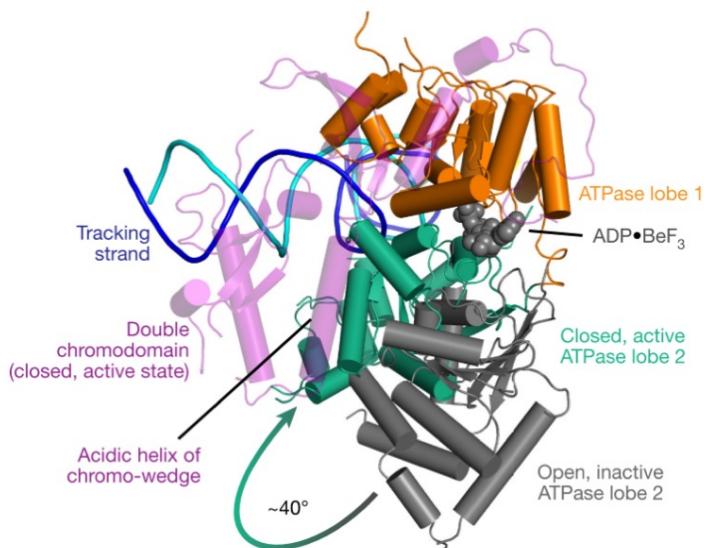
Schematic representation of protein domains found in human CHD remodellers. There are three sub-families under the CHD family and are characterised by the difference in their C-terminal domains. Members in the subfamily I, II and III are characterised by its DNA-binding domain, paired PHD zinc-finger like domains and BRK domains, respectively.

#### 1.3.3.4 Nucleosome Remodelling by CHD Remodellers

The mechanisms of how these CHD remodellers reposition nucleosomes through 150 bp of double-stranded DNA and determine nucleosome repeat lengths remain ambiguous. Nucleosome sliding by CHD remodellers is achieved by the highly conserved SF2-like ATPase motor with two core domains (lobes) that open and close in a nucleotide-dependent motion (Hauk and Bowman, 2011). The more well-accepted model, called twist diffusion model, postulated that nucleosomes are repositioned through an alteration of twist at the histone-DNA contact due to the transient accommodation of an additional base pair. This twist alteration process (called twist defect) will be transferred to the next DNA segment and hence the defect would travel around the nucleosome (Winger *et al.*, 2018; Bowman, 2019; Sabantsev *et al.*, 2019). Recent atomic resolution structures of yeast Chd1-nucleosome complex and human CHD4-nucleosome complex solved by cryo-electron microscopy

support the twist diffusion model (Farnung *et al.*, 2017; Farnung, Ochmann and Cramer, 2020).

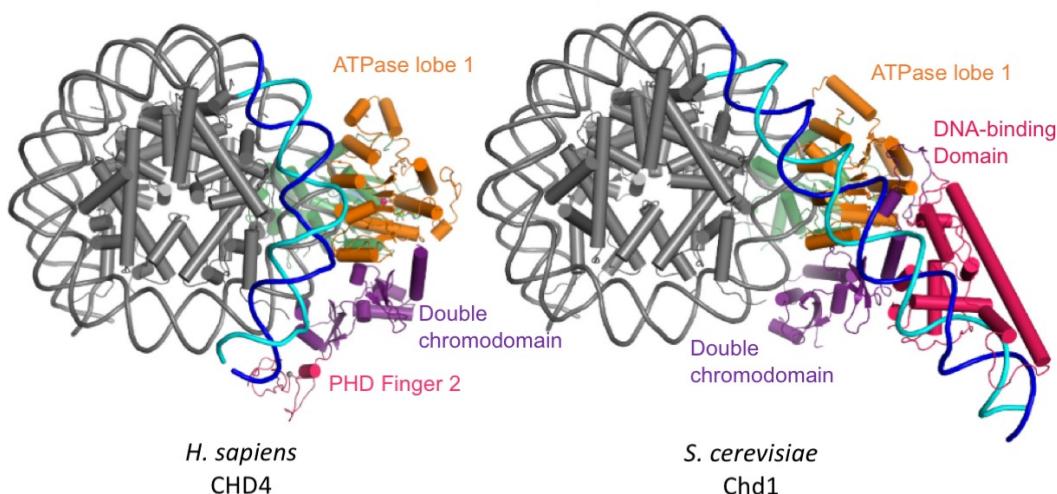
The yeast Chd1-nucleosome cryo-EM structure showed direct connection between the remodeler action and twist defects. The structure of the state where Chd1 was poised for catalytic activity in the presence of an ATP analogue was revealed (Figure 1.6). In the open state (apo or ADP bound), a bulge is formed between the ATPase motor and DNA at super-helix location 2 (SHL2) that shifts the entry DNA towards the nucleosome by one base pair. Next, one of the lobes (lobe 2) in the ATPase motor rotates in the presence of ATP and leads to the complete closure of the active-site cleft. This disfavours the bulged and undertwisted duplex and hence DNA is translocated in steps of one base pair in the 3' to 5' direction to re-establish the canonical twist of nucleosomal DNA. Lastly, ATP hydrolysis resets the ATPase domain to the pre-translocation state at the new DNA position, in preparation for the next round of translocation. In addition, the structure also reveals multiple contact sites between Chd1 and nucleosome, in particular, lobe 2 of the ATPase domain binds to the N-terminal tail of histone H4 (Farnung *et al.*, 2017). A temporal delay between DNA movements at the entry and exit sides of the nucleosomes was observed through single-molecule and cross-linking experiments, hence the translocation of DNA on the entry side is not coordinated to the DNA movement on the exit side (Sabantsev *et al.*, 2019)



### Figure 1.6 Structural Changes of Chd1 During ATPase Activation

The ATPase active-site adopts a closed conformation in the presence of the ATP analogue ADP-BeF<sub>3</sub>. ATPase lobe 2 is rotated by 40° towards lobe 1 to close the active site and hence disfavouring the undertwisted duplex to allow the translocation of DNA. Adapted from Farnung et al., 2017.

CHD4, part of the CHD sub-family group II has also been reported to display nucleosome spacing activities in the twist diffusion fashion, and a mechanism of action of human CHD4 was supported by results from single-molecule assays (Zhong et al., 2020) and cryo-EM analysis (Farnung, Ochmann and Cramer, 2020). Interestingly, the mechanism of action between CHD4 and Chd1 showed some degree of variability. A major difference is that Chd1 induces unwrapping of exit side nucleosomal DNA from the histone octamer, whereas CHD4 was unable to perform this action. This might be due to the different domain compositions between the two sub-families, the DNA-binding region in Chd1 interacts extensively with the terminal DNA, when compared to CHD4, the lack of DNA-binding region impacted its ability to interact and unwrap the substrate (Figure 1.7) (Farnung, Ochmann and Cramer, 2020). Similar to Chd1, a single-molecule assay also reported that the DNA entry and exit processes from the nucleosome by CHD4 are partially decoupled, where DNA enters the nucleosome continuously, but exits in bursts of 4 to 6 bp (Zhong et al., 2020).



### **Figure 1.7 Comparison of Structures of Known Nucleosome-CHD Remodellers**

Due to the difference in the terminal domain compositions between the two CHD remodellers, the mechanism of DNA unwrapping varies. CHD4 (left) does not possess a DNA-binding domain and therefore DNA is not detached from the second gyre of the nucleosome. In contrary, the nucleosomal DNA in the presence of Chd1 (right) unwraps from SHL -7 to -5, and the detached DNA is stabilised via its DNA-binding domain. Adapted from Farnung, Ochmann and Cramer, 2020.

Functional assays have also been carried out on CHD remodellers to understand its nucleosome spacing activity. The results suggested that yeast Chd1 exhibited clamping activity, where the nucleosome spacing is kept constant regardless of the nucleosome density. For instance, in areas with low nucleosome density, these nucleosomal arrays would still be of constant nucleosome repeat length but interspersed with nucleosome-free regions (Lieleg *et al.*, 2015).

To conclude, the first two CHD sub-families have been extensively studied, but the structure and differing mechanisms by which the CHD group III remodellers engage with the nucleosomal substrate differently remain largely unknown. In addition, there are still a lot of open questions that need to be answered, these include: the mechanism behind targeting of chromatin remodelling complexes to specific nucleosomes; and the mechanism CHD proteins use to induce regular nucleosome spacing.

#### **1.3.3.5 Techniques Used to Study Chromatin Remodelling**

Single molecule fluorescence resonance energy transfer (smFRET) is a biophysical technique that has been widely employed for understanding how chromatin remodelling proteins alter nucleosome positioning in a dynamic environment (Sabantsev *et al.*, 2019; Zhong *et al.*, 2020). This technique can be used to measure the distance of a pair of donor and acceptor fluorophores located within 1 to 10 nanometres. In chromatin remodeller studies, the fluorophores are usually placed on the DNA and the histone core, hence as the DNA moves either onto or off the histone, the FRET between the donor-acceptor pair can be recorded (Zhong *et al.*, 2020). In addition to two-colour

FRET experiment, three-colour smFRET was recently developed to monitor nucleosomal exit and entry sites to reveal the coordination of DNA movements on both sides simultaneously. Two different coloured fluorophores were placed on opposite ends of the DNA where the third fluorophore was positioned centrally on the histone core. The central dye acts as the donor for one fluorophore and acceptor for the other one, thereby allowing the measurement for two FRET pairs (Sabantsev *et al.*, 2019).

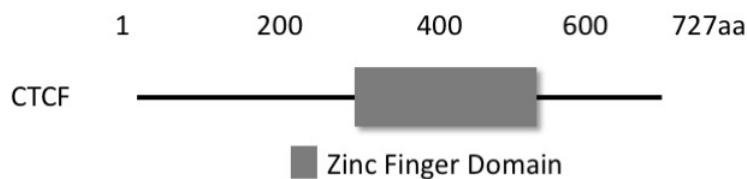
Recent structures of chromatin remodelling proteins bound to nucleosome were probed by cryo-electron microscopy (cryo-EM) (Farnung *et al.*, 2017; Farnung, Ochmann and Cramer, 2020). Cryo-EM is a powerful technique that can reveal the structure of individual protein structures and the resolution revolution of cryo-EM allowed the determination of atomic models of protein complexes (Kühlbrandt, 2014). Resolving the protein structure provides a greater level of understanding of the molecular mechanism of how these protein complexes remodel nucleosomes.

## 1.4 Chromatin Architecture Protein CTCF

### 1.4.1 Domain Architecture

CTCF is a ubiquitously expressed DNA-binding protein that has been identified as a driver gene in several cancer types (Guo *et al.*, 2018; Liu, Wu and Wang, 2018; Fang *et al.*, 2020). It is a highly conserved multidomain protein that influences the global chromatin architecture through sequence-specific DNA binding, protein-protein interactions and protein-RNA interactions (Phillips and Corces, 2009). It is best known for its role as an architectural protein where it acts as the “master weaver” of the genome through establishing gene expression patterns in higher eukaryotes (Phillips and Corces, 2009). CTCF serves as a transcriptional activator, a repressor and an insulator protein to help modulate the genome environment through facilitating enhancer-promoter loops within TAD regions (Splinter *et al.*, 2006) and acts as enhancer-blocking elements at TAD boundaries (Bell, West and Felsenfeld, 1999; Krivega and Dean, 2017).

CTCF consists of three domains (Figure 1.8), its central DNA-binding domain, comprised of 11 Cys<sub>2</sub>-His<sub>2</sub> (C2H2) zinc fingers (ZFs), recognises a diverse range of DNA sequences through combinatorial usage of its zinc fingers and is critical for its multiple roles within 3D genome organisation (Hashimoto *et al.*, 2017). The zinc finger domain is flanked by unstructured N- and C-terminal domains (Martinez and Miranda, 2010). The N-terminal domain of CTCF is involved in transcriptional activation and the activity is modulated by SUMOylation (Kitchen and Schoenherr, 2010), in particular, the N terminus is required to retain cohesin and stabilise the formation of chromatin loops (Li *et al.*, 2020; Nora *et al.*, 2020). It has also been suggested that CTCF forms homodimers *in vivo* through the N-terminal domain to mediate chromatin loop formation based on yeast two-hybrid and co-immunoprecipitation assays (Bonchuk *et al.*, 2020).



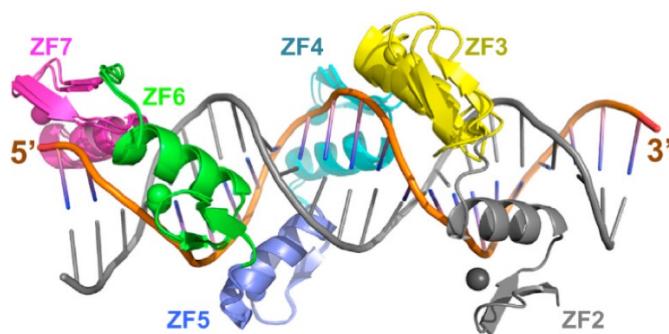
**Figure 1.8 Schematic Representation of Human CTCF Domains**

The domain architecture of human CTCF. The zinc finger domain is comprised of eleven zinc fingers.

#### 1.4.2 Mechanisms of CTCF Binding to DNA

CTCF is located at 55,000-80,000 sites in the mammalian genome (Ong and Corces, 2014), however the mechanism behind how CTCF recognises and binds to many degenerate DNA sequences with high affinity remains unclear. A CTCF binding motif comprised of 12-15 bp consensus sequence 5'-NCA-NNA-G(G/A)N-GGC-(G/A)(C/G)(T/C)-3' was revealed using chromatin immunoprecipitation exonuclease experiments (ChIP-exo) and this sequence is present at 80% of CTCF binding sites (Nakahashi *et al.*, 2013; Hashimoto *et al.*, 2017). The structure of CTCF ZF2-9 bound to DNA was solved by X-ray crystallography (Hashimoto *et al.*, 2017) and demonstrated that only CTCF ZF3-7 make base-specific contacts with the 15 bp consensus sequence and

recognise one strand of the double-stranded DNA in a linear polarity from 3' to 5' (Figure 1.9). This observation is in accordance to the DNA binding ability of C2H2 zinc fingers where each zinc finger interacts mainly with three adjacent DNA base pairs (Choo and Klug, 1997). At the same time specific residues within ZF3-7 are able to adopt alternative conformations to form versatile hydrogen bonds with some bases without altering the binding affinity to achieve high sequence adaptability (Hashimoto *et al.*, 2017). Furthermore, no additional sequence-specific binding to the 15 bp sequence was observed after ZF8 (Hashimoto *et al.*, 2017), but it has been suggested that ZF9-11 recognise a 10 bp long upstream motif that is present in 13% of CTCF binding sites (Nakahashi *et al.*, 2013). The lack of DNA specificity for the terminal ZFs leads to postulations that these ZFs together with the C-terminal domain might be involved in RNA-binding (Saldaña-Meyer *et al.*, 2019) or interactions with other proteins such as CHD8 (Ishihara, Oshimura and Nakao, 2006).



**15 bp CTCF consensus sequence:**

5'	ZF7			ZF6			ZF5			ZF4			ZF3			3'
	N	C	A	N	N	A	G	G/A	N	G	G	C	A/G	C/G	T/C	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	

**Figure 1.9 CTCF ZF3-7 Form Base-Specific Contacts**

The structure of ZF2-7 of human CTCF solved by X-ray crystallography by Hashimoto *et al.*, 2017. Each DNA base has been shown to interact with a particular amino acid in ZF3-7 in CTCF. The 15 bp CTCF consensus sequence is also illustrated where the top line indicates the zinc fingers involved in the interaction, the second line indicated the consensus DNA sequence, whereas the third line corresponds to the base pair positions (1-15).

CTCF-DNA binding can be modulated via DNA methylation at CpG sites within CTCF binding sites, and in particular, 40% of cell-type specific CTCF binding

sites are linked to differential DNA methylation (Wang *et al.*, 2012). Two cytosine methylations occur in mammalian cells, the affinity between CTCF and the consensus DNA can either increase or decrease depending on which cytosine is methylated (Hashimoto *et al.*, 2017). Methylation of CTCF loop anchor sites has been speculated to disable interactions with enhancers and impact gene expressions, therefore abnormal DNA methylation patterns are associated with the dysregulation of tumour suppressor or oncogenic genes in several type of cancers (Liu, Wu and Wang, 2018; Fang *et al.*, 2020).

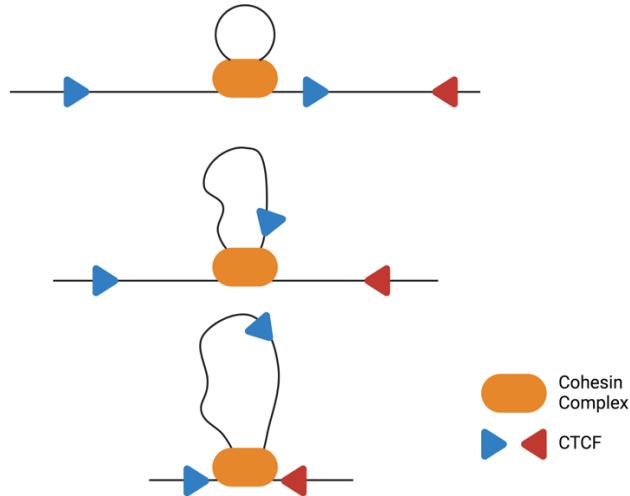
#### 1.4.3 CTCF-Mediated Enhancer-Promoter Looping Within TADs

Although CTCF was first identified as an insulator protein that predominantly blocks enhancer activities in eukaryotes, studies revealed that CTCF also mediates long-range interactions between distal elements and promoters that are not related to its insulator function (Sanyal *et al.*, 2012; Ong and Corces, 2014). In addition, among the 55,000 - 80,000 CTCF sites across the mammalian genome, 85% of the CTCF-binding sites are present within TAD boundaries and approximately ~ 15% of CTCF sites are located near promoters supporting an additional role of CTCF in mediating the interaction between distal regulatory elements (Ong and Corces, 2014; Rao *et al.*, 2014; Ren *et al.*, 2017). CTCF-mediated topological organisation has been confirmed in several individual loci, including transcription activation of the major histocompatibility complex class II and synaptotagmin 8 (Ong and Corces, 2014).

Activation of transcription is mediated by chromatin interactions. It requires the assembly of specific activators and cohesin is one of the most prominent CTCF interacting partner that is also involved in long-range chromosomal interactions (Schoenfelder and Fraser, 2019). CTCF was originally known as a cohesin loading factor because genome-wide ChIP-seq analysis showed that CTCF and cohesin colocalise extensively throughout the genome (Merkenschlager and Nora, 2016; Song and Kim, 2017). Cohesin is a ring-shaped complex comprised of two structural maintenance of chromosomes (SMC) molecules, SMC1 and SMC3, the kleisin subunit RAD21, and either

stromal antigen (STAG) STAG1 or STAG2 (Vietri Rudan and Hadjur, 2015). CTCF mediates the recruitment of cohesin to CTCF sites, whereas the presence of CTCF in the genome is independent of cohesin (Rubio *et al.*, 2008; Song and Kim, 2017). In particular, the N-terminal domain and the first two zinc fingers of CTCF are essential for cohesin positioning at CTCF sites and chromatin loop formation (Li *et al.*, 2020; Nora *et al.*, 2020; Pugacheva *et al.*, 2020). Cohesin together with CTCF are enriched at the base of enhancer-promoter chromatin loops and TAD boundaries and are required to establish the topological domain structure of the genome (Dixon *et al.*, 2012; Rao *et al.*, 2014).

The mechanism of how CTCF and cohesin form chromatin loops remains elusive. The loop extrusion model was developed to describe the loop formation process (Sanborn *et al.*, 2015; Fudenberg *et al.*, 2016) (Figure 1.10). This model proposed that both cohesin and CTCF are loaded onto the chromatin, cohesin translocates along the DNA and facilitates the extrusion of the flexible chromatin fibre. The growing loop is stalled once it reaches directional barriers. These directional barriers are CTCF homodimers bound onto CTCF consensus sequences pointed towards each other in a convergent orientation (de Wit *et al.*, 2015; Nora *et al.*, 2020). A recent study confirmed that the inversion of the orientation of the consensus sites can alter enhancer-promoter interactions and reshape the TAD domains (Guo *et al.*, 2015) and this specific CTCF orientation appears to be conserved in mammalian cells (Gómez-Marín *et al.*, 2015). Cohesin-mediated loop extrusion not only brings the enhancers and their target promoters to close proximity but also stabilises the enhancer-promoter interaction. It has been proposed that these loops are ‘quasi-stable’ and are further stabilised by a positive feedback loop by RNA and CTCF-interacting protein YY1 (Ren *et al.*, 2017).



**Figure 1.10 The Loop Extrusion Model**

Model of DNA loop extrusion by cohesin and CTCF. Cohesin translocates along the DNA and begins extruding a loop in a symmetrical fashion. The process is stalled until cohesin encounters convergently oriented CTCF molecules (one red and one blue arrow), resulting in a chromatin loop with cohesin and two copies of CTCF present at its anchor.

It is also important to note that enhancer-promoter contacts are more dynamic than CTCF loops because CTCF-loop anchors were not detected at some of these contacts even in high-resolution Hi-C contact maps. In addition, CTCF loops might not only direct enhancer-promoter contacts, but also restrict the range of action of enhancers to sequences located within the loops (Schoenfelder and Fraser, 2019).

#### 1.4.4 CTCF-Mediated Insulation at TAD Boundaries

CTCF was first known as an insulator protein that can block enhancer activity in eukaryotes as it has been shown to bind to known insulator sequences around the chicken  $\beta$ -globin locus (Bell, West and Felsenfeld, 1999) and within the imprinted control region of the mammalian H19/Igf2 locus (Bell and Felsenfeld, 2000). These observations were linked to the potential enhancer-blocking role of CTCF, where CTCF mediates chromatin loop formation and prevents the enhancer from making contact with the promoter of a target gene (Felsenfeld *et al.*, 2004) (Figure 1.2D). The evidence supporting CTCF as a genome-wide insulator was provided in 2007 when conserved regulatory

elements were analysed across the genome (Xie *et al.*, 2007). This study identified 15,000 conserved CTCF sites where the level of expression of nearby gene pairs is substantially reduced if they were separated by these CTCF sites (Xie *et al.*, 2007). Moreover, the insulator function of CTCF was correlated to TAD boundaries as seen in the enrichment of CTCF at boundaries between active and repressive chromatin domains (Cuddapah *et al.*, 2008), in particular, over 75% of TAD boundaries contain CTCF binding sites (Dixon *et al.*, 2012) confirming the importance of CTCF in defining boundaries. Together with cohesin, CTCF form chromatin loops at TAD anchors and in contrast to the convergent CTCF orientation sites as seen in enhancer-promoter loops, CTCF sites are positioned in diverging orientations at TAD boundaries (Nanni, Ceri and Logie, 2020).

#### 1.4.5 Nucleosome Positioning

The local nucleosome repeat length (NRL - average distance between centres of adjacent nucleosomes) is determined by the different transcriptional states of the chromatin, and the genome-wide NRL changes during cell differentiation (Nikitina *et al.*, 2017). CTCF binding sites are usually surrounded by well-positioned nucleosomes (Cuddapah *et al.*, 2008), and recent study showed that CTCF is able to displace nucleosomes from its binding sites to organise phased nucleosomal arrays (Owens *et al.*, 2019). In addition, CTCF possesses a nucleosome positioning signal as suggested by the observation of an approximately 10 bp decrease in the NRL near CTCF sites when compared to the genome-wide NRL (Clarkson *et al.*, 2019). Firstly, the decrease in NRL length is dependent on the CTCF binding strength towards its consensus site, with stronger binding strengths corresponding to shorter NRL distances. Next, the nucleosome arrangement near CTCF motifs is asymmetric due to the sequence of the DNA region (~ 200 bp) near CTCF motifs. In addition, nucleosome-depleted regions are present upstream of CTCF motifs, and these regions correspond to the increase in transcription of transposon elements (Clarkson *et al.*, 2019). Lastly, the nucleosome position near CTCF is further rearranged by the presence of chromatin remodelers,

where CHD4 and CHD8 were identified as two of the main regulators of NRL near CTCF (Clarkson *et al.*, 2019).

#### 1.4.6 Influences of CTCF Depletion

The presence of CTCF is not exclusive to TAD boundaries (Dixon *et al.*, 2012), it also plays an important role in intra-TAD loop formations (Splinter *et al.*, 2006). This is supported by the evidence where the depletion of CTCF reduces the intradomain interactions and alters gene expression patterns (Zuin *et al.*, 2014). Deletion of CTCF at the human *X inactive specific transcript* locus and *HoxA* locus in *Drosophila melanogaster* result in a shift of the boundary position and lead to the fusion of two adjacent TADs, hinting a critical function in genome spatial segregation (Nora *et al.*, 2012; Narendra *et al.*, 2015). The recent advancement in the auxin-inducible degron (AID) system has been employed to study the effect of depleted CTCF levels in the 3D genome. The studies conducted by Nora *et al.* in mouse embryonic stem cells suggest the depletion of CTCF levels leads to insulation defects at ~80% of TAD boundaries and the chromatin loop formation at CTCF sites reduces drastically (Nora *et al.*, 2017). This study confirmed that CTCF is required for chromatin loops and the deletion of the underlying CTCF sites can lead to the dramatic loss of physical insulation between two adjacent TADs. These chromatin loops and TAD boundaries can be restored upon the removal of auxin. Furthermore, the effect of CTCF disruption is restricted to the sub-megabase level as it does not affect higher-order chromosome folding or compartmentalisation into active and inactive compartments (Nora *et al.*, 2017).

Apart from depletions, inversion of the CTCF convergent binding sites also induces TAD reorganisation and redirects the enhancer-promoter interactions. Single CTCF element inversion has been shown to encapsulate two TAD domains (de Wit *et al.*, 2015; Guo *et al.*, 2015) since chromatin loops between enhancer-promoter pair requires convergent CTCF orientation (Nora *et al.*, 2020). These results altogether demonstrated that CTCF is a major determinant of the mammalian chromosome folding.

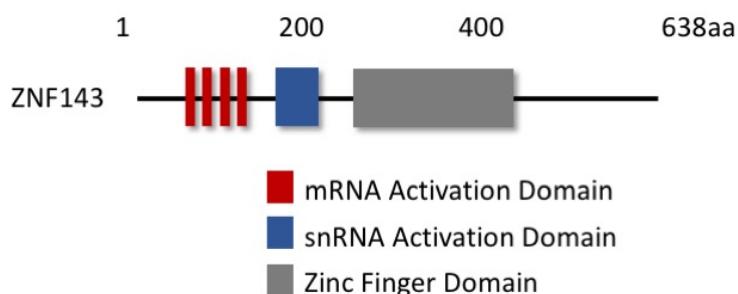
## 1.5 Transcription Factor ZNF143

### 1.5.1 Domain Architecture

ZNF143 is a ubiquitously expressed transcription factor conserved in all vertebrates. Over 2000 mammalian promoters are regulated by ZNF143 (Ngondo-Mbongo *et al.*, 2013) and it is particularly important for regulating genes involved in primary metabolism and cell growth (Izumi *et al.*, 2010). ZNF143 is highly expressed in tumours, such as oesophageal carcinoma, gastric and lung adenocarcinoma (Izumi *et al.*, 2010; Kawatsu *et al.*, 2014; Ye *et al.*, 2020). It was first identified as a transcription activator, but recently there is growing evidence suggesting that ZNF143 has a distinct role in coordinating chromatin interactions and establishing higher order structure within the genome (Ye *et al.*, 2020; Huning and Kunkel, 2021). Most studies on ZNF143 have been carried out on the *Xenopus laevis* homolog Staf (selenocysteine tRNA gene transcription activating factor), where human ZNF143 and *Xenopus* Staf are highly conserved with 84% sequence similarity (Myslinski, Krol and Carbon, 1998).

There are three main domains in human ZNF143 / *X. laevis* Staf (Figure 1.11). The N-terminal domain is comprised of two separable activation domains where the first activation domain is made up of four repeated motifs that activate mRNA promoters controlled by RNA polymerase II. The second activation domain is smaller and has been shown to activate small nuclear RNA (snRNA) controlled by RNA polymerase III type III promoters (Schuster *et al.*, 1995; Schuster, Krol and Carbon, 1998). These promoter activator motifs are conserved across human ZNF143, mouse ZNF143 and *X. laevis* Staf (Myslinski, Krol and Carbon, 1998). The third domain is referred to as the DNA-binding domain, it is comprised of seven tandemly repeated zinc finger DNA binding motifs. In each of the C2H2 type zinc fingers (ZFs), the zinc ion is coordinated by two histidine and two cysteine residues. In Staf, ZF3-6 account for the minimum binding region to retain specific and high affinity binding to DNA (Schaub, Krol and Carbon, 2000). Although no known function has been attributed to the C-terminal domain, the evolutionary conserved

sequence suggests it might have an important role that is yet to be determined (Myslinski, Krol and Carbon, 1998).



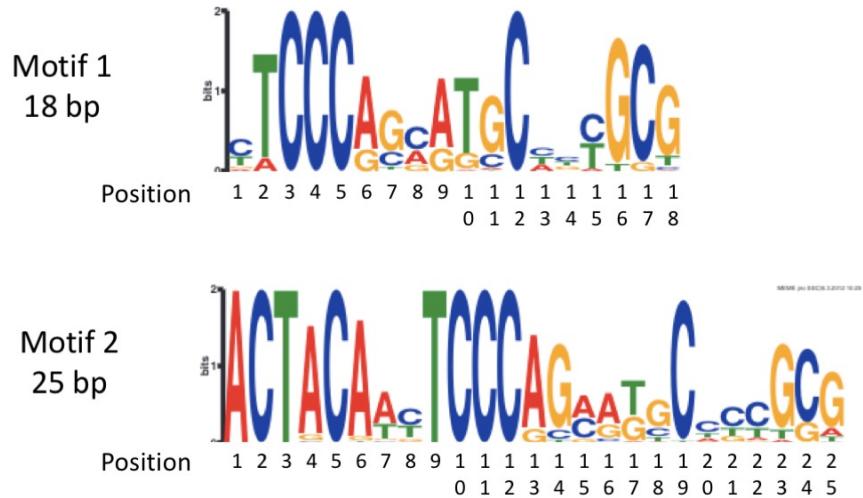
**Figure 1.11 Schematic Representation of Human ZNF143 Domains**

The domain architecture of ZNF143 illustrated in a bar diagram. The domains and the sequences that have been shown to have specific functions are colour coded. There are seven zinc fingers in the zinc finger domain.

Although ZNF143 plays a critical role in regulating eukaryotic genes, the molecular mechanism of how ZNF143 binds and activates two different polymerases still remains ambiguous. Furthermore, the atomic structure for both ZNF143 and Staf are yet to be determined, which is likely due to its intrinsically disordered regions.

### 1.5.2 Mechanisms of ZNF143 Binding to DNA

The seven zinc fingers in the DNA binding domain of ZNF143 bind to a degenerate consensus sequence of 18 bp in proximal promoter regions, where these elements are usually located within 200 base pairs of the transcriptional start site of specific gene targets (Myslinski *et al.*, 2006). This 18 bp ZNF143 consensus sequence is comprised of GC rich regions, and specifically the core of this motif is a conserved CCCA sequence (Ngondo-Mbongo *et al.*, 2013) (Figure 1.12). A second 25 bp motif containing an extended ACTACA accessory sequence located immediately upstream of the 18 bp motif was also observed in ZNF143 binding events. The presence of the 18 bp motif and/or 25 bp motif were detected in 74% of all binding events (Ngondo-Mbongo *et al.*, 2013). The variability in the 5' and 3' ends of the ZNF143 consensus sequence provides the basis for flexible and versatile DNA binding activities (Huning and Kunkel, 2021).



## Figure 1.12 ZNF143 Consensus Sequence

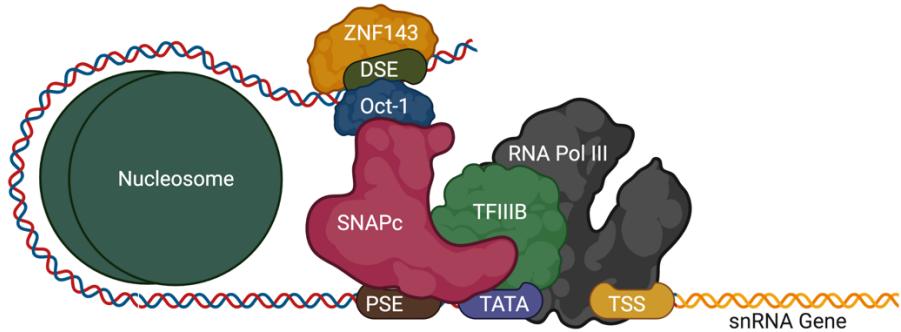
ZNF143 binding motifs discovered by Ngondo-Mbongo *et al.*, 2013 *de novo* at ZNF143 peaks in human genome. Two sequences were detected where motif 1 comprised of 18 bp was identified in 32% of all peaks, and motif 2 with an extended ACTACA sequence was identified in 23% of all peaks. The presence of the 18 bp motif and/or 25 bp motif were detected in 74% of all binding events.

DNA binding assays revealed that the zinc finger domain of *X. laevis* Staf makes extensive contacts with the DNA major groove and it associates with the non-template strand more closely (Schaub, Krol and Carbon, 2000). The number of zinc fingers required for DNA binding might be promoter dependent, where Staf ZF2-7 are required for human *U6* snRNA gene binding, whereas mutations with the ZNF143 ZF1 impacted the binding of ZNF143 to *Xenopus laevis* tRNA<sup>sec</sup> promoters (Schaub, Krol and Carbon, 1999, 2000). Specifically, ZF3-6 represent the minimal zinc finger region required for high DNA binding affinity, yet it is important to note that the binding of Staf to DNA might not be restricted to its zinc finger domain (Schaub, Krol and Carbon, 2000). Although ZF7 does not induce base specific contact, it might be involved in potential intramolecular protein-protein interactions. The non-specific base contact of ZF7 and the flexible use of ZF1 in Staf illustrated how the protein can bind to divergent target sequences with high affinity to accommodate its ability to influence gene transcription from a variety of different promoters (Schaub, Krol and Carbon, 1999, 2000).

### 1.5.3 Regulation of ZNF143 at Gene Promoters

#### 1.5.3.1 RNA Polymerase III Promoters

ZNF143 binding sites are observed in at least 70% of vertebrate snRNA and snRNA-like promoter regions that are transcribed by RNA polymerase II or RNA polymerase III type III promoters, including *U1*, *U6*, *7SK* snRNA promoters (Huning and Kunkel, 2021). snRNA promoters are regulated by two sequence regulatory elements, including the proximal sequence element (PSE) and distal enhancer-like sequence element (DSE). PSE is required for transcription initiation of snRNA genes and is usually located within 100 bp of the transcriptional start site (TSS), whereas DSE is located close to 200 bp upstream of TSS and it stabilises the transcription initiation complex and stimulates transcription activity (Huning and Kunkel, 2021). ZNF143 binding sites for snRNA gene promoters are usually comprised of the canonical 18 bp motif and are nested within the DSE region. Along with ZNF143, another highly characterised transcriptional activator protein Oct-1 is also recruited to the DSE to stabilise the transcription initiation complex. With the help of the nucleosome, Oct-1 gets into close proximity to the SNAPc complex which is located on the PSE, and thus activates transcription through RNA polymerase III (James Faresse *et al.*, 2012; Ramsay and Vannini, 2018) (Figure 1.13). Although ZNF143 and Oct-1 are located in close proximity, no cooperative binding has been observed between these two factors suggesting that they function independently (Schaub, Krol and Carbon, 1999; Huning and Kunkel, 2021). Due to the sequence similarities in the snRNA activating domain of ZNF143 and the POU domain of Oct-1, it has been speculated that ZNF143 might also associate with the SNAPc complex directly or indirectly to activate transcription (Schuster, Krol and Carbon, 1998). How ZNF143 facilitates the formation of the stable transcription initiation complex and whether it interacts with any of the other proteins in the pre-initiation complex require further evidence.



**Figure 1.13 Transcription Pre-Initiation Complex at snRNA Promoters**

snRNA genes are transcribed by RNA polymerase III type III *U6* promoter. ZNF143 and Oct-1 are present at the distal sequence element (DSE). A nucleosome is bound to the intervening region (~150 bp) that allows Oct-1 and ZNF143 to be in close proximity with SNAPc to stabilise the binding of SNAPc at the proximal sequence element (PSE). SNAPc is also associated with TFIIIB that binds to the TATA box. RNA polymerase III is then recruited via TFIIIB to transcription start site (TSS) to activate transcription.

ZNF143 might also possess nucleosome positioning activity at snRNA promoters as seen in the study where ZNF143 was able to activate transcription on a chromatinised *U6* promoter and colocalise with a chromatin remodeller CHD8 on the *U6* promoter (Yuan *et al.*, 2007). Moreover, the co-immunoprecipitation of ZNF143 and CHD8 implies a possible interaction and that ZNF143 might stimulate transcription activation through recruiting CHD8 to the promoter (Yuan *et al.*, 2007).

#### 1.5.3.2 RNA Polymerase II Promoters

In addition to stimulating transcription for snRNA genes, ZNF143 is also involved in mRNA encoding gene promoters transcribed by RNA polymerase II and a significant set (93%) of RNA pol II bidirectional promoters are dependent on ZNF143 (Anno *et al.*, 2011). It has been reported that ZNF143 is the key transcription regulator for cell cycle progression genes such as the *BUB1B* gene involved in spindle-checkpoint activation (Myslinski *et al.*, 2007) and the knockdown of ZNF143 has been correlated to the G2/M arrest in prostate cancer cells (Izumi *et al.*, 2010). ZNF143 binding sites in these promoters are a combination of the 18 bp ZNF143 consensus sequence and

the expanded 25 bp motif with the secondary accessory ACTACA sequence (Ngondo-Mbongo *et al.*, 2013). Mutation within the ACTACA sequence has been shown to impact the RNA polymerase II density recruited to the promoter and reduces the amount of active histone markings around ZNF143 bound promoter regions (Huning and Kunkel, 2021).

#### 1.5.4 Potential Role as a Chromatin-Looping Factor

Increasing evidence has suggested that distal enhancer elements are brought into close proximity to its cognate promoter pair through the formation of chromatin loops and that CTCF and cohesin are the two main architecture proteins involved in this chromatin interaction (Ong and Corces, 2014). The colocalisation of ZNF143 and CTCF and cohesin was first reported in 2013 (Xie *et al.*, 2013), followed by the discovery of the distinct role of ZNF143 through ChIA-PET experiments in human leukaemia cell lines where it showed that ZNF143 is involved in long-range chromatin interactions (Heidari *et al.*, 2014).

With the advancement of molecular crosslinking-based assays, further publications observed the extensive enrichment of ZNF143 at CTCF binding sites (Bailey *et al.*, 2015; Ye *et al.*, 2016). Moreover, knockdown of ZNF143 has been associated with the elimination and destabilisation of chromatin loops (Wen *et al.*, 2018; Zhou *et al.*, 2021). Together with motif analysis, these results proposed that ZNF143 cooperates with architecture proteins to establish chromatin interactions (Bailey *et al.*, 2015; Ye *et al.*, 2016; Wen *et al.*, 2018). However, it is interesting to note that the ZNF143 consensus motif is not always present at ZNF143 enriched chromatin anchors, and the colocalization of ZNF143 with CTCF-cohesin tends to involve weaker ZNF143 binding sites (Bailey *et al.*, 2015). Although ZNF143 has been shown to be the key regulator of CTCF-bound promoter-enhancer loops (Wen *et al.*, 2018; Zhou *et al.*, 2021), ZNF143-CTCF-cohesin complexes mainly utilise CTCF motifs instead of ZNF143 sequences hinting that ZNF143 might act as a cofactor instead of a pioneer factor in chromatin loop formation (Wen *et al.*, 2018). Lastly, in addition to the positive influence on the level of chromatin

loops, ZNF143 has also displayed insulator properties similar to other architectural proteins (El Mourad and Cuvier, 2017).

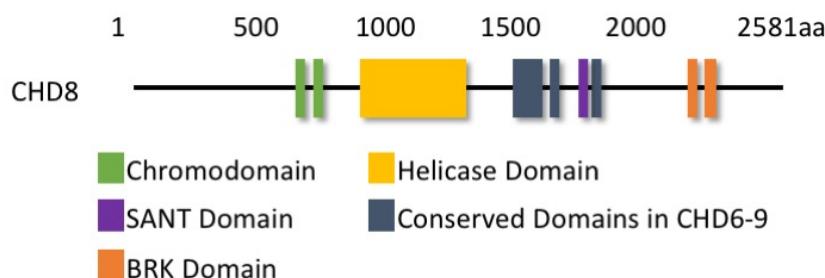
In summary, ZNF143 is an eukaryotic transcription factor that performs a broad range of distinctive roles within the genome, including the stabilisation of transcription initiation complex at a variety of gene promoters; and its architectural role through the interaction with CTCF and cohesin to form chromatin loops and insulator barriers within the chromatin (Ye *et al.*, 2020; Huning and Kunkel, 2021). Yet, the molecular mechanisms behind each of these functions are still unclear and therefore further investigations are required to reveal the structural and functional details of ZNF143.

## 1.6 Chromatin Remodelling Protein CHD8

### 1.6.1 Domain Architecture

ATP-dependent chromatin remodelling proteins play an important role in the regulation of gene transcription. They modulate the contacts between histones and DNA. CHD8 (Chromodomain helicase DNA-binding protein 8) belongs to group III of CHD remodelers. It has a strong link with gastric, prostate and breast cancers (Mills, 2017). All group III CHD proteins share a high degree of sequence identity, but these paralogs play non-redundant roles in the cell (Manning and Yusufzai, 2017). A lot of studies on CHD8 have been on its role associated with autism spectrum disorders as it is essential for the gene expression pathways during brain development (Sugathan *et al.*, 2014; Barnard, Pomaville and O’Roak, 2015). In addition, CHD8 has also been shown to regulate the Wnt signalling pathway through its direct interaction with  $\beta$ -catenin (Thompson, Lin and Bochar, 2008). CHD8 also interacts with RNA polymerase II to control cell proliferation (Rodríguez-Paredes *et al.*, 2009), associates with ZNF143 for efficient RNA polymerase III transcription (Yuan *et al.*, 2007), and functions at active insulator sites through its interaction with CTCF (Ishihara, Oshimura and Nakao, 2006).

CHD8 is composed of two chromodomains, a SF2 helicase domain, three regions conserved in CHD6 through CHD9 (CR1-CR3), a SANT domain, and two BRK domains (Yuan *et al.*, 2007) (Figure 1.14). The N-terminal chromodomains are highly conserved and it has been suggested that the chromodomains in CHD family are able to bind methylated histones and act as an interaction surface for DNA (Clapier *et al.*, 2017). The SF2 helicase domain applies an ATP-dependent torsional strain to the DNA, which provides the force to remodel nucleosomes (Ryan and Owen-hughes, 2014). The SANT domain contributes to DNA binding (Aasland, Stewart and Gibson, 1996) and interacts primarily with unmodified histone tails (Marfella and Imbalzano, 2007). The function of the two BRK domains are unknown, but there are speculations that these two domains act as transcriptional coactivators and corepressors through the interaction with chromatin substrates unique to higher eukaryotes (Marfella and Imbalzano, 2007). Although CHD8 does not target specific DNA sequences or genes, CHD8 shows preference for binding to AT-rich substrates (Marfella and Imbalzano, 2007).



**Figure 1.14 Schematic Representation of Human CHD8**

The domain architecture of CHD8 is illustrated in a bar diagram. The domains and sequences that have been shown to have specific functions are colour coded. The SANT, BRK and conserved domains in CHD6-9 are unique to CHD subfamily III.

The overall structure of CHD8, how it is able to remodel the chromatin through its interaction with nucleosomes and how it associates with various transcription factors are yet to be determined.

### 1.6.2 Role in Chromatin Remodelling

The current understanding of how CHD remodellers modulate the chromatin environment and translocate nucleosomes have been discussed in detail in section 1.3.3, thereby only results and observations specific to CHD8 will be mentioned in this section.

Contacts with DNA are crucial for enhancing the binding of CHD8 to nucleosomes, and it has been reported that CHD8 does not bind well to DNA substrates that are shorter than 40 bp long and thereby the interaction between CHD8 and nucleosome core (without linker) is detectably weaker than the interaction between CHD8 and nucleosome with 50 bp linker on the exit site (Manning and Yusufzai, 2017). CHD8 does not only interact with the linker DNA, but also binds to the N-terminus of histone H3 potentially through its chromodomains (Yuan *et al.*, 2007). In addition, CHD8 shows robust nucleosome sliding activity with the use of ATP, however it has been suggested that ATP hydrolysis is correlated to nucleosome binding to a higher extent than to nucleosome sliding. Yet, it has also been speculated that ATP hydrolysis between nucleosome binding and sliding might be uncoupled (Manning and Yusufzai, 2017).

### 1.6.3 The Interplay with 3D Architectural Proteins

Although CHD8 is involved in a range of different activities, a handful of proteins has been shown to associate with CHD8, including β-catenin, RNA polymerase II, ZNF143 and CTCF. In particular, ZNF143 and CTCF are two architecture proteins involved in the higher-order organisation genome and therefore there are increasing speculations on the importance of nucleosome remodelling at gene promoters and chromatin loop anchors. Little is known about the interaction between CHD8 and ZNF143, and CHD8 and CTCF, hence there are still a vast number of open questions in this field that need to be answered.

CHD8 has been shown to be involved in ZNF143-associated RNA polymerase III U6 promoters. The interaction between ZNF143 and CHD8 was identified

through an immunoprecipitation assay (Yuan *et al.*, 2007). Further studies reported that CHD8 is detected and localised on the *U6* promoter, implying that *U6* promoter transcription might involve chromatin remodelling (Yuan *et al.*, 2007).

CHD8 occupies known CTCF binding sites *in vivo* to play an important role in insulation and epigenetic regulation. Direct binding between the C-terminal domain of mouse CHD8 and zinc finger domain of mouse CTCF has been observed through pull-down analysis and immunoprecipitation assays (Ishihara, Oshimura and Nakao, 2006). ChIP-seq analysis revealed that the colocalisation of CHD8 at selected known CTCF insulator sites is directly involved in the insulator activity and CTCF mediates the recruitment of CHD8 to these sites (Ishihara, Oshimura and Nakao, 2006). Furthermore, it has been speculated that the CTCF-CHD8 complex might induce epigenetic remodelling in some insulators as seen by the histone acetylation and CpG hypermethylation at CTCF binding sites due to the loss of CHD8 (Ishihara, Oshimura and Nakao, 2006). In addition, a recent nucleosome positioning study reported a negative correlation between nucleosome repeat length and the presence of CHD8 at CTCF binding sites (Clarkson *et al.*, 2019).

The only protein that has been shown to interact with both ZNF143 and CTCF is the chromatin remodeller CHD8 (Ishihara, Oshimura and Nakao, 2006; Yuan *et al.*, 2007), and thereby a potential mechanism involving the chromatin remodelling activity of CHD8 might be employed at enhancer-promoter loop anchors to create nucleosome-free regions (Huning and Kunkel, 2021). Yet the mechanism behind this remains elusive.

## 1.7 Perspectives and Aims

In summary, the role of chromatin in eukaryotic transcription activation is fundamental and the interplay of different chromatin proteins are essential for gene expression regulation. Several previous studies have underlined the importance of CTCF in the regulation of chromatin architecture, specifically

revealing how CTCF is involved in chromatin loop formation at TAD boundaries and intra-TADs. However interestingly, recent discoveries of ZNF143 and CHD8 colocalising at CTCF-binding sites hinted the potential interaction between these three proteins, generating a more complex picture, with these additional proteins potentially bringing a regulatory role at gene promoters and chromatin loop anchors.

Most of our understanding of the overall landscape of chromatin organisation and transcription regulation has been probed by *in vivo* studies, with the mechanistic features which underlie the role of CTCF, ZNF143 and CHD8 to regulate transcription and genome architecture remaining elusive. Thereby, the main aim of this study is to explore the biochemical and biophysical basis of the interaction between CTCF, ZNF143 and CHD8. With this objective in mind, this study aims to address several questions:

- Does CTCF, ZNF143 and CHD8 genome organisation role depend on the formation of direct protein-protein interactions?
- What is the structural basis behind the genome organisation function?
- Which are the molecular determinants of this mechanism?

To answer these questions, CTCF, ZNF143 and CHD8 will be expressed and purified separately. Detailed functional assays will be carried out on the individual proteins to explore how they act on the genome and to provide functional insights into the mechanistic and structural features of the chromatin protein complexes. If protein interaction is observed, the protein complexes will then be subjected to biochemical and biophysical assays and analysed by cryo-electron microscopy.

## 2 Materials and Methods

### 2.1 Molecular Cloning

#### 2.1.1 Overview

The genes used in this thesis were purchased from Genscript and rare codon optimised for expression in *E. coli* (*Escherichia coli*) regardless of the expression system. The purchased genes were synthesized into pUC57 vector and subsequently cloned into other vectors for expression. The first step of cloning required the use of restriction enzymes to linearise the expression vector. Secondly, the gene of interest was amplified from the pUC57 vector using polymerase chain reaction (PCR). A set of primers was used to create an overhang on the amplified gene of interest to anneal with the digested vector. Next, the PCR product and linearised expression vector were ligated using the Gibson Assembly® Cloning Kit (New England Biolabs). The ligated construct was subsequently transformed into Stellar™ competent cells (Takara Bio). Single colonies were picked after overnight incubation on antibiotic agar plates at 37 °C and colony PCR was carried out to determine the presence of the insert. DNA of the colonies with the correct gene insert was isolated and purified using GeneJet Plasmid Mini Prep Kit (Thermo Fisher Scientific). DNA was then sent to Source Bioscience for sequencing to confirm the cloned construct.

#### 2.1.2 Plasmids constructs

All vectors used for cloning and expression are shown in table 2.1. CTCF, ZNF143, CHD8 and other constructs used in this thesis are listed in tables 2.2, 2.3, 2.4 and 2.5 respectively. The DNA and protein sequence for the affinity tags and cleavage sites are mentioned in table 2.6.

Vector	Antibiotic Resistance	Tag and Cleavage Site	Vector size (bp)	Expression system
pLIB	Ampicillin, Gentamycin	/	4,970	Insect cells
pGEX-6P-1	Ampicillin	N-GST, 3C	4,984	<i>E. coli</i>

<b>POPINJ</b>	Ampicillin	N-His <sub>6</sub> -GST, 3C	5,457	<i>E. coli</i> / Insect cells
---------------	------------	--------------------------------	-------	----------------------------------

**Table 2.1 List of Vectors Used for Cloning Different Protein Constructs**

#	Construct	Length (aa)	Vector	Tag and Cleavage Site
<b>1</b>	FL CTCF	1-727	pLIB	N-HA, 3C C-StrepII, TEV
<b>2</b>	FL CTCF <sub>ybbR</sub>	1-727	pLIB	N-HA, 3C C-ybbR-StrepII
<b>3</b>	CTCF <sub>61-727</sub>	61-727	pLIB	N-His <sub>8</sub> , 3C C-StrepII, TEV
<b>4</b>	CTCF <sub>1-240</sub>	1-240	pLIB	N-HA, 3C C-StrepII, TEV
<b>5</b>	CTCF <sub>241-590</sub>	241-590	pLIB	N-HA, 3C C-StrepII, TEV
<b>6</b>	CTCF <sub>591-727</sub>	591-727	pLIB	N-HA, 3C C-StrepII, TEV

**Table 2.2 List of CTCF Constructs**

#	Construct	Length (aa)	Vector	Tag and Cleavage Site
<b>7</b>	FL ZNF143	1-638	pOPINJ	N-His <sub>6</sub> -GST, 3C
<b>8</b>	FL ZNF143	1-638	pLIB	C-StrepII, TEV
<b>9</b>	ZNF143 <sub>1-466</sub>	1-466	pOPINJ	N-His <sub>6</sub> -GST, 3C
<b>10</b>	ZNF143 <sub>103-466</sub>	103-466	pLIB	C-StrepII, TEV
<b>11</b>	ZNF143 <sub>NTD</sub>	1-232	pGEX-6P-1	N-His <sub>6</sub> -GST, 3C C-StrepII, TEV
<b>12</b>	ZNF143 <sub>ZNF</sub>	233-444	pOPINJ	N-His <sub>6</sub> -GST, 3C C-StrepII, TEV
<b>13</b>	ZNF143 <sub>CTD</sub>	445-638	pOPINJ	N-His <sub>6</sub> -GST, 3C C-StrepII, TEV
<b>14</b>	ZNF143 <sub>229-353</sub>	229-353	pOPINJ	N-His <sub>6</sub> -GST, 3C
<b>15</b>	ZNF143 <sub>296-418</sub>	296-418	pOPINJ	N-His <sub>6</sub> -GST, 3C
<b>16</b>	ZNF143 <sub>326-448</sub>	326-448	pOPINJ	N-His <sub>6</sub> -GST, 3C

**Table 2.3 List of ZNF143 Constructs**

#	Construct	Length (aa)	Vector	Tag and Cleavage Site
<b>17</b>	FL CHD8	1-2581	pLIB	N-StrepII, TEV
<b>18</b>	FL CHD8	1-2581	pLIB	N-StrepII, TEV C-3xFLAG, 3C
<b>19</b>	CHD8 <sub>K842R</sub>	1-2581, K842R	pLIB	N-StrepII, TEV
<b>20</b>	CHD8 <sub>1-800</sub>	1-800	pLIB	N-StrepII, TEV C-3xFLAG, 3C
<b>21</b>	CHD8 <sub>801-1450</sub>	801-1450	pLIB	N-StrepII, TEV C-3xFLAG, 3C
<b>22</b>	CHD8 <sub>1451-2050</sub>	1451-2050	pLIB	N-StrepII, TEV C-3xFLAG, 3C
<b>23</b>	CHD8 <sub>2051-2581</sub>	2051-2581	pLIB	N-StrepII, TEV C-3xFLAG, 3C
<b>24</b>	CHD8 <sub>BRK</sub>	2280-2450	pOPINJ	N-His <sub>6</sub> -GST, 3C C-StrepII, TEV

**Table 2.4 List of CHD8 Constructs**

#	Construct	Length (aa)	Vector	Tag and Cleavage Site
<b>25</b>	FL CHD4	1-1912	pLIB	C-HA, TEV

**Table 2.5 List of Other Protein Constructs**

Tag / Protease Cleavage	DNA Sequence	Protein Sequence
<b>TEV protease</b>	GAAAACCTGTACTTCCAGTC ATG	ENLYFQS
<b>3C protease</b>	CTGGAAGTGCTGTTCAAGG GCCCG	LEVLFQGP
<b>3x FLAG tag</b>	GATTATAAAGATCATGATGG CGATTATAAAGATCACGATA TTGATTATAAAGATGATGAC GATAAA	DYKDHDG DYKDHDIDYKDD DDK
<b>Twin-Strep (StrepII) tag</b>	TGGAGCCATCCGCAATTG AAAAAGGTGGCGGCTCCGG CGGAGGTAGCGGCAGG TTCTGGTCTCACCCCTCAGT TCGAGAAG	WSHPQFEKGGGSGGSG GGSWSHPQFEK

<b>3x HA tag</b>	CCGTATGATGTTCTGACTA TGCAGGGCTACCCCTATGAT GTCCCAGATTACGCCGGGT CCTATCCATACGATGTCCCC GATTACGCCAGCC	PYDVPDYAGYPYDVPDYAG SYPYDVPDYAPA
<b>ybbR tag</b>	GGAGGCGATTCTCTTGAATT TATTGCTAGTAAACTTGCG	GGDSLEFIASKLA

**Table 2.6 Sequences of Protein Tags and Cleavage Sites Used**

### 2.1.3 Polymerase Chain Reaction (PCR)

Polymerase chain reaction was used to amplify the target DNA through thermal cycling. Forward and reverse primers were specifically designed for each PCR reaction. The primers were designed to anneal at around 65 °C with the gene of interest and to include an overhang for annealing between joint fragments. For each PCR reaction, 2.5 µl of 10 µM forward and reverse primers, 1 µl of 100 ng/µl template DNA, 1X Phusion GC buffer (NEB), 200 µM deoxyadenosine triphosphate (dNTP) and 1 µl Phusion High-Fidelity DNA Polymerase (NEB) were used. Nuclease free water was added to top up the final reaction volume to 50 µl.

There were four steps to the polymerase chain reaction. The first step required the template DNA to denature through heating the reaction mixture to 98 °C to break the hydrogen bonds between complementary bases. The temperature was lowered to around 60 °C to allow primers to anneal to the single-stranded DNA template. Next, the temperature was raised to 72 °C to allow the Phusion Polymerase to synthesize the complementary daughter strand. Phusion Polymerase moves along in a 3' – 5' direction and synthesises complementary DNA in 5' – 3' direction. The time required for this elongation process depended on the size of the template DNA. The denaturation, annealing and elongation steps were repeated for 30 cycles. A final elongation step at 72 °C was included before the reaction went into a final hold at 4 °C.

#### 2.1.4 Cloning using the Gibson Assembly cloning system

Gibson assembly® cloning system allows two to six fragments to assemble with one single isothermal step. The reaction was set up using the Gibson Assembly® kit from NEB. 10 µl of 2x Gibson Assembly Master Mix was mixed with 100 ng of vector and 3-fold molar excess of each insert, and subsequently topped up to 20 µl final volume. The reaction was incubated at 50 °C for 15 minutes when 2 or 3 fragments were being assembled. T5 exonuclease in the Gibson assembly master mix removes nucleotides from 5'-ends of the double stranded DNA molecules to allow complementary single stranded DNA overhangs to anneal. Phusion DNA polymerase filled the gaps in the annealed DNA fragments and Taq DNA ligase covalently joined the fragments and sealed the nicks to create a continuous DNA fragment.

#### 2.1.5 Transformation into Stellar Competent Cells

2 µl of DNA construct assembled with Gibson Assembly ® kit was incubated with 20 µl of Stellar™ competent cells (Takara Bio) on ice for 30 minutes. The reaction mix was heat-shocked at 42 °C for 45 seconds and back on ice for 5 minutes. 500 µl of SOC outgrowth medium (NEB) were added to the cells and allowed to recover in a shaking incubator at 37 °C for 1 hour. 250 µl of the transformation was plated onto a 10 cm LB agar plate containing the appropriate antibiotic as a selection marker at 37 °C overnight. Single colonies were picked and analysed by colony PCR. Colony PCR is a high throughput method to identify the presence of the DNA inserts. The DNA inserts are regions where the two primers annealed and selectively amplified, hence only successful colonies would show the correct expected length, allowing for rapid screening of all transformants. The single colonies were touched with a pipette tip and resuspended in 10 µl of nuclease-free water. For each colony PCR reaction, 1 µl of colony template, 10 µl OneTaq Quick Load 2X Master Mix (NEB), 0.5 µl of 10 µM forward and reverse primer were used, and 8 µl of nuclease-free water was added to top up to a final reaction volume of 20 µl. The PCR machine protocol used was mentioned in section 2.1.3. The amplified DNA was analysed using agarose gel electrophoresis (section 2.1.6). Colonies with the correct DNA inserts were grown in 5 ml LB media with the

appropriate antibiotic overnight. They were subsequently purified using GeneJET Plasmid Miniprep Kit (Thermo Fisher Scientific) and sent to Source Bioscience for DNA sequencing.

### 2.1.6 DNA Purification from Agarose Gel

Amplified DNA was purified using agarose gel to remove excess primers, polymerases, nucleotides and/or restriction enzymes. The DNA samples were mixed with 4X DNA gel loading dye (NEB) and loaded onto a 1% agarose 1X TAE (40mM Tris base, 20 mM acetic acid, 1 mM EDTA pH 8.0) gel. Appropriate amount of ethidium bromide was added to the gel to intercalate into DNA to allow its visualisation under UV light. DNA ladder (100 bp / 1 kb DNA ladder, NEB) was loaded to identify the DNA bands of interest. The gel was run at 100 V for 30 minutes. Bands with the correct DNA sizes were cut and extracted from the agarose gel with the use of Zymoclean Gel DNA recovery kit (Cambridge Bioscience).

### 2.1.7 QuikChange Mutagenesis

QuikChange Site-Directed Mutagenesis Kit (Agilent) was used to introduce single amino acid mutation in the CHD8 construct at residue 842 (lysine to arginine). Two primers with the desired mutation flanked by unmodified nucleotides were designed. The reaction was prepared using 5 µl of 10X reaction buffer, 50 ng of dsDNA template, 2.5 µl of 100 µM of forward primer, 2.5 µl of 100 µM reverse primer, 1 µl of dNTP mix, 1 µl of *PfuTurbo* DNA polymerase and nuclease-free water to a final volume of 50 µl. The PCR reaction (denaturation: 98 °C for 30 seconds, annealing: 55 °C for 1 minute and elongation: 68 °C for 10 minutes) was cycled sixteen times prior to transformation into Stellar competent cells.

	<b>Sequence</b>
<b>Forward Primer</b>	GGGTCTGGGCC <b>CGT</b> ACCATCCAAAG
<b>Reverse Primer</b>	ATTCGTCCGCCAGAATG

Table 2.7 Primers Used for CHD8 Mutagenesis

## 2.2 Protein Expression

### 2.2.1 Protein expression in *E. coli*

Most of the ZNF143 constructs were expressed in *E. coli*. The recombinant plasmid was transformed into competent *E. coli* BL21-Codon Plus(DE3)-RIL cells (Stratagene). Colonies were picked from the LB Agar plate and grown overnight in 100 ml LB (lysogeny broth: 1% w/v tryptone, 0.5% w/v yeast extract, 170 mM NaCl) media with 100 µg/ml ampicillin. 15 ml of the overnight culture were added to 1 L of LB media with ampicillin. 6 litres of LB media were used for each expression and the cells were grown in a shaking incubator at 200 rpm at 37 °C until the OD<sub>600</sub> reached 0.8. 0.5 mM IPTG and 50 µM ZnCl<sub>2</sub> were added to each 1 L flask to induce expression overnight at 18 °C. The cells were then harvested via centrifugation at 4,000 g for 15 minutes and the pellets were stored at -80 °C until further use.

### 2.2.2 Protein Expression in Insect Cells

#### 2.2.2.1 Bacmid Generation Using DH10EMBacY Cells

Proteins expressed in insect cells were post-translationally modified. Bacmids for protein expression in insect cells were generated from pLIB DNA construct. The constructs were transformed into DH10 EMBacY cells. The EMBacY backbone contained a constitutively expressing YFP expression cassette to monitor the viral titres via fluorescence. These cells were kanamycin resistant and have the Tn7 attachment site within LacZ reporter, and a helper plasmid encoding for the transposase with tetracycline resistance. Through transposition, the gene of interest was inserted into the Tn7 attachment site within the LacZ alpha peptide sequence that encodes the protein β-galactosidase. This insertion disrupts the expression of LacZ peptide hence functional β-galactosidase could not be formed, producing white colonies on LB-Agar plates. Active β-galactosidase is able to hydrolyse X-Gal and yields galactose and 5-bromo-4-chloro-3-hydroxyindole. The latter dimerises and oxidizes to form 5,5'-dibromo-4,4'-dichloro-indigo, therefore forming blue colonies if the transposition failed.

2 µl of pLIB DNA construct was transformed into 75 µl DH10 EMBacY cells. The DNA-cell mixture was incubated on ice for 30 minutes, heat-shocked at 42 °C for 45 seconds and recovered on ice for 5 minutes. 500 µl of SOC media was added to the cells and were allowed to recover for 6 hours at 37 °C in a shaking incubator. 50 µl of the DNA-cell mixture was plated on a Bacmid Transposition plate containing X-Gal, IPTG, gentamycin, kanamycin and tetracycline. White colonies were picked and grown overnight in 5 ml of LB medium containing the same antibiotics. The cells were pelleted and resuspended in 300 µl resuspension buffer (GeneJet Plasmid Miniprep Kit, Thermo Fisher Scientific). 100 µl of lysis buffer was added to lyse the cells followed by the addition of 300 µl sodium acetate at pH 5.5 to neutralise the reaction. The samples were incubated on ice for 10 minutes and centrifuged at 17,000 g for another 10 minutes. Next, the supernatant was incubated with 800 µl of isopropanol on ice and centrifuged again to remove the supernatant. 500 µl of 70% ethanol was added to wash the pellet. After another round of centrifugation, the supernatant was removed to allow the pellet to air-dry at room temperature. The dried pellet was resuspended in 40 µl of ultra-filtered water and stored at -20 °C ready for transfection into Sf9 insect cells.

#### 2.2.2.2 Transfection into Sf9 Cells

Lonza Insect Xpress media (Lonza Bioscience) was used for insect cell growth and expression. 10 µl of Bacmid was mixed with 200 µl of growth media and 10 µl of Cellfectin™ II Reagent (Thermo Fisher Scientific). 100 µl of the mixture was added in a dropwise fashion into 2 ml of Sf9 cells at a density of  $5 \times 10^5$  cells/ml adhered to the bottom of a six-well plate. The plate was incubated at 27 °C for three days and the transfection was monitored using a light microscope. The supernatant was harvested and stored with 5% of fetal bovine serum (FBS) (Gibco) at 4 °C. This was the P1 virus.

To amplify the viral stock, 2 ml of P1 virus was added to 50 ml of Sf9 cells (density:  $5 \times 10^5$  cells/ml) and incubated in a shaking incubator at 130 rpm at 27 °C. The viability and fluorescence of the cells were monitored over 3 to 5 days and harvested when over 85% of Sf9 cells showed fluorescence. The

cells were centrifuged and the supernatant containing the virus (P2 virus) was collected and stored with 5% of FBS at 4 °C.

#### 2.2.2.3 Protein Expression

3 ml of P2 virus was added into 500 ml of High5 cells (density:  $5 \times 10^5$  cells/ml) and incubated at a shaking incubator at 130 rpm at 27 °C. The viability and fluorescence were monitored for 3 days. The cells were harvested once the viability dropped to below 85%. They were harvested through centrifugation at 1,000 g for 15 minutes and the pellet was stored at -80 °C until purification.

### 2.3 Protein Purification

#### 2.3.1 Overview

In general, the purification protocol of human recombinant proteins followed a three-step protocol: an affinity purification to trap the specific affinity tagged protein, an ion exchange chromatography that removed contaminating proteins or DNA based on their charges and lastly size exclusion chromatography to remove contaminants and aggregates based on the hydrodynamic radius ( $R_h$ ) of the proteins. All columns were controlled by Akta Pure System (GE Healthcare). The purity of the expected proteins was confirmed by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) after each step. The purification steps shown in this section are the optimised protocols. The purification optimisation trials are mentioned in the results section.

#### 2.3.2 Purification Protocol of Human CTCF

##### 2.3.2.1 Purification Protocol of HA-CTCF-StrepII

Frozen cell pellets (corresponding to a 500 ml culture of High5 cells) were thawed on ice and resuspended in 200 ml of *buffer A* (25 mM HEPES pH8.0, 300 mM NaCl, 10% (v/v) glycerol, 1 mM MgCl<sub>2</sub>, 1 mM DTT) with 2 EDTA-free protease inhibitor tablets. The cells were lysed by sonication on ice (3 seconds on, 8 seconds off, 3 minutes at 60% amplitude), and incubated with 25 µl benzonase for an hour on a roller. Next, the lysate was supplemented with 5

M NaCl to increase the total concentration of NaCl in the lysate to 600 mM to remove DNA contaminants. The lysate was then centrifuged for 40 minutes at 37,000 g. The supernatant was filtered and applied to StrepTrap HP column pre-equilibrated with *buffer A*. The bound protein was washed with 5 CV of *buffer A* and eluted with 5 CV of *buffer A* supplemented with 5 mM D-dethiobiotin (Sigma Aldrich). The eluted  $A_{280}$  peak fractions were collected and diluted to 150 mM NaCl with *buffer B* (25 mM HEPES pH8.0, 10% (v/v) glycerol, 1 mM DTT) before loading it onto HiTrap Heparin HP column pre-equilibrated with 7.5% *buffer C* (25 mM HEPES pH8.0, 2 M NaCl, 10% (v/v) glycerol, 1 mM DTT). After sample application, the column was washed with 5 CV of 10% *buffer C* and the protein was eluted in a gradual gradient (12% - 100%) against *buffer C* over a course of 20 CV. Fractions were pooled and concentrated to less than 5 ml total volume using a 50 kDa molecular weight cut-off (MWCO) VivaSpin concentrator (Generon). The 5 ml concentrated protein was passed over a HiLoad 16/600 Superose 6 prep grade gel filtration column (GE Healthcare) pre-equilibrated in 1.5 CV of ultra-filtered water and 1.5 CV of *buffer D* (25 mM HEPES pH8.0, 150 mM NaCl, 10% (v/v) glycerol, 1 mM DTT). Peak fractions were analysed by SDS-PAGE. Fractions were pooled and concentrated, flash-frozen in liquid nitrogen and stored at -80 °C.

### 2.3.2.2 Purification Protocol of ybbR-tagged CTCF

The ybbR-tag was incorporated into the C-terminus of CTCF to allow it to conjugate with coenzymeA (CoA)-fluorophore to create a fluorescently labelled CTCF protein (Yin *et al.*, 2005, 2006). CTCF<sub>ybbR</sub> was purified using StrepTrap and Heparin columns following the protocol described in section 2.3.2.1. Post heparin, CTCF<sub>ybbR</sub> was concentrated to around 300 µl at around 5 µM concentration and kept on ice whilst the CoA-conjugated fluorophore was prepared.

The fluorophore Alexa488 (IDT Technologies) was resuspended in DMSO and CoA (purified by Dr. Erin Cutts) was resuspended in sodium phosphate pH 7.0. 125 µM of CoA was incubated with five times excess of Alexa488 fluorophore at room temperature for one hour in a final reaction buffer of 100 µl. The

reaction was quenched with 5 µl of 1 M DTT. Following the conjugation of CoA and fluorophore Alexa488, CTCF<sub>ybbR</sub> was incubated with CoA-Alexa488 at a molar ratio of 1 : 4, and supplemented with 0.1 µM of Sfp phosphopantetheinyl transferase (purified by Dr. Erin Cutts) at room temperature for 30 minutes. The reaction volume was topped up to a final volume of 500 µl with 10 mM MgCl<sub>2</sub> and 50 mM HEPES pH 7.5. The reaction sample was filtered before loading onto Superose 6 Increase 10/300 GL column (GE Healthcare) pre-equilibrated with 25 mM HEPES pH 8.0, 200 mM NaCl, 10% (v/v) glycerol, 2 mM MgCl<sub>2</sub> and 1 mM DTT. The eluted fractions were analysed by SDS-PAGE and the fractions were pooled and concentrated, flash-frozen in liquid nitrogen and stored in small aliquots at -80 °C.

### 2.3.3 Purification Protocol of Human ZNF143 constructs

#### 2.3.3.1 Purification of GST-Tagged ZNF143 Constructs

This is the general purification protocol for most GST-tagged ZNF143 constructs unless otherwise specified in the results section. Furthermore, all steps were performed at 4 °C unless otherwise specified. Pellets (6 L of *E. coli* cells) were thawed on ice and resuspended in 200 ml of *buffer A* (50 mM HEPES pH8.0, 625 mM NaCl, 10 µM ZnCl<sub>2</sub>, 10% (v/v) glycerol, 1 mM MgCl<sub>2</sub>, 200 µM EDTA, 2 mM DTT) with 2 Pierce™ EDTA-free protease inhibitor tablets (Thermo Fisher Scientific) and 15 µl benzonase (Sigma Aldrich). The cells were lysed by sonication on ice (3 seconds on, 8 seconds off, 3 minutes at 60% amplitude), and centrifuged at 37,500 g for 40 minutes. The lysate was filtered and applied to GST HiTrap column (GE Healthcare) pre-equilibrated with *buffer A* and washed with *buffer A* before elution with 5 CV of *buffer A* supplemented with 25 mM reduced glutathione (Sigma Aldrich). The eluted fractions were treated with 3C protease overnight (protease : protein ratio = 1:50 (w/w)) to remove the His<sub>6</sub>-GST tag at the N-terminus. The cleaved protein was diluted by 3-fold with *buffer B* (50 mM HEPES pH8.0, 10% (v/v) glycerol, 2 mM DTT) to reduce the concentration of NaCl in the buffer. The diluted sample was loaded onto HiTrap Heparin HP column (GE Healthcare) pre-equilibrated with 20% *buffer C* (50 mM HEPES pH8.0, 1 M NaCl, 10% (v/v) glycerol, 2 mM DTT). The column was washed with 5 CV of 20% *buffer C* and

the protein was eluted in a gradual gradient (20% -100%) of *buffer C* over a course of 15 CV. Fractions were pooled and concentrated to less than 5 ml total volume using a 30 kDa molecular weight cut-off (MWCO) VivaSpin concentrator (Generon) and loaded onto a HiLoad 16/600 Superdex 200pg gel filtration column (GE Healthcare) pre-equilibrated in 1.5 CV of *buffer D* (50 mM HEPES pH8.0, 200 mM NaCl, 10% (v/v) glycerol, 2 mM DTT). Peak fractions, as determined by absorbance at 280 nm ( $A_{280}$ ) were analysed with SDS-PAGE and relevant fractions were pooled and concentrated, flash-frozen in liquid nitrogen and stored in aliquots at -80 °C.

#### 2.3.3.2 Purification of StrepII-tagged ZNF143 constructs

For Strep-tagged ZNF143 constructs, the pellet from 500 ml of High5 cells was thawed, resuspended and lysed using the same protocol as mentioned in section 2.3.3.1. The lysate was applied to StrepTrap HP column (GE Healthcare) pre-equilibrated with *buffer A* (50 mM HEPES pH8.0, 625 mM NaCl, 10% (v/v) glycerol, 1 mM MgCl<sub>2</sub>, 2 mM DTT). The bound protein was washed with *buffer A* and eluted with 5 CV of *buffer A* supplemented with 5 mM D-desthiobiotin (Sigma Aldrich). The eluted peak fractions were collected and diluted to 200 mM NaCl with *buffer B* (50 mM HEPES pH8.0, 10% (v/v) glycerol, 1 mM DTT) before loading it onto HiTrap Heparin column pre-equilibrated with 20% *buffer C* (50 mM HEPES pH8.0, 1 M NaCl, 10% (v/v) glycerol, 1 mM DTT). The column was washed with 5 CV of 20% *buffer C* and the protein was eluted in a gradual gradient (20% - 100%) against *buffer C* over a course of 15 CV. The peak fractions were pooled, concentrated to 5 ml and loaded onto HiLoad 16/600 Superdex 200pg gel filtration column in *buffer D* (25 mM HEPES pH8.0, 200 mM NaCl, 10% (v/v) glycerol, 1 mM DTT). StrepII-tagged ZNF143 was concentrated again, flash-frozen and stored at -80 °C.

#### 2.3.4 Purification Protocol of Human CHD8

Frozen cell pellets (corresponding to a 500 ml culture of High5 cells) were thawed on ice and resuspended in 200 ml of *buffer A* (25 mM HEPES pH8.0, 500 mM NaCl, 10% (v/v) glycerol, 1 mM MgCl<sub>2</sub>, 1 mM DTT) with two EDTA-

free protease inhibitor tablets and 15 µl benzonase. The cells were lysed by sonication on ice (3 seconds on, 8 seconds off, 3 minutes at 60% amplitude), and centrifuged at 37,500 g for 40 minutes. The lysate was applied to StrepTrap HP column pre-equilibrated with *buffer A* and eluted with 5 CV of *buffer A* supplemented with 5 mM D-desthiobiotin. The eluted A<sub>280</sub> peak fractions were collected and diluted to 150 mM NaCl with *buffer B* (25 mM HEPES pH8.0, 10% (v/v) glycerol, 1 mM DTT) before loading it onto HiTrap Q HP column (GE Healthcare) pre-equilibrated with 15% *buffer C* (25 mM HEPES pH8.0, 1 M NaCl, 10% (v/v) glycerol, 1 mM DTT). After sample application, the column was washed with 5 CV of 15% *buffer C* and the protein was eluted in a gradual gradient (20% - 100%) against *buffer C* over a course of 20 CV. Fractions were pooled and concentrated to less than 5 ml total volume using a 100 kDa molecular weight cut-off (MWCO) VivaSpin concentrator (Generon). The 5 ml concentrated protein was passed over a HiLoad 16/600 Superose 6 prep grade gel filtration column pre-equilibrated in 1.5 CV of *buffer D* (25 mM HEPES pH8.0, 150 mM NaCl, 5% (v/v) glycerol, 1 mM DTT). Peak fractions were analysed by SDS-PAGE. Fractions were pooled and concentrated, flash-frozen in liquid nitrogen and stored in small aliquots at -80 °C.

### 2.3.5 Nucleosome Reconstitution

#### 2.3.5.1 Overview

Nucleosomes used in the experiments described in this thesis were prepared with the help of Dr. Thangavelu Kaliyappan. The expression and purification protocol was performed as described previously in Luger, Rechsteiner and Richmond, 1999, whereas the nucleosome reconstitution followed the protocol published by Dyer *et al.*, 2003. Recombinant human histones H2A, H3, H4 and *Xenopus laevis* H2B were used for histone octamer assembly. These four histone proteins were expressed and purified individually, followed by the assembly of the histone octamer and DNA to reconstitute the nucleosome.

### 2.3.5.2 Protein Expression of Histones

All histone constructs were cloned into pET vector and transformed into competent *E. coli* BL21-Codon Plus(DE3) pLysS (Agilent). Colonies were picked from the LB Agar plate and grown overnight in 100 ml LB media supplemented with 100 µg/ml ampicillin and 25 µg/ml chloramphenicol. 15 ml of the overnight culture was added to 1 L of 2x TY media (16% (w/v) tryptone, 10% (w/v) yeast extract and 5% (w/v) NaCl) with ampicillin and chloramphenicol. 6 litres of 2x TY media were used for each expression and the cells were grown in a shaking incubator at 200 rpm at 37 °C until the OD<sub>600</sub> reached 0.6. At which point, 0.4 mM IPTG was added to each 1 L flask to induce expression. The cells were incubated for another 3 hours at 37 °C and harvested via centrifugation at 4,000 g for 15 minutes. Each litre of pellet was resuspended in 5 ml *lysis buffer A* (50 mM Tris pH 7.5, 100 mM NaCl and 1 mM β-mercaptoethanol) and stored at -80 °C until further use.

### 2.3.5.3 Purification Protocol of Histones

The purification protocols for all histone proteins were similar. The histones were purified in collaboration with Dr. Thangavelu Kaliyappan from inclusion bodies using cation and anion ion-exchange chromatography, then subsequently lyophilised for long term storage at -20 °C.

The histone pellet was thawed, supplemented with protease inhibitors and treated with homogeniser to reduce viscosity. The lysate was centrifuged at 20,000 g for 20 minutes at 4 °C. The supernatant was discarded, and the pellet was washed with *triton buffer B* (*lysis buffer A* supplemented with 1% (v/v) Triton X-100) three times, followed by three washes with *lysis buffer A*, each with 20 ml of buffer and centrifuged at 20,000 g for 20 minutes at 4 °C. Supernatant was discarded, all pellets of the same histone protein were pooled into the same tube and 1 ml of DMSO was added for a 30 minutes incubation at room temperature.

Next, the pellet that was incubated with DMSO was mixed with 10 ml of *unfolding buffer C* (6M guanidine-HCl, 20 mM sodium acetate pH 5.2 and 5

mM DTT) and incubated for 1 hour and followed by centrifugation at 23,000 g for 20 minutes at room temperature. The resulting supernatant was filtered and loaded onto HiPrep Sephacryl S200 size exclusion column (GE Healthcare) pre-equilibrated with *urea buffer D* (7M urea, 20 mM Tris pH 7.5, 200 mM NaCl, 2 mM β-mercaptoethanol and 1 mM EDTA). The fractions were analysed by SDS-PAGE and fractions with the target histone protein were pooled and loaded onto HiPrep SP FF (GE Healthcare) and HiPrep Q HP (GE Healthcare) connected in tandem. Histone was eluted over a gradient (15% - 100%) of *urea buffer E* (*urea buffer D* supplemented with 1 M NaCl). Fractions were pooled and dialysed against 5 L of *dialysis buffer E* (1 mM DTT) for 16 hours at 4 °C. Following dialysis, the protein was plunge frozen with liquid nitrogen and lyophilised overnight. The lyophilised histone proteins were stored at -20 °C.

The histone octamer was assembled by Dr. Thangavelu Kaliyappan. It was prepared by mixing individual lyophilised histones at 1.2 fold excess of H2A and H2B to H3 and H4. The mixture was resuspended in the *lysis buffer A* for 45 minutes at room temperature and dialysed against *refolding buffer F* (10 mM Tris pH 7.5, 2 M NaCl, 1 mM EDTA and 1 mM DTT) for 24 hours at 4 °C with three buffer exchanges. The sample was concentrated to 1 ml and subjected to size-exclusion chromatography on a HiLoad 16/600 Superdex 200 pg gel filtration column (GE Healthcare) pre-equilibrated in *refolding buffer F*. The fractions were analysed with SDS-PAGE and the corresponding fractions were pooled, concentrated and stored with 50 % (v/v) glycerol for storage at -20 °C

#### 2.3.5.4 Purification Protocol of DNA

Widom 601 DNA sequence was used as a template for *in vitro* nucleosome reconstitution. Nucleosomal DNA were generated using large-scale PCR with 0.12 µM of in-house purified *Pfu* polymerase, forward and reverse primers, reaction buffer, dNTP and nuclease-free water topped up to 10 ml. Four 96-well plates were set up for this reaction following the PCR protocol mentioned in section 2.1.3. The obtained PCR products were pooled and precipitated with

25 ml ice-cold ethanol and centrifuged at 20,000 g for 20 minutes at 4 °C. The pellets were resuspended in *buffer A* (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0) and loaded onto HiTrap Q HP 1ml ion-exchange column (GE Healthcare) and eluted with a salt gradient from 10% - 100% of *buffer B* (10 mM Tris-HCl pH 8.0, 2 M NaCl, 1 mM EDTA pH 8.0). Fractions corresponding to the DNA fragments were assessed by 4-12% Native-PAGE, pooled, followed by ethanol precipitation. Lastly, the pellet was re-dissolved in *buffer A*, and stored at -20 °C prior to use.

#### 2.3.5.5 Nucleosome Reconstitution

Nucleosomes were reconstituted with the help of Dr. Thangavelu Kaliyappan. The nucleosomes were reconstituted by salt dialysis method as described previously (Dyer *et al.*, 2003). The purified DNA and histone octamer were mixed at 1 : 1.2 ratio of DNA : octamer in a *salt buffer A* (10 mM Tris-HCl pH 7.5, 2 M NaCl, 1 mM EDTA pH 8.0 and 1 mM DTT) and transferred to Slide-A-Lyzer dialysis unit 10,000 MWCO (Thermo Fisher Scientific). The sample was dialysed gradually against *salt buffer B* (10 mM Tris-HCl pH 7.5, 200 mM NaCl, 1 mM EDTA pH 8.0, 1 mM DTT) over 24 hours at 4 °C. The nucleosomes were further dialyzed in a *salt buffer C* (20 mM Tris-HCl pH 7.5, 50 mM NaCl, 1 mM DTT) overnight at 4 °C, concentrated and stored at 4 °C.

### 2.4 Protein Analysis and Quantification

#### 2.4.1 Theoretical Molecular Weights and Extinction Coefficients

The protein sequences of all proteins were submitted to ExPASy ProtParam (Gasteiger *et al.*, 2003) to estimate their molecular weights (kDa) and extinction coefficients ( $M^{-1} cm^{-1}$ ) for protein quantification.

	<b>Protein Construct</b>	<b>NTD Tag</b>	<b>CTD Tag</b>	<b>MW (kDa)</b>	<b>Ext. Co (<math>M^{-1} cm^{-1}</math>)</b>	<b>pI</b>
<b>1</b>	FL CTCF	HA	StrepII	91.3	58355	6.11
<b>2</b>	FL CTCF	HA	ybbR-StrepII	92.6	58355	6.11

<b>3</b>	CTCF <sub>61-727</sub>	His <sub>8</sub>	StrepII	82.5	44945	6.93
<b>4</b>	CTCF <sub>1-240</sub>	HA	StrepII	35.2	37485	4.22
<b>5</b>	CTCF <sub>241-590</sub>	HA	StrepII	49.7	45280	9.27
<b>6</b>	CTCF <sub>591-727</sub>	HA	StrepII	23.9	24410	4.47
<b>7</b>	FL ZNF143			70.1	30675	5.56
<b>8</b>	FL ZNF143		StrepII	73.0	41675	5.62
<b>9</b>	ZNF143 <sub>1-466</sub>			51.0	27695	6.12
<b>10</b>	ZNF143 <sub>103-466</sub>		StrepII	44.4	37205	6.54
<b>11</b>	ZNF143 <sub>NTD</sub>			25.5	8940	4.38
<b>12</b>	ZNF143 <sub>ZNF</sub>			25.2	20245	9.36
<b>13</b>	ZNF143 <sub>CTD</sub>			21.7	2980	4.18
<b>14</b>	ZNF143 <sub>229-353</sub>			14.3	7950	9.30
<b>15</b>	ZNF143 <sub>296-418</sub>			14.4	12420	9.37
<b>16</b>	ZNF143 <sub>326-448</sub>			14.3	12420	9.17
<b>17</b>	FL CHD8	StrepII		294.5	245970	6.04
<b>18</b>	FL CHD8	StrepII	3xFLAG	298.5	250440	5.93
<b>19</b>	CHD8 <sub>K842R</sub>	StrepII		294.5	245970	6.04
<b>20</b>	CHD8 <sub>1-800</sub>	StrepII	3xFLAG	90.9	41495	5.93
<b>21</b>	CHD8 <sub>801-1450</sub>	StrepII	3xFLAG	79.7	79165	5.87
<b>22</b>	CHD8 <sub>1451-2050</sub>	StrepII	3xFLAG	73.9	99530	6.10
<b>23</b>	CHD8 <sub>2051-2581</sub>	StrepII	3xFLAG	62.8	34170	5.07
<b>24</b>	CHD8 <sub>BRK</sub>		StrepII	23.7	25105	9.26
<b>25</b>	FL OCT1			76.5	12615	6.34
<b>26</b>	OCT1 <sub>POU</sub>			18.4	12615	9.12
<b>27</b>	FL CHD4		HA	223.5	242045	5.47

**Table 2.8 Molecular Weights, Isoelectric Point and Extinction Coefficients of All Constructs**

#### 2.4.2 Protein Concentration Quantification

Absorbance at 280 nm was used to estimate protein concentration using Nanodrop ND-1000 spectrophotometer (Thermo Fisher Scientific) during and after protein purification. The absorption at 260 nm was also measured to analyse the amount of DNA contamination. A260/A280 ratio lower than 0.57 corresponded to no DNA contamination whereas higher ratios indicated DNA contamination. This is because amino acids tyrosine and tryptophan

specifically absorbs at 280 nm, whereas the maximum absorbance of DNA occurs at 260 nm due to its purine and pyrimidine bases. The protein concentration was calculated using the extinction coefficients (table 2.8) and applying the values into the Beer-Lambert law (Equation 1). According to the Beer-Lambert law, A = absorbance,  $\epsilon$  = extinction coefficient, c = concentration and  $x$  = path length.

$$c = \frac{A}{\epsilon * x} \quad \text{Eq.1}$$

#### 2.4.3 SDS-PAGE Analysis

Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) analysis is an electrophoretic separation technique. The linearised protein chains are negatively charged in the SDS loading buffer. The proteins were separated by mass as they migrate from the cathode to the anode. Protein samples were mixed with SDS gel loading buffer (400 mM DTT, 8% SDS, 40% glycerol, 0.4% bromophenol blue, 200 mM Tris-HCl pH 6.8), boiled at 100 °C for 5 minutes and loaded onto SDS-PAGE gels. The samples were loaded in commercial 4-12% NuPAGE SDS-PAGE (Life Technologies) and ran in 1 x NuPAGE MES buffer (Invitrogen). 4  $\mu$ l of Color Prestained Protein Standard (P7712/P7719, NEB) was also loaded to identify molecular weights of the samples. SDS-PAGE gels were ran at 200 V for 40 minutes and stained with Coomassie Instant Blue (Abcam).

#### 2.4.4 Western Blot Analysis

The SDS-PAGE gel was transferred to a nitrocellulose membrane (Amersham Protran Premium Blotting membrane, nitrocellulose 0.45, VWR) using XCell II BlotModule (Invitrogen) on ice for 1 hour at 100 V in 1X Transfer Buffer (25 mM Tris Base, 10% (v/v) methanol, 192 mM glycine). The membrane was first blocked in 5% (w/v) milk (Sigma) in TBS-T (TBS with 0.05% Tween20) at room temperature for 1 hour. The membrane was washed with TBS-T three times and incubated with the primary antibody overnight at 4 °C. The primary antibody was diluted in a 1:1000 dilution in 10 ml of 5% milk in TBS-T. The membrane was washed with TBS-T and secondary antibody (1:1000)

conjugated to a fluorescence probe diluted in 5 ml of milk TBS-T was applied for 1 hour at room temperature. The membrane was washed with TBS-T three times. The bands of interest were visualised using Odyssey® CLx Imaging system (Li-COR).

	<b>Antibody</b>	<b>Supplier</b>
<b>Primary</b>	Rabbit Anti-CTCF #ab70303	Abcam
	Mouse Anti-His #mAB050-100	R&D Systems
	Mouse Anti-ZNF143 #H00007702	Novus Biologicals
	Rabbit Anti-CHD8 #ab84527	Abcam
<b>Secondary</b>	Anti-Rabbit IgG DyLight 680 4X PEG conjugate	Cell Signalling Technology
	Anti-Mouse IgG DyLight 800 4X PEG conjugate	Cell Signalling Technology

**Table 2.9 Antibodies Used for Western Blots**

## 2.4.5 Silver Stain Analysis

Silver stain analysis was performed in the event of low protein concentrations that could not be observed with Coomassie Blue. Silver Stain Kit (Pierce Thermo Fisher Scientific) was used following the manufacturer's protocol. In brief, the gel was fixed with fixation solution (30% ethanol and 10% acetic acid), followed by water washes, and stained with silver stain solution. The reaction was stopped with 5% acetic acid. The Silver Stain Kit was able to detect proteins as low as of 0.25 ng.

## 2.5 Biochemical and Functional Analysis

### 2.5.1 Limited Proteolysis

#### 2.5.1.1 Trypsin Limited Proteolysis

Limited proteolysis was carried out to explore a more stable domain of ZNF143 for protein crystallisation. A stock of 1 mg/ml solution of trypsin was serially diluted five times by a factor of three in each instance, generating six decreasing concentrations of trypsin for analysis. Each of the six

concentrations of trypsin was incubated with 167 nM of full length ZNF143 in 40 µl of reaction buffer (50 mM HEPES pH8.0, 200 mM NaCl, 10% (v/v) glycerol, 1 mM DTT) at room temperature for 1 hour. The proteolysis reactions were stopped by adding 6X SDS buffer loading dye and boiling the samples at 100 °C for 5 minutes. The boiled samples were run on SDS-PAGE gels to visualise the digested fragments.

#### 2.5.1.2 LC-MS/MS

To identify the region of the protein fragments, the band of interest was cut from the SDS-PAGE and sent for liquid chromatography tandem mass spectrometry (LC-MS/MS) analysis at the Proteomics Facility at the Institute of Cancer Research. The samples underwent tryptic digestion, then the digested peptides were ionised by electrospray ionisation and separated by the Orbitrap Fusion Lumos Tribid mass spectrometer (Thermo Fisher Scientific). The resulting spectra were visualised and analysed using the Scaffold software.

#### 2.5.2 Electrophoretic Mobility Shift Assay (EMSA)

The DNA substrate for EMSA analysis was produced from forward and reverse complementary DNA strands which were synthesized separately by Integrated DNA Technologies and annealed through a temperature gradient from 95 °C to 4 °C at 1 minute interval per degree drop in temperature. In some cases, the forward DNA strands were fluorescently labelled at the 5' end to visualise the interaction. The DNA constructs used for EMSA were mentioned in the respective result sections.

The binding affinity between proteins and DNA were tested using a range of different concentrations of purified proteins against a constant concentration of DNA of 0.1 µM. They were incubated on ice for 30 minutes in a total reaction volume of 20 µl. The interaction buffer was composed of 25 mM HEPES pH 8.0, 200 mM NaCl, 10% (v/v) glycerol and 1 mM DTT unless otherwise specified in the results section.

Agarose gels and commercial 4-12% Native Novex TBE gels (Invitrogen) were used to visualise the bands from electrophoretic mobility shift assays. Different protein-DNA species were separated by their charges and sizes. To make agarose gels, agarose powder (Thermo Fisher Scientific) was dissolved in 0.5X Tris/Borate/EDTA (TBE) buffer to make up to a final concentration of 2%. 5 µl of each reaction sample was loaded into the wells and ran at 100 mA for 30 minutes. Commercial native gels were ran at 130 V in 0.25X TBE buffer for 1.5 hour at 4 °C before sample loading. 10 µl of samples were loaded into the wells and ran at 130 V for 1.5 hour.

For DNA constructs that were not fluorescently labelled, the gel was incubated with 10 µl of ethidium bromide in 10 ml of ultra-filtered water for 20 seconds and washed with water three times. The bands were then visualised using Typhoon FLA-9000 image scanner (GE Healthcare) with an excitation filter for ethidium bromide. For fluorescently labelled DNA constructs, the appropriate filters that match with the corresponding fluorescent dyes of the DNAs were used to observe the gel bands (Table 2.10).

Dye	Excitation Wavelength	Emission Wavelength	Filter
<b>Ethidium Bromide</b>	518 nm	605 nm	LPG
<b>Cyanine-5</b>	649 nm	670 nm	LPR
<b>Alexa 488</b>	495 nm	519 nm	LPB
<b>6-FAM</b>	490 nm	520 nm	LPB

**Table 2.10 Different Typhoon Imaging Filters Used for Dye Visualisation**

### 2.5.3 Insect Cell Co-expression Pulldown Assay

For protein co-expression, 50 ml of insect High5 cells were infected with Baculovirus of two different proteins of interest. These proteins were expressed for three days in a 27 °C shaking incubator. 10 ml of cells were pelleted at 1000 g for 5 minutes and resuspended in 1 ml interaction buffer (25 mM HEPES pH 8.0, 200 mM NaCl, 10% (v/v) glycerol 1 mM DTT, 0.1% (v/v) Tween20). The resuspended pellet was sonicated for 10 seconds and

centrifuged at 14,000 g for 10 minutes. The supernatant was incubated with 30 µl resin for 1 hour at 4 °C. The beads were washed three times with 10 CV of interaction buffer and the supernatant was discarded. Finally, 6X SDS loading buffer was added to the resins and boiled, proteins released from the resins were analysed by SDS-PAGE.

#### 2.5.4 Recombinant Protein Pulldown Assay

Recombinant protein pulldown assays were used to identify direct interactions between the proteins of interests. The bait protein contained an affinity tag was immobilised onto an appropriate affinity resin. Potential interaction partners (prey) that did not contain the affinity tag were incubated with the bait protein and resin. The presence of the interaction partner was detected through SDS-PAGE.

Depending on the bait protein, different affinity resins were used. For all magnetic resins (Dynabeads™ MyOne™ Streptavidin C1 (Thermo Fisher Scientific), Anti-FLAG® M2 Magnetic Beads (Sigma Aldrich) and Pierce™ Anti-HA Magnetic Beads (Thermo Fisher Scientific)), a magnetic rack was used to separate the supernatant from the magnetic beads. 30 µl of resin was washed and pre-equilibrated in interaction buffer (200 mM NaCl, 25 mM HEPES pH 8.0, 10% (v/v) glycerol, 1 mM DTT, 0.1% (v/v) Tween20). 50 µl of 1 µM of bait protein was incubated with the pre-equilibrated resins at 4 °C for 1 hour. The resin was washed three times with 10 CV of interaction buffer. 100 µl of 2 µM prey protein was then added to the resin and incubated for another hour at 4 °C. The resin was washed again, the supernatant was removed and 6X SDS loading buffer was added to the resin and analysed by SDS-PAGE.

#### 2.5.5 Analytical Gel Filtration Chromatography

Analytical gel filtration chromatography was performed to analyse and study the purified protein, and/or its interaction with another protein, DNA or nucleosome. The gel filtration chromatography column is tightly packed with polymer beads at a stationary phase. A constant flow of running buffer is applied to allow the sample travel through the stationary phase after the

injection. The retention volume of the protein samples depends on their hydrodynamic radius. Smaller protein complexes migrate deeper into the polymer bead pores and are retarded more, requiring greater volumes of buffer to elute them when compared to larger proteins and aggregates that are less likely to enter the pores.

Gel filtration column (Superose 6 Increase 10/300 GL column (GE Healthcare) or Superdex 200 Increase 10/300 GL column (GE Healthcare)) was pre-equilibrated with 1.5 column volume of gel filtration buffer prior to sample injection. 250 µg of protein of interest and molar excess of interacting partners incubated on ice for 30 minutes and centrifuged at 10,000 g for 5 minutes to remove aggregates before loading onto the gel filtration column. The gel filtration buffer was composed of 25 mM HEPES pH 8.0, 200 mM NaCl, 10% (v/v) glycerol and 1 mM DTT unless otherwise specified in the results section. The eluted fractions were analysed via SDS-PAGE and/or Native PAGE.

### 2.5.6 Nucleosome Remodelling Assay

Nucleosome remodelling assay was carried out to understand if the protein of interest was able to shift the nucleosome along a piece of DNA and therefore observe a shift through an EMSA Native PAGE gel. 10 nM to 320 nM of protein was incubated with nucleosome at 100 nM in a final reaction volume of 20 µl. The reaction buffer was composed of 25 mM HEPES pH 8.0, 2 mM MgCl<sub>2</sub>, 100 mM NaCl, 10% (v/v) glycerol, 2mM ATP and 1 mM DTT. After 30 minutes incubation at room temperature, the reactions were quenched for 10 minutes on ice with 1 µl of 20 µM DNA and 2 µl of 500 mM EDTA pH 8.0. 5 µl of the reaction mixture was loaded onto a 4-12% Native TBE gel with 0.5X TBE buffer and ran for 90 minutes at 130 V. Finally, the gel was stained with ethidium bromide and visualised using Typhoon image scanner.

### 2.5.7 Co-immunoprecipitation Pulldown Assay

The aim of co-immunoprecipitation (Co-IP) was to identify endogenous interacting partners of ZNF143. Co-IP targets a known protein with a specific

antibody to pull the entire protein complex out of solution to identify its interacting proteins.

#### 2.5.7.1 HEK 293T Cell Growth

$1 \times 10^6$  HEK 293T cells were cultured in 10 ml DMEM (Thermo Fisher Scientific) containing 10% FBS, 50 units/ml penicillin and 50 µg/ml streptomycin in a 10 cm dish and harvested after 3 days or when the plates reached 90 % confluency. The cells were washed with ice-cold PBS and centrifuged at 100 g four times before flash-freezing the pellets and storing them at -80 °C.

#### 2.5.7.2 Nuclear Extraction from HEK293T Cells

For each set of experiment, twelve plates of HEK 293T cells were lysed in 6 ml of ice-cold lysis buffer (10 mM HEPES pH8.0, 85 mM KCl, 0.5% NP-40, Pierce™ protease inhibitor mini tablet (Thermo Fisher Scientific) and Pierce™ phosphatase inhibitor mini tablet (Thermo Fisher Scientific). The lysed cells were centrifuged at 500 g for 10 minutes and the pellets were resuspended in 2 ml of high-salt nuclei lysis buffer (50 mM HEPES pH 8.0, 5 mM MgCl<sub>2</sub>, 0.2% TritonX-100, 20% (v/v) glycerol, 300 mM NaCl, and protease and phosphatase inhibitor mini tablets). The lysates were transferred to a cold homogeniser to lyse the cells. 2 µl of benzonase and 2 mM MgCl<sub>2</sub> was incubated with the lysed cells at 4 °C for 1 hour. After incubation, the samples were centrifuged at 12,000 g for 10 minutes. The supernatant was diluted with no-salt nuclei lysis buffer (50 mM HEPES pH 8.0, 5 mM MgCl<sub>2</sub>, 0.2% TritonX-100, 20% (v/v) glycerol, and protease and phosphate inhibitor mini tablets) to reduce the final sodium chloride concentration to 150 mM.

#### 2.5.7.3 Immunoprecipitation of ZNF143

100 µl of Dynabeads™ Protein G for Immunoprecipitation (Thermo Fisher Scientific) was washed with PBS-T (PBS + 0.01% (v/v) Tween20). 15 µg of antibody (anti-ZNF143 from Novus Biologicals, anti-IgG from Merck) was diluted in PBS-T and incubated with Protein G dynabeads for 15 minutes at room temperature. The supernatant was removed and the dynabeads were mixed with the diluted nuclear extract and incubated for 2 hours with rotation

at 4 °C. The beads were washed five times with low-salt nuclei lysis buffer (50 mM HEPES pH 8.0, 5 mM MgCl<sub>2</sub>, 0.2% TritonX-100, 20% (v/v) glycerol, 150 mM NaCl, and protease and phosphate inhibitor mini tablets) and the proteins captured were digested on-bead with trypsin for mass spectrometry analysis.

#### 2.5.7.4 On-Bead Digestion

The dynabeads were washed with cold 50 mM ammonium bicarbonate (ambic) three times before resuspending in 90 µl of 50 mM ambic and 10 µl of 0.1 µg/µl trypsin stock and incubated overnight at 37 °C. The addition of trypsin digested proteins that were captured on the beads. The supernatant containing digested peptides was collected and acidified with 5 µl of 5% formic acid. The volume was reduced to ~50 µl with SpeedVac Vacuum (Thermo Fisher Scientific). 100 µl of 100% acetonitrile was added to the peptide solution and the samples were filtered through Multiscreen HTS Filter Plate (Merck Millipore). The filtered peptides were dried completely with SpeedVac Vacuum and resolubilised in 20 mM TCEP and 5% formic acid before sending for LC-MS/MS analysis.

#### 2.5.7.5 LC-MS/MS

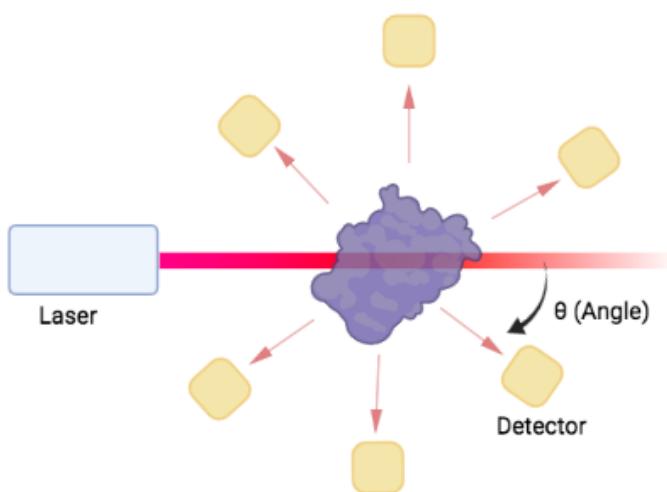
To identify the interacting partners of ZNF143, the digested peptides were sent for liquid chromatography tandem mass spectrometry (LC-MS/MS) analysis at the Proteomics Facility at the Institute of Cancer Research. The peptides were ionised by electrospray ionisation and separated by the Orbitrap Fusion Lumos Tribid mass spectrometer (Thermo Fisher Scientific). The co-IP experiment was repeated 3 times independently and the results were analysed using CRAPome (Mellacheruvu *et al.*, 2013). Spectrum/peptide matches were retained only if they had a normalized difference in cross-correlation scores of at least 0.08.

## 2.6 Biophysical Analysis

### 2.6.1 Size Exclusion Chromatography Coupled to Multi-Angle Light Scattering (SEC-MALS)

#### 2.6.1.1 Background

The intensity of light scattered by a molecule is correlated to its mass, concentration and refractive index (Folta-Stogniew, 2006). The molecular weight can then be derived from the light scattering signals, whereas the change in refractive index ( $d\eta/dc$ ) refers to the polarizability of a material in relation to its change in concentration in a particular solvent. During a SEC-MALS experiment, the protein samples were first passed through a size exclusion chromatography (SEC) column that separates molecules by their hydrodynamic volume. As the molecules exit the SEC column, they enter the multi-angle light scattering detector and are probed by a laser beam. The light scattered by the molecules are detected by an array of eighteen detectors at discrete angles (Figure 2.1). This information allows us to determine the radius of gyration  $R_g$  and molecular weight of the protein samples, which is defined as the mass weighted average distance from the core of the molecule to each mass element in the molecule and is independent of the protein's shape. The light scattering detector is often coupled to a differential refractive index (dRI) detector that is sensitive to the sample concentration. For instance, for a concentration-dependent aggregate, it would have a strong light scattering signal but a small dRI signal.



### **Figure 2.1 SEC-MALS Setup**

The incident beam is polarised and scattered by the protein molecule. Detectors are located at discrete angles to detect the scattered beam.

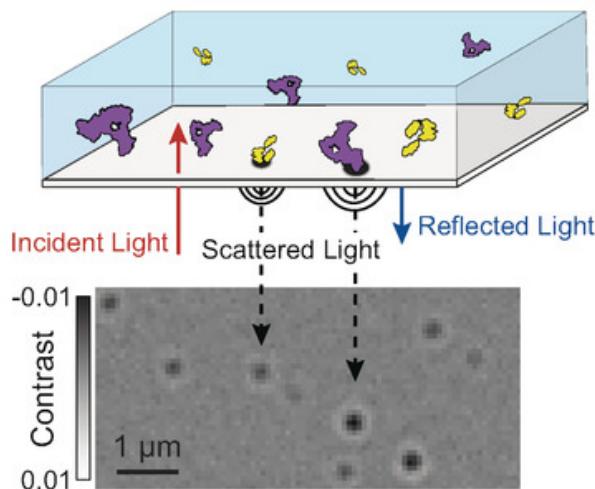
#### **2.6.1.2 Data Collection and Analysis**

The SEC-MALS experiment for ZNF143 was kindly performed by the Seiradake group at the University of Oxford. Purified ZNF143 samples and buffers required were prepared at the ICR and sent to University of Oxford. Superdex 200 10/300 GL column (GE Healthcare) was used for the size exclusion chromatography. 100 µl of full length ZNF143 at 2.5 mg/ml was injected into the column that was equilibrated with running buffer (50 mM HEPES pH 8.0, 250 mM NaCl, 5% (v/v) glycerol, 5 µM ZnCl<sub>2</sub> and 2 mM DTT) at a constant flow of 0.5 ml/min. Light scattering was measured using a DAWN Heleos-II light scattering instrument (Wyatt), and the differential refractive index was measured using an Optilab rEX instrument (Wyatt). The analysis was performed using ASTRA Software from Wyatt. Data for the light scattering, differential refractive index and molecule weight over the elution volume were exported and plotted.

#### **2.6.2 Mass Photometry**

##### **2.6.2.1 Background**

Mass photometry is used to measure protein masses in their native state in solution, monitor protein-protein interactions and identify heterogeneity in the samples. The measurement of molecule mass is based on the interference reflection microscopy and interferometric scattering microscopy. The samples are illuminated, and the individual binding events are monitored over time as the protein molecules adhere to the cover glass surface. Most importantly, the change in refractive index caused by the interference of scattered and reflected light are interferometrically recorded. The amount of light scattered (interferometric contrast) and the particle volume follows a linear relationship. Therefore, the molecular mass of the protein samples can be extrapolated based on the recorded scattering signal (Sonn-Segev *et al.*, 2020).



**Figure 2.2 Principle of Mass Spectrometry**

As incident light is shone through the glass/water interface, each protein binding onto the glass would lead to a change in refractive index (contrast) and mass photometry images the interference between the scattered and reflected light over time to estimate the molecular weights of the protein species. Adapted from Soltermann *et al.*, 2020.

#### 2.6.2.2 Data Collection and Analysis

Microscope coverglass (VMR) and Grace Bio-labs reusable CultureWell™ gaskets (Sigma Aldrich) were washed with isopropanol and ultra-filtered water twice, followed by drying with a clean nitrogen stream. A drop of IMMOIL-F30CC immersion oil for microscopy (Olympus) was added between the mass photometer light source and the clean coverslip.

All measurements were performed with Refeyn One<sup>MP</sup> mass photometer (Refeyn Ltd). Data acquisition was performed using AcquireMP v2.3.0 (Refeyn Ltd). The focal position was identified prior to each measurement using the autofocus system in the AcquireMP software. To find focus, 10 µl of fresh stock *buffer A* (25 mM HEPES pH 8.0, 150 mM NaCl, 10% (v/v) glycerol and 1 mM DTT) was added to one of the flow chambers. After the focus has been identified, protein stocks were diluted in stock *buffer A* and the typical working concentrations of protein samples were between 50 – 100 nM. A new flow chamber was used and the focal position was re-identified for every sample measurement. For each acquisition, 3 µl of diluted protein was added onto the

10 µl buffer A which was already present in the flow chamber that was used for focusing. Movies of 60 second duration were recorded, and each sample was measured at least three times independently. Mass photometry movies were recorded at 1 kHz, with exposure time of 0.6 ms. The time and pixel binning resulted in an effective pixel size of 84.4 nm and effective frame rate of 200 Hz.

Data analysis was performed using Refeyn DiscoverMP v2.3.0 (Refeyn Ltd). Firstly, the static scattering background from the glass/water interface was removed, then the adhering particles were identified, lastly the particles were fitted to extract maximum contrast and the contrast values were compared against the calibration curve.

### 2.6.3 Fluorescence Polarisation

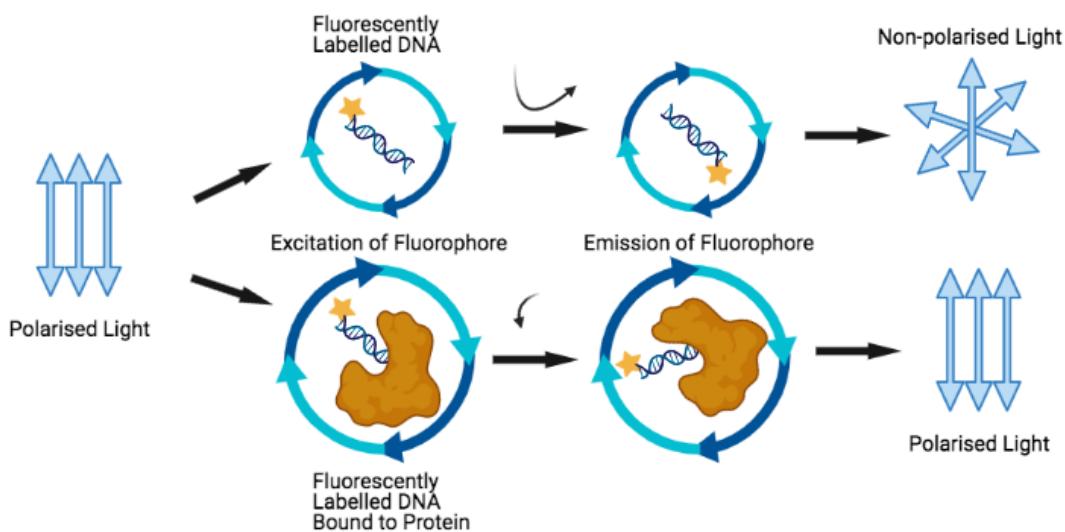
#### 2.6.3.1 Background

Fluorescence polarisation measures the degree of polarisation when a fluorescently labelled molecule is excited by polarised light to study molecular interactions in solution (Moerke, 2009). The degree of polarisation is inversely proportional to the rate of molecular rotation of the particles. The emitted light from the molecule is measured by two detectors located in parallel and perpendicular to the plane of excitation. The difference in fluorescence intensity measurements between the parallel and perpendicular detectors is used to calculate the fluorescence polarisation of the sample (Moerke, 2009). Fluorescence anisotropy ( $A$ ) is closely related to fluorescence polarisation ( $P$ ) where it is defined by the following equation (Piazza *et al.*, 2014):

$$A = \frac{2P}{3 - P} \quad \text{Eq. 2}$$

Smaller molecules rotate more rapidly when excited by polarised light and therefore the emitted light will be largely depolarised due to the rapid reorientation of the fluorophore, giving rise to a low FP signal. When larger molecules are bound to the fluorescent probe, the rotational speeds are reduced and reorient to a smaller degree, hence the emitted light remain

largely polarised, giving rise to a high FP signal (Figure 2.3). The changes in the polarisation and FP signal were used to monitor the binding activity of fluorescently labelled molecules to other non-labelled molecules.



**Figure 2.3 Schematic of Fluorescence Polarisation Assay Principle**

When DNA with fluorescent label attached (yellow star) is excited with polarised light, the emitted light detected is largely depolarised due to the fast rotational correlation time relative to the excited state lifetime of the fluorophore. On the contrary, when the DNA is bound to the protein (orange object), the formation of the complex leads to a slower molecular tumbling rate and hence the emitted light retains a high degree of polarisation.

#### 2.6.3.2 Data Collection and Analysis

The forward and reverse complementary DNA strands were synthesized separately by Integrated DNA Technologies. The forward DNA strands were fluorescently labelled with 6-carboxyfluorescein fluorophore (6-FAM) at the 5' end. The length of the DNA was 28 base pairs, and the complementary DNA strands were annealed through a temperature gradient from 95 °C to 4 °C at 1 minute interval each. ZNF143 was diluted twelve times by two-fold dilution and titrated against a constant concentration of 6-FAM labelled dsDNA of 100 nM in a total reaction volume of 50 µl. The buffer was composed of 25 mM HEPES pH 8.0, 150 mM NaCl, 5% (v/v) glycerol and 1 mM DTT. DNA and ZNF143 were incubated at room temperature for 90 minutes in Molecular Probes® 96-well microplates (Thermo Fisher Scientific). The fluorescence anisotropy signals were visualised with an excitation wavelength of 485 nm

and an emission wavelength of 525 nm using BMG Labtech POLARstar Omega plate reader. Each condition was repeated three times.

The calculation of equilibrium dissociation constants followed the method explained by Piazza *et al.*, 2014. Normalised fluorescence anisotropy ( $\Delta A$ ) was calculated using the following equation, where  $r_n$  is the anisotropy with protein  $r_0$  is the anisotropy without protein, and  $r_{max}$  is the anisotropy at the highest protein concentration.

$$\Delta A = (r_n - r_0) / (r_{max} - r_0) \quad \text{Eq. 3}$$

To estimate the equilibrium dissociation constants for different oligonucleotides, the normalised fluorescence anisotropy was plotted as a function of protein concentration using Graphpad PRISM. A curve was fit using the following quadratic expansion, with error bars indicating standard deviation:

$$\Delta A = \left( \frac{\Delta A_{max}}{2[DNA]} \right) \left( ([ZNF] + [DNA] + Kd) - \sqrt{([ZNF] + [DNA] + Kd)^2 - 4[ZNF][DNA]} \right) \quad \text{Eq. 4}$$

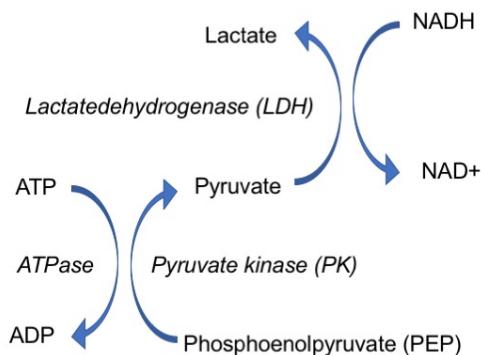
where  $\Delta A_{max}$  is the total anisotropy change after saturation of the curve, [ZNF] and [DNA] are the concentration of ZNF143 and 6-FAM labelled oligonucleotides in  $\mu\text{M}$  at each point in titration, respectively.

## 2.6.4 ATP Hydrolysis Assay

### 2.6.4.1 Background

ATP/NADH coupled hydrolysis assays were used to monitor the ATP hydrolysis rates of CHD8 in the presence of nucleosomes. The hydrolysis of ATP is enzymatically coupled to the oxidisation of NADH to NAD<sup>+</sup> (Nicotinamide adenine dinucleotide). This assay measured the level of NADH absorption at 340 nm and monitored the level continuously over time. First, following each cycle of ATP hydrolysis, the ADP generated is regenerated to ATP by pyruvate kinase (PK). On the other hand, pyruvate kinase (PK) converts one molecule of phosphoenolpyruvate (PEP) to pyruvate.

Subsequently, pyruvate is then reduced to lactate by lactate dehydrogenase (LDH), which catalyses the oxidation of NADH (Radnai *et al.*, 2019). ATP is constantly regenerated as long as PEP is available in the reaction system and this allows the measurement of a steady-state ATP hydrolysis rate. Therefore, the decrease in ATP concentration is proportional to the decrease in NADH concentration which is monitored over the entire course of the assay.



**Figure 2.4 ATP/NADH Coupled Assay**

The hydrolysis of ATP is enzymatically coupled to the oxidation of NADH to NAD<sup>+</sup> through a series of coupled events.

#### 2.6.4.2 Data Collection and Analysis

The assay was performed at room temperature. The ATP mix components (50 mM PEP, 2.5 mM NADH, 2500 units lactate dehydrogenase, 2500 units of pyruvate kinase and 50 mM ATP) were prepared with the ATPase assay buffer (25 mM HEPES pH8.0, 100 mM NaCl, 10% (v/v) glycerol and 1 mM DTT). The reaction was initiated by adding ATP mix to 1 µM of CHD8 (in some conditions supplemented with 1 µM nucleosome) in a 96-well half area plate and monitored at OD<sub>340nm</sub> using POLARstar plate reader for 4000 seconds. The absorbance level over time was plotted using Graphpad PRISM. Each sample in the ATP hydrolysis assay was performed with three replicates. The ATP hydrolysis rate was determined by the linear fit of change in ATP concentration of the reaction.

## 2.6.5 Dynamic Light Scattering

### 2.6.5.1 Background

Dynamic light scattering (DLS) measures the hydrodynamic radius of particles, based on the Brownian motion of dispersed molecules. Smaller molecules tend to move at higher speeds than larger particles, and hence the relationship between the speed of the particles and particle sizes is given by the Stokes-Einstein equation.

$$D = \frac{k_B T}{6\pi\eta R_H} \quad \text{Eq. 5}$$

Temperature ( $T$ ) and viscosity ( $\eta$ ) of the solution are both important parameters that affect the movement of particles. The speed of particles is given by the translational diffusion coefficient  $D$ .  $k_B$  corresponds to the Boltzmann constant and  $R_H$  refers to the hydrodynamic radius of the particles.

An incident laser beam is directed to the sample solution, as the beam hits the protein particles, the scattered light at a certain angle is monitored over time. The smaller the molecules, the more fluctuation is detected by the sensor. This intensity can be used to generate a correlation function that describes how long a protein molecule is at the same spot within the sample. The bigger the particles, the slower the decay of the correlation function. The diffusion coefficient can then be determined from the correlation function, and hence further obtain the hydrodynamic radius of the protein molecules.

### 2.6.5.2 Data Collection and Analysis

The assay was performed at room temperature. Purified CTCF was diluted following a 2-fold dilution three times. Four different concentrations of CTCF were analysed using SpectroLight 610 (Xtal Concepts). 96 well terasaki plate was covered with a thin layer of paraffin oil, and 2  $\mu$ l of protein solution was injected into the well.

## 2.6.6 Cross-Link Mass Spectrometry

The interacting regions of two proteins can be studied through cross-link mass spectrometry. This experiment was carried out in collaboration with Dr. Duccio

Conti from the Max Planck Institute in Germany. Purified CHD8 and purified CTCF samples at 2 µM each were sent to MPI. These two proteins were then cross-linked with disuccinimidyl dibutyric urea (DSBU). The sample was trypsin-digested and sent to mass spectrometry facility. Nano electrospray ion source (Thermo Fisher Scientific) was used to ionise the cross-linked samples, followed by Orbitrap Q-Exactive Plus (Thermo Fisher Scientific) for determination of crosslinks of purified protein complexes. MeroX was used to search for the sequences and results that passed 0% false discovery rate cutoff were visualised using xiVIEW.

## 2.7 X-ray Crystallography

### 2.7.1 Overview

X-ray crystallography is the most widely used structural biology technique that allows us to visualise the protein structure at near-atomic resolution and therefore contributes to our understanding of macromolecular functions. It has advanced greatly since the first structural determination of myoglobin and haemoglobin by Perutz and Kendrew in 1960 (Kendrew *et al.*, 1960; Perutz *et al.*, 1960; de Chadarevian, 2018).

As an incident X-ray beam hits the molecules in a sample, the electrons will scatter the incoming photons, which generates a diffraction pattern that can be observed on a detector as the result of all scattered X-ray waves. The positions and the intensities of the dark and light reflections on the diffraction image (or diffractogram) can then be used to reconstruct the structural information of the protein crystal lattice (Rupp, 2009). The amplitudes and the phases of each diffracted wave are required to interpret the resulting diffraction image and to reveal the molecular structure. The relationship between these factors can be explained using the following equation, where the structure factor ( $f(S)$ ) is the sum of all scattering waves in the crystal lattice and it is related to the amplitudes ( $|f(S)|$ ) and the phases ( $\phi$ ) of the diffracted waves.

$$f(S) = |f(S)|e^{i\phi(S)} \quad \text{Eq. 5}$$

The intensity of the reflections can be directly measured from the diffractogram. The intensity is proportional to the amplitudes of the diffracted waves, and thus proportional to structure factor ( $f(S)$ ) too. However, the phase information of the diffracted waves is lost, which is coined as the ‘phase problem’ (Rupp, 2009). There are several ways to estimate the phases, including isomorphous replacement, which involves introducing heavy atoms into proteins to create scattering with higher densities; or molecular replacement, where an existing structure of a similar macromolecule is used as a model to estimate the phase information. Once the structure factor values for all directions are determined, the electron density map can then be worked out through Fourier Transform. Lastly, the atomic model is performed through an iterative process of manual model building by fitting residue chains into its respective electron density, followed by further refinement.

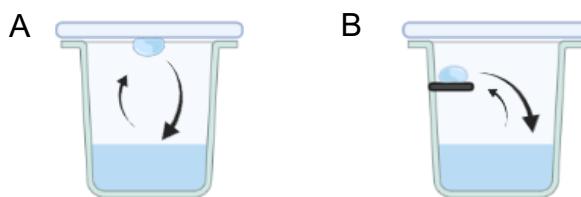
### 2.7.2 Protein Crystal Formation

The X-ray diffraction signal from one single molecule is not sufficient to visualise the intrinsic wave properties, therefore macromolecular crystals, which are ordered three-dimensional arrays of molecules, are needed to amplify the diffraction signal. Protein crystals are repeated assemblies of macromolecules held together by a network of sparsely weak and non-covalent intermolecular interactions. These interactions largely depend on the inherent properties of the protein, and on environmental factors including pH, solvent composition and temperature (Rupp, 2009). Obtaining protein crystals can be challenging as it requires dynamic protein molecules to assemble into an ordered periodic three-dimensional lattice. In addition, these macromolecular crystals are highly sensitive to mechanic stress due to its high solvent composition, which arises from the solvent channels present between the macromolecules.

Crystallisation is an entropy-driven process. The driving force of crystallisation is the slow precipitation of a macromolecule which promotes the formation of specific bonds to overcome the loss of entropy from the more ordered system. There are two main steps during protein crystallisation; nucleation and growth

(Rupp, 2009; McPherson and Gavira, 2014). The mechanism of nucleation is not fully understood, but the current theory suggests that crystal nucleation is a first-order phase transition, where molecules transition from a disordered state to an ordered state with latent heat (McPherson and Gavira, 2014). The free energy barrier between the solution-crystal boundary has to be overcome to form a thermodynamically stable aggregate, which is termed as critical nucleus. This nucleus provides the surface for further crystal growth. Crystal growth is further dependent on the diffusion of macromolecules to the surface of the critical nuclei, followed by an ordered assembly onto the growing crystal (Krauss *et al.*, 2013).

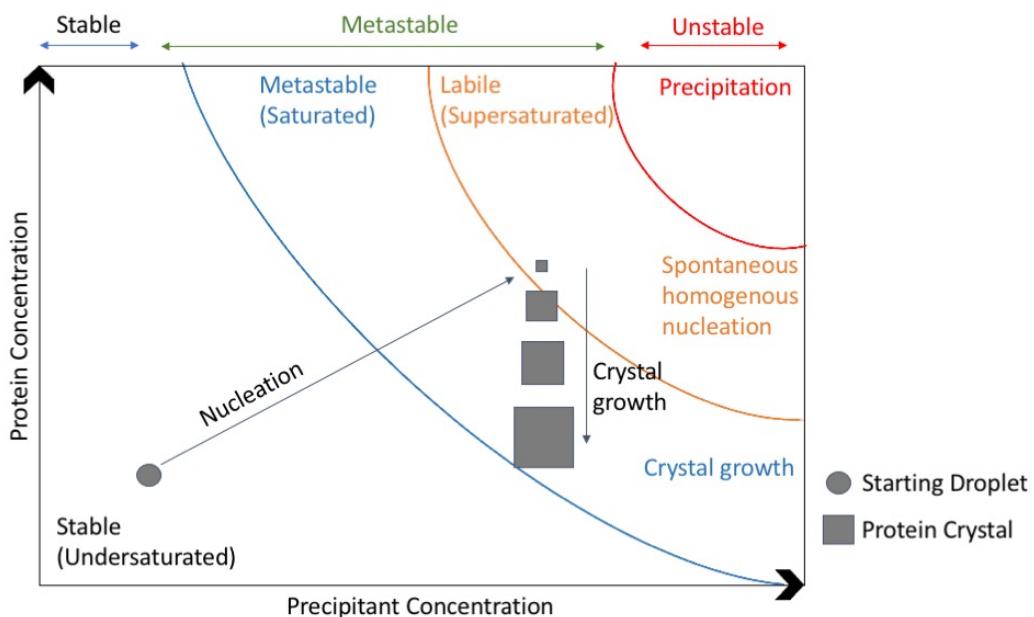
The most popular crystallisation technique is vapour diffusion, either in hanging-drop or sitting-drop crystallisation plates (Figure 2.5). A solution of purified and highly concentrated homogenous protein or protein complex is mixed with precipitants. Once the drops are set up, the system is sealed to allow the establishment of equilibrium between the protein-precipitant droplet and the precipitant reservoir. As water diffuses from the droplet to the reservoir, the concentration of protein and precipitants within the droplet increase, facilitating the shift of macromolecules to the supersaturation state. Due to fewer water molecules present, the ionic strength between the macromolecules increase and hence they tend to interact with other macromolecules to maintain their electrostatic charges, leading to the formation of critical nuclei (Rupp, 2009; Krauss *et al.*, 2013).



**Figure 2.5 Vapour Diffusion Crystallisation Experiment Setup**

A: Illustration of hanging-drop vapour diffusion experiment. B: Illustration of sitting-drop vapour diffusion experiment. Water will diffuse out from the drop to the reservoir and protein concentration in the droplet will increase until an equilibrium is reached between the two entities.

Supersaturation is required for nucleation and crystal growth, and the level of preferred supersaturation vary. Depending on the concentration of protein and precipitants, the level of saturation can be further divided into three zones; precipitation zone where molecules are precipitated or aggregated, labile zone where both nucleation and growth of crystals occur, and lastly the metastable zone where only crystal growth is favoured (Figure 2.6) (Rupp, 2009; Krauss *et al.*, 2013; McPherson and Gavira, 2014). Therefore, it is preferred to induce nucleation at the lowest level of supersaturation within the labile zone. As the concentration of protein in the solution gradually decreases, the equilibrium is then shifted towards the metastable zone where crystal growth occurs slowly. However, these zones are highly related to their kinetic parameters with no clear boundary between each region, therefore extensive crystallisation trials are required to meet the optimal conditions for crystallisation.



**Figure 2.6 Phase Diagram for Crystallisation of Macromolecules**

The starting droplet is a mixture of protein and precipitant solution. Water diffuses out of the droplet and the equilibrium of the system shifts to the labile zone where spontaneous nucleation is reached. As the crystal grows in size, the concentration of protein drops and the system is then shifted into the metastable zone where crystal growth is favoured.

## 2.7.3 Sample Preparation

### 2.7.3.1 Sparse Matrix Screening

MRC 2-well plates were used for sparse matrix screening. The complete list of commercial screens used for crystallisation trials are listed in table 2.11. 50 µl well solution was dispensed into each well using Phoenix liquid handling system. For protein only screens, 100 nl of protein solution (concentration mentioned in results section) was dispensed into one of the wells using a Mosquito® Liquid Handler robot (SPT Labtech), followed by 100 nl of well solution from the commercial screens. Plates were sealed with Qiagen clear seal. The plates were incubated at 4 or 18 °C.

For full length ZNF143-DNA complex crystallisation trial, protein and DNA was incubated on ice for one hour at a ratio of 1:1.2 prior to crystallisation plate set up. 100 nl of protein-DNA mixture was dispensed into the wells using Mosquito robot, followed by 100 nl of well solution. Qiagen clear seal was used to ensure an air-tight environment, and the plates were incubated at 4 or 18 °C.

For ZNF143<sub>103-466</sub>-DNA complex trials, protein was incubated with DNA on ice for 30 minutes, and then passed through Superdex 200 10/300 GL column (GE Healthcare) in the reaction buffer with 25 mM HEPES pH 8.0, 200 mM NaCl, 5% (v/v) glycerol and 1 mM DTT. The fractions from the protein-DNA peak were concentrated before setting up crystallisation plates. The plate settings and droplet dispense were the same as previously described.

Commercial Screen	Manufacturer
Crystal Screen HT-96	Hampton Research
Index HT-96	Hampton Research
JCSG+ HT-96	Molecular Dimensions
Morpheus HT-96	Molecular Dimensions
PACT Premier HT-96	Molecular Dimensions
PEG Suite HT-96	Qiagen
ProPlex HT-96	Molecular Dimensions
Salt-RX HT-96	Hampton Research

<b>Shotgun HT-96</b>	Molecular Dimensions
<b>Stura Footprint HT-96</b>	Molecular Dimensions

**Table 2.11 List of Commercial Screens Used for Crystallisation Trials**

### 2.7.3.2 Optimisation Plate Setup

24-well Linbro style XRL plate (Molecular Dimensions) were used for setting up condition optimisations. The protein-DNA stock used in this experiment was prepared following the protocol mentioned in section 2.7.3.1. Hanging drops were made up mixing 1 µl of protein-DNA solution with 1 µl of well solution, the reservoir volume was 500 µl. Two drops were plated per well. The wells were sealed with CleneGlass™ coverslips (Molecular Dimensions) and Vaseline. Plates were incubated at 18 °C.

### 2.7.4 Data Collection

X-ray diffraction data was collected by Dr. Stephen Hearnshaw and the data was collected at the Diamond synchrotron on the micro-focus beamline I24. The data was not further processed due to the poor diffraction pattern of the tiny crystals.

## 2.8 Electron Microscopy

### 2.8.1 Overview

Electron microscopy is a powerful technique that can reveal the structure of whole cell organelles (electron tomography) to individual protein structures (cryo-EM). Recent advancements in the cryo-electron microscopy field has made it one of the strongest rivals against X-ray crystallography. The resolution revolution of cryo-EM allowed atomic resolutions of protein complexes to be determined (Kühlbrandt, 2014). One of the main limitations of X-ray crystallography is that it is highly dependent on the homogeneity and concentration of the protein samples to create a highly pure crystal to generate a high-resolution electron density map. Whereas for cryo-electron microscopy, the requirement for the homogeneity and concentration of proteins samples are less stringent, and therefore creates a lot more opportunities for samples

that do not readily crystallise. The development of electron microscopes, detectors and software in the recent years have led to the improvement of determining protein structures at the near-atomic resolution.

### 2.8.2 Sample preparation

The electron microscope operates under a tight vacuum environment to avoid electrons being deflected as they hit air particles. As liquid water evaporates immediately in a vacuum, it is important to stabilise the biological specimen on a metal grid prior to imaging (Orlova and Saibil, 2011). The two common techniques used to prepare the biological samples for imaging are negative-staining and cryo-EM. With negative staining, the protein solution is applied to the support metal grid, the excess liquid is blotted off with cellulose filter paper before applying a coat of heavy metal salt solution such as uranyl acetate. The heavy metal salt stains the entire grid except areas where protein samples are present (Brenner and Horne, 1959). Thus, this technique provides good contrast and outlines the shape of the samples. However, the resolution of reconstruction is limited to around 20 Å.

Cryo-EM allows us to visualise the biological specimen under its native state through rapid freezing method to allow reconstruction of the sample to near-atomic resolution. The protein sample is applied to the metal support grid, followed by blotting off excess liquid and plunge-freezing in liquid ethane (Adrian *et al.*, 1984). Liquid ethane is used at a temperature that is closer to its freezing point so it does not evaporate and creates an insulating gas layer (Orlova and Saibil, 2011). This process allows the sample to be in a vitrified state and prevents formation of ice crystals and sample dehydration. Samples prepared through plunge-freezing are stabilised in its native-like, hydrated state and therefore are ideal for structural characterisation.

### 2.8.3 The Transmission Electron Microscope

#### 2.8.3.1 Overview

To visualise objects that are smaller than the wavelength of light, electrons instead of photons are used as a beam source for the transmission electron

microscope (TEM). The TEM consists of an electron gun source, three lens systems and an imaging detector operating under a tight vacuum to avoid electrons interacting with gases or other particles.

#### 2.8.3.2 Electron Gun

The most common sources of electrons come from passing current through tungsten filaments or lanthanum hexaboride ( $\text{LaB}_6$ ) crystals. The field emission gun (FEG) is the most commonly used electron gun where its tungsten filament is heated to 2000 °C to 3000 °C to generate a coherent source of electrons (Orlova and Saibil, 2011). A new method, cold-FEG was recently employed to optimise the energy spread. The gun was subjected to a round of quick heating (known as ‘flashing’) to remove gas particles that were attached, which might reduce the coherence of the emitted beam (Nakane *et al.*, 2020). The emitted electrons are accelerated to voltages between 100 – 300 kV through a series of accelerator stacks before reaching the first set of condenser lens system.

#### 2.8.3.3 Len Systems

Next, the condenser lens converts the diverging beam of electrons into a near-parallel beam to illuminate the sample that is located in the middle of the objective lens surrounded with magnetic field. As the beam passes through the sample, the scattered electrons are focused with the objective lens. The objective lens magnifies the image 20 – 50x and the image is further magnified with the third set of lens system comprising of the intermediate and projector lenses (Orlova and Saibil, 2011). A pair of apertures located below each set of lenses are used to remove electrons that are scattered at high angles. This reduces the amount of inelastic scattering, which might otherwise limit the resolution of the images.

#### 2.8.3.4 Lens Aberrations

There are aberrations caused by the lens system that need to be corrected. For instance, spherical ( $C_s$ ) and chromatic ( $C_c$ ) aberrations are the two most significant factors that affect the imaging. Spherical aberration corresponds to

the situation where electron beams that are further away from the optical axis refract more strongly than the beams that are closer to the centre. On the other hand, chromatic aberration refers to the dependence of ray focus on the wavelength of the electron beams. Electron rays with longer wavelengths focus more readily, forming planes closer to the object than beams with shorter wavelengths. Spherical aberrations can be corrected during image processing whereas chromatic aberrations require the installation of energy filter before the electron beams hit the image detector (Orlova and Saibil, 2011). As the energy of electrons are inversely proportional to their wavelengths, an energy filter can be applied to limit electrons that are within a set range of energy to reach the image detector. The energy slit is usually set to a slit width of 20 ev to select elastically scattered electrons with no energy loss.

#### 2.8.3.5 Image Detectors

Image detectors have advanced greatly in the past decade to allow higher resolution images to be recorded. Photographic film was the conventional way to detect the images, but it has a lot of limitations, such as long imaging and developing time that limited the number of images that could be taken during each session. Charge-coupled detectors (CCD) were then developed. It is a type of indirect electron detector that converts electrons into photons before generating a digital readout. As the incoming electron beams strike the scintillator, the electrons are scattered to produce photons that are transferred through a bed of fibre optics to the detector. The photons that are captured are converted into electric charges. The charges collected in each pixel are amplified and transferred to a digital readout. However, one of the limitations of CCD is that it creates a low signal-to-noise ratio, hence the efficiency of signal recording is low (Faruqi and McMullan, 2018). To overcome this challenge, a new type of detector known as direct electron detection devices (DDD) has been developed. DDD are comprised of radiation-hardened complementary metal oxide semiconductors (CMOS) coupled to monolithic active pixel sensors (MAP). CMOS can tolerate high-energy electrons and the detected electron beams are converted to charge directly without the need of photon conversion. This technology improves the signal-to-noise ratio as the

charge counts are more localised and support high frame rates of up to 400 frames per second (McMullan, Faruqi and Henderson, 2016). Thereby the detective quantum efficiency (DQE) of these detectors improved significantly. DQE measures the extent of detectors affect the signal-to-noise ratio. It can be expressed as a function of spatial frequency, with a DQE of 1 being a perfect detector where all electrons are detected and converted to signal with no noise (Faruqi and McMullan, 2018).

As a beam of electron illuminates the sample, the biological specimen that sits within a thin layer of vitreous ice inside a perforated carbon support film shifts by up to a few nanometers due to the deformation of the carbon support film. The changes in position or orientation of the particles, termed as beam-induced motion (BIM), cause significant blurring in the collected images and hinders the potential to achieve near-atomic resolution (Brilot *et al.*, 2012; Campbell *et al.*, 2012). With the advancement of image detectors that support high frame rates, multi-frame movies in addition to images are recorded over the exposure time during a data collection. This, in turn allows for a major improvement in the quality of imaging as frames can be aligned during a subsequent motion-correction step, allowing for the recovery of the previously blurred high resolution features (Zheng *et al.*, 2017). Technical advances in grid preparation have also been made to reduce BIM and improve imaging. Gold grids were developed to allow the charge to disperse more quickly and homogenously (Russo and Passmore, 2014), and hexagonal grids with more densely packed holes were recently developed to support ‘movement-free’ imaging (Naydenova, Jia and Russo, 2020). The beam-induced movement correction algorithm in image processing software suites takes into consideration the different weighting of frames throughout the collected movies, where the early and last few frames are down-weighted due to the effect of BIM being strongest in early frames and accumulated radiation damage starts to degrade the quality of the particles in the last few frames. Thereby the largest contribution to the overall motion-corrected image is from the intermediate frames which have minimal effects from both BIM and radiation damage (Scheres, 2014).

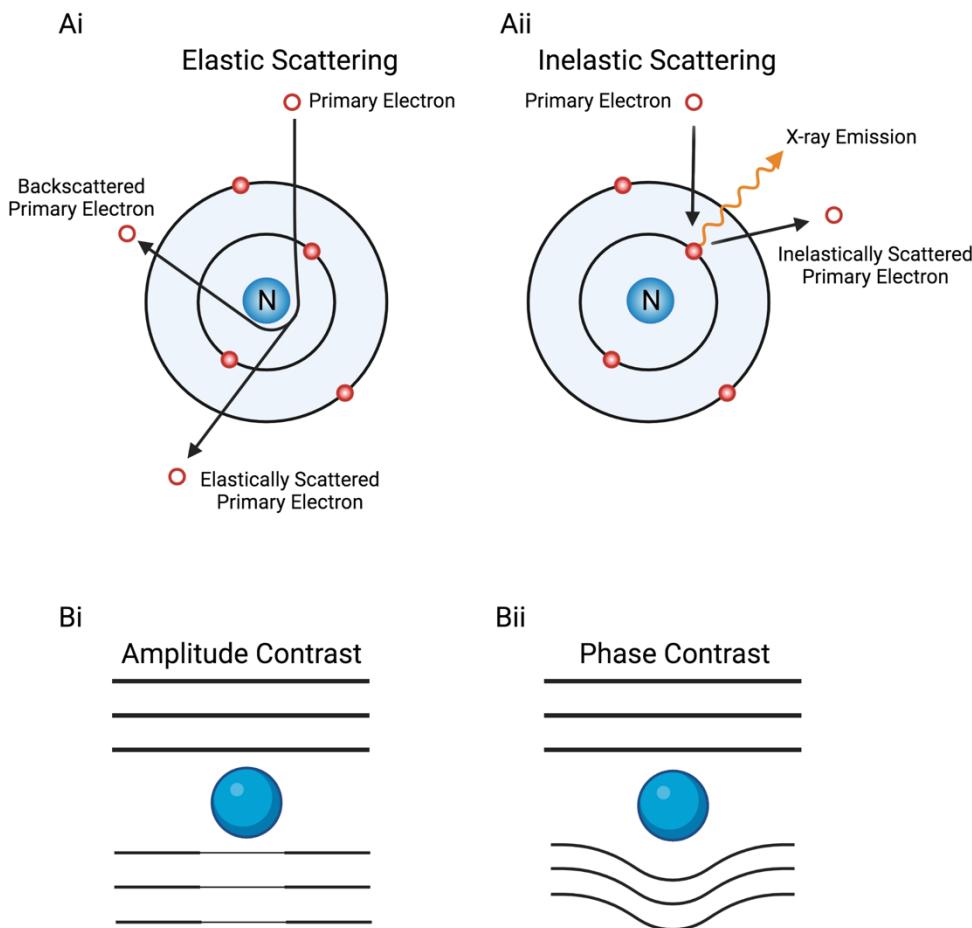
#### 2.8.4 Electron Scattering and Image Formation

As an incident beam of electron reaches the sample stage, the majority of electrons pass straight through the sample with no interaction, whereas some electrons interact with the sample, leading to either of the two types of scattering interactions. The first type of interaction is inelastic scattering (Figure 2.7Aii), where the energy from an incoming electron is transferred onto the sample. Inelastic scattering can be damaging to the samples because it leads to ionisation damage including chemical bond rearrangements, X-ray emission and generates free radicals, thereby limiting the resolution (Orlova and Saibil, 2011). Nonetheless, electrons that are inelastically scattered have lower energy in random directions and do not encode any structural information. Hence, biological specimens are usually imaged using lower doses of electrons ( $\sim 1 - 40$  electrons /  $\text{\AA}^2$ ) to reduce inelastic scattering events (Orlova and Saibil, 2011).

The second type of interaction is termed as elastic scattering (Figure 2.7Ai), where no energy is transferred to the specimen. As the incoming electron interacts with the atomic nuclei, the resulting scattered electron trajectory is deviated from its original path but retains the same amount of energy (Orlova and Saibil, 2011). Elastic scattering forms the basis of image formation on the transmission electron microscope. The scattering angles are inversely proportional to the distance between two scattering centres and these angles provide information of the spatial frequencies of the sample. High scattering angle reveals the high frequency detail present in that sample (Orlova and Saibil, 2011).

The scattered and unscattered electron rays result in the image contrast in electron microscopy. Amplitude contrast is observed when energy of the incident beam is partially absorbed during inelastic interactions (Figure 2.7Bi), whereas phase contrast refers to the change in phases after the incident beam of electrons collide with the specimen without affecting its intensity (Figure 2.7Bii). For instance, during negative stain imaging, the uranyl acetate stain inelastically scatters electrons, these electrons hit stained regions at high

angles which are removed by the objective aperture. As uranyl acetate is a heavy element which scatters significantly, it causes the stained regions of the grid to appear black. However, the amplitude contrast of biological molecules is very weak because its light atomic composition (H, O, N and C) rarely absorbs electrons but deflect them.



**Figure 2.7 Electron Scattering and Image Contrast**

Ai: Electron elastic scattering of the incident primary electrons with specimen. Aii: Inelastic scattering events between the incident primary electron and specimen. Bi: Schematics illustrating amplitude contrast where the particle absorbs part of the incoming beam. Bii: Schematics illustration phase contrast where the particle changes the phase of the incoming beam.

As the exit wave of electrons are deflected at different angles, the objective lens focuses all unscattered and scattered beams, creating different path lengths that in turn produce phase contrasts and improve image features. The phase contrast can be further increased through spherical aberration or

defocusing of the microscope. Defocusing causes further phase shift at the detector and lead to the constructive and destructive interference between scattered and unscattered components of the electron beam (Orlova and Saibil, 2011).

The phase contrast can be visualised by combining the wavelengths of the scattered and unscattered waves to generate amplitude contrast. Together with the spatial frequencies of the electron beams, they provide information of how much signal has been transferred into the image. However, information of some spatial frequencies is missing due to the constructive and destructive interference between waves with different path lengths. In addition, the amplitudes of the resulting waves alternate between positive and negative phase contrasts. The sinusoidal function of contrast plotted against spatial frequency is known as the contrast transfer function (CTF) (Orlova and Saibil, 2011). Information at higher spatial frequencies describe the finer details in the image whereas low spatial frequencies describe the overall shape of the sample. The CTF can be represented by a power spectrum or Fourier Transform of a TEM micrograph. The effect of CTF can be seen in the Thon rings that represent the relationship between contrast and spatial frequency where light ring corresponds to maximum contrast transfer, and the dark Thon rings represent no transfer of signal. The CTF of an electron microscope is damped at high spatial frequencies, which is known as the envelope function. This is due to a range of imperfections in the microscope, such as partial spatial coherence of electron gun, lens aberrations, and microscope instabilities. Defocusing the microscope provides greater information transfer at lower spatial frequencies, therefore during data collection, a range of different defocus steps are collected to cover a larger spectrum of spatial frequencies. Further CTF correction can be performed by “phase-flipping” where the negative contrast frequencies are inverted into positive amplitudes (Orlova and Saibil, 2011). Newer version of processing software like RELION-3 determines and refines the CTF parameters and corrects for CTF during this refinement process (Zivanov *et al.*, 2018). Overall, the CTF describes the

modifications of the image of a sample due to the aberrations of the electron microscope.

## 2.8.5 Image Processing and Analysis

### 2.8.5.1 Alignment and Classification of Particles

Following the collection of movies and images from the electron microscope and alignment of individual frames to produce correct micrograph images. The particles present in each image must be selected, either manually or automatically by various software packages, such as SPHIRE-crYOLO (Wagner *et al.*, 2019) and RELION (Scheres, 2012b). The particles are then classified and averaged in order to improve the signal-to-noise ratio. Poor quality particles and contamination can be removed during the classification processes, allowing for more accurate analysis of the protein structural features.

In earlier approaches, particle alignment, following particle selection or ‘picking’, is performed using cross-correlation function (CCF), where the consistency or dependency between two values are compared to determine the similarity between two particles. The particles are normalised and aligned with the reference image so that the particles are rotated and shifted to a position where they are in similar orientations (Orlova and Saibil, 2011).

Particles can be further grouped into different stacks (class averages) based on their features to improve signal-to-noise ratios. However, pairwise comparison of each coordinate is very computationally intensive due to high dimensionality. This can be avoided using a technique from multivariate statistical analysis (MSA) known as the principal component analysis (PCA). This technique was employed by earlier software programs such as IMAGIC (Van Heel *et al.*, 1996) and EMAN (Ludtke, 2016). In this MSA approach, the pixels of an image form  $n$ -dimensional space vectors. As each pixel is a new axis in  $n$ -dimensional space, this reduces each image down to a point. PCA is used to extract a low dimensional feature set (major components) from these reduced data, which accounts for most of the variations in the data. This

feature set is then used to compare and sort the particles. The sorted particles can then be grouped either by hierarchical classification, where all the data is initially considered as one cluster, and separated into smaller groups based on their dissimilarity; or by K-means clustering, where some data are randomly selected as seeds, so the remaining particles are assigned to their nearest clusters and iteratively refined to minimise the sum of intra-cluster variations (Orlova and Saibil, 2011). These approaches allowed for the definition of different classes present in the data set, with averaging of each class producing class sum images with had a higher SNR and allowed for analysis, in two-dimensions, of the protein structure.

Hierarchical classification and K-means clustering were the conventional methods used before the implementation of the maximum-likelihood (ML) classification in RELION (Scheres, 2012a, 2012b). In RELION, during classification, a random subset of reference particles is initially used to generate the first reference images to which the whole dataset is aligned to. Importantly, the ML approach allows for the probability of class assignment to each of the given reference images to be calculated for each particle, with the particle then contributing to each class weighted by this calculated probability during alignment and averaging. This allows for a more robust data processing as, particularly in the early stages of classification where references are often of poorer quality, particles are not committed to any one class following comparison, as compared to the CCF approach employed in earlier software suites such as EMAN. CCF was used to compare particles and references, and the particle is assigned to the class with the highest CCF value. This is an important distinction as by spreading the particles over multiple classes (termed marginalisation), using the ML approach can reduce the inappropriate commitment of a particle to a particular class caused by noise and poor reference quality in early stages of the classification. This can help to improve resolution of the final classes and prevent incorrect classes being calculated due to error introduced at earlier stages in the classification. Following each round of classification, the particle images are then averaged, generating two-dimensional class averages, which are used as a new set of references for a

new round of classification in an iterative process which iterates several times until the production of the final classes before proceeding to 3D reconstruction (Sigworth, 2015).

#### 2.8.5.2 Three-Dimensional Reconstruction and Refinement

The relative orientations of each of the generated 2D class averages need to be determined to reconstruct a three-dimensional model. This can be achieved either through an experimentally based approach of taking images in pairs at two different angles, or through computational based approach by using the projection theorem. The projection theorem states that the Fourier transform (FT) of a 2D projection of an object is a slice through the centre of a 3D FT of the same object (Nogales and Scheres, 2015). Therefore, for each pair of 2D projections in Fourier space, there will be at least one 1D line in common in the reciprocal space. Once three images and three common lines are identified, the angles between each pair of 2D projections and their orientation in 3D FT can be assigned. The assigned orientation angles are called Euler angles. The original real-space 3D object can then be reconstructed by computing the inverse Fourier transform (Orlova and Saibil, 2011). However, there are also increasing numbers of *ab initio* model builders such as cryoSPARC which use a stochastic gradient descent with small subsets of particles to produce the initial model for refinement (Punjani *et al.*, 2017).

One of the major advances in modern EM data processing is the treatment of discontinuous heterogeneity in the particle set. To account for heterogeneity in the particle population and to improve the resolution of the 3D object, 3D classification is implemented. The principle of 3D classification is similar to that of classification of 2D images. Using maximum-likelihood classification,  $n$ -many initial 3D classes are created using the initial 3D reference and different random subsets of data. These different 3D classes, which at the first round differ only randomly, are projected and compete for the particles. The dataset is being clustered into distinct classes over an iterative process, allowing for the sorting of discontinuous heterogeneity such as partial occupancy of a particular subunit, or alternative conformations of the protein of interest

(Sigworth, 2015). To account for continuous heterogeneity, new approaches like multi-body refinement has been introduced to treat the flexible structure as rigid body models for focused refinement on multiple independently moving parts of the complex to correct for and characterise continuous motion in EM datasets (Nakane *et al.*, 2018).

3D reconstruction can be further refined in an iterative process through projection matching. Projection matching involves projecting the initial model across multiple different angles to generate reprojections, these reprojections are used as reference images to align the particles, producing updated angular assignments and hence improve the model and achieve the maximal resolution of the EM structure. The resolution of the 3D reconstruction can be calculated using the Fourier Shell Correlation (FSC). Each dataset in RELION is analysed using the ‘gold-standard’ method where the particles are split into two independent half sets and are refined against the same model independently. These two maps are then compared in the 3D Fourier space using Fourier shells. The resolution is estimated by the correlation between the two refined models, and the resolution of the map is determined when the correlation between them drops below 0.143 (Rosenthal and Henderson, 2003). This is a global measure of resolution and does not account for local differences in resolution, which can be estimated directly from the models using programs such as ResMap (Kucukelbir, Sigworth and Tagare, 2014).

## 2.8.6 Negative-Stain Electron Microscopy Data Collection

### 2.8.6.1 CTCF-CHD8 Complex Sample Preparation

1  $\mu$ M of CHD8 was incubated with 2  $\mu$ M CTCF for 30 minutes on ice in a total reaction volume of 200  $\mu$ l in the interaction buffer with 25 mM HEPES pH 8, 150 mM NaCl, 1 mM DTT and 5 % (v/v) glycerol. The complex was crosslinked with 0.05% glutaraldehyde (final concentration) for 10 minutes at room temperature and quenched with 100 mM Tris pH 8 (10  $\mu$ l). The sample was centrifuged at 12,000 g for 10 minutes and loaded onto Superose 6 Increase 10/300 GL column (GE Healthcare) pre-equilibrated with the interaction buffer. The fractions were analysed by SDS-PAGE and Native PAGE.

#### 2.8.6.2 CHD8-Nucleosome Complex Sample Preparation

1 µM of CHD8 was incubated with 2.2 µM nucleosome (with 52 bp linker) for 30 minutes on ice in a total reaction volume of 200 µl in the interaction buffer with 20 mM HEPES pH 7.5, 75 mM NaCl, 1 mM DTT, 3 mM MgCl<sub>2</sub> and 5 % (v/v) glycerol. Following incubation, the complex was added to a glycerol gradient where the glycerol content for each layer ranged from 10 % to 40 % (v/v) (Table 2.12) diluted in the interaction buffer. The sample was loaded onto an ultracentrifuge and spun for 18 hours at 6900 g at 4 °C. This step separated different species based on their weight into different glycerol layers. The glycerol gradient was carefully fractionated into twenty-five 200 µl aliquots. The aliquots were analysed using SDS-PAGE and Native PAGE. The fraction that showed homogenous CHD8-nucleosome sample was crosslinked with 0.05% glutaraldehyde for 5 minutes prior to loading onto the grids.

Glycerol Content (v/v)	Volume (µl)
10%	500
15%	500
20%	1000
25%	1000
30%	1000
40%	700

**Table 2.12 Glycerol Percentage Used for the Glycerol Gradient Experiment**

#### 2.8.6.3 Grid Preparation

Copper Quantifoil grids (R1.2/1.3) coated with a carbon film (Quantifoil Microtool GmbH) were glow-discharged with the easiGlow Glow Discharge Cleaning System (PELCO) for 60 seconds at 15 mA. This step removes surface contamination, and the grid becomes negatively charged and hydrophilic to allow the biological specimen to spread evenly across the grid. Next, 3 µl of sample was pipetted onto the grid and incubated for 1 minute. The excess liquid was removed with a cellulose blotting paper and the grid was washed three times in water and once in 2% uranyl acetate stain before staining for an additional 30 seconds in 2% uranyl acetate. Excess stain was

removed with the blotting paper and the grid was air dried before it was placed into the electron microscope.

#### 2.8.6.4 Screening and Data Collection

Negative stain grids were screened at the Tecnai T12 with a LaB<sub>6</sub> filament at an accelerating voltage of 120 kV. The images were collected with a 4K x 4K CCD F415 detector (Tietz). Large set of data collections were performed by Dr. Fabienne Beuron at the Tecnai F20 with a field emission gun and an accelerator voltage of 200 kV. 4K x 4K CCD F415 detector (Tietz) was used and the particles were detected at a magnification of 50,000 x with an electron dose of 100 e<sup>-</sup>/ Å<sup>2</sup>. The resulting pixel size was 1.73 Å. The collection was done automatically with EM-Menu (TVIPS GmbH). Micrographs were analysed using the Relion 3.0 software (Scheres, 2012b).

### 2.8.7 Cryo-Electron Microscopy Data Collection

#### 2.8.7.1 CTCF-CHD8 Complex Sample Preparation

1 µM of CHD8 was incubated with 2 µM CTCF for 30 minutes on ice in a total reaction volume of 200 µl in the interaction buffer with 25 mM HEPES pH 8, 150 mM NaCl and 1 mM DTT. The complex was crosslinked with 0.05% glutaraldehyde (final concentration) for 10 minutes at room temperature and quenched with 100 mM Tris pH 8 (10 µl). The sample was centrifuged at 12,000 g for 10 minutes and loaded onto Superose 6 Increase 10/300 GL column (GE Healthcare) pre-equilibrated with the interaction buffer. The fractions were analysed by SDS-PAGE and Native PAGE.

#### 2.8.7.2 Grid Preparation

Quantifoil Copper R2/2 400 mesh (Quantifoil Microtool GmbH) grids were coated with a carbon film (Quantifoil Microtool GmbH). The grids were glow-discharged with the easiGlow Glow Discharge Cleaning System (PELCO) for 20 seconds at 15 mA. The grid was mounted inside the Vitroblot Mark IV (FEI) and 2.5 µl of sample were pipetted onto the grid, incubated for 60 seconds and blotted for 1.5 seconds. The grids were then plunge frozen into liquid ethane and transferred to liquid nitrogen tank for storage.

### 2.8.7.3 Data Collection

Data was collected using the Glacios™ Cryo-TEM (Thermo Fisher Scientific) in house operated at 200 kV connected to the Falcon 3EC Direct Detector (Thermo Fisher Scientific). The magnification was set to 150,000 x and the pixel size was 0.94 Å. The defocus range was set from -1.5 to -3.5 µm at 0.5 µm incremental steps. 850 micrographs were taken with a total electron dose of 33.75 e<sup>-</sup>/Å<sup>2</sup>. Relion 3.1.1 (Scheres, 2012b) and crYOLO 1.5.6 (Wagner *et al.*, 2019) were used to process and analyse the dataset.

# 3 Purification and Characterisation of Human CTCF

CCCTC-binding factor (CTCF) is a ubiquitously expressed and highly conserved zinc finger protein. Its role as a transcription factor has been widely explored (Phillips and Corces, 2009). Recent publications mapped the CTCF-binding sites and revealed the importance of CTCF at chromatin domain boundaries (Ong and Corces, 2014; Rao *et al.*, 2014; Ren *et al.*, 2017). Through recent advancement in ChIP-seq analysis, CTCF has been shown to be one of the most fundamental chromatin architecture protein that mediates chromatin looping and organises the three-dimensional eukaryotic genome (Ong and Corces, 2014). This chapter discusses the purification trials of CTCF, and biochemical studies that were carried out to explore its diverse role in chromatin regulation.

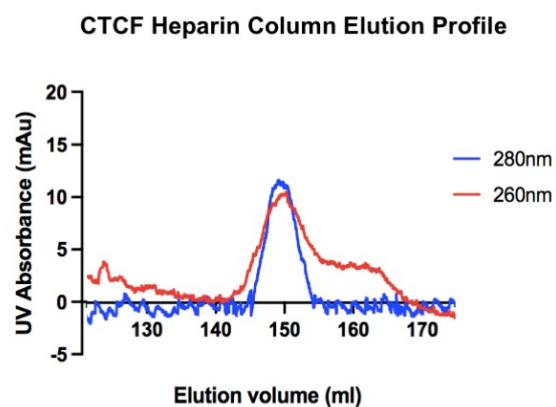
## 3.1 Purification of Recombinant Human CTCF

### 3.1.1 Purification Optimisation Trials of CTCF

To date, a protocol for the expression and purification of full length CTCF that is free of DNA contamination has not been established. As one of the most important transcription factors and genome architectural protein (Pugacheva *et al.*, 2020), it is crucial to further understand its role in genome organisation through *in vitro* assays. Therefore a functional, recombinantly purified CTCF is the foundation of these assays. The main reasons why it has been challenging to express and purify CTCF are because of its highly flexible and unstructured N- and C-terminal domains, and its strong binding affinity towards DNA (Martinez and Miranda, 2010).

N-terminal HA-tagged and C-terminal twin-Strep tagged human CTCF was expressed for baculovirus in High5 cells for four days prior to harvesting. In the first purification trial, filtered lysate was loaded onto Strep affinity column, and then passed through heparin column to remove DNA contaminants.

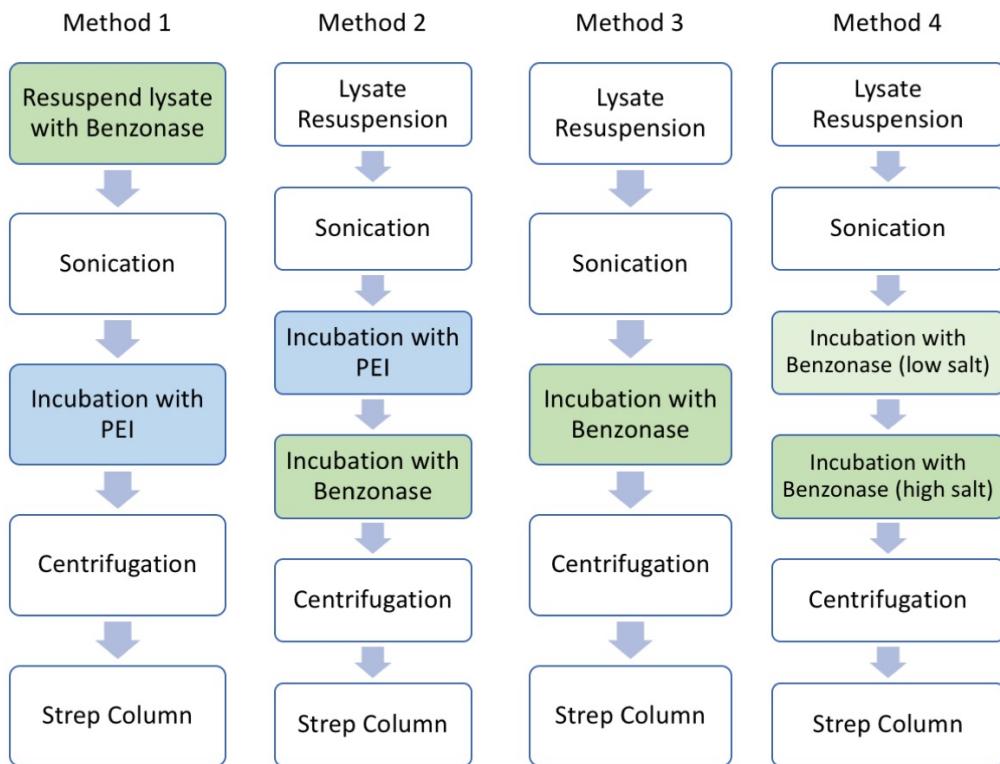
However, by comparing the ratio between the UV absorbances at 280 nm and 260 nm post heparin column, DNA was still present in the eluted CTCF-containing fractions (Figure 3.1). This suggested a more rigorous DNA removal method was needed.



**Figure 3.1 First CTCF Purification Trial**

CTCF elution profile of heparin column without any DNA removal treatments. The UV absorbance ratio between 260 nm and 280 nm was close to 1 hinting a strong level of DNA contamination.

To remove excess DNA contaminants, the lysate was subjected to four different protocols before loading onto the StrepTrap HP affinity column and the level of DNA contamination was analysed by EMSA or the ratio between 280 nm and 260 nm of the protein peak from heparin column to identify which method was the most efficient (Figure 3.2). The first two methods involved the use of polyethylenimine (PEI), which has been shown to precipitate DNA and therefore eliminates DNA in the downstream purification steps (Cordes, Sims and Glatz, 1990). Whereas the other two methods involved incubation with benzonase only.



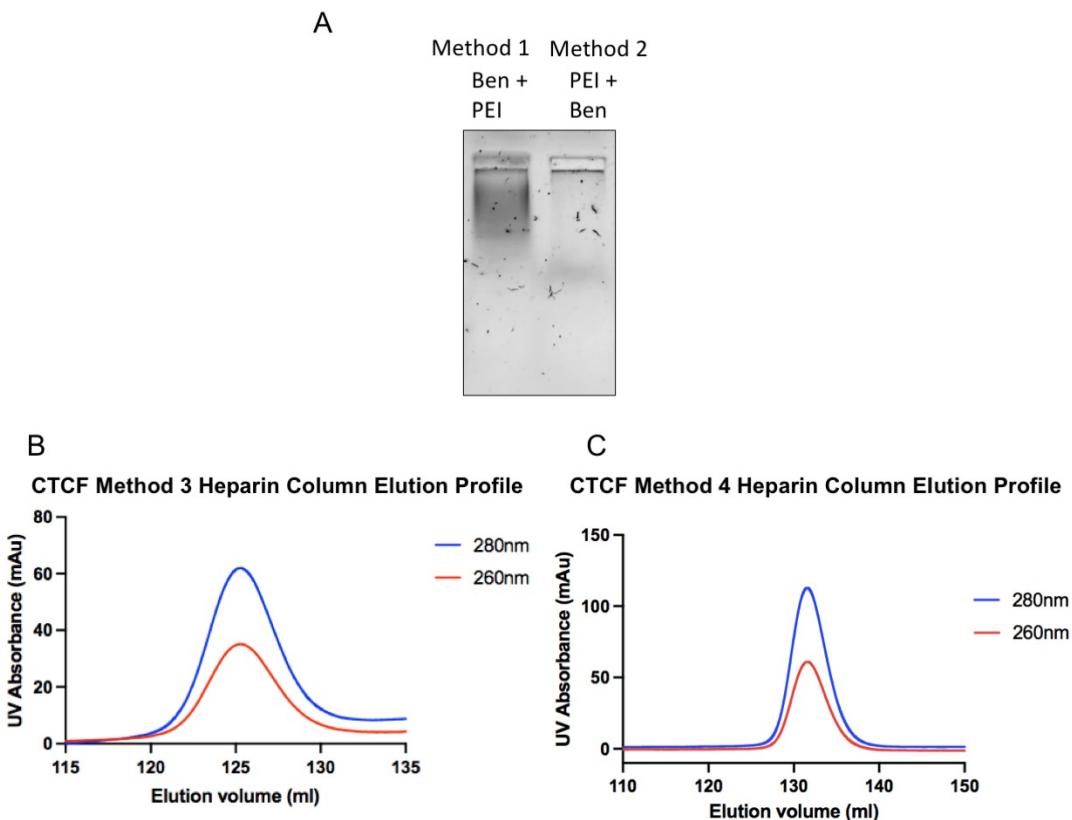
**Figure 3.2 Four Methods Used to Remove DNA Contamination**

A flow diagram illustrating the steps for each of the four methods used to remove DNA contamination during CTCF purification. Steps involving benzonase and polyethylenimine (PEI) are highlighted in green and blue, respectively.

In the first method, the cell pellet was resuspended in lysis buffer supplemented with benzonase. The lysate was sonicated and centrifuged. 0.4% (v/v) PEI was added to the supernatant and incubated for one hour to condense DNA before centrifugation to remove precipitants. The filtered lysate was then loaded onto the affinity column. The second method was similar to the first method, except benzonase was not added prior to sonication, but rather after PEI treatment. The sample was incubated with benzonase for an extra hour at 4 °C on a roller. Post-benzonase treatment, the sample was subjected to centrifugation and the filtered lysate was analysed to examine the level of DNA contamination. The reason why the sequence of benzonase and PEI treatment might affect the DNA removal efficacy was because of their specificity towards different lengths of DNA. Benzonase nuclease digests nucleic acids to 2-5 bases in length whereas PEI condenses DNA that are over hundreds of base pairs long. Therefore, PEI might be more effective before

benzonase has digested long nucleic acids into smaller fragments. Samples of the treated lysates were loaded onto an agarose gel and stained with ethidium bromide to understand the level of DNA contamination post PEI and benzonase incubation. The results showed that the sequence of incubation affected the level of DNA remained in the sample and incubating with PEI first was more effective in removing excess DNA (Figure 3.3A). However, a substantial portion of CTCF was also precipitated by PEI, hence the amount of protein eluted from the affinity column was very low (data not shown).

In the third method, after sonication, benzonase was incubated with the sample for 1 hour at 4 °C on a roller before the lysate was centrifuged and filtered to be loaded onto the column, the salt concentration was kept constant at 500 mM NaCl throughout the incubation. In the last method, the pellet was resuspended in a lysis buffer with lower salt content (300 mM NaCl). The sample was sonicated and benzonase was added to the sample and incubated for 30 minutes at 4 °C on a roller. The salt concentration was increased to 600 mM NaCl by supplementing with 5 M NaCl and incubated further for 30 minutes. Increasing the salt concentration might disrupt the non-specific interactions between DNA and CTCF and hence reduce the amount of DNA contaminants. Comparing the last two methods, the heparin column elution profiles revealed that both methods were efficient in reducing contaminant DNA as the ratios between UV absorbances at 260 nm and 280 nm was reduced to ~ 0.6 (Figure 3.3B, C). Although both methods were able to reduce DNA contamination, method four was used for future purifications because the activity of benzonase has been reported to be more efficient in lower salt concentrations, thus the level of DNA contamination can be further reduced.



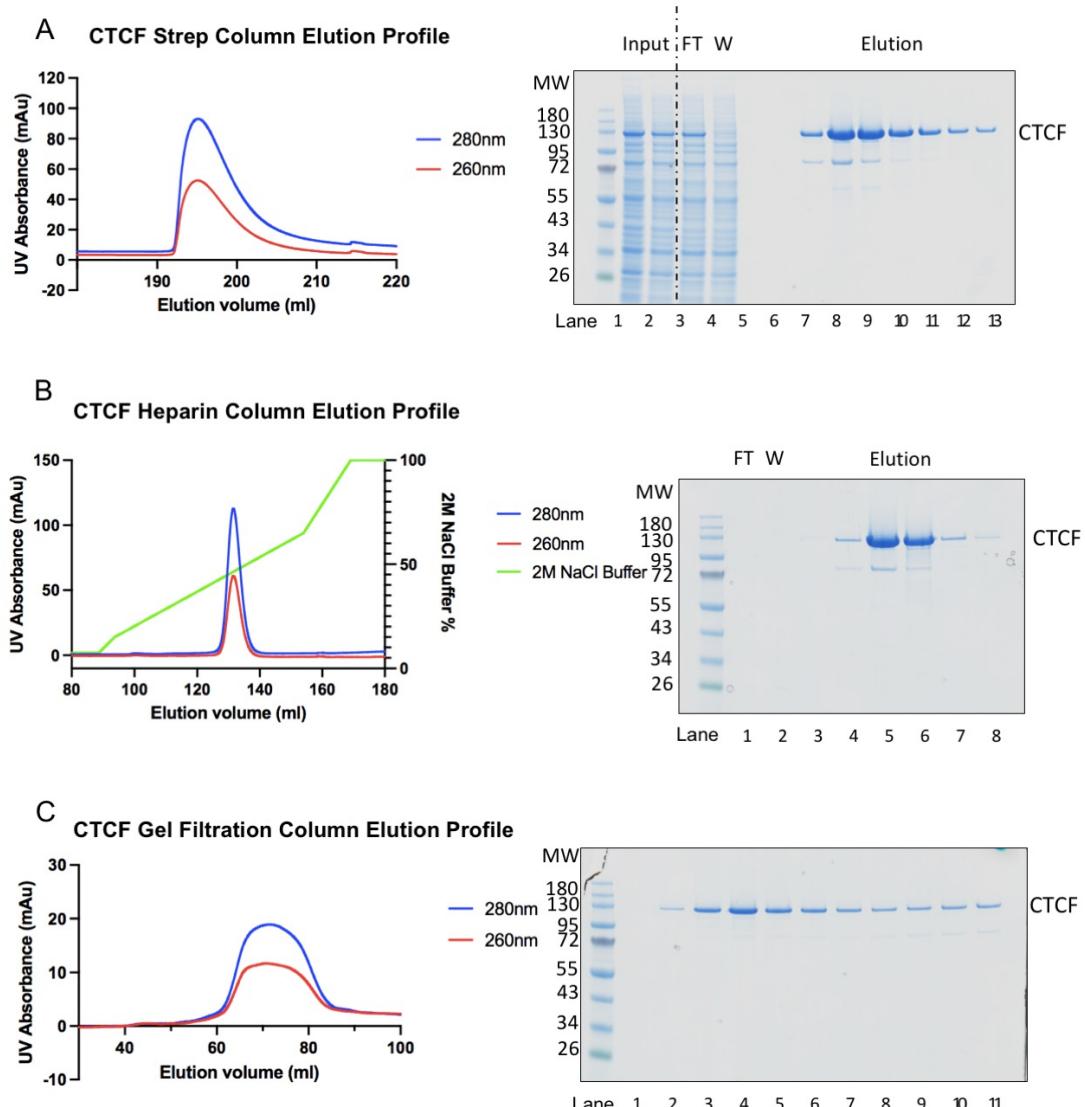
**Figure 3.3 CTCF Purification Optimisation**

A: EMSA gel showing the effectiveness of PEI and benzonase in removing excess DNA contaminants in lysates with over-expressed CTCF. B: Heparin column of CTCF treated with method 3, where it was incubated with benzonase prior to purification. C: Elution profile of the heparin column of CTCF treated with method 4, where it was incubated with benzonase in two different salt concentrations.

### 3.1.2 Optimised Purification Protocol of Human CTCF

The cell pellet was resuspended in lysis buffer (salt concentration: 300 mM NaCl) and underwent sonication. The sample was incubated with benzonase (~ 2500 U) for 30 minutes and supplemented with 5 M NaCl to bring the final salt concentration to 600 mM NaCl. The filtered supernatant was loaded onto a StrepTrap column (Figure 3.4A), followed by a Heparin column to remove nucleic acid contaminants (Figure 3.4B). The protein was eluted in a gradual gradient up to 2 M NaCl. Target protein was eluted close to 800 mM NaCl. The fractions were pooled and purified via Superose 6 16/600 column. A large peak eluted across 60 to 80 ml was observed (Figure 3.4C). It has been suggested

that the terminal domains of CTCF are monomeric in solution, however these residues are predominantly extended and disordered and thus resulting in large hydrodynamic radius (Martinez and Miranda, 2010). *In vitro* studies have also shown that full length CTCF forms dimers or multimers (Pant *et al.*, 2004; Yusufzai *et al.*, 2004), hence explaining why the CTCF protein eluted over a large volume in the gel filtration chromatography. The respective fractions were analysed by SDS-PAGE and pooled, concentrated to 12 mg/ml, flash frozen and stored at -80 °C. The yield was estimated to be 8 mg per 1 L of High5 insect cells.



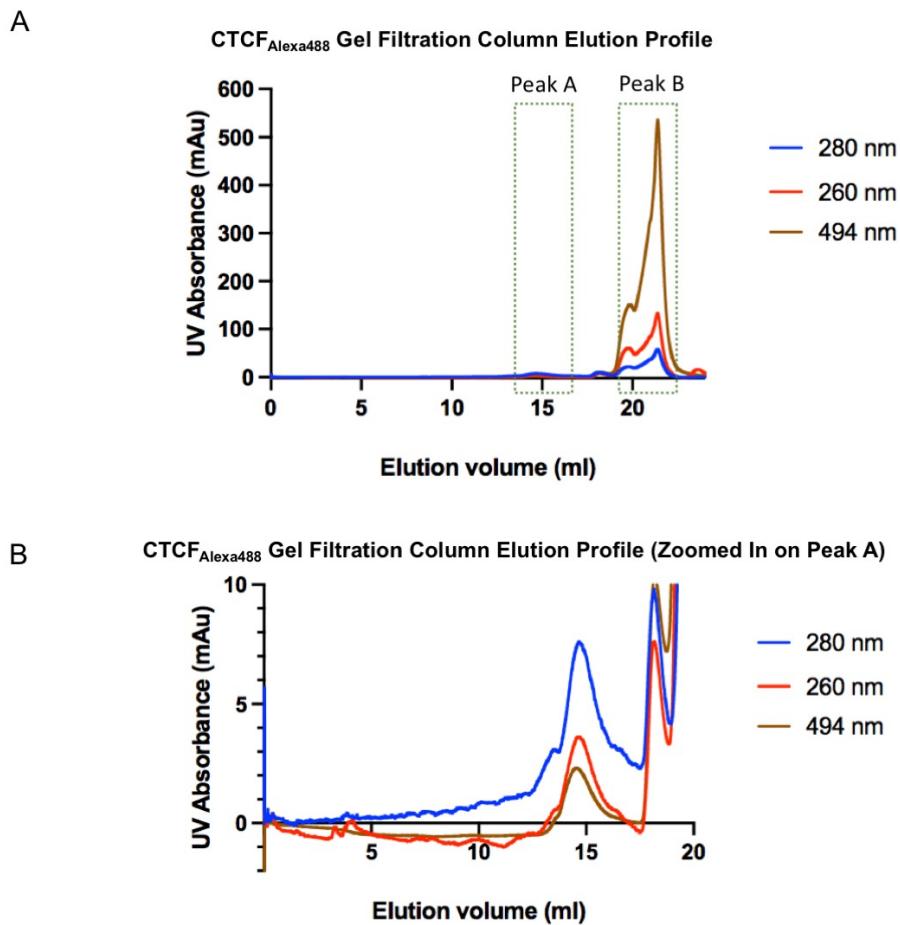
#### **Figure 3.4 Optimised Purification Protocol of CTCF**

A: Elution profile and SDS-PAGE analysis of the Strep column. Lane 1 corresponded to the lysate after incubation with benzonase and 600 mM NaCl, and lane 2 corresponded to the supernatant after centrifugation. FT = flow through, W = wash. B: Elution profile and SDS-PAGE analysis of the fractions from Heparin column. FT = flow through, W = wash. C: Gel filtration chromatography profile and SDS-PAGE analysis of the eluted fractions.

#### **3.1.3 Purification of ybbR-tagged CTCF**

N-terminal HA tagged and C-terminal ybbR-StrepII tagged CTCF was conjugated to a fluorophore for easy visualisation in *in vitro* experiments. ybbR is a 11-residue peptide tag that acts as a substrate for Sfp phosphopantetheinyl transferase and labels the protein with coenzymeA (CoA)-fluorophore conjugates (Yin *et al.*, 2005).

Human CTCF<sub>ybbR</sub> was expressed in High5 cells for four days prior to harvesting. CTCF<sub>ybbR</sub> lysate was treated with benzonase as mentioned in section 3.1.2 and purified using StrepTrap and Heparin columns. Post-heparin column, the protein was kept on ice while the sulphhydryl group of coenzymeA was conjugated with fluorophore Alexa488 via a maleimide reactive group. Alexa488 was incubated with CoA at room temperature for an hour and the reaction was quenched with an excess of dithiothreitol. The conjugated CoA-Alexa488 was then mixed with purified CTCF<sub>ybbR</sub> with the addition of Sfp (provided by Dr. Erin Cutts) to catalyse the modification. The reaction mixture was loaded onto a gel filtration column Superose 6 10/300 Increase to remove Sfp and excess unconjugated CoA-Alexa488. The gel filtration was monitored at three different UV wavelengths at 260 nm, 280 nm and 494 nm. Two peaks were observed (Figure 3.5A), where the first peak (eluted near 14.5 ml) corresponded to CTCF conjugated to Alexa488 and excess unlabelled CTCF (Figure 3.5B), whereas the second peak (eluted close to 22 ml) was the excess CoA-Alexa488 fluorophore and Sfp. The efficiency of the conjugation was determined by comparing their absorbances at 280 nm and 490 nm using the Nanodrop. The labelling efficiency was around 30%.



**Figure 3.5 Gel Filtration Elution Profile of CTCF<sub>Alexa488</sub>**

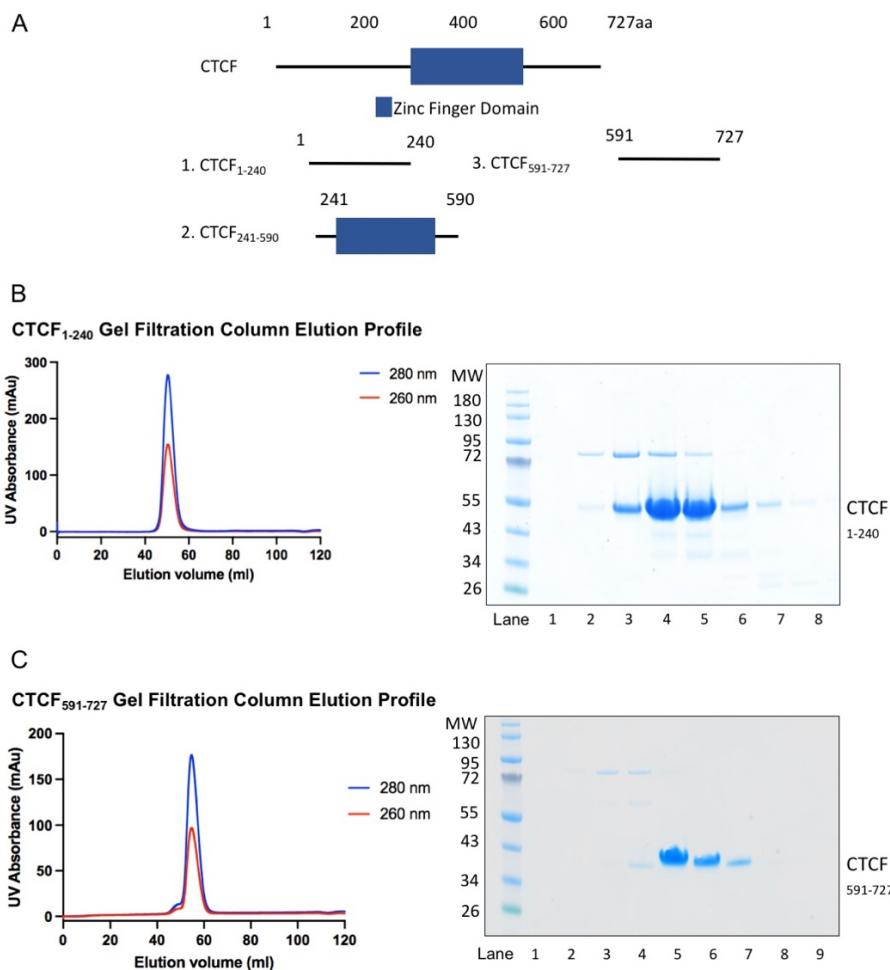
A: Gel filtration elution profile of CTCF conjugated to CoA-Alexa488. Two peaks were observed in this purification. B: Elution profile of CTCF conjugated to CoA-Alexa488 that was zoomed in on the peak that corresponded to the purified CTCF<sub>Alexa488</sub> protein (peak A).

### 3.1.4 Purification of N- and C-terminal Domains

Previous publications showed successful purification of the N- and C-terminal domains of CTCF (Martinez and Miranda, 2010). In addition, the structure of the zinc fingers domain (residues 294 – 518) were solved by X-ray crystallography (Hashimoto *et al.*, 2017). Expression and purification of these domains were designed to understand how each of the domains might contribute towards protein interactions involved in genome organisation.

All three constructs (CTCF<sub>1-240</sub>, CTCF<sub>241-590</sub>, CTCF<sub>591-727</sub>) were HA-tagged at the N-terminus and StrepII-tagged at the C-terminus (Figure 3.6A). They were

expressed in High5 insect cells for four days prior to harvesting. Unfortunately, the construct that comprised of the eleven zinc fingers CTCF<sub>241-590</sub> was not well-expressed which might be due to the insufficient flanking regions around the zinc fingers and therefore could not proceed to the purification stage. For the remaining two constructs, CTCF<sub>1-240</sub> and CTCF<sub>591-727</sub>, the insect cell pellets were resuspended in lysis buffer supplemented with protease inhibitors and benzonase, the lysate was sonicated and centrifuged. The supernatant was filtered and applied to StrepTrap HP column, followed by Q HP column and passed onto HiLoad Superdex 75 pg 16/600 column (Figure 3.6B, C). The eluted fractions were analysed with SDS-PAGE. CTCF-containing fractions were pooled, concentrated and stored at -80 °C. The yield of CTCF<sub>1-240</sub> and CTCF<sub>591-727</sub> were 12 mg and 8 mg, respectively, from 1 L of High5 insect cells. These two constructs were used to probe the region of interaction with CHD8 in chapter 6.1.2.2.



### **Figure 3.6 Purification of CTCF<sub>1-240</sub> and CTCF<sub>591-727</sub>**

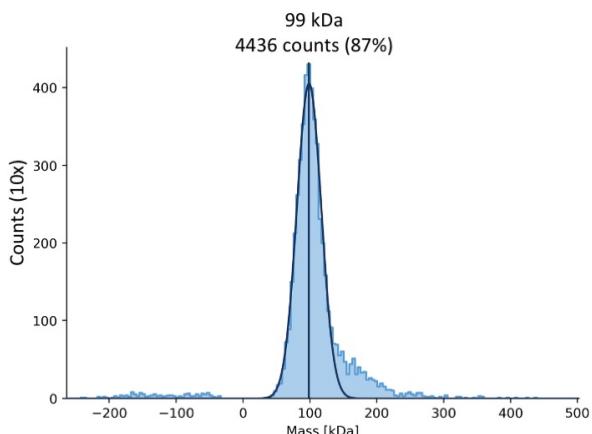
A: Schematic representation of the three new CTCF constructs. B: Elution profile of CTCF<sub>1-240</sub> from gel filtration column. The fractions of the peak were analysed by SDS-PAGE. C: Gel filtration chromatogram of CTCF<sub>591-727</sub> and SDS-PAGE analysis of the peak fractions.

## 3.2 CTCF Shows Concentration-Dependent Oligomerisation

### 3.2.1 Mass Photometry Analysis

CTCF dimerisation has been speculated to be the organising principle behind genome organisation because CTCF dimers mediate chromatin loops together with cohesin, another crucial genome architectural protein (Phillips and Corces, 2009; Pugacheva *et al.*, 2020). While self-association of CTCF has been observed in yeast two-hybrid and co-immunoprecipitation experiments, the domain which CTCF self-associates with was not conclusive and some publications suggested that the dimerisation of CTCF was DNA / RNA dependent (Pant *et al.*, 2004; Yusufzai *et al.*, 2004; Bonchuk *et al.*, 2020).

To understand the oligomeric state of CTCF in solution *in vitro*, CTCF was analysed using mass photometry. Mass photometry is a technique that measures protein masses in their native state through interference reflection microscopy. CTCF was diluted 30-fold from stock solution (12 mg/ml). The estimated molecular weight of CTCF (with HA- and StrepII- tag) from ExPASy ProtParam (Gasteiger *et al.*, 2003) was 93 kDa. Based on the results from the mass photometry, the measured molecular weight of the CTCF was around 99 kDa, hence suggesting that CTCF at 0.4 mg/ml was monomeric in solution (Figure 3.7). It was also important to point out the trail of the peak that was close to 200 kDa. Although the number of counts at this molecular weight were not substantial to fit a Gaussian curve, it showed that a small portion of CTCF might exhibit as dimers at the measured concentration. To ensure accurate measurement of protein samples in mass photometry, the concentration of samples used must be kept low and therefore it was not possible to explore whether or not CTCF might dimerise in a concentration-dependent manner using mass photometry.

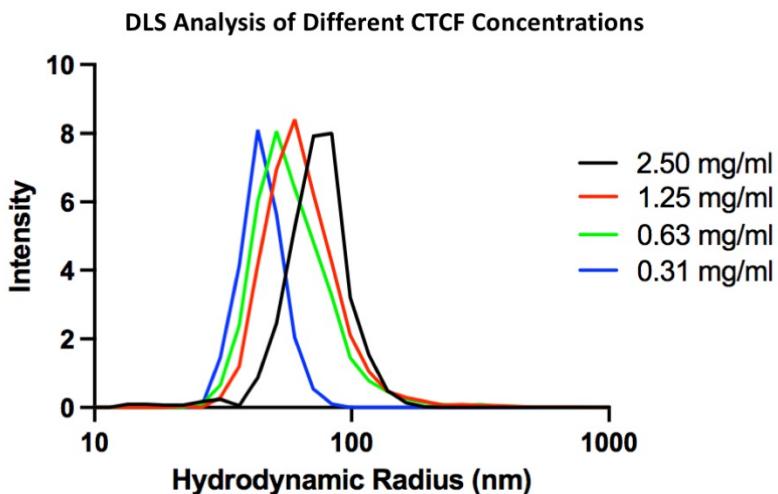


**Figure 3.7 Mass Photometry Analysis**

Full length CTCF at 0.4 mg/ml was analysed by mass photometry. The measured size of each molecule adhered to the glass cover slip was recorded. The number of counts was plotted against the measured molecular weight. 87% of molecules measured had a molecular weight of 99 kDa.

### 3.2.2 Dynamic Light Scattering Analysis

To understand if CTCF oligomerises in a concentration-dependent manner, dynamic light scattering (DLS) was performed on four different concentrations of CTCF to observe the change in the species' hydrodynamic radius. DLS measures the scattered light from the protein molecules, and the change in the light scattering is correlated to the hydrodynamic radius of the protein molecules. From the DLS results, it showed that as the concentration of CTCF decreased, the hydrodynamic radius of the species decreased, suggesting smaller species were present at lower concentrations. Furthermore, the radius distributions plots (Figure 3.8) and the polydispersity (Table 3.1) confirmed that only one species was present at each concentration. This demonstrated that CTCF might be able to interact with itself. Unfortunately, it was not possible to correlate the measured hydrodynamic radius to its oligomeric state. Nonetheless, whether or not the dimeric or oligomeric forms of CTCF are present in the physiological environment requires further investigation.



**Figure 3.8 Dynamic Light Scattering Analysis**

Four different concentrations of CTCF in 2-fold serial dilution was analysed using the DLS machine. The hydrodynamic radius of CTCF at different concentrations were measured and plotted against its frequency of occurrence.

CTCF Concentration	Polydispersity	Hydrodynamic Radius
2.5 mg/ml	17.7%	77.15 ± 4.19 nm
1.25 mg/ml	21.3%	60.71 ± 8.10 nm
0.625 mg/ml	19.5%	53.6 ± 7.06 nm
0.313 mg/ml	16.5%	45.89 ± 0.56 nm

**Table 3.1 Hydrodynamic Radius and Polydispersity of Different CTCF Species**

The hydrodynamic radius and the polydispersity of each CTCF concentration were recorded.

### 3.3 DNA Binding Studies of CTCF

#### 3.3.1 Purified CTCF binds to DNA

To investigate whether the purified CTCF was functional, purified CTCF was incubated with CTCF DNA consensus sequence and analysed through EMSA on an agarose gel. The 37 bp DNA construct used in this experiment was comprised of the consensus sequence for CTCF binding predicted by Persikov and Singh (Persikov and Singh, 2014) and the 17 bp sequence used in the CTCF zinc finger crystallisation experiments (Hashimoto *et al.*, 2017) (Table 3.2). This sequence has high G/C rich regions. The forward and reverse

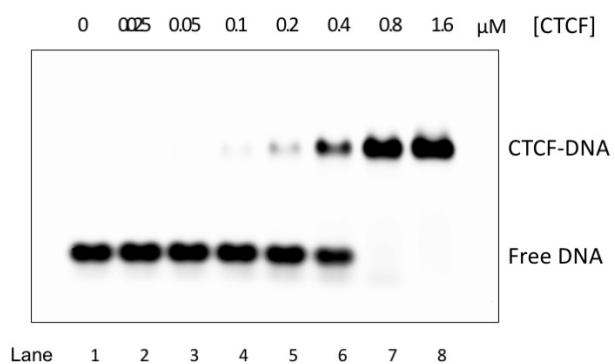
complementary DNA strands were synthesized separately and annealed through a temperature gradient. The forward strand was labelled with Cyanine-5 (Cy-5) fluorophore to allow better visualisation of the binding properties between CTCF and DNA.

	<b>Construct Name</b>	<b>DNA Sequence</b>
<b>12</b>	CTCF (37bp)	CCTCAC TAGCGCCCCCTGCTGGCCTCTGTGGGCAC TG

**Table 3.2 DNA Construct with CTCF-Consensus Binding Site**

37 bp DNA construct with CTCF consensus site was designed for CTCF DNA binding studies. Regions highlighted in green were the predicted CTCF binding sequences reported by Persikov and Singh (Persikov and Singh, 2014) and region in black was the sequence used in crystallisation studies by the Hashimoto group (Hashimoto *et al.*, 2017).

Increasing concentrations of CTCF were incubated with 0.1  $\mu$ M 37 bp consensus DNA, the samples were run on 0.5x TBE agarose gels and imaged with Typhoon imager. At 2:1 CTCF:DNA molar ratio, traces of CTCF-DNA complex were observed. As the concentration of CTCF increased, the intensities of the free DNA bands decreased, and a complete shift was observed when CTCF was in eight molar excess over the labelled DNA (Figure 3.9, lane 7). This experiment confirmed that the recombinantly purified CTCF was functional and able to bind to DNA.



**Figure 3.9 EMSA of CTCF-DNA Complex**

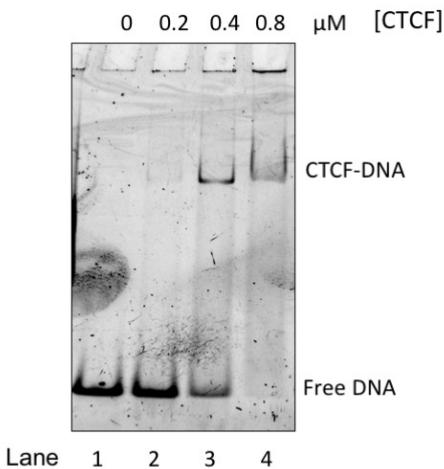
Different concentrations of CTCF were incubated with 0.1  $\mu$ M CTCF-consensus DNA. The numbers above the gel indicated the concentration of CTCF used in the respective reactions. The gel was visualised using Cy-5 filter on the Typhoon imager.

The DNA-binding activity of CTCF was also explored with another 37 bp DNA construct that did not include the CTCF consensus sequence or the crystallised sequence (Table 3.3). EMSA assay was performed with different concentrations of CTCF against 0.1  $\mu$ M DNA. The results suggested weak binding starting from two-fold molar excess of CTCF against DNA, and complete shift was observed at eight-fold molar excess similar to the observation from binding to consensus DNA (Figure 3.10). Although this showed that CTCF was able to bind to non-consensus DNA, further experiments will be required to corroborate the difference in the binding affinities between the consensus and non-consensus DNA constructs.

	Construct Name	DNA Sequence
13	Non-CTCF (37bp)	GGAAATCACCATAACGTGAAATGTCTTGGATTGG

**Table 3.3 DNA Construct Without CTCF-Consensus Sequence**

A new dsDNA construct was designed to understand the DNA binding properties of CTCF.

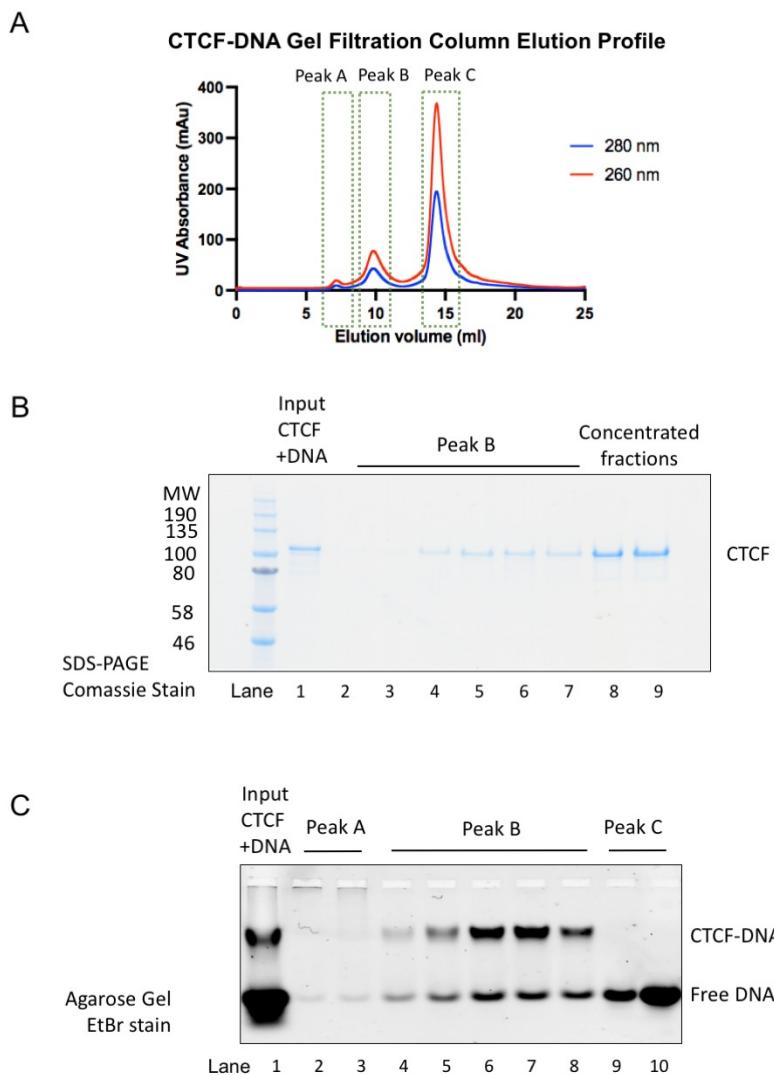


**Figure 3.10 CTCF Binds to Non-Consensus DNA**

Different concentrations of CTCF were incubated with 0.1  $\mu$ M DNA lacking the CTCF consensus site. The numbers above the gel indicated the concentration of CTCF used in the respective reactions. The gel was stained with ethidium bromide and visualised on Typhoon Imager.

### 3.3.2 CTCF-DNA Forms a Stable Complex Over Gel Filtration

Studies have shown that CTCF binds to DNA with strong affinity (Hashimoto *et al.*, 2017), therefore the next step was to test whether the purified CTCF formed a stable complex with DNA. CTCF was incubated with DNA for 30 minutes on ice prior to loading onto the Superdex 200 10/300 gel filtration column. Four molar excesses of DNA were incubated with CTCF to ensure all CTCF molecules were bound to DNA. Three peaks were observed from the gel filtration chromatogram (Figure 3.11A). Fractions from each of these peaks were analysed with SDS-PAGE (Figure 3.11B) and loaded onto an agarose gel, stained with ethidium bromide and visualised using a Typhoon scanner (Figure 3.11C). Peak A eluted before 7 ml, corresponding to the void volume and the results from the agarose gel showed that most of the protein was stuck in the well, suggesting it was mostly aggregated CTCF. Peak B eluted close to 10 ml, the gel filtration chromatogram suggested the presence of DNA as the ratio between the UV absorbance at 260 nm and 280 nm was close to 2. The EMSA gel showed the presence of CTCF-DNA complex as well as free DNA. However, DNA might be dissociated from the complex during the electrophoresis step which might be due to the shift of equilibrium and hence free DNA was also observed. The third peak, eluted at around 15 ml, corresponded to the excess of DNA used in the reaction, as only free DNA was observed in the EMSA gel. CTCF alone ran at near 12 ml (data not shown). Therefore, this analytical gel filtration experiment confirmed that CTCF and DNA formed a stable complex which was stable enough to remain bound over the gel filtration chromatography experiment.



**Figure 3.11 CTCF-DNA Formed a Stable Complex over Gel Filtration**

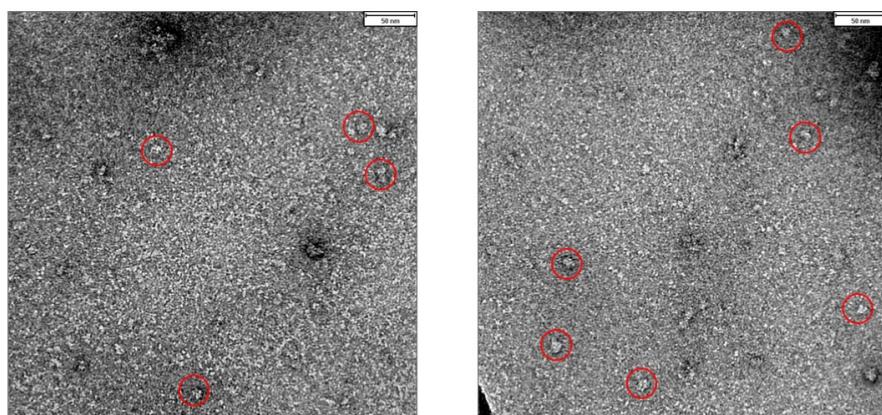
A: Elution profile of the CTCF-DNA analytical gel filtration. Three peaks were indicated with green boxes. B: SDS-PAGE analysis of the peak fractions eluted from the analytical gel filtration. Lanes 8 and 9 corresponded to two of the fractions pulled from peak B that were concentrated for negative stain analysis (section 3.3.3). C: Fractions of each peak were analysed with agarose gels and stained with ethidium bromide to visualise DNA.

### 3.3.3 CTCF-DNA Negative Stain Studies

Since CTCF and DNA (with CTCF consensus site) were able to form a stable complex over gel filtration, the complex was analysed using negative stain microscopy. The fractions eluted from the analytical gel filtration in section 3.3.2 were concentrated and analysed using negative stain electron microscopy. Two fractions from peak B (Figure 3.11A, B) were concentrated,

stained with 2% uranyl acetate and screened at the FEI T12 electron microscope.

The micrographs showed particles of globular shape that varied slightly in size and most of them were approximately 10 nm long in diameter (Figure 3.12). Potential CTCF-DNA samples were circled in red. However, the CTCF/DNA complex may have dissociated after it was stained with uranyl acetate. In addition, even after concentration, the number of particles on each micrograph was still quite low and the contrast between the protein particles and the background was relatively low, and so further structural characterisation was not possible. The poor contrast might be due to poor uranyl acetate staining. The presence of small molecules, contaminant and protein degradation products in the buffer might have also contributed to the high background.



**Figure 3.12 Negative Stain EM Micrographs of CTCF-DNA Complex**

Negative stain micrographs taken at the FEI T12 microscope. Potential CTCF-DNA particles were marked in red.

Further optimisation will be required to generate high-contrast micrographs. In addition to analysing the CTCF-DNA complex, it will also be intriguing to explore other CTCF complexes to reveal its structural information and its role in genome organisation.

## 3.4 Nucleosome Binding Studies

### 3.4.1 CTCF Interacts with Nucleosomes

Studies have suggested that CTCF is required to maintain the local nucleosome repeated array and some papers confirmed that the presence of CTCF and its binding sites alter the average distance between adjacent nucleosomes (Clarkson *et al.*, 2019; Owens *et al.*, 2019). Given the relationship between CTCF and nucleosome positioning, it was intriguing to explore if CTCF interacts with nucleosomes directly and potentially displaces or slides the nucleosomes into their respective new positions.

To investigate if CTCF interacts with nucleosomes, electrophoretic mobility shift assays were carried out with CTCF conjugated to Alexa488 fluorophore and end-positioned nucleosomes. The nucleosomes were purified in collaboration with Dr. Thangavelu Kaliyappan. Different concentrations of CTCF were incubated with 0.1 μM nucleosome on ice for 15 minutes. The nucleosome used for this study had a 52 bp linker after the 147 bp Widom 601 sequence (Table 3.4). The 52 bp linker did not have a CTCF consensus motif. Following incubation, the samples were loaded onto a commercial Novex Native 4-12% TBE gel and scanned using the appropriate filters.

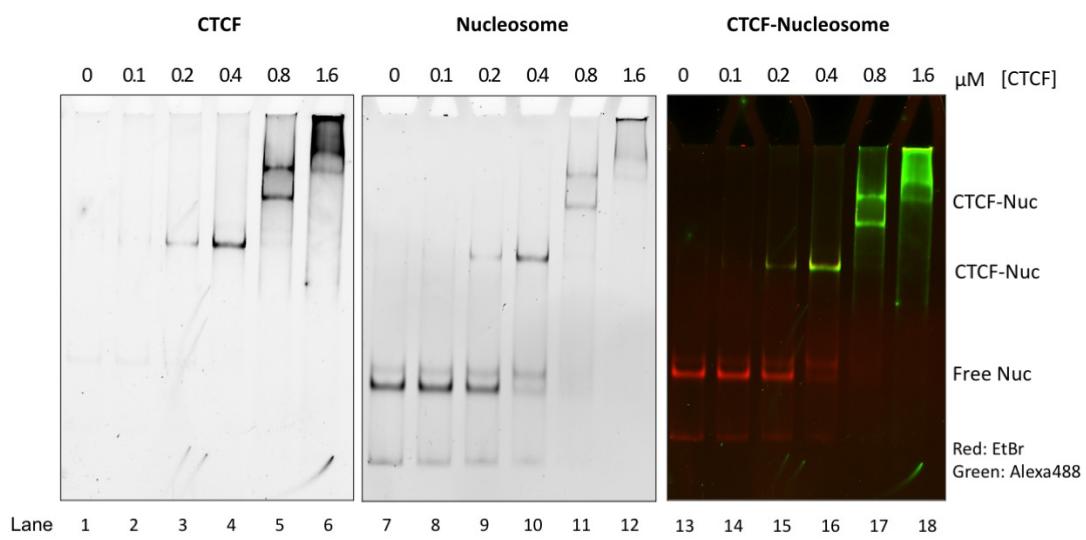
	<b>Construct Name</b>	<b>DNA Sequence</b>
<b>14</b>	Nucleosome _52bp	<b>CTGGAGAATCCCGGTCTGCAGGCCGCTCAATTGGTCGT AGACAGCTCTAGCACCCTAAACGCACGTACCGCGCTG TCCCCCGCGTTTAACCGCCAAGGGGATTACTCCCTAG TCTCCAGGCACGTGTCAGATATATACATCCTGTGGCAT AAGCGTGGTTCAATACCGGCGATAAAGGGTAAATTGAAAA CAGCG</b>

**Table 3.4 DNA Sequence Used for Nucleosome Assembly**

DNA comprised of the Widom 601 sequence and a 52 bp linker at the exit site was used to assemble the nucleosome for CTCF-nucleosome binding studies. Sequence highlighted in red refers to the Widom 601 sequence discovered by Lowary and Widom (Lowary and Widom, 1998).

CTCF showed binding to nucleosomes. Interaction between CTCF and nucleosome was observed starting from 2:1 CTCF:nucleosome molar ratio

and all nucleosomes were occupied when CTCF was in eight molar excess (Figure 3.13). As the concentration of CTCF increased to eight molar excess over nucleosomes, two new bands were observed. These two bands were positioned above the original CTCF-nucleosome shift observed in the previous lane 4. The higher band might be due to the oligomerisation of CTCF at higher concentrations, or it might also be possible that a few binding sites were present on the nucleosome.



**Figure 3.13 Interaction Between CTCF and Nucleosome**

Different concentrations of CTCF were incubated with 0.1  $\mu\text{M}$  nucleosome lacking CTCF consensus motif. The EMSA gel was visualised using two different filters. The individual channels for CTCF (lanes 1-6) and nucleosome (lanes 7-12) were displayed in black and white. The overlay of the two filters was shown in lanes 13-18, red bands corresponded to ethidium bromide-stained nucleosomes, whereas the green bands highlighted CTCF that were conjugated to Alexa488.

### 3.5 Summary

CTCF has an important role as an architecture protein and as a transcription factor to regulate genome organisation and gene expression (Ong and Corces, 2011). However, it has been challenging to express and purify human CTCF due to its unstructured N- and C-terminal domains, and its strong binding affinity towards DNA (Martinez and Miranda, 2010). An optimised purification protocol for human CTCF was presented in this chapter. The purification involved an incubation with benzonase at two different salt concentrations to

remove DNA contamination, followed by Strep affinity column, Heparin ion exchange chromatography and gel filtration. In addition, a CTCF construct conjugated to a fluorophore via a ybbR tag was also purified for visualisation in EMSA experiments.

Following the successful purification of CTCF, functional assays were carried out to understand the intrinsic properties of CTCF. Firstly, through dynamic light scattering analysis, it suggested that CTCF oligomerises in a concentration-dependent manner that might explain the large elution volume from the gel filtration purification. Secondly, the recombinantly purified CTCF was shown to be functional through DNA-binding studies with CTCF consensus DNA, yet the DNA binding activity of CTCF is not limited to sequences comprised of the CTCF motif. Next, CTCF and DNA (with CTCF consensus site) were able to form a stable complex over gel filtration and the complex was analysed using negative stain microscopy. Nonetheless the high background and poor contrast of the negative stain micrographs restricted the possibility of further structural characterisation and thereby further optimisation will be required to generate high-contrast micrographs. Lastly, CTCF has been reported to possess nucleosome positioning activity (Clarkson *et al.*, 2019) and therefore nucleosome binding studies were performed. The EMSA assay showed interaction between CTCF and nucleosome, and a supershift was observed hinting that several CTCF molecules might be bound to one nucleosome. In addition to nucleosomes, further investigation on the interactions between CTCF and other chromatin factors such as chromatin remodellers will be addressed in the following chapters.

# 4 Purification and Characterisation of Human ZNF143

Human ZNF143 is a zinc finger protein that was first identified as the transcriptional activator for RNA polymerases II and III. It is overexpressed in various types of cancers. Recent studies suggested that ZNF143 is enriched at the boundaries of topological-associating DNA domains (TADs) where it co-localises with known chromatin organisational proteins such as CTCF. It is recruited to gene promoters involved in chromatin interactions, therefore ZNF143 could contribute to genome organisation and regulation. To this end, as detailed in this chapter, ZNF143 was purified and subject to biochemical characterisation in order to gain a deeper understanding of its functions.

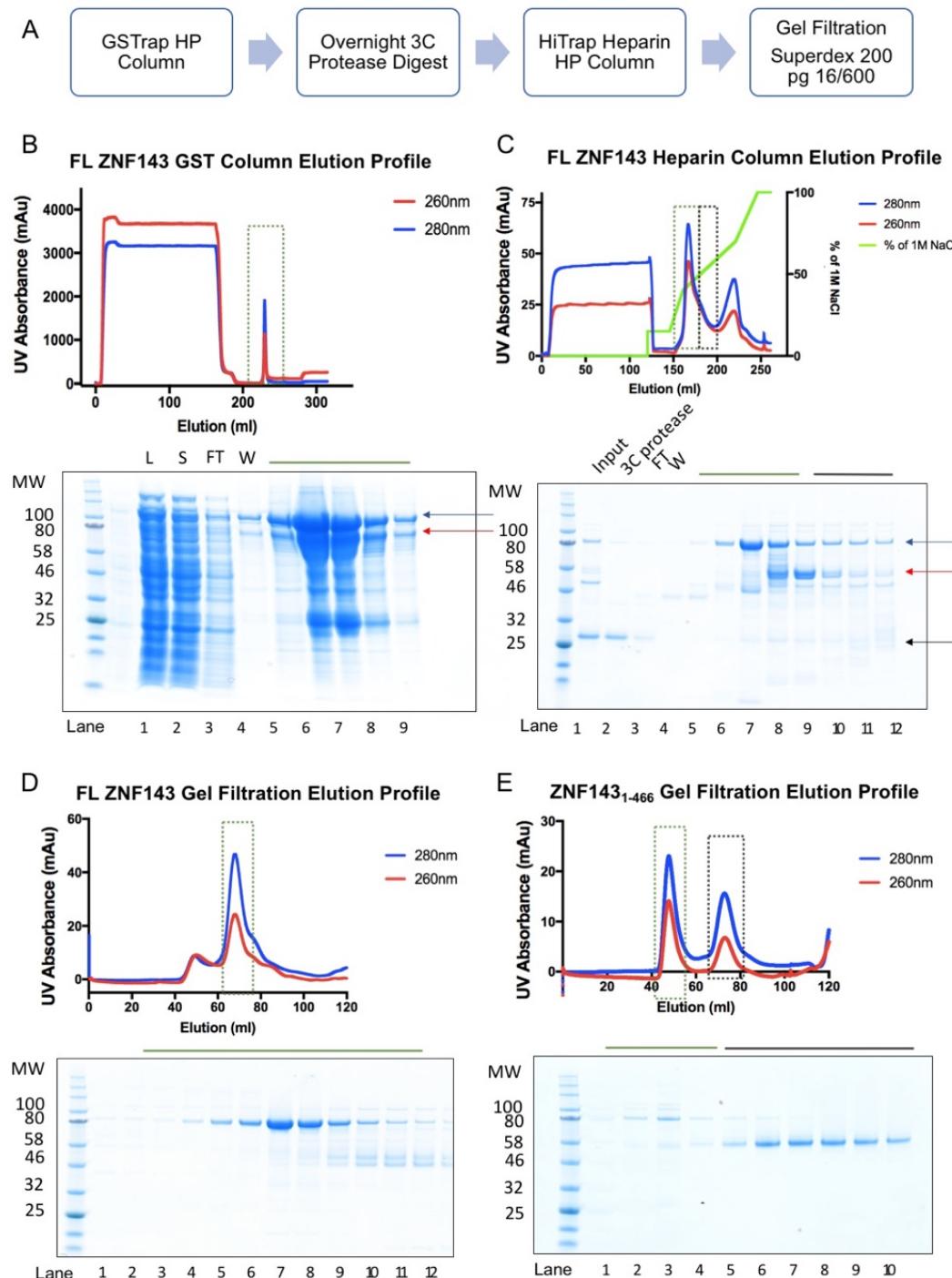
## 4.1 Purification of Recombinant Human ZNF143

### 4.1.1 Purification of *E. coli* Expressed ZNF143

N-terminal His<sub>6</sub>-GST tagged full-length human ZNF143 was expressed in *E. coli* in LB supplemented with ZnCl<sub>2</sub>. The filtered lysate protein was purified via GSTrap HP column (Figure 4.1B). The His<sub>6</sub>-GST tag was cleaved with 3C protease overnight. Cleaved ZNF143 was passed through HiTrap Heparin HP (Figure 4.1C) and eluted between 450-600 mM NaCl. This step helped separate full-length protein from degradation products and remove protein and DNA contaminants. Fractions without contaminants were pooled, concentrated and purified further with gel filtration chromatography using a HiLoad Superdex 200 16/600 pg column (Figure 4.1D).

During the gel filtration, the ZNF143 protein peak eluted at around 65 ml. When the elution volume was compared to a gel filtration standard (BioRad, #1511901) that was run on the same column, it suggested a molecular weight of approximately 158 kDa (Appendix). The theoretical molecular weight of a cleaved full-length monomeric ZNF143 is 70 kDa, suggesting that ZNF143 might exist as a dimer in solution. The discrepancy might also arise from the extended and disordered N- and C-terminals of ZNF143, resulting in larger

hydrodynamic radius than a globular protein of the same molecular weight. The oligomeric state of full-length ZNF143 in solution was investigated by size exclusion multi angle light scattering (SEC-MALS) and explained in section 4.1.4. The fractions from gel filtration column were analysed by SDS-PAGE. ZNF143 fractions were pooled, concentrated to 6 mg/ml, flash frozen and stored at -80 °C. The yield was estimated to be 0.25 mg per 1L of *E. coli* cells.



#### **Figure 4.1 Purification of *E. coli* Expressed Full-length ZNF143**

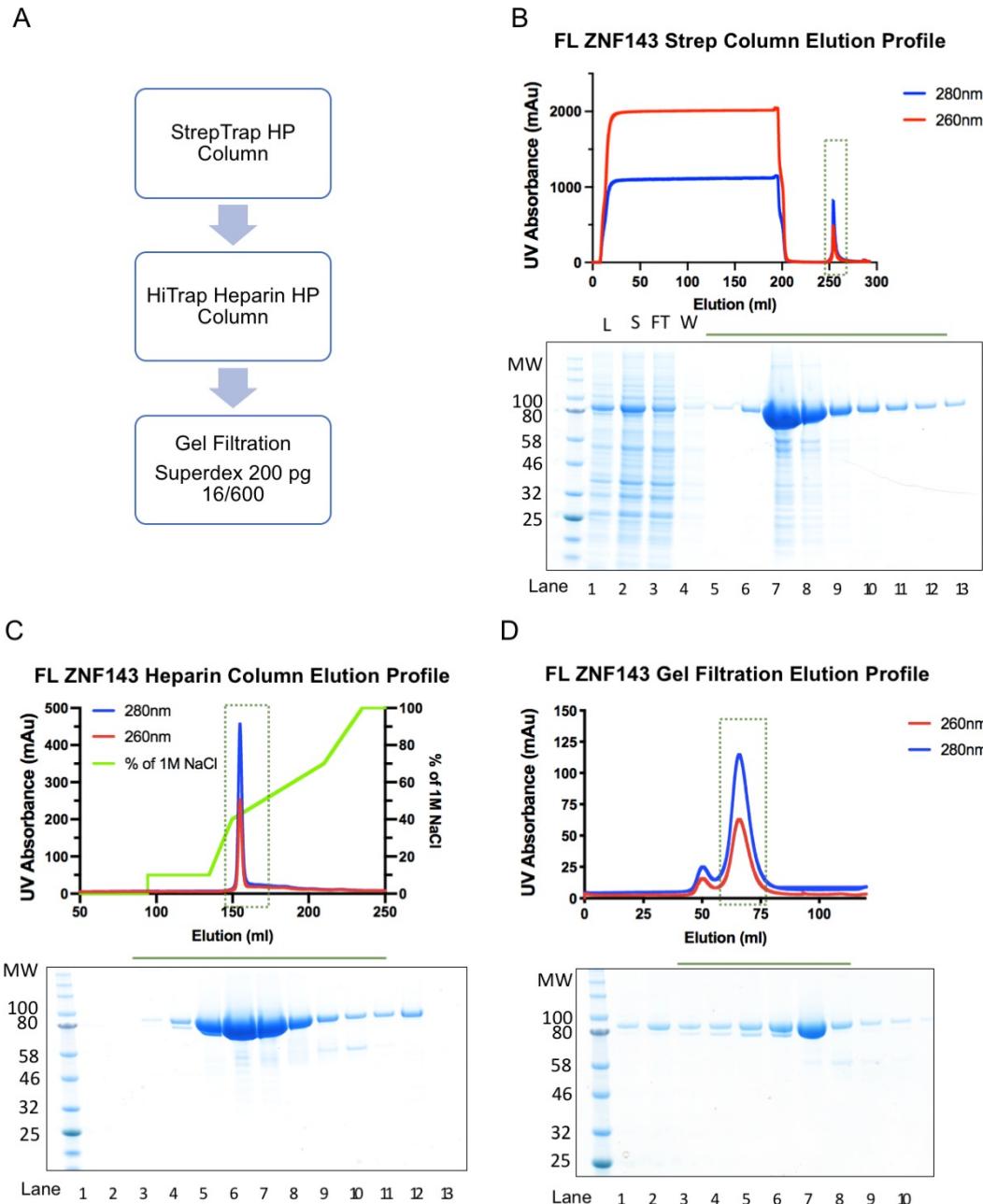
A: Flow chart showing the purification steps for full-length ZNF143. B: Elution profile of GSTrap HP column. Fractions in the box were analysed by SDS-PAGE. His<sub>6</sub>-GST-ZNF143 is 93 kDa and the possible degradation product is 80 kDa, indicated by the blue and red arrows, respectively. L: lysate after sonication, S: supernatant after centrifugation, FT: flow through, W: wash. C: Elution profile of Heparin HP column. Most of the full-length protein was eluted first, indicated by the green box. ZNF143<sub>1-466</sub> was eluted immediately afterwards forming the shoulder of the peak, indicated by the black box. Cleaved ZNF143 is 70 kDa and ZNF143<sub>1-466</sub> has an expected molecular weight of 51 kDa, these two proteins were indicated by the blue and red arrows, respectively. The black arrow corresponded to the cleaved GST tag (~ 28 kDa), which had a similar molecular weight to the 3C protease. FT: flow through, W: wash. D: Elution profile of full-length ZNF143 gel filtration. Fractions within the green box were analysed by SDS-PAGE. E: Elution profile and SDS-PAGE analysis of ZNF143<sub>1-466</sub> gel filtration. Fractions within the green and black boxes corresponded to full length ZNF143 and ZNF143<sub>1-466</sub>, respectively.

#### **4.1.2 Contaminant Observed During Purification**

During the purification of full-length ZNF143 protein, a contaminant of around 80 kDa was observed after the GST affinity step (Figure 4.1B, red arrow). The molecular weight of the contaminant was reduced by around 25 kDa post 3C protease cleavage (Figure 4.1C, red arrow), this reduction corresponded to the molecular weight of His<sub>6</sub>-GST. Since ZNF43 was purified with an N-terminal GST tag, the contaminant might be a truncated version missing the C-terminal region. The protein band that corresponded to the speculated truncated ZNF143 was extracted from the SDS-PAGE gel, digested with trypsin and analysed with liquid chromatography mass spectrometry (LC-MS/MS). Peptides were mapped to the first part of the protein with no peptides found after residue 466 (data not shown), confirming this protein was a stable degradation product of ZNF143. The first 466 amino acids in ZNF143 includes the N-terminal domain, zinc finger domain and 22 amino acids from the C-terminus. From the FL ZNF143 heparin purification step, fractions with 1-466 aa were pooled (lanes 9 to 11 in Figure 4.1C), further purified with gel filtration (Figure 4.1E) to homogeneity and stored at -80 °C.

#### 4.1.3 Optimised Full-length ZNF143 Purification

Since one of the applications of ZNF143 was for crystallography, it was important to design a construct that was more stable and generated higher yield. To try improving the yield and stability of full-length ZNF143, a new construct was designed to express ZNF143 in Eukaryotic High5 insect cells instead of *E. coli*. A baculovirus expressing C-terminal StrepII-tagged ZNF143 was used to infect High5 cells. Cells were harvested four days post infection. The cells were lysed and full-length ZNF143 was purified with a StrepTrap HP column, Heparin HP column and gel filtration with a HiLoad Superdex 200 16/600 pg column (Figure 4.2A). As the expression in *E. coli* showed degradation from the C-terminus, adding a StrepII tag to the C-terminus and purifying the protein using Strep affinity column would deplete the undesired degradation products during purification. Any remaining contaminants eluted from the StrepTrap column were removed over a shallow salt concentration elution gradient in the subsequent Heparin chromatography step. Slight aggregation was observed and removed in the void volume during the gel filtration step. The gel filtration elution volume of insect cell expressed ZNF143 was also ~65 ml, similar to the *E. coli* expressed protein. ZNF143 expressed from insect cells that were purified with C-terminal affinity tag appeared to be more stable since less degradation products were observed. Purified ZNF143 was concentrated to 25 mg/ml and stored at -80 °C. The estimated final yield from 1 L of High5 cells was 4 mg.



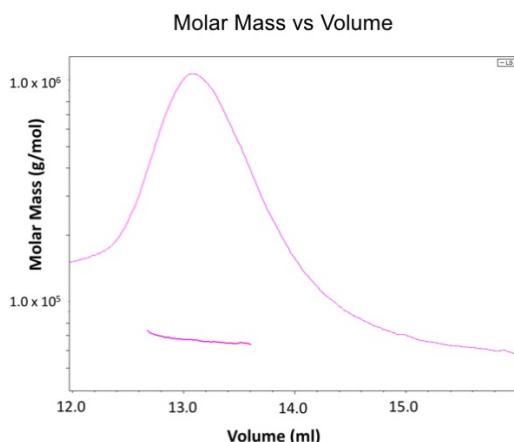
**Figure 4.2 Purification of Insect Cell Expressed Full-length ZNF143**

A: Flow chart showing the purification steps for insect cell expressed full-length ZNF143. B: Elution profile of StrepTrap HP column. Fractions in the green box were analysed by SDS-PAGE. ZNF143-StrepII predicted from sequence was expected to be 73 kDa. C: Elution profile and SDS-PAGE analysis of Heparin HP column. ZNF143 eluted at around 450 mM NaCl. D: Elution profile of gel filtration column. Fractions within the box were analysed by SDS-PAGE.

#### 4.1.4 Full-length ZNF143 Exists as Monomer in Solution

ZNF143 was characterized biophysically to understand whether it formed oligomers in solution. SEC-MALS experiment was performed in collaboration with the Seiradake group at University of Oxford. ZNF143 at 2.25 mg/ml was injected into Superdex 200 10/300 column coupled to a MALS detector. One major peak was eluted, with a fitted molecular weight of 66.8 kDa, which matched with the estimated molecular weight of 70.1 kDa (Figure 4.3). The results showed low polydispersity, indicating high homogeneity in the sample. The measured molecular weight was in accordance with the expected monomeric molecular weight from sequence, suggesting full-length ZNF143 existed as monomer in solution at the measured concentration.

	EXPECTED MW	CONC.	COLUMN	MEASURED MW	S.D.	POLYDISPERSITY
<b>FL ZNF143</b>	70.1 kDa	2.25 mg/ml	Superdex 200 10/300	66.8 kDa	±1.75%	1.001 (±2.5%)



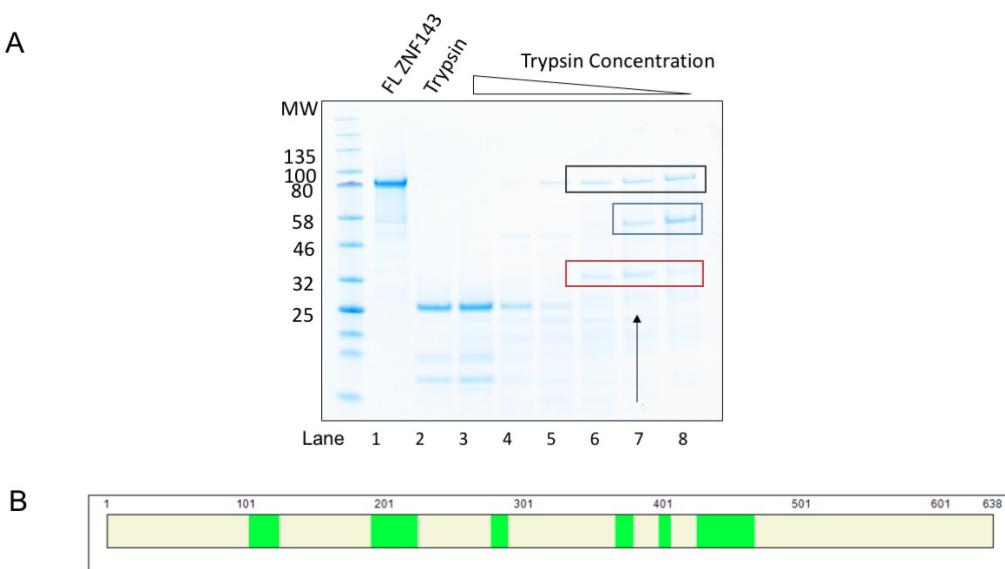
**Figure 4.3 ZNF143 Behaves as a Monomer in Solution**

Molar mass versus elution volume of full-length ZNF143 was obtained from size exclusion chromatography combined with multi-angle light scattering detection. The curve showed the UV absorbance of ZNF143 at 280nm. The measured molecular weight was plotted as a horizontal line, suggesting a molecular weight of 66.8 kDa.

#### 4.1.5 Design of a More Stable Construct ZNF143<sub>103-466</sub>

As flexible regions of protein could compromise crystallisation, a new construct of ZNF143 was designed. The new construct excluded the flexible N- and C-terminal regions but maintain its core function as a DNA binding protein.

In order to experimentally identify such stable candidate regions of ZNF143 for purification and structural analysis, limited trypsin proteolysis of purified full-length ZNF143 was carried out (Figure 4.4). 0.2 mg/ml of full-length ZNF143 was incubated with a range of concentrations of trypsin (0.004 mg/ml to 1 mg/ml) at room temperature for an hour. SDS-PAGE was run to visualise the digested products. As expected, products of lower molecular weight were observed at higher concentrations of trypsin due to higher digestion activities. At 0.01 mg/ml trypsin concentration, three distinct bands were detected (Figure 4.4A, lane 7 indicated with a black arrow). The upper band (black box) corresponded to the full-length construct, whereas the second band (blue box) resembled the 1-466aa construct observed during the purification of N-terminally tagged full-length ZNF143 from *E. coli*. The lowest band (red box) was around 40 kDa, and it was analysed by mass spectrometry. The result revealed that this construct was comprised of amino acids between residues 103 and 466 (Figure 4.4B). The first half of the N-terminal domain and most of the C-terminal domain were deleted in this new construct. This coincided with the fact that the N- and C-terminal domains of ZNF143 were predicted to be unstructured and flexible, therefore by removing these regions, the resulting construct could potentially be more stable.

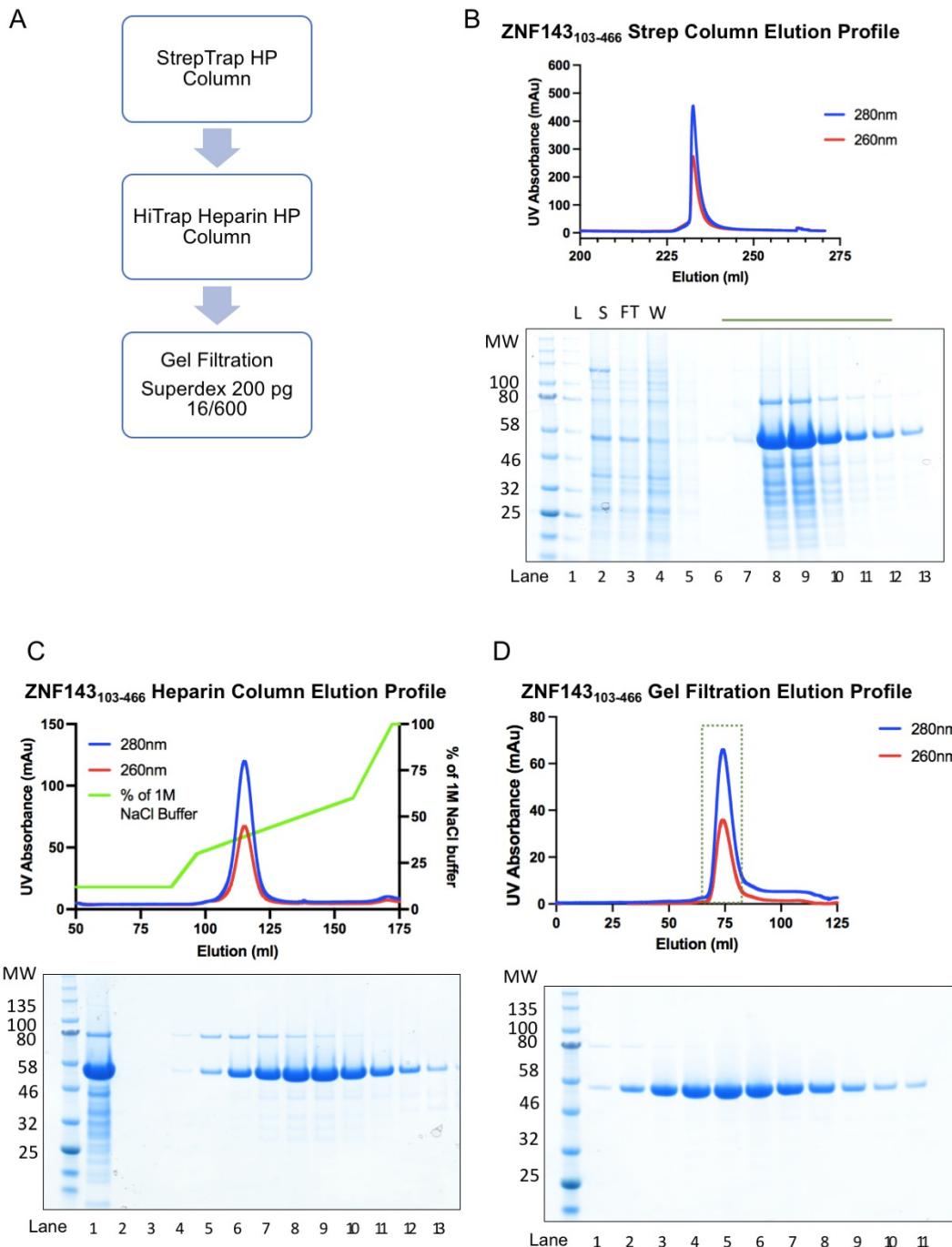


#### **Figure 4.4 Trypsin Digestion of ZNF143**

A: Trypsin limited proteolysis with ZNF143. Black arrow indicates the lane where the trypsin concentration was 0.1 mg/ml. The three major bands observed after proteolysis represented full-length ZNF143, ZNF143<sub>1-466</sub> and ZNF143<sub>103-466</sub> (later confirmed through mass spectrometry), these bands were indicated with black, blue and red boxes, respectively. B: Sequence analysis from mass spectrometry. Green boxes represent sequences that were detected.

#### **4.1.6 Purification of ZNF143<sub>103-466</sub>**

ZNF143<sub>103-466</sub> was cloned into pLIB with C-terminal twin-Strep tag and used to generate baculovirus. High5 insect cells were infected for four days prior to harvest. The ZNF143<sub>103-466</sub> construct was purified in a three-step protocol detailed in Figure 4.5A. Following affinity chromatography (Figure 4.5B), the eluted fractions were diluted to a final salt concentration of 150 mM NaCl and loaded onto a Heparin column. ZNF143<sub>103-466</sub> was passed through a salt gradient and eluted at 350 mM NaCl (Figure 4.5C). Finally, the protein was concentrated to 5 ml and loaded onto Superdex 200 16/600 column (Figure 4.5D). One symmetrical peak was observed at 75 ml, this elution volume corresponded to a molecular weight of 66 kDa, which was slightly larger than the expected molecular weight of monomeric ZNF143<sub>103-466</sub>-StrepII (44 kDa), suggesting that ZNF143 may not be a globular protein but rather an extended molecule with a larger hydrodynamic radius than a globular protein with the same molecular weight, as previously observed for the full-length construct (Figures 4.1 and 4.2). The purified protein appeared to be stable throughout the purification steps and purified to homogeneity with no visible degradation observed by SDS-PAGE. Purified ZNF143 was concentrated to 11 mg/ml and the estimated yield from 1 L of High5 cells was 6 mg of protein which was slightly higher than the yield from full-length ZNF143 expressed in High5 cells. The fractions were stored at -80 °C.



**Figure 4.5 Purification of ZNF143<sub>103-466</sub>**

A: Flow chart showing the purification steps for insect cell expressed ZNF143<sub>103-466</sub>. B: Elution profile of StrepTrap HP column. Peak fractions were analysed in SDS-PAGE. ZNF143<sub>103-466</sub>-StrepT was 44 kDa. C: Elution profile of Heparin HP column. ZNF143 eluted at around 450 mM NaCl. D: Elution profile and SDS-PAGE analysis of gel filtration.

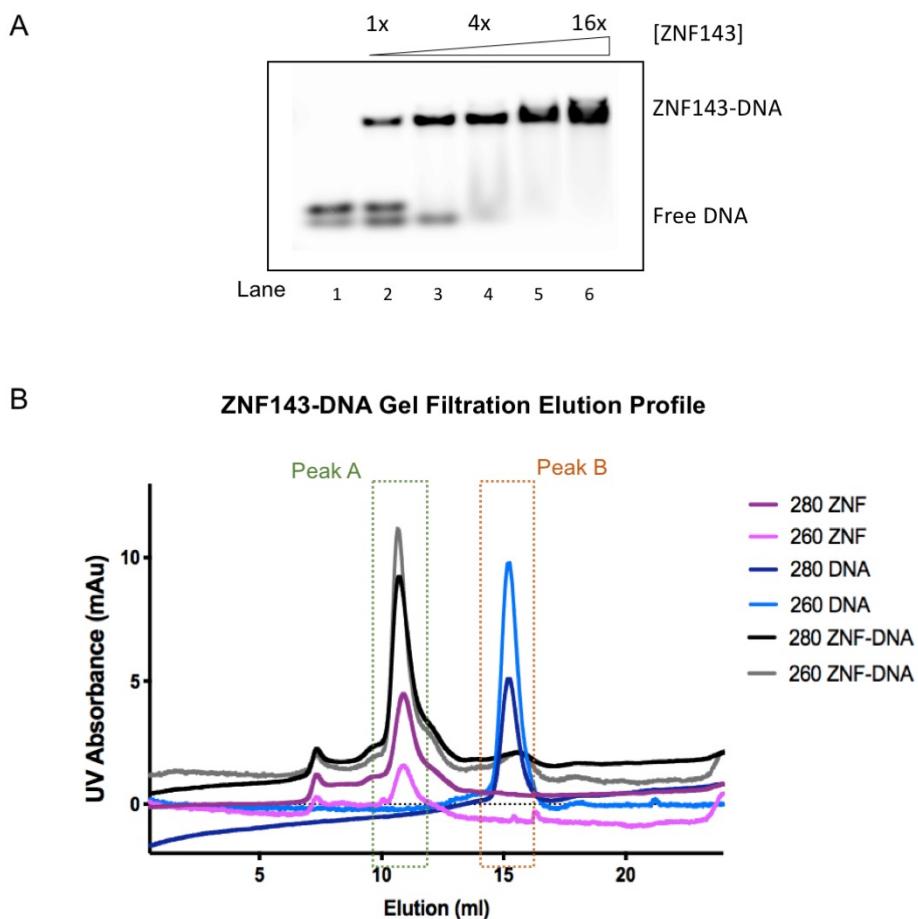
## 4.2 DNA Binding Studies of ZNF143

### 4.2.1 ZNF143-DNA Forms a Stable Complex Over Gel Filtration

One of the well documented functions of ZNF143 is its role as a DNA-binding protein (Myslinski, Krol and Carbon, 1998; Schaub, Krol and Carbon, 2000). Thereby, electrophoretic mobility shift assay (EMSA) was used to test if the recombinantly purified ZNF143 bound DNA. Increasing concentrations of ZNF143 were titrated against 0.1  $\mu$ M consensus DNA, incubated on ice for 20 minutes and the samples were run on 0.5X TBE gels (Figure 4.6A). At 1:1 ratio, ZNF143 was able to interact with DNA and a ZNF143-DNA shift was observed. As the concentration of ZNF143 increased, the intensities of the free DNA bands decreased, until its disappearance at a 4:1 protein:DNA ratio. This experiment confirmed that ZNF143 was able to bind to DNA and a complete DNA shift can be achieved at 4:1 ratio.

Next, ZNF143 and DNA were passed through gel filtration column to test whether they form a stable complex since a stable protein-DNA complex will be essential for structural studies. Four molar excess of DNA was used in this analytical gel filtration experiment to saturate the protein with DNA substrate, allowing for maximal protein:DNA complex formation. Using a Superdex 200 10/300 increase GL column, three runs were carried out (DNA only, ZNF143 only, ZNF143 and DNA) to investigate the stability of protein-DNA complex (Figure 4.6B). For the run where ZNF143 and DNA were both present, ZNF143 and 2-fold molar excess DNA were incubated on ice for 30 minutes before loading onto the gel filtration column to enable the sample to reach equilibrium. Unbound ZNF143 eluted at around 10.9 ml, with the UV absorbance 260/280 at around 0.6. DNA alone eluted around 15.2 ml with the ratio between the 260 and 280 nm of ~2.0. In the ZNF143-DNA run, two peaks were observed. The first peak eluted at around 10.9 ml (peak A), with 260/280 ratio of ~1.3 indicating an increased presence of DNA in this peak and hinting at the formation of a ZNF143-DNA complex. The second peak comprised of excess DNA eluted at 15.2 ml (peak B), matching with the free DNA peak in the previous run. The molecular weight of the 28 bp DNA construct was small, therefore the shift between the ZNF143-DNA complex peak and the unbound

ZNF143 peak might not be observable. However, the UV absorbance difference between the two peaks and the shift in 260 nm to 280 nm UV absorbance ratio suggested an increased presence of DNA. In conclusion, this experiment confirmed that ZNF143 and DNA formed a stable complex over gel filtration.



**Figure 4.6 ZNF143 Forms a Stable Complex with DNA**

A: EMSA between full-length ZNF143 and 28bp DNA construct. The concentration of DNA was kept constant. B: Elution profile of three runs: ZNF143 only, DNA only and ZNF143-DNA run. The UV absorbance at 280 nm and 260 nm for all runs were plotted against the elution volume. Two major peaks were indicated with the green and orange boxes respectively.

#### 4.2.2 N- and C-terminal region of ZNF143 do not bind DNA

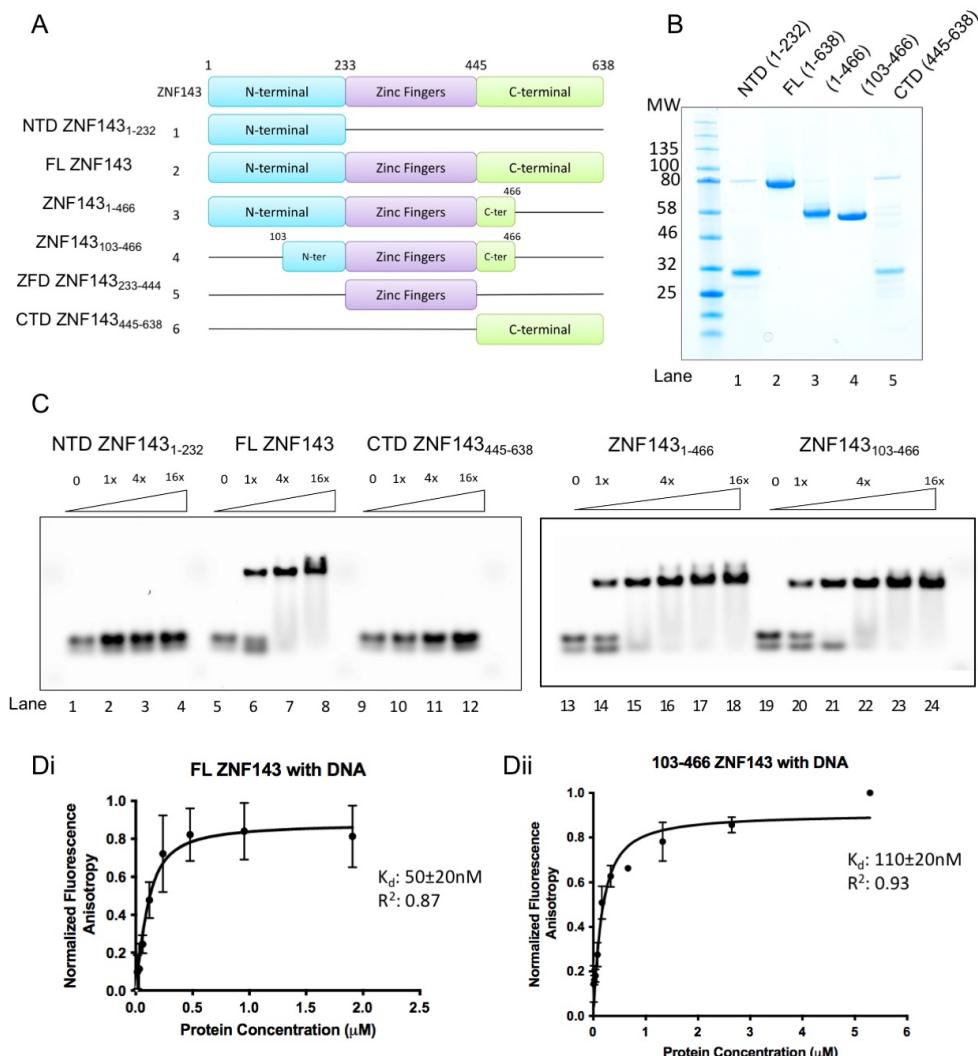
There are seven tandemly repeated zinc fingers in the DNA binding domain in ZNF143. Although functional studies have been carried out on the zinc finger domain of the *Xenopus* homolog Staf (Schuster *et al.*, 1995; Myslinski, Krol

and Carbon, 1998; Schaub, Krol and Carbon, 2000), no experimental results have shown whether or not the three human ZNF143 domains bind DNA or modulate binding indirectly. An EMSA was consequently carried out using purified ZNF143 constructs comprised of various domains and lengths to confirm the DNA binding region (Figure 4.7A, B). A 28bp dsDNA construct with the extracted sequence from the ZNF143 binding motif located at the RNA polymerase III type III U6\_2 promoter was used in this experiment. The results showed that ZNF143<sub>1-466</sub> (truncated by-product from full-length purification, see section 4.1.2), ZNF143<sub>103-466</sub> and full-length ZNF143 were able to bind DNA, whereas the N-terminal domain alone and C-terminal domain alone lacked the ability to form a complex with DNA (Figure 4.7C). The construct containing only the seven zinc fingers could not be purified, however the other constructs used in this experiment were sufficient to show that the DNA binding region in ZNF143 was the zinc finger domain between residues 233-444.

To understand if the N- and C-terminal domains modulate the DNA binding affinities of the zinc finger domain, fluorescence polarisation was carried out with full-length ZNF143 and ZNF143<sub>103-466</sub>. Fluorescence polarisation is a useful tool to measure the binding affinity of fluorescently labelled molecules. Polarisation is dependent only on the molecular volume of the fluorescent probe, hence changes in the volume when the fluorescent probe binds to a macromolecule is recorded, resulting in a bind curve to which a  $K_d$  can be fit. In this experiment, DNA was 6-carboxyfluorescein (6-FAM) labelled and its binding to ZNF143 would impede the rate of molecular rotation. This measurement was then used to calculate the binding affinities between ZNF143 and DNA.

Various concentrations of ZNF143 constructs were titrated against a constant concentration (0.1  $\mu$ M) of 6-FAM labelled 28 bp DNA and incubated at room temperature. The fluorescence polarisation signal was measured using POLARstar plate reader. The signal was analysed and plotted against protein concentration. Comparative DNA binding affinities for full-length ZNF143 and

ZNF143<sub>103-466</sub> were similar (Figure 4.7D). The  $K_d$  of full-length ZNF143 binding to its consensus sequence was  $50 \pm 20$  nM and the  $K_d$  of ZNF143<sub>103-466</sub> to the same DNA sequence was  $110 \pm 20$  nM. These binding affinities were in reasonable agreement with published measurements of zinc finger transcription factors binding to consensus sequences (Jung *et al.*, 2018). This experiment demonstrated that the N- and C-terminal domains did not contribute to the DNA binding property of ZNF143.



**Figure 4.7 N- and C-terminal Domains Do Not Affect ZNF143 DNA-Binding**

A: All ZNF143 constructs that were designed and created. The first and last residues of each construct were listed. The domain architecture of ZNF143 was illustrated in a bar diagram. B: All purified ZNF143 constructs on the same SDS-PAGE gel. ZNF143<sub>233-444</sub> was insoluble and therefore was not included. C: DNA binding EMAS with purified ZNF143 constructs. 1 to 16 times ZNF143 was titrated against a 0.1  $\mu\text{M}$  DNA. Di: Fluorescence polarisation of full-length ZNF143 against DNA. The normalised fluorescence anisotropy was plotted against protein concentration. Dii: Fluorescence polarisation of ZNF143<sub>103-466</sub> against DNA.

#### 4.2.3 ZNF143 Binds to A Range of DNA Constructs

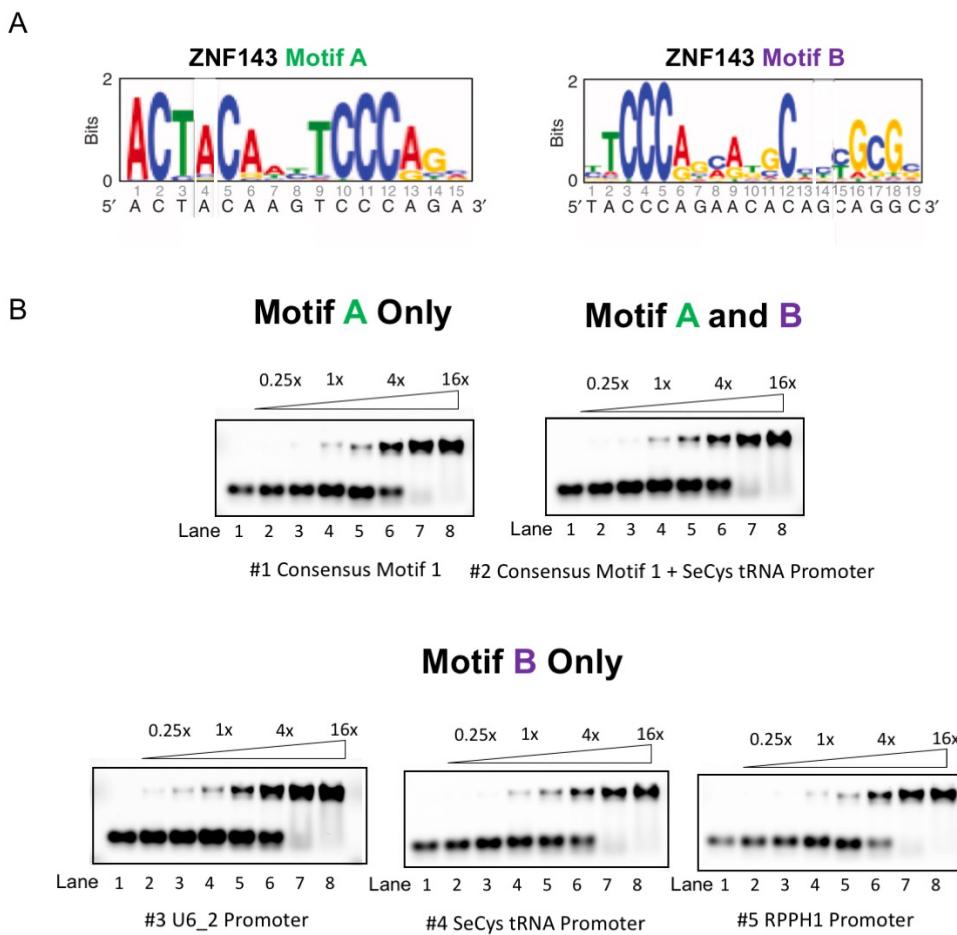
Although consensus ZNF143 binding motifs have been identified (Ngondo-Mbongo *et al.*, 2013; Bailey *et al.*, 2015), it was intriguing to compare the binding affinities of ZNF143 across various DNA sequences that were comprised of its consensus sites to identify ZNF143's preferred sequence. Five 28bp dsDNA constructs were designed based on the two predicted consensus ZNF143 binding motifs (motif A and motif B) observed at gene promoters (Bailey *et al.*, 2015) (Table 4.1, Figure 4.8A). ZNF143 consensus motif B is the downstream extension of consensus motif A, with the sequence TCCCA present in both motifs. 43% and 76% of gene promoters bound by ZNF143 were found to contain consensus motif A or motif B, respectively (Bailey *et al.*, 2015). DNA construct 1 included only the sequence of predicted consensus motif A, whereas DNA construct 2 was comprised of the predicted consensus motif A and the exact endogenous sequence of the ZNF143 motif at the SeCys tRNA promoter (James Faresse *et al.*, 2012). DNA constructs 3, 4 and 5 were the extracted sequences of ZNF143 motifs at the human RNA polymerase III type III promoters U6\_2, SeCys tRNA and RPPH1 promoters respectively. The RPPH1 promoter can also be occupied by RNA Polymerase II (James Faresse *et al.*, 2012). These three promoters showed high sequence similarity to consensus motif B.

Different concentrations of ZNF143 were titrated against Cyanine-5 (Cy5) labelled DNA constructs and visualised through TBE agarose gels. The agarose gel was scanned by Typhoon scanner with a filter at the Cy5 excitation wavelength. The results suggested that ZNF143 was able to bind to all sequences strongly and either motif by itself was sufficient for strong ZNF143 binding (Figure 4.8B).

	Construct Name	DNA Sequence
1	Consensus Motif 1	GTTTATA <u>ACTACAAT</u> <b>TCCCAG</b> GCTAATTG
2	Consensus Motif 1 + SeCys tRNA Promoter	GA <u>ACTACAAT</u> <b>TCCCAG</b> AATGCGTG <b>GCGA</b>
3	U6_2 Promoter	TATTGG <u>CACCA</u> CAATCCAC <b>CGCG</b> GCTA
4	SeCys tRNA Promoter	GTTTATT <u>TCCCAG</u> AATGCGTG <b>GCG</b> ATTG
5	RPPH1 Promoter	GTGATT <u>TCCCAG</u> AACACATA <b>GCG</b> ACATG

**Table 4.1 DNA Constructs Used for ZNF143 Binding Studies**

Five dsDNA constructs were used to understand the DNA binding properties of ZNF143. Sequences highlighted in green match with the predicted ZNF143 consensus motif A, whereas sequences in purple match with predicted consensus motif B by Bailey *et al* in 2015. The exact promoter sequences at their respective promoters were underlined.



**Figure 4.8 ZNF143 Binding to its Consensus Motifs**

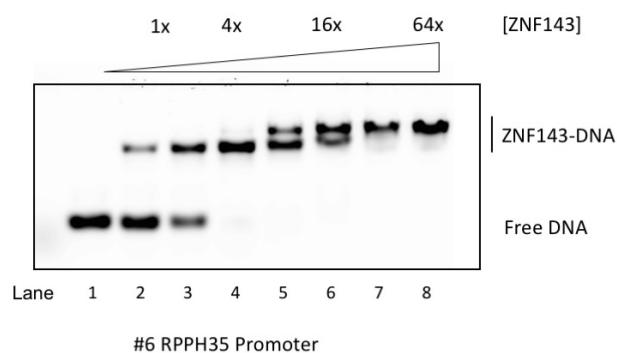
A: ZNF143 binding motifs predicted by Bailey *et al.*, 2015. Both binding motifs were observed at gene promoters. B: EMAS between ZNF143 and five different dsDNA constructs. The DNA sequences used in this experiment were listed in table 4.1.

A longer (84bp) sequence of DNA was used to test the binding properties with ZNF143 further (Table 4.2). This construct was composed of the exact sequence of the RPPH35 promoter with one ZNF143 and one SNAPc binding site. As the concentration of ZNF143 increased, two binding events were observed as the lower ZNF143-DNA band shifted to a higher band (Figure 4.9). One of the possible reasons to account for that shift was that ZNF143 was binding at two different sites on the DNA construct. At low ZNF143 concentrations, only the more preferable binding site (ZNF143 consensus site) was occupied by ZNF143. However, at eight times more ZNF143 than DNA, two shifted bands were present representing two species, and ZNF143 might be bound to regions other than its consensus site. The lower band corresponded to the species with only one ZNF143 bound, whereas the other species might represent two or more ZNF143 bound to one molecule of DNA.

	<b>Construct Name</b>	<b>DNA Sequence</b>
<b>6</b>	RPPH35 Promoter	AAAAGTGGTCTCATACAGAACTTATAAGAT <b>TCCCA</b> AATCCAA AGAC <b>ATTCACGTTATGGTGA</b> TT <b>TCCCAG</b> AACACATA <b>GCGA</b>

**Table 4.2 84bp DNA Used for ZNF143-DNA Binding Study**

A new 84 bp DNA construct was designed to test ZNF143's DNA binding activity further. Sequences highlighted in purple represents the ZNF143 consensus motif B. The orange text refers to the sequence that might be the potential second binding site to ZNF143. Sequences highlighted in blue corresponded to the binding site of SNAPc.



**Figure 4.9 Additional ZNF143 Binding Sites Present**

EMSA of ZNF143 and RPPH35 DNA construct (84bp). 0.1  $\mu$ M of DNA was incubated with increasing concentrations of ZNF143. A higher shift was observed when ZNF143 was eight times in excess.

However, since there was only one ZNF143 binding motif present in this DNA construct, this suggested that ZNF143 might be binding to non-consensus sequences. Although the exact location of the second binding site on this DNA construct could not be confirmed, it might be possible that ZNF143 was binding to the rich cytosine regions (highlighted in orange in table 4.2) as the consensus motifs predicted showed high GC rich preference (Bailey *et al.*, 2015).

#### 4.2.4 ZNF143 Binds to Constructs with Non-Consensus Sequence

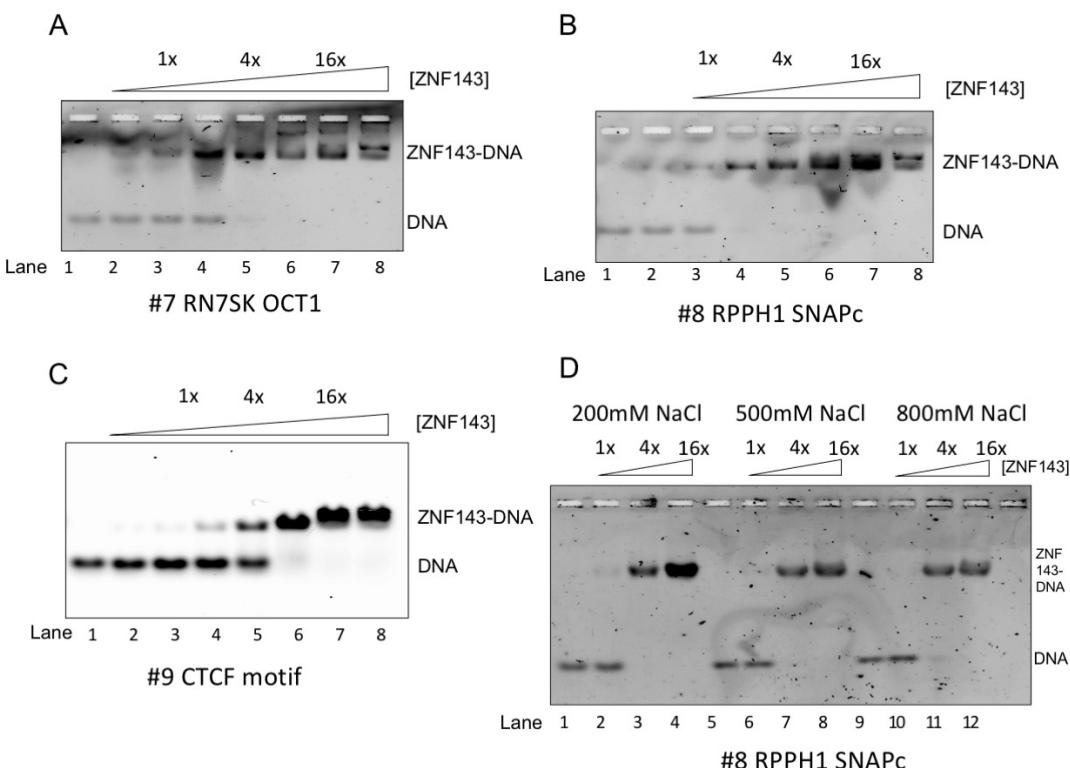
Following up on the hypothesis that ZNF143 binds to non-consensus sequences, three other non-ZNF143 motif DNA constructs were designed (Table 4.3). All double stranded DNA constructs were 28bp long. The sequence of construct 7 was extracted from the RN7SK promoter with OCT-1 motif site, construct 8 was the sequence of the SNAPc motif at RPPH1 promoter, lastly, construct 9 was comprised of the CTCF consensus sequence (Table 4.3). 0.1 µM of DNA constructs were incubated with full-length ZNF143 and the samples were run on agarose gels. As shown in figures 4.10A-C, ZNF143 was able to bind to all constructs starting from two molar excess of protein to DNA.

To explore whether the interaction between ZNF143 and non-consensus DNA was due to non-specific electrostatic interaction, the EMSA experiment was repeated in higher salt concentrations. High salt concentrations would increase the ionic strength of the buffer and shield possible non-specific interactions arising from opposing molecular charges. As seen in Figure 4.10D, the interaction between ZNF143 and non-consensus DNA was not disrupted when the salt concentration was increased from 200 mM to 800 mM NaCl. This confirmed that ZNF143 was able to bind to DNA without the consensus Z-motif.

	Construct Name	DNA Sequence
7	RN7SK Promoter (OCT-1)	AATAAATGAT <b>ATTTGCTA</b> TGCTGGTTAA
8	RPPH1 Promoter (SNAPc)	GGAAA <b>TCACCCATAAACGTGAAAT</b> GTCTT
9	CTCF Consensus Motif	CCACAGAGG <b>CCAGCAGGGGGCGCT</b> AGTG

**Table 4.3 DNA Constructs without ZNF143 Consensus Sites**

Three more dsDNA constructs were designed to understand the DNA binding properties of ZNF143. Sequences highlighted in blue corresponded to the binding sites of the respective transcription factors mentioned in the construct names.



**Figure 4.10 ZNF143 Binds to Non-Consensus DNA Sequences**

A: EMSA of ZNF143 and DNA construct with the intrinsic sequence of OCT-1 on RN7SK promoter. B: EMSA of ZNF143 binding to the SNAPc sequence on RPPH1 promoter. C: EMSA of ZNF143 and DNA construct with CTCF consensus motif. D: EMSA between ZNF143 and RPPH1 SNAPc DNA construct under three different salt concentrations, at 200 mM, 500 mM and 800 mM respectively.

#### 4.2.5 ZNF143 Prefers Z-motif Over Non-Consensus DNA

Since ZNF143 was able to bind to DNA sequences lacking the consensus site, it was intriguing to understand if ZNF143 shows stronger affinity towards constructs with the Z-motif.

A competition assay was carried out with three 28 bp DNA constructs:

1. Unlabelled non-consensus DNA (construct 8 in table 4.3)
2. Unlabelled Z-motif DNA (construct 3 in table 4.1)

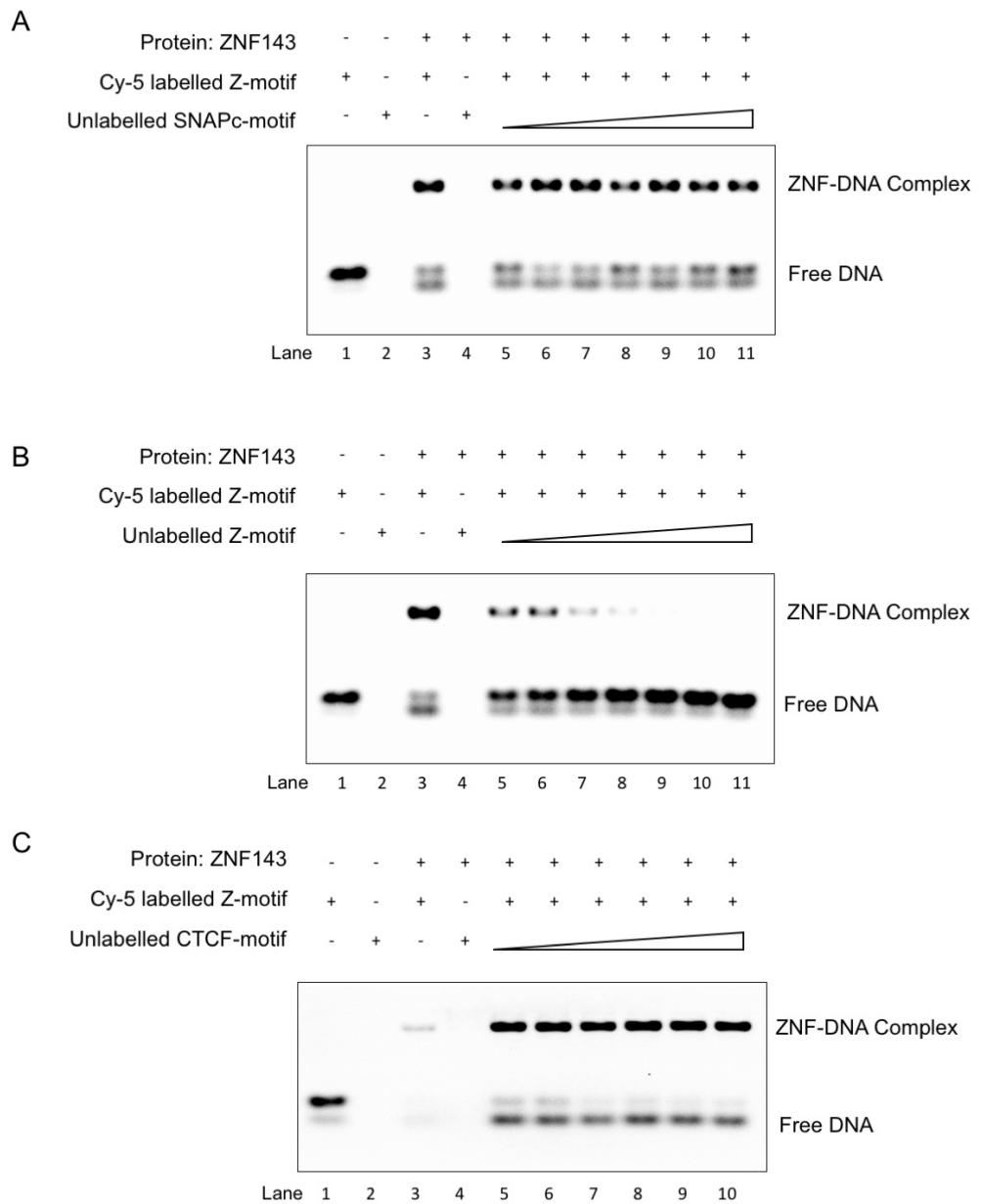
### 3. Cy-5 labelled Z-motif DNA (construct 3 in table 4.1).

Cy-5 labelled Z-motif was first incubated with purified ZNF143 at a 1:4 ratio on ice for 20 minutes. Next, increasing concentration of unlabelled non-consensus DNA (up to 16-fold excess of labelled Z-motif) was then added to compete with the bound Z-motif DNA. The samples were run on agarose gels and visualised using Typhoon imager with the Cy-5 filter. Hence, only Cy-5 labelled DNA was visualised and shown in figure 4.11A. As the concentration of non-consensus DNA increased, the intensity of the ZNF143- (Z-motif) DNA shift was not reduced. This suggested that ZNF143 showed stronger affinity to Z-motif DNA over non-consensus DNA, hence the unlabelled non-consensus DNA was not able to displace the Cy-5 labelled Z-motif (Figure 4.11A).

As a control, the unlabelled non-consensus DNA was replaced with unlabelled Z-motif DNA (Figure 4.11B). Because ZNF143 interacts with unlabelled Z-motif and Cy-5 labelled Z-motif with the same affinity, the labelled DNA could exchange with the unlabelled, resulting in the reduction in the intensities of ZNF-DNA detected as the concentration of unlabelled Z-motif DNA increased.

It has been suggested that ZNF143 and CTCF co-localise extensively to form enhancer-promoter chromatin loops (Bailey *et al.*, 2015; Ye *et al.*, 2020). Although CTCF and ZNF143 have different consensus sites, their consensus sequences are highly similar in terms of their high GC content. Therefore, it was interesting to investigate whether ZNF143 binds to both the Z-motif and CTCF-motif DNA with similar affinity. The competition EMSA assay was carried out with unlabelled CTCF-motif (construct 9 in table 4.3) and Cy-5 labelled Z-motif DNA, of which all of them were 28 bp long. The results showed that CTCF-motif was not able to displace Z-motif DNA, showing that ZNF143 preferred Z-motif DNA over CTCF-motif (Figure 4.11C). Although the consensus sequences for ZNF143 and CTCF are highly similar, this may imply that the zinc fingers in ZNF143 recognise specific features in its consensus

sequence and make specific contacts that might explain the preferential behaviour of ZNF143.



**Figure 4.11 ZNF143 Prefers DNA Sequences with Z-motif**

A: Competition assay between Z-motif DNA and SNAPc-motif DNA. ZNF143 was incubated with Cy-5 labelled Z-motif DNA at 4:1 ratio, prior to incubation with increasing concentration of unlabelled SNAPc-motif DNA. EMSA gels were visualised with Typhoon imager using Cy5 filter. B: Control competition assay between Cy-5 labelled Z-motif DNA and unlabelled Z-motif DNA. C: Competition assay between Z-motif DNA and CTCF-motif DNA. Unlabelled CTCF-motif was titrated against pre-formed ZNF143-(Z-motif) DNA complex.

#### 4.2.6 Binding Affinities Between ZNF143 and Different DNA Constructs

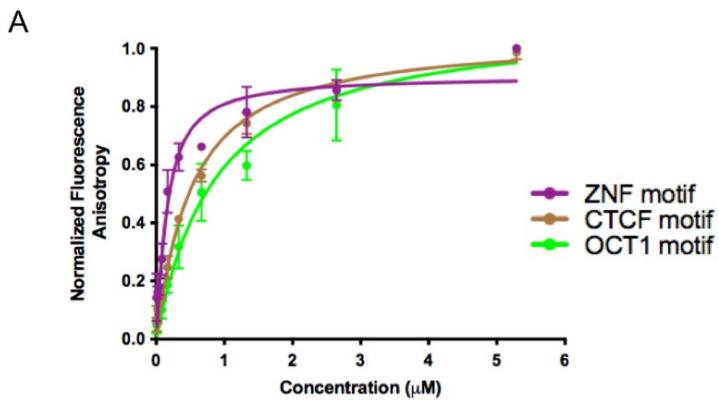
ZNF143 showed binding to a range of different DNA constructs even in the absence of its preferred consensus sites. It has been shown in the previous section (section 4.2.5) that ZNF143 interacted with sequences with Z-motif more strongly compared to sequences lacking the Z-motif. Fluorescence polarisation was carried out to understand how the binding affinities might differ in the absence of ZNF143 consensus sequences. Three different DNA constructs were incubated with titrating concentrations of ZNF143<sub>103-466</sub>. One DNA construct contained the ZNF143 consensus sequence (construct 3 from table 4.1). The other two constructs were comprised of CTCF consensus site (construct 9 from table 4.3) and the endogenous binding site for OCT-1 at the RN7SK promoter (construct 7 from table 4.3), respectively.

ZNF143<sub>103-466</sub> showed strong binding affinities towards all sequences, with binding affinities at the nanomolar range (Figure 4.12A). It was unsurprising that ZNF143<sub>103-466</sub> showed highest binding affinity towards its consensus sequence, with  $K_d$  of approximately  $110 \pm 20$  nM. However, it was also interesting to note that although the binding affinity for DNA with CTCF consensus site ( $470 \pm 40$  nM) was lower than its preferred Z-motif DNA, it was slightly higher than the other construct with a non-consensus sequence from the RN7SK promoter ( $800 \pm 130$  nM). This might be explained by the high similarity in CTCF and ZNF143 consensus sites, where both sequences are high in GC content, and therefore the zinc fingers in ZNF143 might recognise this feature and displayed higher binding affinity towards DNA sequence with CTCF-motif.

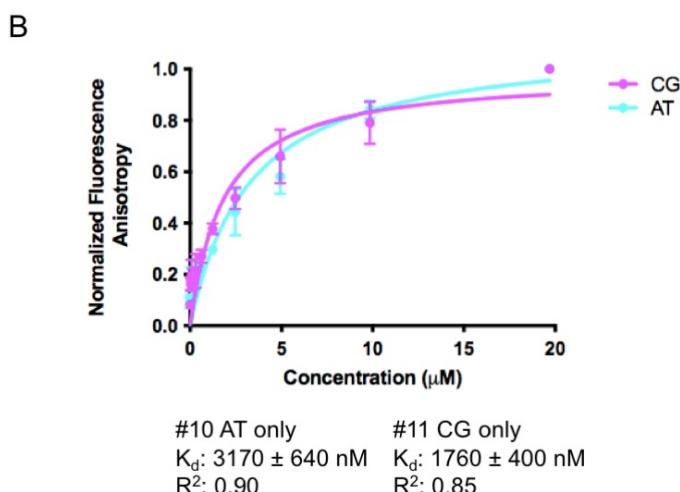
To show that ZNF143 might prefer sequences with high GC content, two new DNA constructs were designed. The first construct (construct 10 from table 4.4) was comprised of randomly allocated adenine and thymine, whereas the second construct (construct 11 from table 4.4) was comprised of randomly allocated cytosine and guanine. Fluorescence polarisation was carried out with these two constructs and ZNF143<sub>103-466</sub> under the same conditions as previously described. The results suggested that ZNF143 was able to bind to

both sequences, but with 10-fold lower binding affinities compared to DNA with Z-motif (Figure 4.12B). In addition, ZNF143 showed higher binding affinity towards the sequence comprised of only CGs, with  $K_d$  at 1760 nM when compared to the AT sequence, with  $K_d$  at 3170 nM, suggesting that the zinc fingers in ZNF143 potentially recognise and prefer sequences with high GC content. This hypothesis will need to be further explored through structure determination.

To conclude, these fluorescence polarisation experiments demonstrated that ZNF143 binds strongly to a wide range of DNA sequences with or without its consensus site. Nonetheless, ZNF143 showed higher binding affinities towards its consensus motif and to sequences with high GC content.



#3 ZNF143 motif      #7 OCT-1 motif      #9 CTCF motif  
 $K_d$ : 110 ± 20 nM     $K_d$ : 800 ± 130 nM     $K_d$ : 470 ± 40 nM  
 $R^2$ : 0.92                 $R^2$ : 0.95                 $R^2$ : 0.98



#### **Figure 4.12 Fluorescence Polarisation of ZNF143 and Different DNA Sequences**

A: Fluorescence polarisation of ZNF143 and DNA constructs with ZNF-motif, CTCF-motif and OCT-1-motif plotted on the same graph. Increasing concentrations of ZNF143 was incubated with DNA and the fluorescence polarisation was detected and recorded using POLARstar plate reader. Normalised fluorescence anisotropy was plotted against protein concentration. B: Fluorescence polarisation of ZNF143 and DNA constructs with either AT or CG only. Normalised anisotropy was plotted against ZNF143 concentration.

	<b>Construct Name</b>	<b>DNA Sequence</b>
<b>10</b>	AT	ATATTATTAATTATAATATAAAAATTTAT
<b>11</b>	CG	CGGGCCCGCGGCCGCCCCGGGGCCGCCG

**Table 4.4 Artificial DNA Constructs for ZNF143 DNA Binding Studies**

Two constructs were designed and created to understand if ZNF143 was able to bind to DNA with lacking CG or AT. Construct 10 was comprised of randomly allocated adenines and thymines, whereas construct 11 was comprised only guanines and cytosines.

#### **4.2.7 ZNF143 Interacts with Nucleosomes**

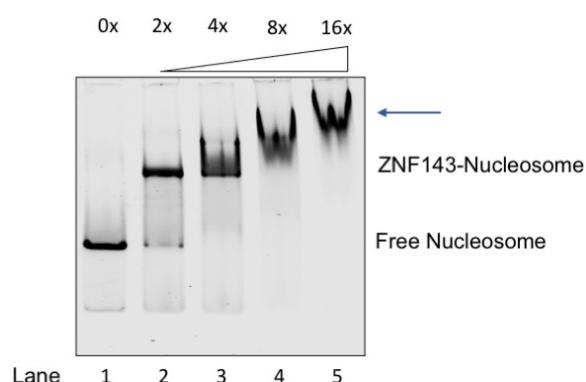
In addition to understanding the DNA-binding activities of ZNF143, it was also intriguing to investigate if ZNF143 interacts with nucleosomes since nucleosomes play an important role in transcription initiation (Kujirai and Kurumizaka, 2020).

To investigate this, EMSAs were carried out. Different concentrations of purified ZNF143 were incubated with 0.1 µM nucleosome on ice for 30 minutes. 147 bp DNA with Widom 601 sequence was used to wrap around the histones and there was a 52 bp DNA linker on one end of the nucleosome that did not have a ZNF143 consensus motif. The samples were then loaded onto a commercial Novex Native 4-12% TBE gel, stained with ethidium bromide and visualised using Typhoon image scanner.

ZNF143 showed strong binding to nucleosomes. Complete shift was observed when four-fold excess of ZNF143 was present. As the concentration of nucleosomes increased to eight-fold excess of ZNF143, the ZNF143-nucleosome complex band shifted to a higher position (Figure 4.13). One of the possible explanations for this observation was that two or more ZNF143

might be binding to the same nucleosome substrate. There might be three potential binding sites on nucleosome for ZNF143. Firstly, ZNF143 might interact directly with the histones. Secondly, ZNF143 might bind to the 52 bp linker at the end of the nucleosome. Alternatively, it might interact with the Widom 601 DNA that is wrapped around the histones. Hence, several ZNF143 molecules might be binding to the same nucleosome substrate and in return displaying a second binding event. It will require further analysis to identify the preferential binding site. Similar binding events were also observed when high concentrations of ZNF143 was incubated with DNA constructs (Figure 4.9).

Given the strong interaction between ZNF143 and nucleosomes, this complex could be amenable to future structural investigation with Cryo-EM, which has been used to obtain the structure of numerous nucleosome-bound complexes.



**Figure 4.13 ZNF143 Binds to Nucleosomes**

Electrophoretic mobility shift assay between ZNF143 and 0N52 nucleosome. As the concentration of ZNF143 increased, a second binding event was observed (blue arrow), that could be explained by the possibility of several ZNF143 molecules binding to the same nucleosome substrate.

### 4.3 ZNF143 Crystallography Trials

Structural analysis was undertaken after the characterisation of DNA interaction behaviour of ZNF143. The aim was to solve the structure of apo and DNA-bound ZNF143 to explore the specificity of its zinc fingers and their roles in DNA recognition. Since the full-length ZNF143 is 70 kDa, X-ray

crystallography was used to obtain high resolution structural information of ZNF143.

#### 4.3.1 Crystallisation Trial with Apo ZNF143

Zinc finger domains bound to DNA molecules have been successfully crystallised in the past and therefore ZNF143 crystallisation was tested against commercial screens that included some of the conditions (Table 4.5) at which those zinc finger domain crystals were formed.

PDB Entry	Name	Method	Conditions	Reference
<b>1AAY</b>	ZIF268 Zinc Finger-DNA complex	Vapor Diffusion, Hanging Drop	2.5-5% PEG400, 500-700 mM NaCl, 25 mM Bis-Tris Propane pH 8.0	(Elrod-Erickson <i>et al.</i> , 1996)
<b>6ML4</b>	BTB24 Zinc Fingers 4-8 with DNA	Vapor Diffusion, Sitting Drop	25% PEG3350, 0.1 M HEPES pH 7.5, 0.2 M Ammonium Acetate	(Ren <i>et al.</i> , 2019)
<b>5K5H</b>	CTCF Zinc Fingers 4-7 with DNA	Vapor Diffusion, Sitting Drop	25% PEG 3350, 0.2 M ammonium acetate, 0.1 M Bis-Tris HCl pH 5.5	(Hashimoto <i>et al.</i> , 2017)
<b>4GZN</b>	ZFP57 Zinc Fingers with methylated DNA	Vapor Diffusion, Sitting Drop	25% MPD, 15% PEG8000, 100 mM CaCl <sub>2</sub> , 100 mM acetate pH 4.6	(Liu <i>et al.</i> , 2012)
<b>4QF3</b>	BAZ2B Zinc Finger in Free Form	Vapor Diffusion, Sitting Drop	2.2 M Na/K phosphate pH 7.8	(Tallant <i>et al.</i> , 2015)

**Table 4.5 Previous Zinc Finger Domain Crystal Formation Conditions**

Five examples of conditions in which zinc finger domains were crystallised and structures were solved and deposited on PDB.

Purified full-length ZNF143 was prepared for crystallisation trials in commercial sparse-matrix screens (sitting-drop, 96-well 2-drop MRC plates from Hampton Research). Eight commercial screens (with previous successful zinc finger crystals conditions) were set up with apo ZNF143, with 200 nl sitting drops and 1:1 mix of protein and well solution (Table 4.6). The screens were tested at 4 °C and 18 °C, and with two different protein concentrations (7 mg/ml and 14 mg/ml) on each plate. The plates were monitored for six months. No crystals were formed, clear and light precipitations were observed in approximately 60% of the wells.

Screening Kit	Protein Concentration	Temperature
<b>Stura Footprint HT-96</b>	7 mg/ml, 14 mg/ml	4 °C, 18 °C
<b>PACT Premier HT-96</b>	7 mg/ml, 14 mg/ml	4 °C, 18 °C
<b>Index HT-96</b>	7 mg/ml, 14 mg/ml	4 °C, 18 °C
<b>JCSG+ HT-96</b>	7 mg/ml, 14 mg/ml	4 °C, 18 °C

**Table 4.6 Screening Kits Used for Full-length ZNF143 Crystallisation**

Eight plates were set up with apo ZNF143, comprising of four different commercial screening kits and tested at two different temperatures.

#### 4.3.2 Crystallisation Trial with ZNF143-DNA

A possible cause of the lack of crystallisation may arise from the zinc finger domain, which might be flexible in the absence of DNA and hindering the formation of crystals. It was speculated that ZNF143 might require the presence of DNA to stabilise its zinc finger regions to promote crystal formation. As shown in previous sections, the binding affinity between ZNF143 and DNA was strong, and the interaction was stable over gel filtration chromatography (Figure 4.6).

The DNA constructs used for ZNF143 crystallisation were based on the DNA construct 3 as shown in table 4.1. This sequence was chosen not only because it is one of the endogenous sites for ZNF143 at the U6\_2 promoter site, but also it resembles the consensus motif B where ZNF143 showed strong binding affinity. It has been shown that each zinc finger contacts 3 base-pairs of DNA (Desjarlais and Berg, 1992). Since ZNF143 has seven zinc fingers, the DNA

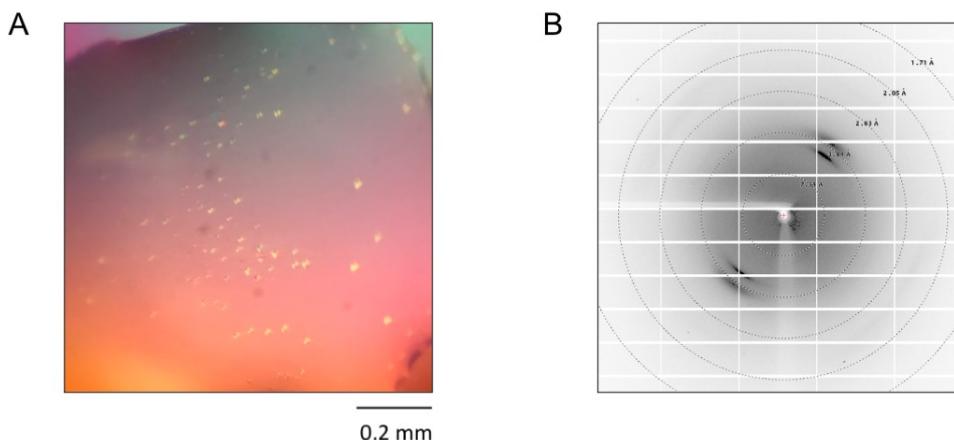
construct used in the crystallisation trials was 25 base-pairs long (TATCACCCACAATCCACCGCGGATA).

ZNF143 concentrated to 4.5 mg/ml and 9 mg/ml were used to set up the crystallisation plates. ZNF143 was incubated with DNA at 1:1.2 ratio on ice for an hour prior to plate set up. Excess DNA was used to saturate all binding sites on the protein (Ruigrok *et al.*, 2012). Twelve commercial screens were set up with 200 nl sitting drops and 1:1 mix of protein/DNA and well solution at 4°C and 18°C (Table 4.7). Approximately 50% of the wells showed precipitation. Tiny crystals were observed in one of the wells with the condition (0.2 M calcium chloride dihydrate, 0.1 M MES pH 5.5 and 45% (w/v) MPD) after two weeks. The condition was reproduced and optimised in 24-well plates, with 2 µl hanging drops (1 µl protein solution + 1 µl well solution). Although new crystals were formed, the size of the crystals did not improve (Figure 4.14A). They were fished and sent to synchrotron and the results suggested that only DNA was present in the crystals (Figure 4.14B).

Screening Kit	Protein Concentration	Temperature
<b>Stura Footprint HT-96</b>	4.5 mg/ml, 9 mg/ml	4 °C, 18 °C
<b>PACT Premier HT-96</b>	4.5 mg/ml, 9 mg/ml	4 °C, 18 °C
<b>Index HT-96</b>	4.5 mg/ml, 9 mg/ml	4 °C, 18 °C
<b>JCSG+ HT-96</b>	4.5 mg/ml, 9 mg/ml	4 °C, 18 °C
<b>Proplex HT-96</b>	4.5 mg/ml, 9 mg/ml	4 °C, 18 °C
<b>Crystal Screen HT-96</b>	4.5 mg/ml, 9 mg/ml	18 °C
<b>PEG Suite HT-96</b>	4.5 mg/ml, 9 mg/ml	18 °C

**Table 4.7 Screening Kits Used for Full-length ZNF143-DNA Crystallisation**

Twelve plates were set up with ZNF143-DNA (25bp) complex, comprising of seven different commercial screening kits and tested at two different temperatures.



**Figure 4.14 Full-length ZNF143-DNA Crystal Leads**

A: Crystals were observed in one of the conditions (0.2 M calcium chloride dihydrate, 0.1 M MES pH 5.5 and 45% (w/v) MPD). B: Diffraction pattern of one of the fished crystals.

#### 4.3.3 Crystallisation Trial with Apo ZNF143<sub>103-466</sub>

Due to the lack of progress in the crystallisation trials with full-length ZNF143 in the presence or absence of DNA, ZNF143<sub>103-466</sub> was designed to remove the flexible N- and C-terminal regions and aiming that a more rigid structure would allow the protein to crystallise more readily.

Purified ZNF143<sub>103-466</sub> concentrated to 12 mg/ml was used for crystallisation trials. Eight commercial screens were set up with 200 nl sitting-drop in 2-drop MRC plates (Table 4.8). ZNF143 was mixed with well solution at 1:1 and the screens were incubated at 18°C, at two different protein concentrations (6 mg/ml and 12 mg/ml). The plates were monitored for six months. No crystals were formed, clear and light precipitations were observed in approximately 60% of the wells.

Screening Kit	Protein Concentration	Temperature
<b>Stura Footprint HT-96</b>	6 mg/ml, 12 mg/ml	4 °C, 18 °C
<b>PACT Premier HT-96</b>	6 mg/ml, 12 mg/ml	18 °C
<b>Index HT-96</b>	6 mg/ml, 12 mg/ml	18 °C
<b>JCSG+ HT-96</b>	6 mg/ml, 12 mg/ml	18 °C
<b>PEG Suite HT-96</b>	6 mg/ml, 12 mg/ml	18 °C
<b>Morpheus HT-96</b>	6 mg/ml, 12 mg/ml	18 °C
<b>ProPlex HT-96</b>	6 mg/ml, 12 mg/ml	18 °C

**Table 4.8 Screening Kits Used for ZNF143<sub>103-466</sub> Crystallisation Trials**

Eight plates were set up with ZNF143<sub>103-466</sub>, seven different commercial screening kits were used.

#### 4.3.4 Crystallisation Trial with ZNF143<sub>103-466</sub>-DNA

ZNF143<sub>103-466</sub> was incubated with DNA constructs to reduce the flexibility of the zinc finger domain. ZNF143<sub>103-466</sub> at 4 mg/ml and 10 mg/ml concentrations were incubated with DNA at 1:1.2 ratio on ice for an hour prior to setting up twelve commercial crystallisation screens (Table 4.9). In addition to the 25 bp DNA construct used for full-length ZNF143-DNA crystallisation trials, two shorter pieces of DNA (21 bp and 23 bp) with the same sequence were also used in ZNF143<sub>103-466</sub>-DNA crystallisation trials. Tiny crystals were observed under the same condition (0.2 M calcium chloride dihydrate, 0.1 M MES pH 5.5 and 45% (w/v) MPD) where DNA crystals were observed during full-length ZNF143-DNA screening. No crystals were observed in other conditions, and approximately 50% of the wells were precipitated.

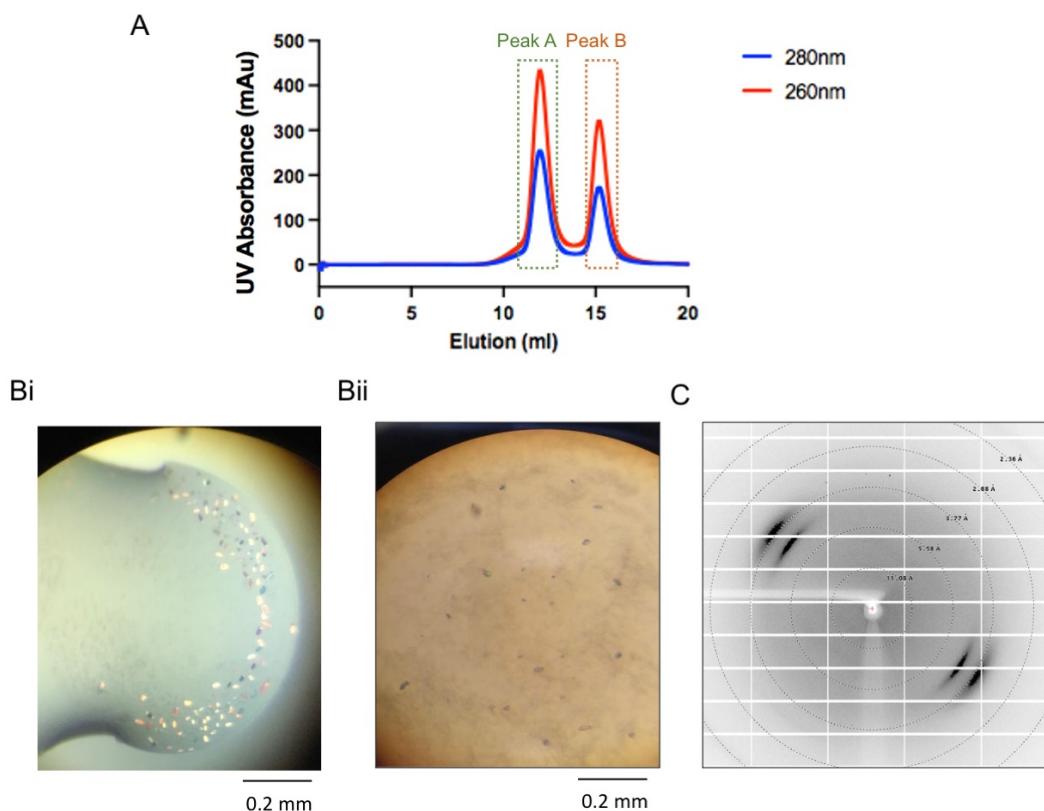
Screening Kit	Protein Concentration	Temperature	DNA Length (bp)
<b>Stura Footprint HT-96</b>	4 mg/ml, 10 mg/ml	4 °C, 18 °C	23
<b>Salt Rx HT-96</b>	4 mg/ml, 10 mg/ml	18 °C	23
<b>Index HT-96</b>	4 mg/ml, 10 mg/ml	18 °C	23
<b>JCSG+ HT-96</b>	4 mg/ml, 10 mg/ml	18 °C	23, 25
<b>PEG Suite HT-96</b>	4 mg/ml, 10 mg/ml	18 °C	23
<b>Morpheus HT-96</b>	4 mg/ml, 10 mg/ml	18 °C	23
<b>ProPlex HT-96</b>	4 mg/ml, 10 mg/ml	18 °C	23
<b>Shotgun 1 HT-96</b>	4 mg/ml, 10 mg/ml	18 °C	21, 23, 25

**Table 4.9 Screening Kits Used for ZNF143<sub>103-466</sub>-DNA Crystallisation Trials**

Twelve 96 well plates were set up with ZNF143<sub>103-466</sub> incubated with DNA sequences of three different lengths, eight different commercial screening kits were used.

To reduce DNA only crystals and to increase the chances of nucleation of the ZNF143-DNA complex, ZNF143<sub>103-466</sub> was incubated with 23 bp DNA on ice for 30 minutes and then applied through an analytical gel filtration column.

Based on the gel filtration chromatogram, ZNF143-DNA complex was well separated from free DNA (Figure 4.15A). Only fractions that showed complex formation (peak A) were concentrated and used to set up more crystallisation screens. Four crystallisation screens were set up following the same settings as previously described. Tiny crystals were observed in more conditions (Table 4.10) after two or three weeks. Although these conditions were optimised in 24-well plates, with 2 µl hanging drops, the optimisation did not improve the sizes of crystals. These crystals were fished and sent to synchrotron for analysis (Figure 4.15B). The results suggested that these crystals were DNA only crystals also (Figure 4.15C). Therefore, further optimisation of the ZNF143 constructs and DNA sequences will be required to initiate nucleation and form strong and stable protein-DNA complex crystals.



**Figure 4.15 Crystallisation Trials with ZNF143<sub>103-466</sub>-DNA Complex**

A: Elution profile of the analytical gel filtration of ZNF143<sub>103-466</sub> and DNA (23 bp). Peak A refers to ZNF143<sub>103-466</sub>-DNA complex whereas peak B corresponds to the excess free DNA. Fractions from peak A were concentrated and used to set up four crystallisation screens. Bi: Example of crystals observed in the condition 0.1 M Tris pH 8.5 + 40% MPD. Bii: Example of crystals observed in the condition 0.1 M MES pH 6.5 + 12% PEG 20000. C: Diffraction pattern of crystal fished from the condition 0.1 M Tris pH 8.5 + 40% MPD.

Screening Kit	Condition
<b>Crystal Screen HT-96</b>	0.1 M MES pH 6.5 + 12% PEG 20000
<b>Crystal Screen HT-96</b>	1.6 M Sodium citrate tribasic pH 6.5
<b>JCSG+ HT-96</b>	0.2 M Sodium chloride + 0.1 M Sodium phosphate pH 6.2 + 50% PEG200
<b>JCSG+ HT-96</b>	0.1 M Tris pH 8.5 + 40% MPD
<b>Index HT-96</b>	0.2 M Calcium chloride + 0.1M MES pH 5.5 + 45% MPD

**Table 4.10 Crystallisation Conditions with Potential Crystal Leads**

Five conditions showed potential crystal leads and optimisation around these conditions were carried out.

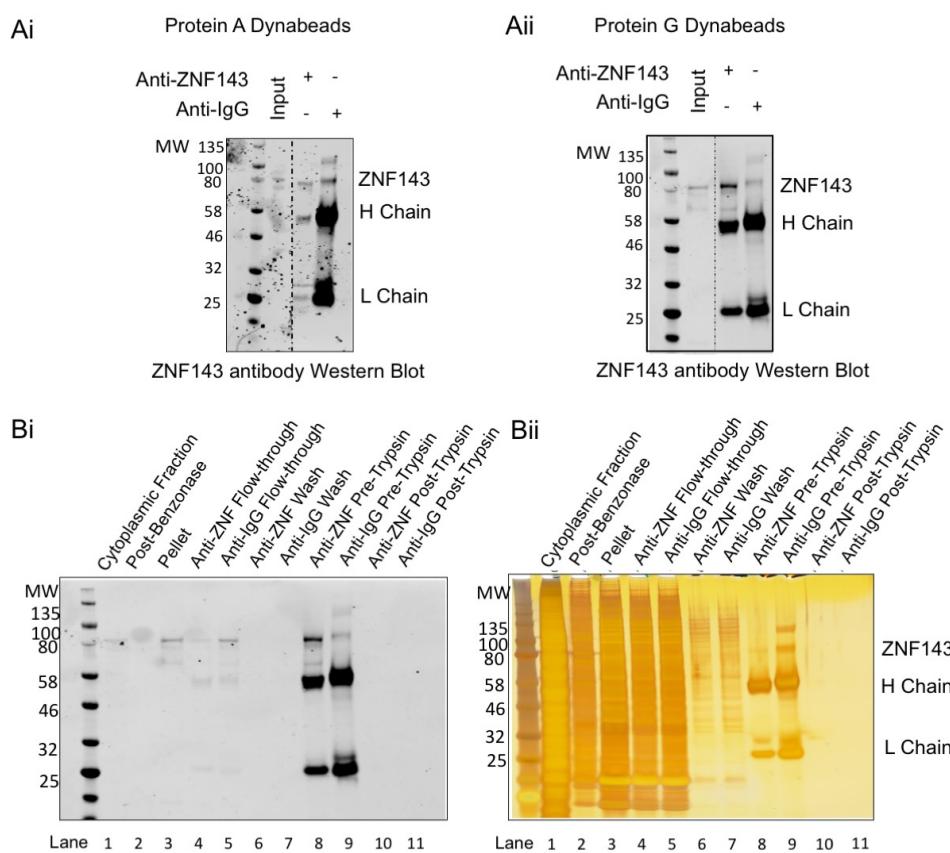
#### 4.4 ZNF143 and Its Potential Interacting Partners

Crystallisation of ZNF143 with and without DNA was challenging and therefore it was intriguing to explore potential interacting partners to create a stable complex for structural studies using cryo-electron microscopy. To date, no known proteins have been shown to interact with ZNF143 directly. As one of the transcription factors that is heavily involved in genome architecture and organisation, it is crucial to explore potential interacting proteins and how this relationship might contribute to chromatin regulation.

To identify more potential interacting partners of ZNF143, a co-immunoprecipitation pulldown assay (Co-IP) with endogenous ZNF143 was carried out. Co-IP targets a known protein with a specific antibody to pull the protein complexes that contain that protein out of solution. These unknown interaction partners can then be identified with mass spectrometry. Nuclear extract from HEK293T cells were used because HEK293T cells have the highest ZNF143 expression level when compared to other cell lines (Pontén, Jirström and Uhlen, 2008). The nuclear lysate was treated with benzonase to remove DNA to reduce the influence of chromatin acting as a mediator between ZNF143 and other chromatin proteins. Dynabeads™ Protein A were used to capture ZNF143 antibody and immunoglobulin G antibody. Immunoglobulin G antibody acted as a negative control. HEK293T nuclear extract was subsequently incubated with the beads. The proteins captured by

dynabeads were trypsin digested and sent for LC-MS/MS analysis to identify the potential interacting partners. Various samples were taken in intermediate steps and analysed via western blot and SDS-PAGE to verify the presence of ZNF143 throughout the experiment. From the western blot analysis, comparing the heavy and light chains of the two antibodies (anti-ZNF143 and anti-IgG), it suggested that anti-ZNF143 antibody was not well-captured by Protein A Dynabeads<sup>TM</sup> (Figure 4.16Ai). Hence the amount of endogenous ZNF143 pulled down was scarce and required optimisation.

To improve the amount of endogenous ZNF143 captured, differences between Protein A and Protein G Dynabeads<sup>TM</sup> were compared. The anti-ZNF143 antibody used was a mouse monoclonal IgG1 that showed stronger binding to Protein G Dynabeads<sup>TM</sup> (Figure 4.16Aii). The same co-immunoprecipitation was performed with Protein G Dynabeads<sup>TM</sup> and significant improvement in the amount of ZNF143 captured can be observed from the western blots. Lanes 10 and 11 were the post-trypsin samples and therefore the absence of full-length ZNF143 band in lane 10 confirmed the digestion (Figure 4.16B).



#### Figure 4.16 Co-Immunoprecipitation of Endogenous ZNF143

Ai: Western blot of pull-down of anti-ZNF143 antibody and anti-IgG antibody with Protein A Dynabeads<sup>TM</sup>. Anti-ZNF143 antibody was not well captured by Protein A Dynabeads<sup>TM</sup>. Input corresponding to the cytoplasmic fraction was also included in the gel. Aii: Western blot of pull-down of anti-ZNF143 antibody and anti-IgG antibody with Protein G Dynabeads<sup>TM</sup>. Endogenous ZNF143 was captured by anti-ZNF143 antibody. Bi: Western blot of ZNF143 immunoprecipitation with anti-ZNF143 antibody. Bii: Silver stain of ZNF143 immunoprecipitation.

The co-IP experiment was repeated three times and LC-MS/MS results were analysed using CRAPome (Mellacheruvu *et al.*, 2013). Spectrum/peptide matches were retained only if they had a normalized difference in cross-correlation scores of at least 0.08. A total of 400 non-redundant proteins were identified and the top 30 proteins that associated with ZNF143 were listed in table 4.11. The interacting potential value correlates to the confidence of which the specific protein was pulled down by endogenous ZNF143. The closer the value to 1, the higher the confidence.

	Gene	Protein Score	Function
1	ZNF143	1.00	Transcriptional regulation
2	ZNF76	1.00	Transcription regulation
3	DDX20	1.00	SMN complex - spliceosome
4	SMN1	1.00	SMN complex - spliceosome
5	GEMIN5	1.00	SMN complex - spliceosome
6	HSD17B4	1.00	Oxidation pathway for fatty acids
7	SNRNP70	0.99	Spliceosome
8	ZNF626	0.98	Transcription regulation
9	ANXA2	0.77	Membrane binding protein
10	GEMIN4	0.67	SMN complex - spliceosome
11	RPL21	0.66	Ribosomal subunit
12	RIF1	0.65	Double-strand break repair
13	PABPN1	0.64	mRNA polyadenylation
14	ADAR	0.58	mRNA splicing
15	PKM	0.58	Glycolytic enzyme
16	RPL19	0.57	Ribosomal subunit

<b>17</b>	MAP4	0.57	Microtubule assembly
<b>18</b>	TUBB4A	0.56	Microtubule assembly
<b>19</b>	P0DMV9	0.55	Molecular chaperone
<b>20</b>	TUBB4B	0.54	Microtubule assembly
<b>21</b>	KRT14	0.49	Keratin
<b>22</b>	ENO1	0.49	Glycolytic enzyme
<b>23</b>	SERBP1	0.47	mRNA regulation
<b>24</b>	RBM4	0.43	mRNA splicing
<b>25</b>	TUBA1A	0.43	Microtubule assembly
<b>26</b>	RBM25	0.43	mRNA splicing
<b>27</b>	TUBB2B	0.42	Microtubule assembly
<b>28</b>	TUBB	0.40	Microtubule assembly
<b>29</b>	DDX39B	0.39	Spliceosome
<b>30</b>	YWHAQ	0.38	Regulates signalling pathway
...			
<b>275</b>	CTCF	0.25	Transcription factor, genome architecture

**Table 4.11 Immunoprecipitation Results of ZNF143**

Top 30 protein hits from immunoprecipitation with ZNF143 were listed in descending order of specific protein score. The closer the protein score to 1, the higher the confidence of interaction. Proteins involved in the transcription regulation, mRNA splicing pathway and microtubule assembly are coloured in red, blue and purple, respectively. Interesting protein target CTCF was also included in the table.

ZNF143 had a score of 1, suggesting ZNF143 had the highest peptide hits and sequence coverage, confirming that the anti-ZNF143 antibody was effective in capturing endogenous ZNF143. Interestingly, ZNF76 also scored 1, this might arise from the high sequence identity between ZNF143 and ZNF76 (63%). It was likely that anti-ZNF143 antibody recognised a region of ZNF76 that was very similar to the epitope of the antibody. On the other hand, ZNF76 might be a potential interacting partner of ZNF143, however this interaction has not been observed in previous publications and will require further investigation.

In addition to ZNF73, four other proteins scored 1 in the analysis, including DDX20, SMN1 and GEMIN5. All of these proteins were involved in mRNA

splicing via the spliceosome. The other top ZNF143-associating proteins were also related to microtubule assembly such as MAP4 and TUBB4A etc. These interactions suggested that ZNF143 might also be involved in other crucial pathways other than chromatin architecture and regulation. Further experiments will be needed to confirm these interactions.

#### 4.5 Summary

ZNF143 was first identified as a transcription factor that is crucial for regulating genes involved in primary metabolism and cell growth (Izumi *et al.*, 2010), but growing evidence suggested that ZNF143 has a distinctive role in establishing higher order structure within the genome (Ye *et al.*, 2020; Huning and Kunkel, 2021). Since most studies on ZNF143 have been carried out on the *Xenopus* homolog Staf, this chapter focused on the functional characterisation of the human ZNF143 to understand how ZNF143 acts on the chromatin.

Full length human ZNF143 was purified following a three-step purification protocol, including a Strep affinity column, followed by Heparin ion exchange and gel filtration chromatography. In addition to the full-length construct, a shorter, but more stable construct ZNF143<sub>103-466</sub> was designed based on the results from the limited trypsin proteolysis experiment. ZNF143<sub>103-466</sub> was purified following the purification protocol for full length ZNF143. Both constructs have been shown to be functional as seen in the DNA binding studies, and this also suggested that the DNA binding region in ZNF143 was nested in the zinc finger domain. The similar DNA binding affinities of both constructs measured by fluorescence polarisation demonstrated that the N- and C-terminal domains of human ZNF143 did not contribute to DNA binding.

Next, further EMSA studies revealed that ZNF143 was capable of binding to a range of different DNA sequences including constructs lacking its ZNF143 motif. The difference in binding affinities measured from the fluorescence polarisation assays hinted a possible hypothesis where the zinc finger region in ZNF143 potentially recognise and prefer sequence with high GC content.

Although structural characterisation of ZNF143 would shed light into the mechanism behind DNA binding, crystallisation trials with apo full length ZNF143, full length ZNF143-DNA, apo ZNF143<sub>103-466</sub>, ZNF143<sub>103-466</sub>-DNA did not succeed.

Lastly, most of the ZNF143 interacting proteins discovered by co-immunoprecipitation assay were involved in the mRNA splicing via the spliceosome and microtubule assembly hinting potential roles of ZNF143 in these pathways. It was interesting to note that the two proteins (CTCF and CHD8) that have been shown to associate with ZNF143 in previous publications (Ishihara, Oshimura and Nakao, 2006; Yuan *et al.*, 2007a) were not scored high through this co-IP assay which might be due to reasons that will be discussed further in the discussion. The association between ZNF143 and CHD8, and CTCF, will be further addressed in the next chapters.

# 5 Purification and Characterisation of Human CHD8

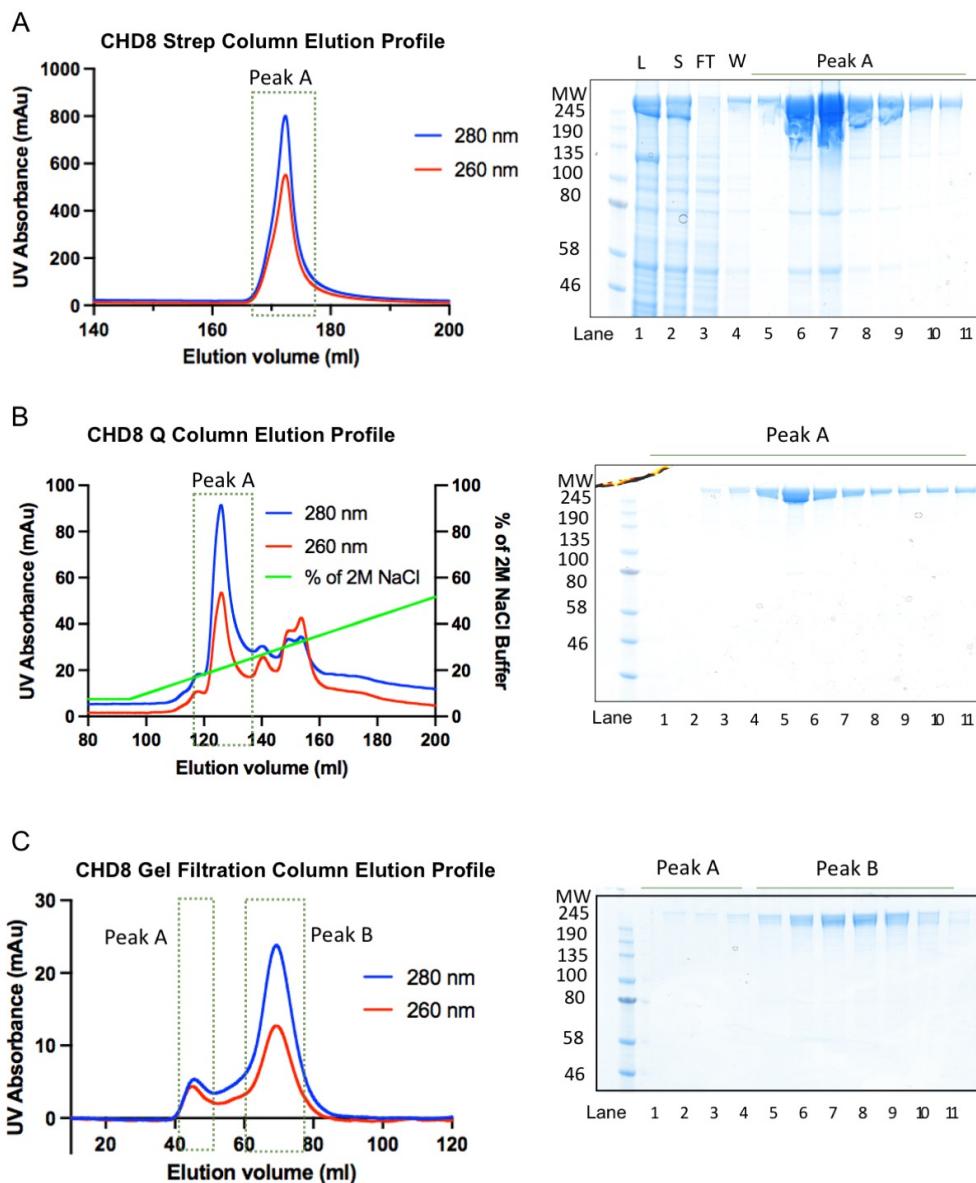
Based on previous publications, chromodomain helicase DNA binding protein 8 (CHD8) is the only protein that has been shown to associate with both ZNF143 and CTCF and hence CHD8 could be the mediator that facilitates the interaction between the two transcription factors (Ishihara, Oshimura and Nakao, 2006; Yuan *et al.*, 2007). CHD8 is an ATP-dependent chromatin remodelling protein that binds to nucleosome cores and facilitates nucleosome remodelling. It is also an important regulator of the chromatin structure and interacts with transcription factors to modulate the chromatin environment. This chapter focuses on the purification, functional characterisation of CHD8 alone and together with its nucleosome substrate. This will allow further functional and structural characterisation between CHD8, ZNF143 and CTCF.

## 5.1 Purification of Recombinant Human CHD8

### 5.1.1 Purification Protocol of CHD8

N-terminal twin-Strep tagged human CHD8 was expressed in High5 insect cells for four days prior to harvesting. The filtered lysate was loaded onto a StrepTrap column, followed by a HiTrap Q HP anion exchange column and passed onto a HiLoad 16/600 Superose 6 prep grade gel filtration column (Figure 5.1). Two peaks were observed from the Q column elution chromatogram (Figure 5.1B). The two peaks eluted at around 500 mM NaCl and 700 mM NaCl, by comparing the ratio between UV absorption at 260nm and 280 nm, these peaks corresponded to CHD8 alone and CHD8 bound to a DNA contaminant, respectively. Therefore, only fractions that were not contaminated by DNA were pooled and concentrated before passing through the gel filtration column. One peak was observed near the void volume (~40 ml) suggesting that a small portion of CHD8 was aggregated or formed a large oligomeric complex that was not resolved by the gel filtration column (Figure 5.1C). The second peak corresponded to the purified and non-aggregated CHD8 protein. The theoretical molecular weight of monomeric StrepII-CHD8

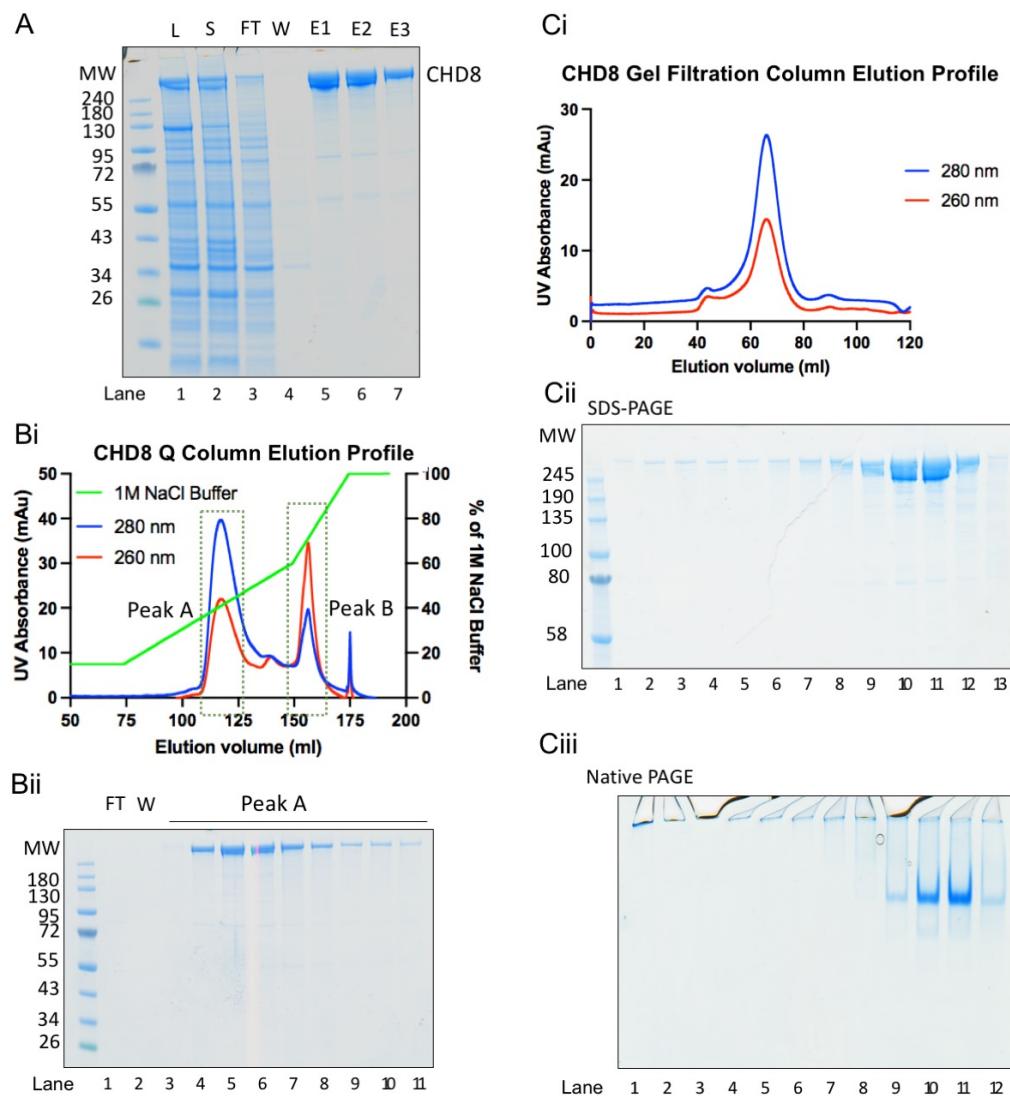
predicted from its protein sequence is 292 kDa, however the peak was eluted at around 70 ml from the gel filtration column, which was expected for a protein with the molecular weight of roughly 660 kDa. The discrepancy between the expected molecular weights might be due to mostly unstructured N- and C-terminal domains of CHD8 increasing the hydrodynamic radius of the particle and inflating molecular mass estimates derived from gel filtration. Another potential explanation is that CHD8 might form a dimer, increasing the measured molecular mass in solution, however the oligomeric state of CHD8 has not been previously characterised. The respective fractions were analysed by SDS-PAGE and pooled and stored at -80 °C. The yield was estimated to be 4 mg per 1 L of High5 insect cells.



### **Figure 5.1 Purification of StrepII-CHD8**

A: Elution profile and SDS-PAGE analysis of the fractions from the Strep affinity column. L: lysate (pre-centrifugation), S: supernatant (post-centrifugation), FT: flow-through, W: wash. B: Elution profile and SDS-PAGE analysis of Q column. C: Gel filtration chromatogram and SDS-PAGE analysis of the peak fractions.

During the purification of StrepII-CHD8, a few bands were observed near the expected CHD8 band in SDS-PAGE gel which might be due to degradation (Figure 5.1C). To improve the quality of the protein and to reduce non-specific degradation of CHD8 at the C-terminal domain, a new construct with an additional C-terminally 3x FLAG-tagged construct was designed and purified. Anti-FLAG beads were used in the first purification step as opposed to the StrepTrap HP column (Figure 5.2A). The following steps were identical to the purification protocol for StrepII-CHD8. Eluted CHD8 protein was diluted in low salt concentration buffer and passed through HiTrap Q HP column (Figure 5.2B) and Superose 6 16/600 column. The void volume peak was significantly reduced suggesting that C-terminally tagged CHD8 was less likely to aggregate (Figure 5.2Ci). This might be because CHD8 lacking the C-terminal tag was prone to degradation and the degraded product was less stable and tend to aggregate. Fractions from the main peak were analysed by SDS-PAGE and a few bands were still observed near the expected CHD8 molecular weight (Figure 5.2Cii). However, the results from the Native PAGE showed distinct bands confirming the homogeneity of the purified CHD8 protein (Figure 5.2Ciii). This result also suggested that although CHD8 eluted earlier than its expected volume, it might be largely due to the disordered regions rather than oligomerisation.



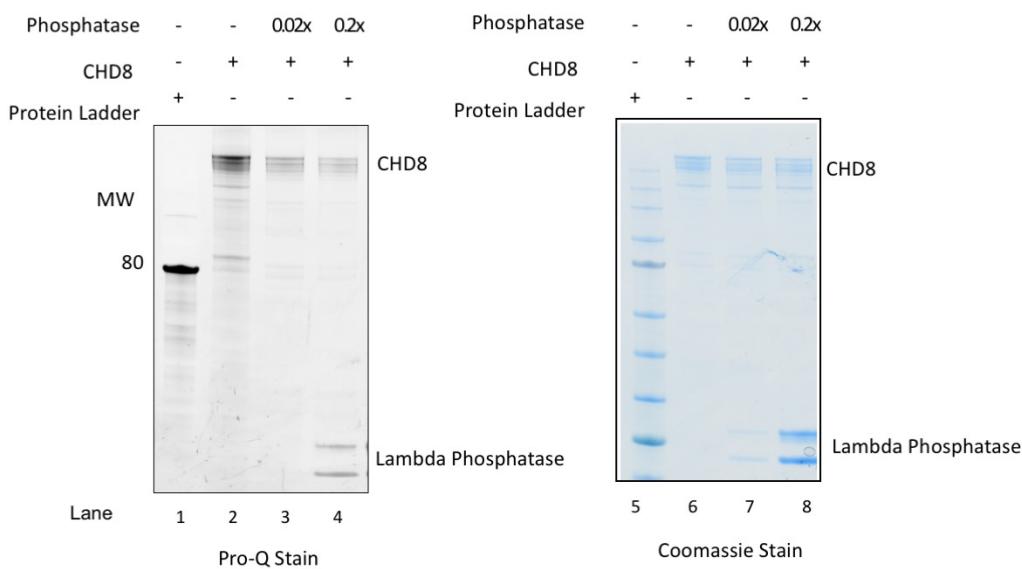
**Figure 5.2 Purification of StrepII-CHD8-FLAG**

A: C-terminally FLAG-tagged CHD8 was purified with anti-FLAG resins using gravity flow columns. L: Lysate before centrifugation, S: Supernatant after centrifugation, FT: Flow through, W: Wash, E: Elution. Bi: Q column chromatogram of CHD8. Bii: SDS-PAGE analysis of Peak A. Ci: Gel filtration column chromatogram of CHD8 purification. Cii: SDS-PAGE analysis of peak fractions. Ciii: Native PAGE analysis of peak fractions from gel filtration column.

### 5.1.2 Purified CHD8 is Phosphorylated

CHD8 was expressed from insect cells, and therefore the purified product was post-translationally modified. The laddering effect observed (Figure 5.2Cii) might be due to the presence of different phosphorylation states of CHD8 in the purified preparation. There are over 70 potential phosphorylation sites in CHD8 predicted from its protein sequence (Hornbeck *et al.*, 2015). Therefore,

a phosphorylation test was carried out using the Pro-Q Diamond phosphoprotein gel stain kit (Thermo Fisher Scientific). Proteins that are phosphorylated will be stained by the Pro-Q Diamond stain in the gel. CHD8 treated with the Pro-Q phosphoprotein kit was ran on SDS-PAGE and visualised using Typhoon imager with a filter at 580 nm. Since Lambda phosphatase releases phosphate groups from phosphorylated serine, threonine and tyrosine residues, it was also added to the CHD8 samples to confirm the bands that were stained by Pro-Q were due to phosphorylation. As the concentration of lambda phosphatase increased, the level of CHD8 band intensities decreased, confirming that the bands detected were due to phosphorylation (Figure 5.3). In addition, the laddering effect of the CHD8 bands might be due to the various phosphorylation states of CHD8. Yet, the function of different CHD8 phosphorylation states will require further investigation.

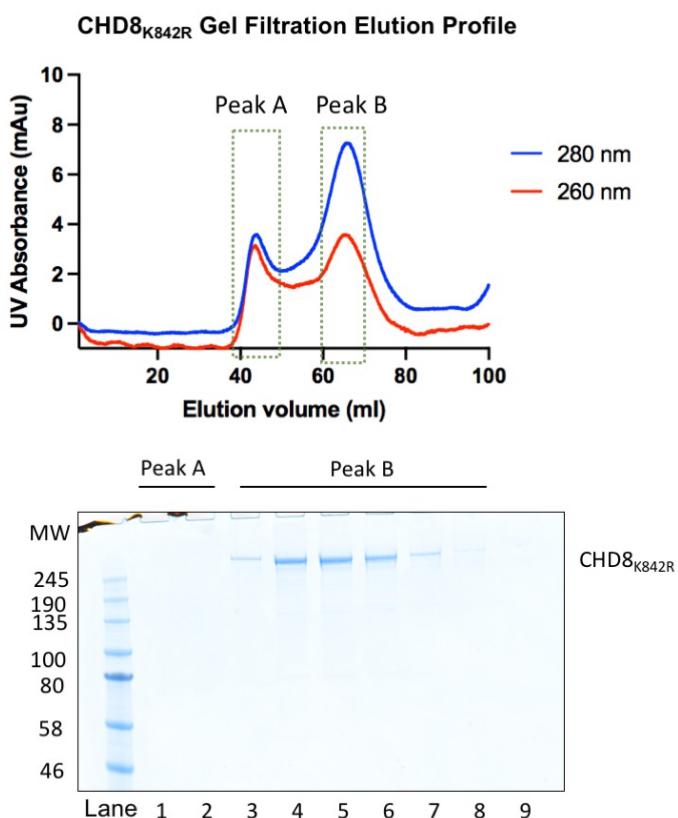


**Figure 5.3 CHD8 Treated with Pro-Q Diamond Stain**

0.7 mg/ml CHD8 was incubated with Pro-Q Diamond stain and SDS-PAGE was used to visualise the results. Two different concentrations of lambda phosphatase were added to the reaction. In lanes 3 and 4, the concentration of CHD8 was in 50-fold excess and 5-fold excess of lambda phosphatase, respectively. This gel was stained with Pro-Q diamond stain, followed by Coomassie blue. Protein standards were loaded onto lane 1, and the marker at 80 kDa was the only phosphorylated protein standard.

### 5.1.3 Purification of the ATPase-Deficient CHD8 K842R Mutant

As CHD8 is an ATP-dependent chromatin remodelling protein that displaces the nucleosome to expose the underlying DNA, it was intriguing to understand more about the ATP hydrolysis activity of CHD8 and to confirm that wild type CHD8 was functional and active. CHD8 with a point mutation at lysine 842 was also designed and purified as the mutation of lysine 842 to arginine (K842R) has been shown to impair the ATPase activity of CHD8 (Thompson, Lin and Bochar, 2008). The site-directed amino acid mutation was created by QuikChange mutagenesis kit. StrepII-CHD8<sub>K842R</sub> was expressed in High5 insect cells and purified using a StrepTrap HP column, followed by a Q column and Superose 6 16/600 gel filtration column. Purified CHD8<sub>K842R</sub> was separated from aggregated protein through the gel filtration column (Figure 5.4). The respective fractions were analysed by SDS-PAGE, pooled, concentrated and stored at -80 °C. The yield was estimated to be 0.4 mg per 1 L of High5 insect cells. The yield of mutant CHD8 was 10-fold lower than wild type CHD8 which might be due to instability at the ATPase domain induced by the mutation.



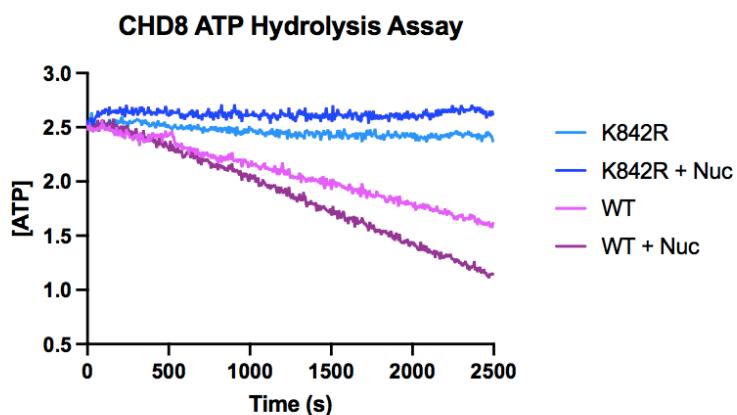
#### Figure 5.4 Purification of CHD8<sub>K842R</sub>

Gel filtration elution profile of CHD8<sub>K842R</sub> purification. Two peaks were observed in the chromatogram and the respective fractions were analysed by SDS-PAGE.

#### 5.1.4 Purified Wild Type CHD8 is Functional and Active

ATP/NADH coupled assay was used to measure the rate of ATP hydrolysis to understand more about CHD8 ATP hydrolysis activities and confirm the wildtype CHD8 was functional and active. Since the hydrolysis of ATP is enzymatically coupled to the oxidisation of NADH to NAD<sup>+</sup>, this assay measured the level of NADH absorption at 340nm over time.

Two CHD8 constructs were used for this assay, wild type CHD8 and CHD8<sub>K842R</sub> with a point mutation at lysine 842 that has been shown to impair the ATPase activity of CHD8 (Thompson, Lin and Bochar, 2008). Previous studies reported that wild type CHD8 showed higher ATPase activity in the presence of nucleosomes (Manning and Yusufzai, 2017) and therefore in addition to analysing ATPase rates of the two CHD8 constructs on its own, the ATP hydrolysis rate of CHD8 constructs incubated with nucleosomes were also measured. The results confirmed that CHD8<sub>K842R</sub> was not able to hydrolyse ATP in the presence or absence of nucleosome, as expected (Figure 5.5). Moreover, wild type CHD8 exhibited 2-fold higher ATPase activity in the presence of nucleosomes, which was in accordance with the previous study (Table 5.1) (Manning and Yusufzai, 2017). This experiment not only confirmed that the purified wild type CHD8 was functional, but also suggested that CHD8 exhibited nucleosome-stimulated ATPase activity.



### Figure 5.5 CHD8 ATPase Activity

ATP hydrolysis assay of wildtype CHD8 (WT) and mutant CHD8<sub>K842R</sub> (K842R) with and without nucleosome. The concentration of NADH absorption was converted to ATP concentration. The level of ATP remained was plotted against time. (Nuc: Nucleosome).

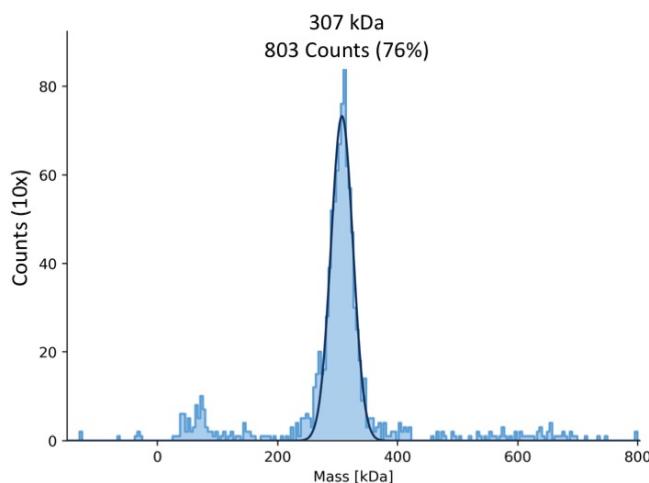
	ATP Hydrolysis Rate (ATP/sec/molecule)
<b>WT CHD8</b>	$0.304 \pm 0.018$
<b>WT CHD8 in the presence of Nuc</b>	$0.589 \pm 0.018$
<b>CHD8<sub>K842R</sub></b>	$0.057 \pm 0.021$
<b>CHD8<sub>K842R</sub> in the presence of Nuc</b>	$0.014 \pm 0.018$

Table 5.1 ATP Hydrolysis Rate of CHD8 Under Different Conditions

## 5.2 Biophysical Analysis of Purified CHD8

### 5.2.1 CHD8 Is Monomeric in Solution

Following the purification of functional CHD8, CHD8 was characterised using mass photometry. Although it has not been shown that CHD8 formed oligomers, it was intriguing to examine its oligomeric state in solution. The estimated molecular weight of CHD8 (with StrepII- and 3xFLAG-tag) based from its protein sequence (Gasteiger *et al.*, 2003) was 299 kDa. The measured molecular weight of CHD8 at 0.1 mg/ml was 307 kDa (Figure 5.6), suggesting that CHD8 behaved as a monomer in solution at the measured concentration.

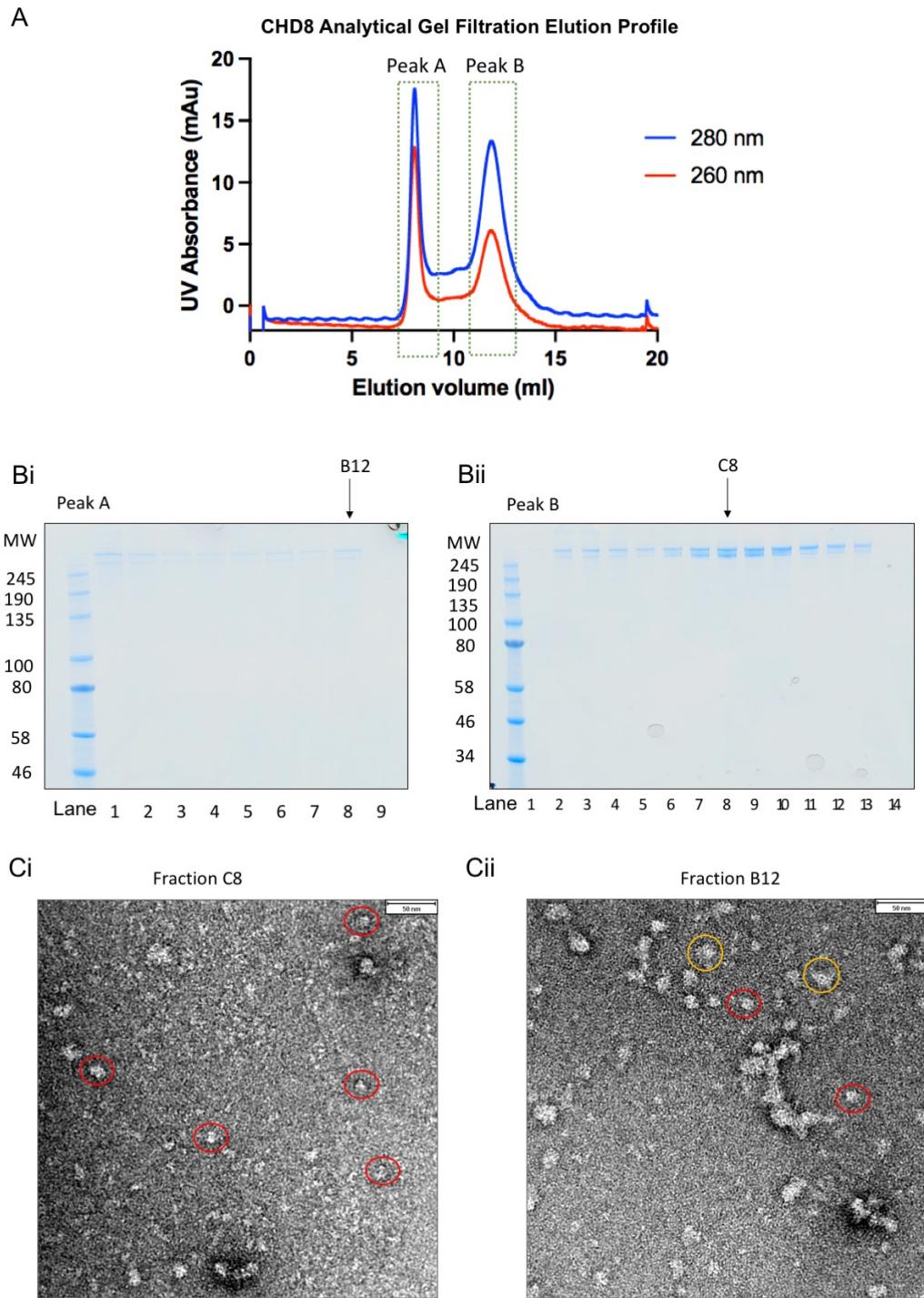


### **Figure 5.6 CHD8 Mass Photometry Analysis**

Full length CHD8 at 0.1 mg/ml was analysed by mass photometry. The measured size of each molecule adhered to the glass coverslip was recorded. The number of counts was plotted against the measured molecular weight. 76% of molecules measured had a molecular weight of around 307 kDa.

### **5.2.2 CHD8 Negative Stain Analysis**

Following in solution biophysical analysis, CHD8 was analysed via negative stain to characterise the sample further. Purified CHD8 was passed through another round of size exclusion chromatography to reduce the aggregates that might have formed during storage and the freeze-thaw cycle. Two peaks were observed from the elution profile, where the first peak eluted near the void volume (~7 ml) and the second peak corresponded to purified CHD8 (Figure 5.7A). Traces of CHD8 were observed near the void volume suggesting that a portion of CHD8 might be aggregated during the freeze-thaw cycle. Two fractions were screened using FEI T12 microscope. The micrograph with samples from fraction C8 showed very few particles and the particles were mostly globular in shape (~10 nm in diameter, red circles in Figure 5.7Ci)) with a few that were a lot bigger potentially due to aggregation. Fraction B12 that was eluted closer to the void volume was also analysed. The particles were larger in size (~20 nm diameter) (Figure 5.7Cii, orange circles) and aggregates were also observed. Due to the heterogeneity of the sample, large scale data collection and further structural analysis were not possible. Substrates such as nucleosomes might be required to stabilise CHD8 and therefore the next step was to investigate potential substrates that would form a stable complex with CHD8 to improve the homogeneity and stability of the particles for further structural analysis.



**Figure 5.7 CHD8 Negative Stain Analysis**

A: Elution profile of CHD8 loaded onto a Superose 6 10/300 gel filtration column. Two peaks were observed and labelled. Bi: SDS-PAGE analysis of the peak A. Bii: SDS-PAGE analysis of peak B. Ci: Negative stain micrograph of fraction C8 from peak B. Sample CHD8 particles were circled in red. Cii: Negative stain micrograph of fraction B12 from peak A. Sample CHD8 particles were circled in red and orange circles highlighted CHD8 particles that might be aggregated.

### 5.3 Nucleosome Binding Studies of CHD8

#### 5.3.1 CHD8 Binds to A Range of Different DNA Lengths

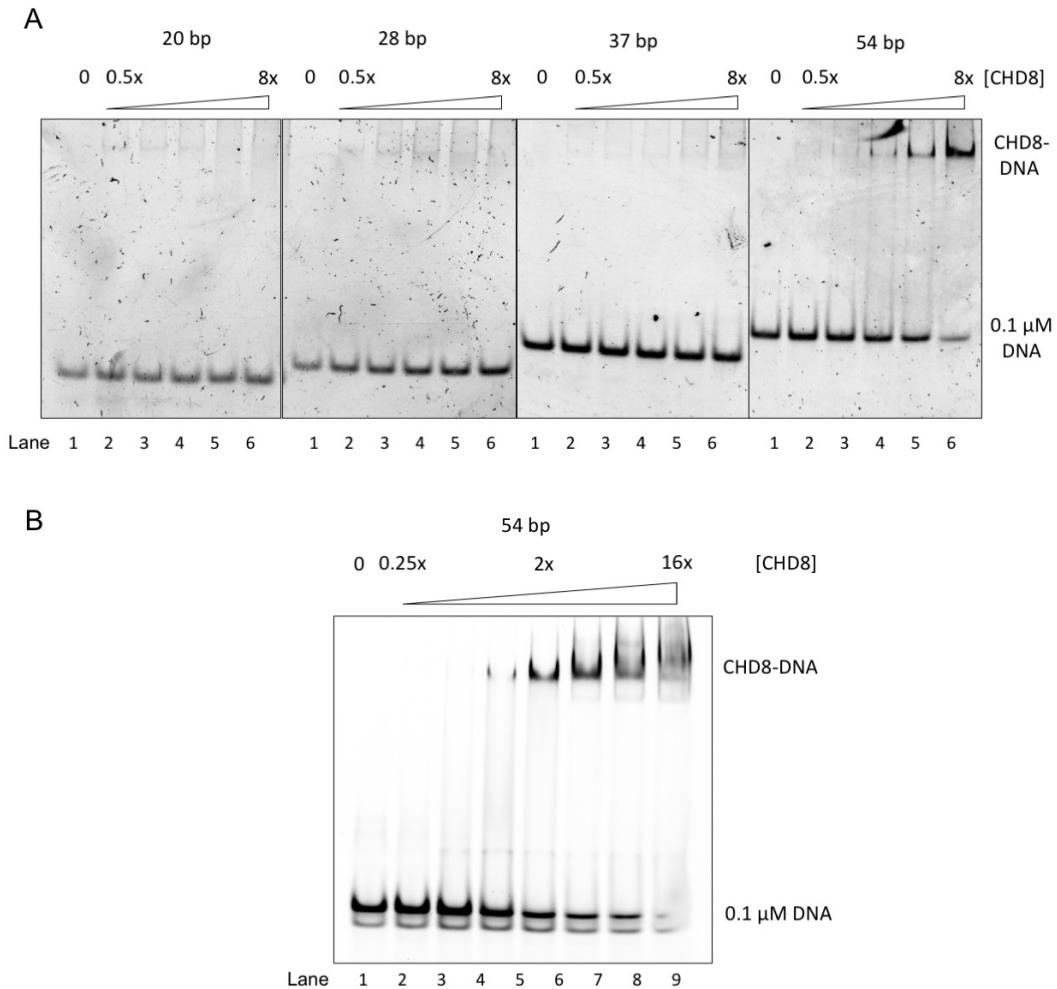
The two chromodomains in CHD8 allow it to bind to DNA without the requirement for a specific consensus sequence and previous study highlighted the importance of the contact between CHD enzymes and DNA to enhance the binding of CHD-nucleosome (McKnight *et al.*, 2011). However, another study demonstrated that CHD8 did not bind to DNA constructs that were under 40 base pairs long with high affinity (Manning and Yusufzai, 2017). Electrophoretic mobility shift assays (EMSA) were carried out to ensure the purified CHD8 interacts with DNA as previously described.

Purified CHD8 was incubated with four different double-stranded DNA constructs of varied lengths, ranging from 20 to 54 base pairs long (Table 5.2). The reaction samples were analysed on Native TBE acrylamide gels, stained with ethidium bromide and visualised using the appropriate filter on the Typhoon Image scanner. Weak interaction between CHD8 and DNA sequences shorter than 37 bp was observed, while the binding between CHD8 and the 54bp DNA was stronger based on the intensities of the CHD8-DNA bands (Figure 5.8A).

Length	DNA Sequence
20 bp	GCCAGCAGGGGGCGCTAGTG
28 bp	TATTGGCACCCACAATCCACCGCGGCTA
37 bp	CAGTGCCCACAGAGGCCAGCAGGGGGCGCTAGTGAGG
54 bp	GCTATGTGTTCTGGAAATCACCATAAACGTGAAATGTCTTGGATTTGGGAAT

**Table 5.2 DNA Sequences Used for CHD8 Binding Studies**

The EMSA experiment was repeated with CHD8 and the 54 bp DNA construct with a wider range of CHD8 concentrations to allow better understanding of the preferred binding ratio between CHD8 and DNA. Cyanine-5 (Cy5) labelled DNA was incubated with up to 16-fold molar excess of CHD8. The Native gel was visualised using the Cy-5 filter (Figure 5.8B). Binding between CHD8 and DNA can be observed starting 1:1 molar ratio, and complete DNA shift was observed when CHD8 was in 16-fold molar excess relative to the DNA.



**Figure 5.8 CHD8 DNA-Binding Study**

A: EMSA of CHD8-DNA binding. Four DNA constructs of different lengths were incubated with CHD8 and analysed using Native PAGE. Concentrations of CHD8 followed a 2-fold serial dilution. B: EMSA of CHD8 incubated with Cy-5 labelled 54 bp DNA analysed on Native PAGE.

### 5.3.2 Nucleosome Sequence Design

Nucleosomes are building blocks in various genome transactions such as DNA transcription, replication and repair. Yet, the DNA sequence that wraps around the histone octamers can vary (Lai and Pugh, 2017). As a result, correct design of the DNA sequence was required to ensure efficient formation of the CHD8-nucleosome-DNA complex. It has been shown that different sequences displayed different binding affinities towards the histone octamer and hence exhibited different nucleosome positioning abilities. A defined DNA sequence (Widom 601) designed by Lowary and Widom has been used to reconstitute nucleosomes *in vitro* (Lowary and Widom, 1998). This sequence

showed higher affinity towards the histones than its natural sequences. Moreover, it can robustly position nucleosomes on a DNA sequence, and therefore has been widely used for studies of chromatin structure and function (Lowary and Widom, 1998).

To examine the ability of CHD8 to remodel nucleosomes, a DNA linker of 52 or 63 base pairs was added to the exit side of the Widom 601 sequence (Table 5.3). These two linker sequences were designed based on endogenous sequences at RNA polymerase III type III promoters since CHD8 has been shown to be involved in transcription initiation at these promoter regions. Nucleosome-CHD1 (Farnung *et al.*, 2017) and nucleosome-CHD4 (Farnung, Ochmann and Cramer, 2020) structures solved through Cryo-EM showed that CHD proteins bind between the extranucleosomal DNA and the DNA gyre to form multiple DNA interactions, together with the DNA-binding results in section 5.3.1, a long DNA linker will be essential for strong interaction between CHD8 and nucleosome. These nucleosomal sequences were subsequently used for CHD8-nucleosome complex formation and investigation.

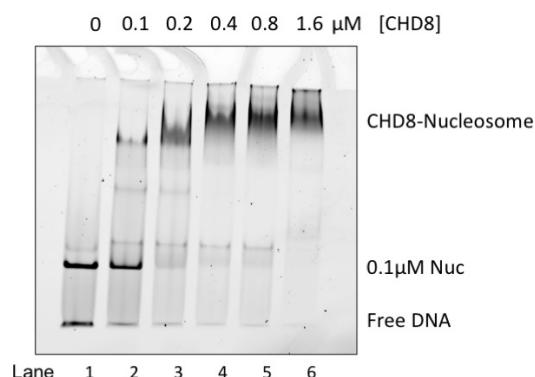
	<b>Construct Name</b>	<b>DNA Sequence</b>
<b>14</b>	Nucleosome _52bp	<b>CTGGAGAATCCCGGTCTGCAGGCCGCTCAATTGGTCGT AGACAGCTCTAGCACCCTAAACGCACGTACCGCGCTG TCCCCCGCGTTTAACCGCCAAGGGGATTACTCCCTAG TCTCCAGGCACGTGTCAGATATACATCCTGT</b> GGCATA AGCGTGGTTCAATACCGGCGCATAAAGGGTAAATTGAAAA CAGCG
<b>15</b>	Nucleosome _63bp	<b>CTGGAGAATCCCGGTCTGCAGGCCGCTCAATTGGTCGT AGACAGCTCTAGCACCCTAAACGCACGTACCGCGCTG TCCCCCGCGTTTAACCGCCAAGGGGATTACTCCCTAG TCTCCAGGCACGTGTCAGATATACATCCTGT</b> TAACGA TGCTGGGCATAAGCGTGGTTCAATACCGGCGCATAAAGGGTA AATTGAAAACAGCG

**Table 5.3 DNA Sequence Used for Nucleosome Assembly**

DNA comprised of the Widom 601 sequence and a 52 bp and 63 bp linker at the exit site were used to assemble the nucleosome for CHD8-nucleosome binding studies. Sequence highlighted in red refers to the Widom 601 sequence discovered by Lowary and Widom (Lowary and Widom, 1998).

### 5.3.3 CHD8 Binds to Nucleosomes

Different concentrations of CHD8 were incubated with nucleosomes with 63 base pairs positioned at the end. The nucleosomes were purified in collaboration with Dr. Thangavelu Kaliyappan (see methods, section 2.3.5). All nucleosomes were bound to CHD8 starting from two-fold molar excess of CHD8, suggesting a strong interaction between CHD8 and the nucleosome (Figure 5.9). Moreover, comparing the results with the CHD8-DNA EMSA (Figure 5.8B), it suggested that nucleosomes might be a better substrate for CHD8 since complete shift was observed at lower CHD8 concentrations.



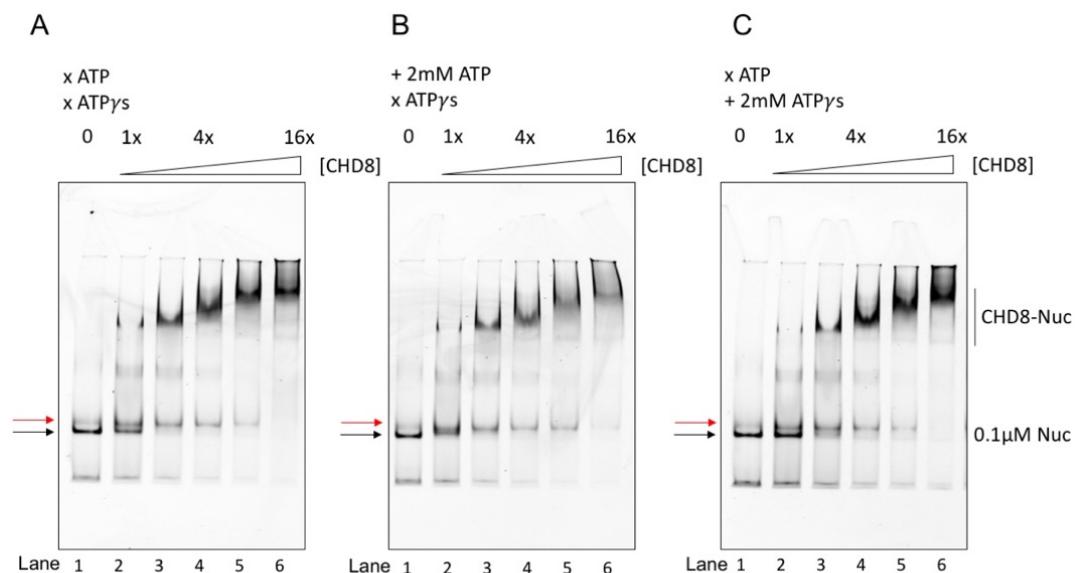
**Figure 5.9 CHD8-Nucleosome Binding Assay**

EMSA of CHD8 and Cy-5 labelled nucleosome with 63 bp linker at the exit side. The reaction samples were loaded onto Native PAGE and visualised with the ethidium bromide filter on the Typhoon imager.

### 5.3.4 Binding of CHD8 to Nucleosome is not ATP-Dependent

Following confirmation of CHD8-nucleosome complex formation in our system, further characterisation was carried out. CHD8 is an ATP-dependent chromatin remodelling protein, hence EMSA assays were performed to understand the importance of ATP in the observed binding of CHD8 to nucleosomes. In the first assay, titrating concentrations of CHD8 were incubated with nucleosomes (63bp linker) in the absence of ATP and the non-hydrolysable ATP analogue. In the second and third assays, the interaction buffer was supplemented with 2 mM ATP or 2 mM ATP $\gamma$ s, respectively. The intensities of CHD8-nucleosome band shifts were similar regardless of the type of nucleotide substrates, suggesting comparable binding affinities across different conditions (Figure 5.10A, B, C).

It was interesting to note that two bands were observed at where free nucleosomes ran in the Native PAGE. These bands refer to nucleosomes that were differently positioned along the DNA. Although the wrapping DNA was designed to position nucleosome at the end, a small fraction of nucleosomes might have positioned differently and could not be separated during purification. End-positioned nucleosomes (black arrows) tend to migrate faster than middle-positioned nucleosomes (red arrows) in a Native PAGE gel. In the presence of 2 mM ATP, the intensities of end-positioned nucleosomes were reduced when compared with the other two conditions (especially in lane 2 in figure 5.10B), hinting a possible nucleosome remodelling activity of CHD8, which will be further discussed in the next section.



**Figure 5.10 Nucleosome Binding is not ATP-dependent**

A: Native PAGE analysis of the interaction between CHD8 and nucleosome in the absence of nucleotides. Red and black arrows corresponded to the middle-positioned, and end-positioned nucleosomes, respectively. B: Native PAGE analysis of the interaction between CHD8 and nucleosome in the presence of 2 mM ATP. C: Native PAGE analysis of the interaction between CHD8 and nucleosome in the presence of 2 mM ATP $\gamma$ s.

### 5.3.5 CHD8 Remodels Nucleosomes

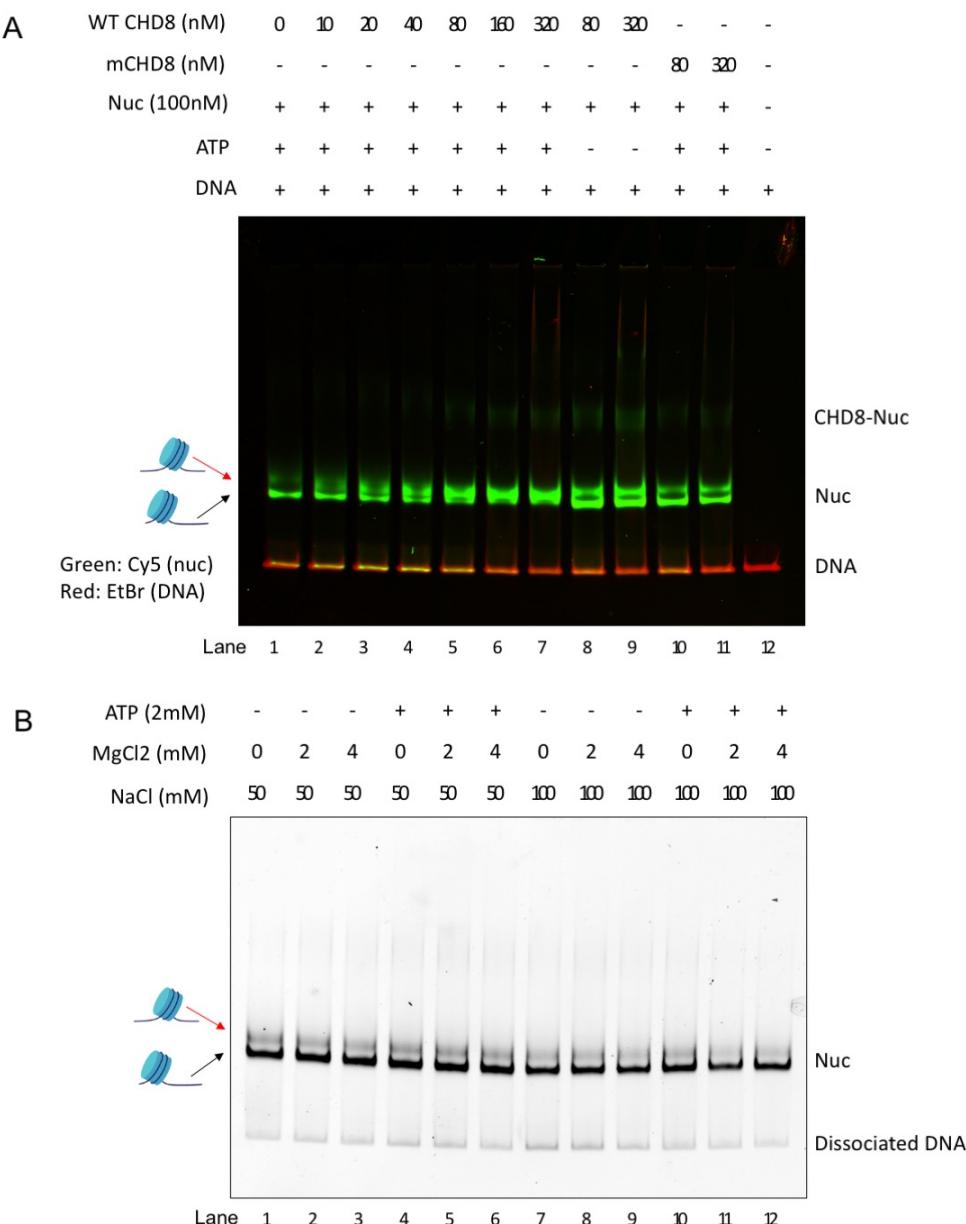
CHD8 has been reported to slide nucleosomes along DNA (Manning and Yusufzai, 2017), and therefore nucleosome remodelling assays were performed to confirm the observation. The assay was carried out with end-

positioned nucleosomes containing 52-bp linked DNA on the exit side incubated with increasing concentrations of CHD8 in the presence of ATP. The reaction was quenched with competitor 199-bp DNA and EDTA, then resolved using Native PAGE (Figure 5.11). As the end-positioned nucleosome slid across the DNA in the presence of CHD8, the nucleosome shifted to the middle of the DNA fragment, and this new middle-positioned nucleosome ran at a lower-mobility in a Native PAGE compared to the end-positioned nucleosome. In the presence of ATP, as the concentration of CHD8 increased, the intensity of the middle-positioned nucleosomes increased, suggesting that CHD8 was able to slide nucleosomes containing a 52-bp linked DNA (Figure 5.11A, lanes 1-7). In the absence of ATP, while a middle-positioned nucleosome band was visible at high CHD8 concentrations, it was significantly less than the equivalent CHD8 concentration in the presence of ATP (Figure 5.11A, lanes 7 vs lane 9), indicating that CHD8 was not able to remodel nucleosomes as efficiently.

Similar experimental setup was repeated with an ATPase deficient CHD8 mutant, the mutant CHD8<sub>K842R</sub> has been reported to lack the ATPase activity of CHD8 (Thompson, Lin and Bochar, 2008). Two concentrations of CHD8<sub>K842R</sub> were used. As shown in lanes 10 and 11 in figure 5.11A, all nucleosomes were converted to middle-positioned in WT CHD8 in the presence of ATP, whereas a marked population of end-positioned nucleosomes were observed for reactions without ATP or with mutant CHD8, suggesting a lower remodelling activity. This finding supported the previously reported hypothesis in Manning and Yusufzai where ATP hydrolysis drove the nucleosome remodelling activity in CHD8.

As a negative control, it was important to ensure the nucleosome do not remodel or shift its position on its own in the absence of a chromatin remodelling enzyme. Hence, nucleosomes were incubated with different concentrations of salt and ATP in the absence of CHD8 to verify that the nucleosome remodelling shift observed was not due to buffer conditions nor self-remodelling of nucleosomes (Figure 5.11B). The lower-mobility species

was not generated in this experiment. Thus, the results suggested that the middle-positioned nucleosomes were the product of CHD8 nucleosome remodelling in the presence of ATP, and the salt concentrations ( $MgCl_2$  and NaCl) did not modulate the level of nucleosome remodelling.



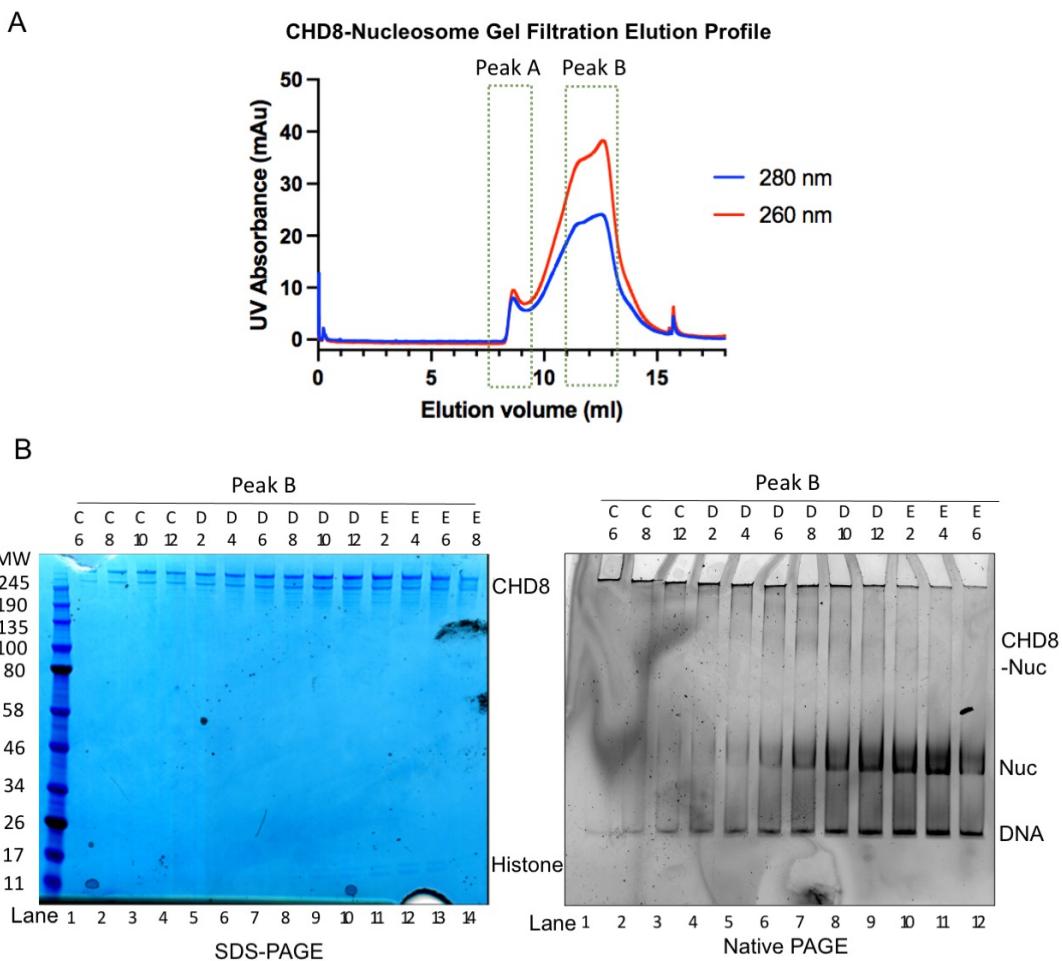
**Figure 5.11 CHD8 Nucleosome Remodelling Assays**

A: EMSA of CHD8 nucleosome remodelling assay. Cy-5 labelled nucleosome was incubated with low concentrations of CHD8 (WT CHD8) and CHD8<sub>K842R</sub> (mCHD8), in the presence and absence of ATP. Cy-5 labelled nucleosomes were showed in green, and the supplemented DNA (stained with EtBr) was shown in red. B: EMSA of nucleosomes under different conditions. Nucleosomes were incubated with different concentrations of  $MgCl_2$ , ATP and NaCl to understand the effect of these elements on the positioning of histones on DNA.

### 5.3.6 Negative Stain Trials and Analysis

Following biochemical analysis of active complex formation between CHD8 and nucleosomes, structural analysis was carried out to determine if nucleosome binding improved sample stability and particle quality sufficiently to allow for CHD8 structural determination. Initial structural analysis was carried out using negative stain EM to assess the overall shape of the complex and particle homogeneity before proceeding with cryo-EM. These experiments were performed in collaboration with Dr. Thangavelu Kaliyappan.

Two different sample preparation protocols were tested to improve the quality of the CHD8-nucleosome complex for negative stain EM analysis. The first protocol involved applying the pre-incubated CHD8-nucleosome complex onto a gel filtration Superose 6 10/300 column. CHD8 and nucleosome (63 bp linker) were mixed at a molar ratio of 2.2:1 and incubated on ice for 25 minutes in the presence of 3 mM ADP and Al(NO<sub>3</sub>)<sub>3</sub>. Two main peaks were observed on the gel filtration chromatogram (Figure 5.12A). The eluted fraction samples were analysed by SDS-PAGE and Native PAGE (Figure 5.12B). Due to the low molecular weight and low level of histones, the histones were not stained as well as CHD8 in the SDS-PAGE. The low level of histones might also be due to sub-stoichiometric complex formation where only a small portion of nucleosome was stably associated with CHD8 at any one time. The Native PAGE was stained with ethidium bromide and suggested that most aggregated CHD8 was present in the first peak which was closer to the void volume as seen by the molecules stuck at the top of the wells. As the elution volume increased to close to 11 ml (Fraction D8), nucleosomes were also present in the fractions as seen by the histone protein bands in the SDS-PAGE. However free nucleosomes and free DNA were present in native PAGE which might indicate the dissociation during the electrophoresis process or that further optimisation was required to improve the stability of the complex. Fractions D8 and D10 were applied to negative stain EM grids. The particles were very poorly stained suggesting that contaminants might be present, and the complex might not be stable, therefore these grids were not suitable for further characterisation (data not shown).

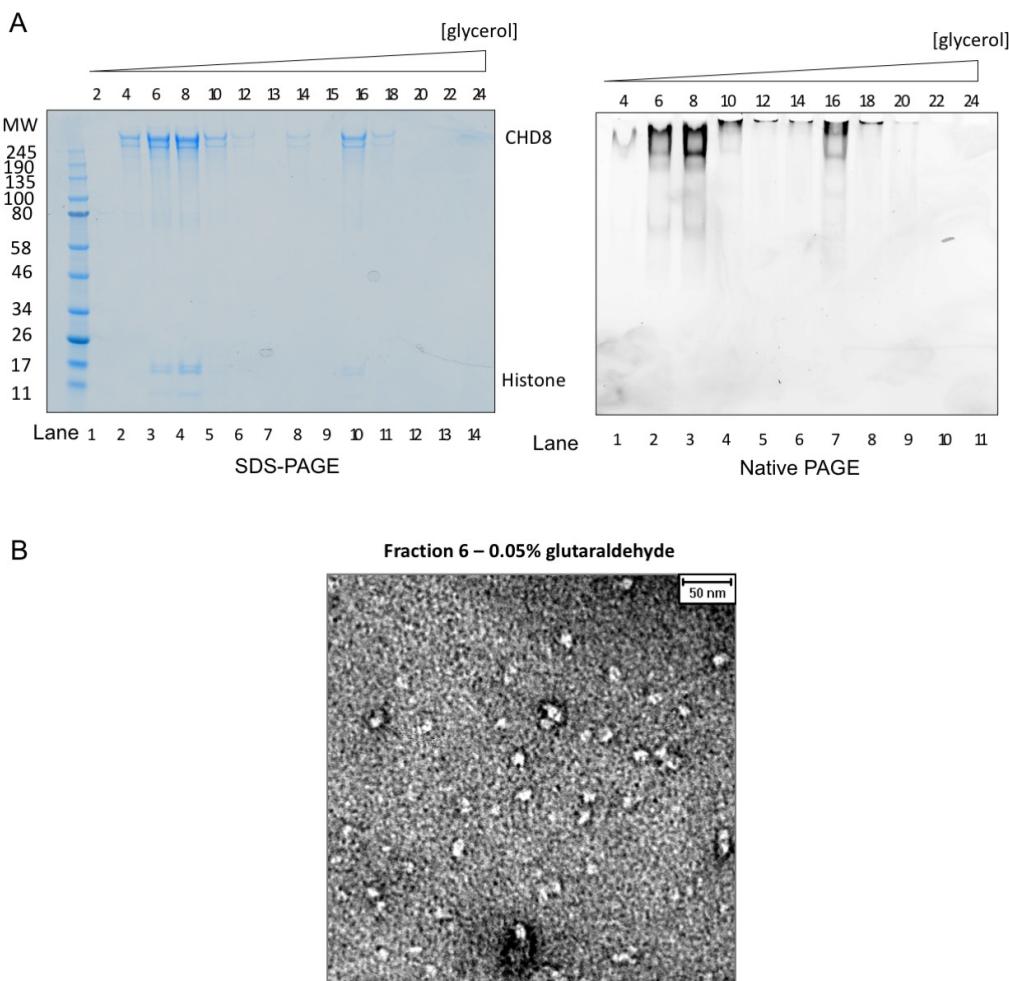


**Figure 5.12 CHD8-Nucleosome Gel Filtration Negative Stain Trial**

A: Elution profile of CHD8-nucleosome from Superose 6 10/300 gel filtration column. 2.2-fold excess of CHD8 was incubated with nucleosome prior to gel filtration, the buffer was supplemented with ADP and Al(NO<sub>3</sub>)<sub>3</sub>. B: SDS-PAGE and Native PAGE of the fractions eluted from the gel filtration.

The second sample preparation method used glycerol gradient to achieve sample homogeneity. Two-fold excess of CHD8 was incubated with nucleosome (63 bp linker) in the presence of 2 mM AMP-PNP. AMP-PNP is a non-hydrolysable ATP-analogue that might reduce the conformational changes from the nucleosome stimulated ATP hydrolysis, trapping the complex in a single state, and thus improve sample homogeneity. The CHD8-nucleosome sample was applied to the tube with a gradient of glycerol concentrations between 20% to 60%. By overnight ultra-centrifugation at 4 °C, the complex and free proteins were separated according to their molecular weights to different glycerol layers. 150 µl fractions were taken from the top of

the glycerol gradient tube and labelled as fraction 1, with 26 fractions collected in total and analysed using SDS-PAGE and Native PAGE (Figure 5.13A). CHD8-nucleosome complex was enriched between fractions 6 to 8 at around 35% glycerol density. However, these samples did not stain well on negative stain grids. To avoid the complex from dissociating, fraction 6 was crosslinked with 0.05% glutaraldehyde and this greatly improved the quality of the particles (Figure 5.13B).



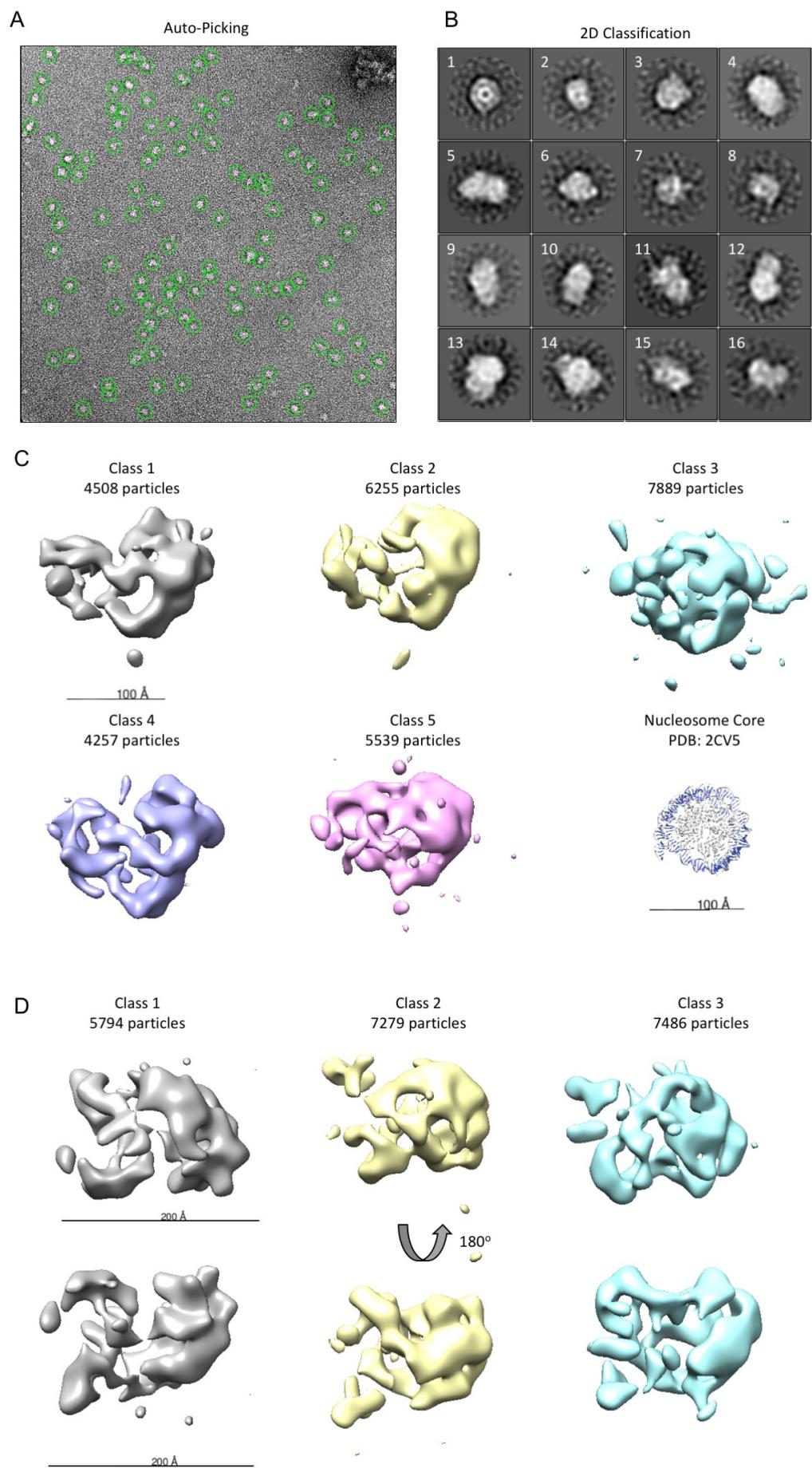
**Figure 5.13 CHD8-Nucleosome Glycerol Gradient Negative Stain Trial**

A: SDS-PAGE and Native PAGE analysis of the glycerol gradient fractions. 2-fold excess of CHD8 was incubated with nucleosomes in the presence of AMP-PNP prior to loading onto a glycerol gradient tube for overnight ultracentrifugation. The smaller the fraction number, the lower the glycerol content in the solution. B: Negative stain micrograph of fraction 6 cross-linked with 0.05% glutaraldehyde.

With the help of Dr. Fabienne Beuron, micrographs were collected at the FEI TF20 microscope. The data set was analysed and processed using Relion. Particles were manually picked, subjected to 2D classification and used as references for auto-picking (Figure 5.14A). The initial particle subset (41752 particles) was 2D classified. 2D class averages were generated after three rounds of 2D classification and subsequent class selection using a mask diameter of 280 Å.

The top 16 most populated 2D classes were shown in figure 5.14B. large portion of the particles were nucleosome only, as nucleosomes have a characteristic circular disc shape, similar to that observed in classes 1, 2, 6, 7, and 8. There were some classes that showed particles which were larger than the size of the particles in class 1, and these might represent different views of the CHD8-nucleosome complex, for example classes 9, 10, 11, 12, 13 and 16 showed extra features along with the nucleosome.

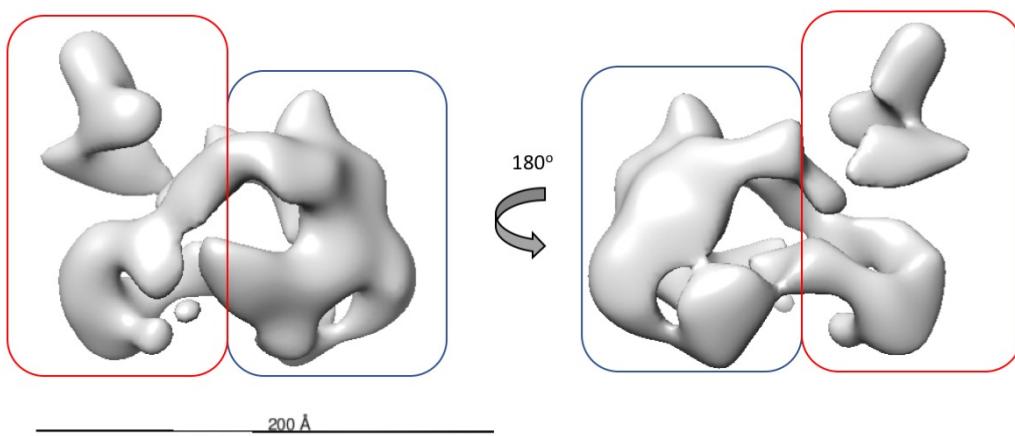
Particles from selected 2D classes were used to build an initial model, and a 3D classification into five different classes was performed. The crystal structure of the nucleosome core particle (PDB: 2CV5, (Tsunaka *et al.*, 2005)) was also included for a comparison between the sizes of different classes. Unsurprisingly the class with the greatest number of particles resembled nucleosome only (class 3, Figure 5.14C). Particles from class 3 were removed for the next round of 3D classification as the other classes were more likely to resemble the CHD8-nucleosome complex. Class 2 was used as the reference model. Three more classes were generated, the particles were evenly distributed across three classes, and their respective 3D maps and number of particles in each class were displayed in figure 5.14D.



### Figure 5.14 CHD8 Negative Stain Analysis

A: Negative stain micrograph with auto-picked particles circled in green. B: Top 16 classes from 2D classification. C: Five 3D models generated from the first round of 3D classification. A nucleosome core particle was also included for scale. D: 3D models generated from the second round of 3D classification.

The 3D map of class 3 in figure 5.14D had been further refined and the final resolution of the structure was 31.5 Å. Due to the low resolution of the map, fitting of any known structure would be ambiguous. However, at low resolution, the structure appeared to be comprised of two components, and the disc shape component might correspond to the nucleosome (blue box in Figure 5.15) (Tsunaka *et al.*, 2005), whereas the remaining component might contain CHD8 (red box). Furthermore, an approximation of the molecular weight using the Chimera Area and Volume measure and a converting factor of 1.21 Å<sup>3</sup> per Dalton, led to a molecular weight of around 570 kDa, which was in accordance with the expected molecular of the CHD8-nucleosome complex.

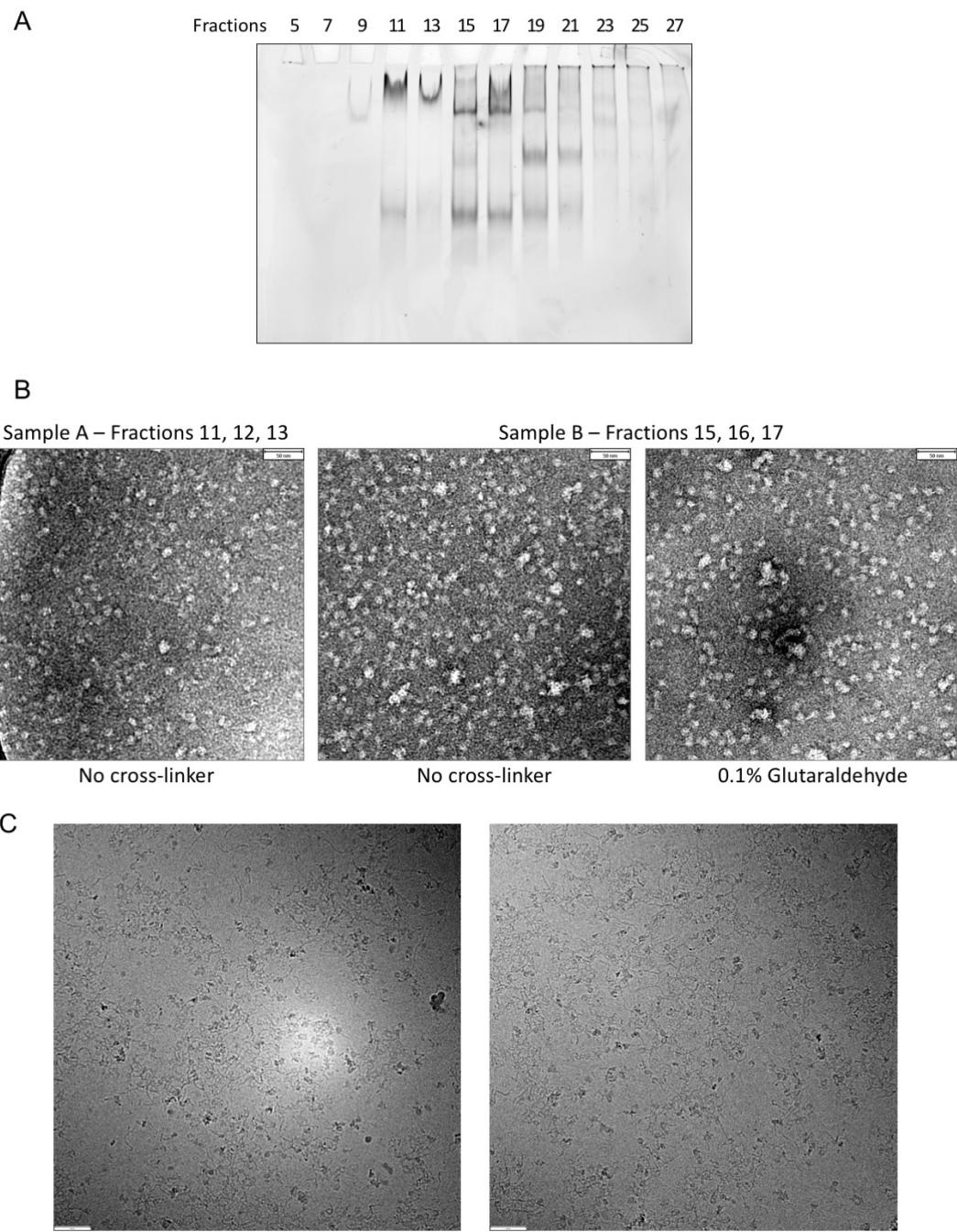


### Figure 5.15 Refined CHD8-Nucleosome 3D Map

A refined 3D model of the CHD8-nucleosome complex that appeared to be comprised of two components. A circular disc-shaped component was observed (blue box) and that might correspond to the nucleosome particle, whereas the other component might refer to the CHD8 molecule (red box). Due to the low resolution, further experiments would be needed to confirm this observation.

### 5.3.7 Cryo-EM Trial

Cryo-EM was the next step to improve the resolution and gain further structural insight into the CHD8-nucleosome complex. For the preparation of cryo grids, CHD8-nucleosome complex was prepared and purified as mentioned in the previous section using a glycerol gradient supplemented with AMP-PNP. Following the glycerol gradient, the complex was dialysed into the same buffer without glycerol. CHD8-nucleosome complex was enriched between 25 - 40% glycerol density (Figure 5.16A). Six fractions were pulled and split to 2 tubes, sample A was comprised of fractions 11, 12 and 13 (~25% glycerol), whereas sample B was comprised of fractions 15, 16 and 17 (~ 40% glycerol). After dialysis, both samples were concentrated and screened with negative stain EM (Figure 5.16B). Sample B that was crosslinked with 0.1% glutaraldehyde was chosen to proceed with cryo-EM since the particles were more homogenous than the other sample. The cryo grid was then screened inhouse, using the FEI TF20 microscope. From the cryo-EM micrographs, free DNA was observed, suggesting that the DNA wrapped around the nucleosome might be disassembled from the histones and therefore the CHD8-nucleosome complex was no longer stable (Figure 5.16C). However, it was uncertain if the DNA dissociated due to the change in buffer conditions during the dialysis process or the stringent cryo conditions during grid application. Therefore, further optimisation will be required to ensure the stability of nucleosomes for cryo-EM analysis.



**Figure 5.16 CHD8-Nucleosome Cryo-EM Trial**

A: Native PAGE analysis of the glycerol gradient experiment of CHD8 and nucleosome (63 bp linker) in the presence of 1 mM AMP-PNP. The glycerol densities for fractions 11 to 13 and fractions 15 to 17 were around 25% and 40%, respectively. B: Negative stain EM screening on sample A and sample B. Sample B was cross-linked with 0.1% glutaraldehyde. C: Cryo-EM screening with crosslinked sample B. DNA was observed suggesting the dissociation of DNA from nucleosomes.

## 5.4 Summary

CHD8 is an ATP-dependent chromatin remodeller that belongs to the CHD family and it is one of the least characterised CHD protein (Tyagi *et al.*, 2016). It has been shown to associate with both ZNF143 and CTCF (Ishihara, Oshimura and Nakao, 2006; Yuan *et al.*, 2007) and hence it was important to explore the functional characteristics of CHD8 to understand the role CHD8 plays in between ZNF143 and CTCF in genome organisation and transcription regulation. The purification protocol was described in this chapter, along with functional and structural characterisation of CHD8 alone and together with its nucleosome substrate. In brief, CHD8 was purified following a three-step purification protocol, with Strep affinity column, followed by Q ion exchange column and gel filtration. Purified CHD8 was heavily phosphorylated and exhibited as monomer in solution based on the results from mass photometry. Following in solution analysis, CHD8 was analysed via negative stain to characterise the sample further, however the samples were heavily aggregated and thereby nucleosomes were added as a substrate to stabilise the complex.

Functional studies were carried out between CHD8 and nucleosome prior to structural characterisation. The CHD8-nucleosome complex was stable and active as seen in the nucleosome binding assays *in vitro* where the binding of CHD8 to nucleosomes did not require ATP, yet the ability to remodel nucleosomes was ATP-dependent and CHD8 exhibited nucleosome-stimulated ATPase activity. Initial structural analysis was performed using negative stain electron microscopy. The CHD8-nucleosome complex was purified through glycerol gradient and the complex was crosslinked in the presence of an ATP analogue AMP-PNP to capture the complex in a single state. Nonetheless due to the dissociation of DNA from the CHD8-nucleosome complex during cryo-EM preparation, it was unable to perform further structural analysis. Therefore, further work will be needed to optimise the stability of the complex for cryo-EM analysis.

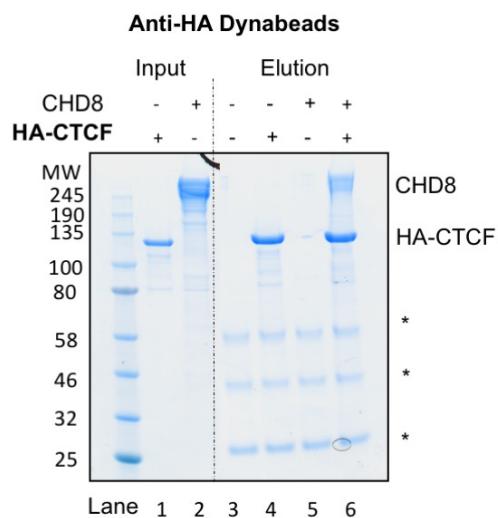
# 6 Interaction Studies of Genome Architecture Proteins

In previous chapters, the purification and functional characterisation of the individual genome architecture proteins, CTCF, ZNF143 and CHD8 was described. This section focuses on the relationships between these three proteins.

## 6.1 Interaction Studies Between CTCF and CHD8

### 6.1.1 CHD8 Interacts Directly with CTCF

CHD8 has been shown to occupy known CTCF binding sites *in vivo* to play an important role in insulation and epigenetic regulation (Ishihara, Oshimura and Nakao, 2006). Furthermore, direct binding between mouse CHD8 and mouse CTCF has been observed through pull-down analysis (Ishihara, Oshimura and Nakao, 2006). To confirm the direct interaction, pull-down assay was performed with recombinantly purified human CTCF and CHD8. Anti-HA magnetic beads were used to capture HA-tagged CTCF, and CHD8 was added subsequently. CHD8 only was added to the beads as a control to confirm that CHD8 did not bind to the beads non-specifically (Lane 5, Figure 6.1). The experiment showed that CTCF immobilised by HA beads was able to pull down CHD8 (Figure 6.1), suggesting a direct interaction.

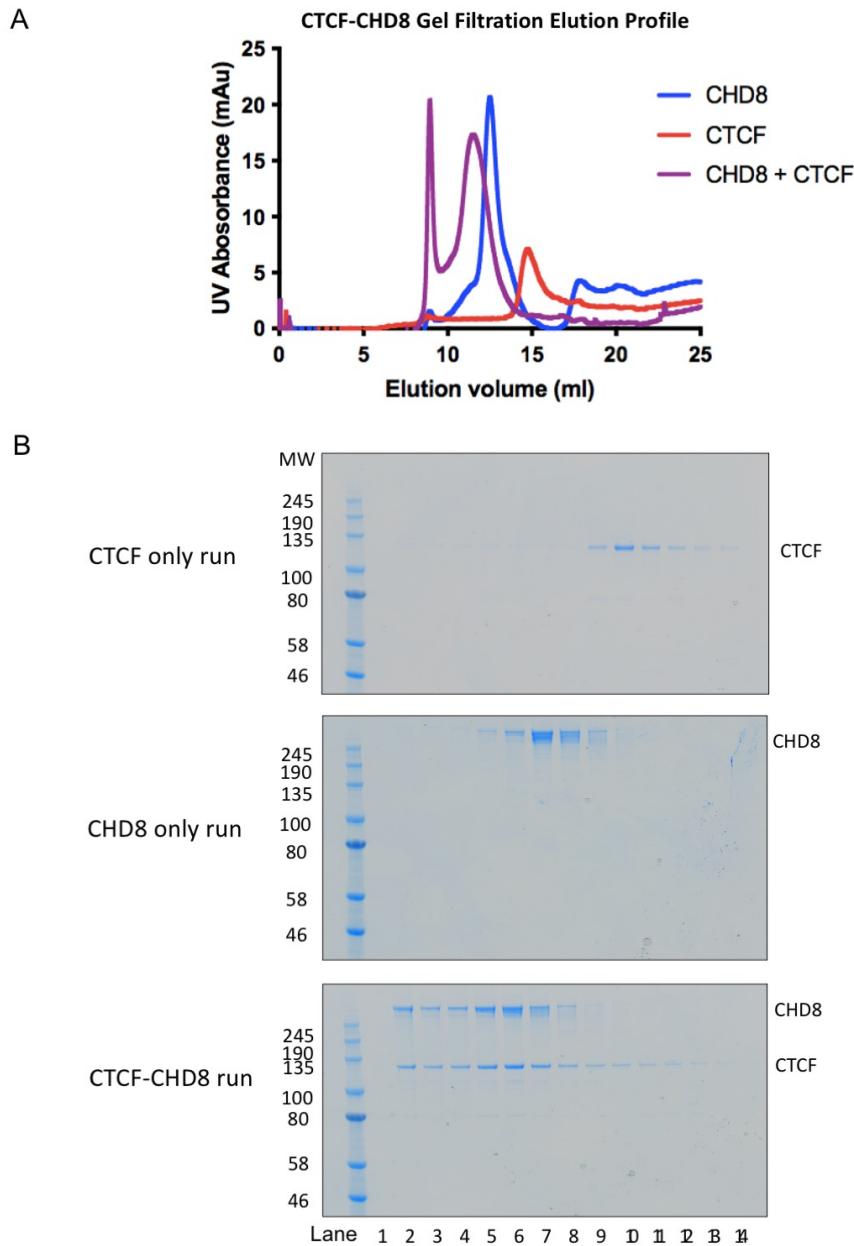


### **Figure 6.1 Direct Interaction Observed Between CTCF and CHD8**

SDS-PAGE analysis of the pull-down assay between recombinantly purified CTCF and CHD8. HA-tagged CTCF was immobilised onto anti-HA dynabeads as the bait. Bands marked with asterisks (\*) corresponded to the heavy and light chains of anti-HA antibodies.

#### **6.1.2 CTCF and CHD8 Form a Stable Complex Over Gel Filtration**

Since CHD8 colocalises with CTCF (Ishihara, Oshimura and Nakao, 2006), it was important to test whether the direct interaction observed was stable or transient. Using a Superose 6 10/300 GL column, three runs were carried out (CHD8 only, CTCF only, CHD8-CTCF complex) to investigate the stability of CHD8-CTCF complex (Figure 6.2A). Four-fold excess of CTCF was incubated with CHD8 to saturate all available CHD8 molecules, CHD8 alone and CTCF alone control runs were also performed to identify the CHD8-CTCF complex peak that was expected to elute at an earlier volume, indicating the presence of a larger species. Fractions were taken across 8 ml to 16 ml from all three runs and analysed on SDS-PAGE to confirm the presence of each protein. CHD8 only eluted at close to 12 ml, whereas CTCF only eluted at 15 ml. The two peaks observed in the CHD8-CTCF complex run, eluted at close to void volume (~7 ml) and 11 ml, respectively. Since the second CHD8-CTCF peak was eluted earlier than the individual protein peaks and SDS-PAGE analysis confirmed both CHD8 and CTCF were present, this experiment suggested a stable and strong interaction between human CHD8 and CTCF. The samples for SDS-PAGE gels for all three runs were taken from the same set of fractions (Figure 6.2B), therefore the gels could be compared directly as each lane of all three gels represented the same elution volume. Together, these results indicated that the CHD8-CTCF complex formed a stable complex over gel filtration.



**Figure 6.2 Analytical Gel Filtration of CTCF and CHD8**

A: Elution profile of three analytical gel filtration runs: CTCF only, CHD8 only and CTCF-CHD8 run. The UV absorbance at 280nm of all runs were plotted against the elution volume. B: SDS-PAGE analysis of the peaks from the three runs. The same set of fractions were analysed across the three runs.

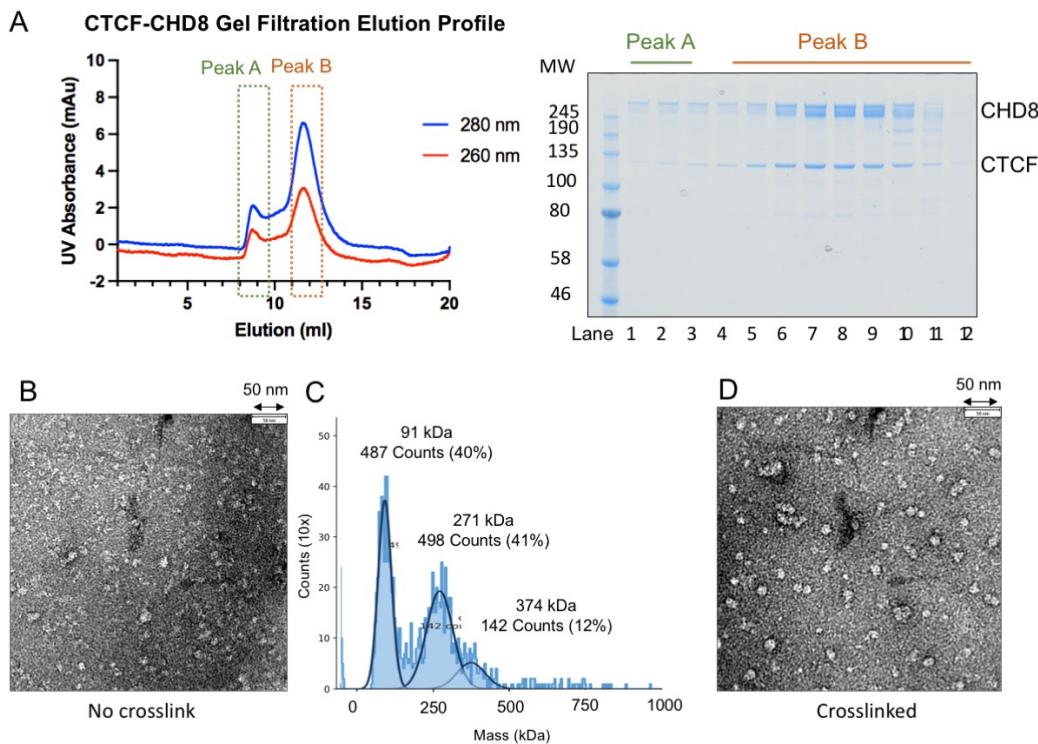
### 6.1.3 Negative Stain EM Trials and Analysis

The stable interaction between CHD8 and CTCF suggested that this complex may have a prominent role in chromatin regulation. The complex was analysed using negative stain to assess at low resolution the sample quality and structure prior to cryo-EM analysis.

In the first negative stain electron microscopy trial, CHD8 was incubated with excess CTCF, followed by gel filtration to separate aggregates and excess substrates from the CTCF-CHD8 complex. The gel filtration chromatogram showed two main peaks, where the first peak A eluted close to the void volume (~ 8 ml), and the second peak B (~ 12 ml) corresponded to the CHD8-CTCF complex as detected on the SDS-PAGE. The peak fraction of peak B (lane 8, Figure 6.3A) was stained with uranyl acetate on a carbon-supported grid for EM analysis and screened on the FEI T12 electron microscope.

The micrographs showed very high background. Moreover, particles of several different sizes were stained, this might be due to the dissociation between CTCF and CHD8, and therefore different species were observed on the micrographs, which might represent CTCF alone, CHD8 alone and the CTCF-CHD8 complex (Figure 6.3B). Due to the heterogeneity of the samples, further structural analysis was not possible. This fraction was also analysed using mass photometry to understand the different species present that gave rise to the heterogeneity observed on the micrographs (Figure 6.3C). Three main peaks were detected, and the measured molecular weights of these species were 91 kDa, 271 kDa, and 374 kDa respectively. These three values were in accordance with the estimated molecular weights of CTCF, CHD8 and CTCF-CHD8 complex based on their protein sequences. This further demonstrated that although CHD8 and CTCF eluted at the same peak, a portion of the complex might have dissociated over time, and the complex may not be sufficiently stable for structural determination.

To overcome the dissociation problem, the same fraction was cross-linked with 0.05% glutaraldehyde for 10 minutes prior to applying onto the copper grids and stained with uranyl acetate. The cross-linked particles were more homogenous when compared to the pre-treated sample. However, different sizes of particles were still present and as a result, the sample was still too heterogeneous for further analysis (Figure 6.3D).

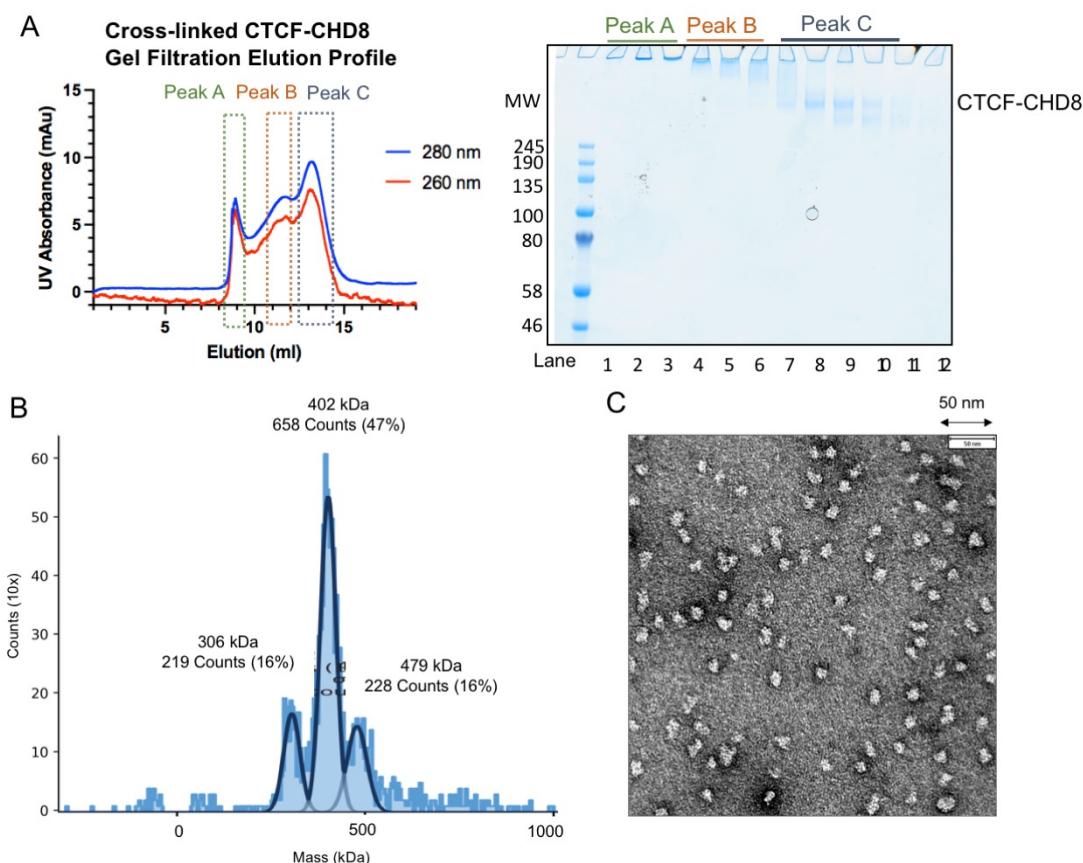


**Figure 6.3 First CTCF-CHD8 Complex Negative Strain Trial**

A: Elution profile and SDS-PAGE analysis of analytical gel filtration of CTCF-CHD8 complex. Two peaks were observed from the gel filtration, where the first peak corresponded to mainly aggregated proteins and the second peak corresponded to the CTCF-CHD8 complex. B: Negative stain micrograph of CTCF-CHD8 complex from the gel filtration chromatography (fraction from lane 8, Figure 6.3A) without cross-linking. Particles of various sizes were observed. C: Mass photometry analysis of the peak fraction in peak B and three species were detected, corresponding to CTCF alone, CHD8 alone and CTCF-CHD8 complex. D: Negative stain micrograph of CTCF-CHD8 complex after crosslinking with 0.05% glutaraldehyde of the same fraction (lane 8 in Figure 6.3A).

The next trial involved crosslinking CTCF and CHD8 with 0.08% glutaraldehyde, and then running the sample through a gel filtration column. Size exclusion chromatography allowed the separation of CHD8-CTCF complex from the over- and under- cross-linked species. Three peaks were observed and analysed using SDS-PAGE. The first two peaks were mostly aggregated and over-crosslinked proteins that were not able to enter the gel matrix. Macromolecule eluted from the third peak corresponded to the cross-linked CTCF-CHD8 complex. To confirm the stoichiometry of the species, the peak fraction (Lane 8, Figure 6.4A) was analysed with mass photometry

(Figure 6.4B). The dominating species in this fraction had a molecular weight of 402 kDa, consistent with the estimated molecular weight of CTCF-CHD8 at a 1:1 molar ratio. Therefore, this sample was imaged following negative staining using an FEI T12 microscope (Figure 6.4C).



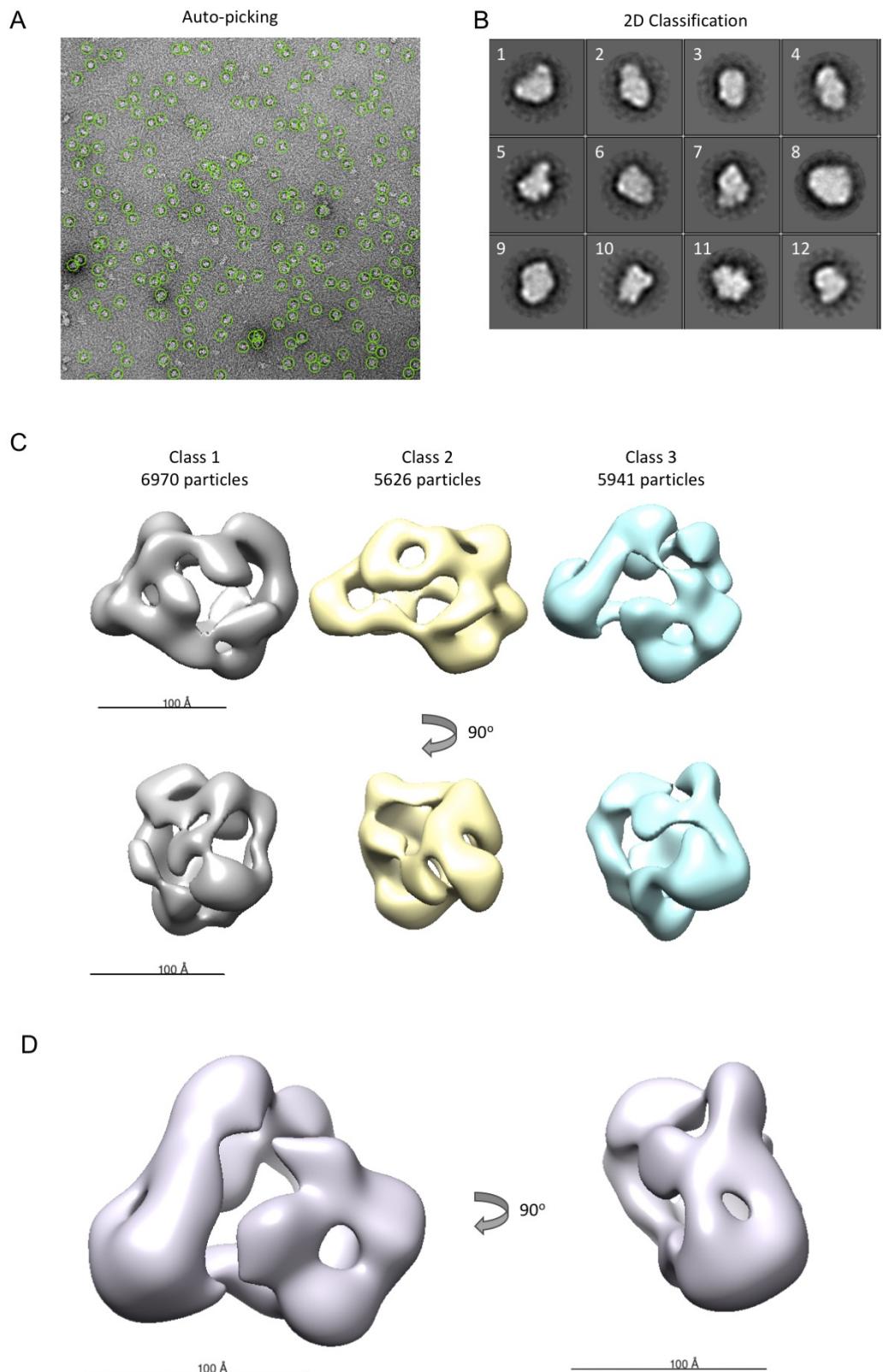
**Figure 6.4 Cross-Linked CTCF-CHD8 Complex Negative Stain Trial**

A: The elution chromatogram and SDS-PAGE analysis of the crosslinked CTCF and CHD8 complex through gel filtration column. The CTCF-CHD8 complex was observed in peak C. B: Mass photometry analysis of the fraction used for negative stain data analysis and data collection (Lane 8, Figure 6.4A). C: Negative stain micrograph of peak fraction from peak C (Lane 8, Figure 6.4A).

Particles of similar shapes and sizes were observed, suggesting that the homogeneity of the particles improved a lot when compared to the previous trials. 399 micrographs were collected and imported into Relion 3.1.1 for analysis. Through a combination of manual and auto-picking (Figure 6.5A), a total of 45,496 particles were picked and went through five rounds of 2D classification. Twelve of the selected 2D classes were shown in figure 6.5B,

using a mask of 260 Å. An *ab initio* model was generated using 18,357 particles and further classified into three 3D classes (Figure 6.5C).

Class 3 was further refined due to its high resemblance of the 2D classes (Figure 6.5D). The final resolution of the 3D map was 28.9 Å. An approximation of the molecular weight using the Chimera Area and Volume measure and a converting factor of 1.21 Å<sup>3</sup> per Dalton, led to a molecular weight of around 400 kDa, which was in accordance with the expected molecular of the CTCF-CHD8 complex. Although the structure for full length CHD8 and CTCF have not been reported, crystal structures of some of the domains have been solved. These include the structure of the zinc finger region of CTCF (Hashimoto *et al.*, 2017) and the BRK domain of CHD8 (PDB: 2CKA, not published) are known. The structure of CHD4, another chromatin remodelling protein has recently been solved cryo-EM (Farnung, Ochmann and Cramer, 2020). Through Phyre2 analysis, the protein structure prediction of the ATPase domain of CHD8 was predicted to be similar to the ATPase domain of CHD4 with high confidence (Kelley *et al.*, 2015). However, due to the low-resolution negative stain structure, the known structures of CHD8 and CTCF could not be fitted.

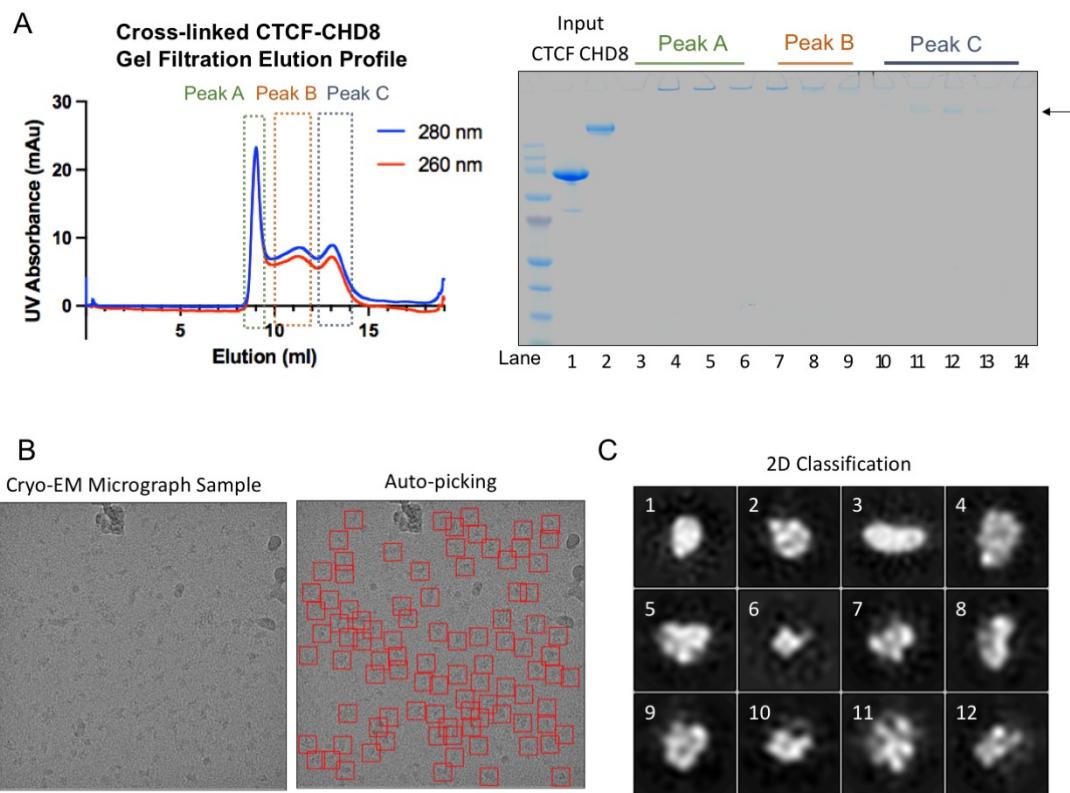


### **Figure 6.5 Negative Stain Analysis of Cross-Linked CTCF-CHD8 Complex**

A: Micrograph with auto-picked particles circled in green. B: Top 12 classes from 2D classification. C: Three 3D models generated after five rounds of 2D classification. D: Refined 3D model of the CTCF-CHD8 complex.

#### **6.1.4 Cryo-EM Trial and Analysis**

The crosslinked CTCF-CHD8 complex was taken forward to cryo-EM analysis to gain further structural insight. Crosslinked CTCF-CHD8 complex was purified through gel filtration with a buffer lacking glycerol (Figure 6.6A). The peak fraction from peak C (Lane 12, Figure 6.6A) was applied onto a copper grid coated with a layer of carbon. The sample was first screened with negative staining (data not shown) before proceeding to inhouse data collection with the Glacios™ Cryo-TEM connected to the Falcon 3EC direct camera. 850 micrographs were collected. The micrographs were motion corrected and the CTF estimation were performed with Relion 3.1.1. The micrographs were then imported into crYOLO for particle picking. Using the general model for auto-picking in crYOLO, approximately 45,000 particles were picked (Figure 6.6B). The picked particles were imported back into Relion for further processing. Following a few rounds of 2D classification, the number of particles was reduced to ~ 20,000 and the top twelve 2D classes were shown in figure 6.6C. Although these particles did not proceed to 3D classification due to the low signal to noise ratio in the 2D classes which might arise from the carbon artefacts, these cryo-EM 2D classes showed similar features to the 2D classes observed in the negative stain data collection, for instance class 1, 7, and 12 in the negative stain analysis (Figure 6.5B) were similar to the classes 5, 4, and 9 in cryo-EM analysis (Figure 6.6C), respectively. Further optimisations will be needed to improve the signal to noise ratios to achieve higher resolution structure, and this will be addressed in the discussion.



**Figure 6.6 Cryo-EM Trial**

A: Elution profile and SDS-PAGE of crosslinked CTCF-CHD8 complex over gel filtration. Three peaks were observed, and the black arrow indicates the presence of CTCF-CHD8 complex observed in peak C. B: Sample of cryo-EM micrograph and the auto-picking performed by crYOLO where picked particles were marked with red boxes. C: Top 12 classes from 2D classification processed with Relion.

### 6.1.5 Domains of CTCF and CHD8 Involved in Their Interaction

#### 6.1.5.1 Cross-link Mass Spectrometry

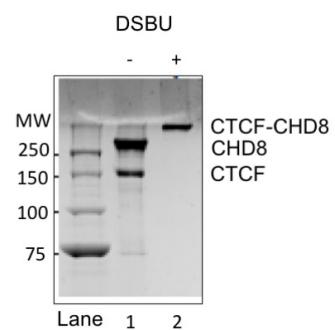
In addition to exploring the interaction between CTCF and CHD8 through electron microscopy techniques, cross-link mass spectrometry was carried out to pinpoint the domains involved in the CTCF-CHD8 interaction. Cross-link mass spectrometry provides an understanding of how the proteins might be packed in the three-dimensional space, hence this experiment would provide great insights into the mechanism of the interaction between CTCF and CHD8. This experiment was performed in collaboration with Dr. Duccio Conti from the Max Planck Institute in Germany. Purified CTCF and CHD8 were cross-linked with disuccinimidyl dibutyric urea (DSBU) with a spacer armlength of 12.5 Å

(Figure 6.7A). The sample was digested to produce a mixture of linear and cross-linked peptides, the enriched cross-linked peptides were sent to the MS facility at MPI for analysis. The results from the experiment were plotted in a circular diagram, where interactions scoring above the threshold of 70 and passed the 0% false discovery rate were shown. The proteins used in this experiment were tagged, and therefore the actual amino acid positions were shifted by the length of the tags as explained in figure 6.7B. Based on the results, one particular region of CTCF at around 450 aa interacted with multiple sites of CHD8, and these sites were mainly concentrated in the CHD8 C-terminal region between 1750 aa and 2550 aa (Figure 6.7B). The region near 450 aa in CTCF corresponded to the zinc-finger domain (around zinc finger 7), and the SANT and BRK domains were present in between 1750 and 2250 aa of CHD8. Lysine and serine residues were distributed throughout both proteins, and hence the interactions observed were not due to the bias of lysine positions.

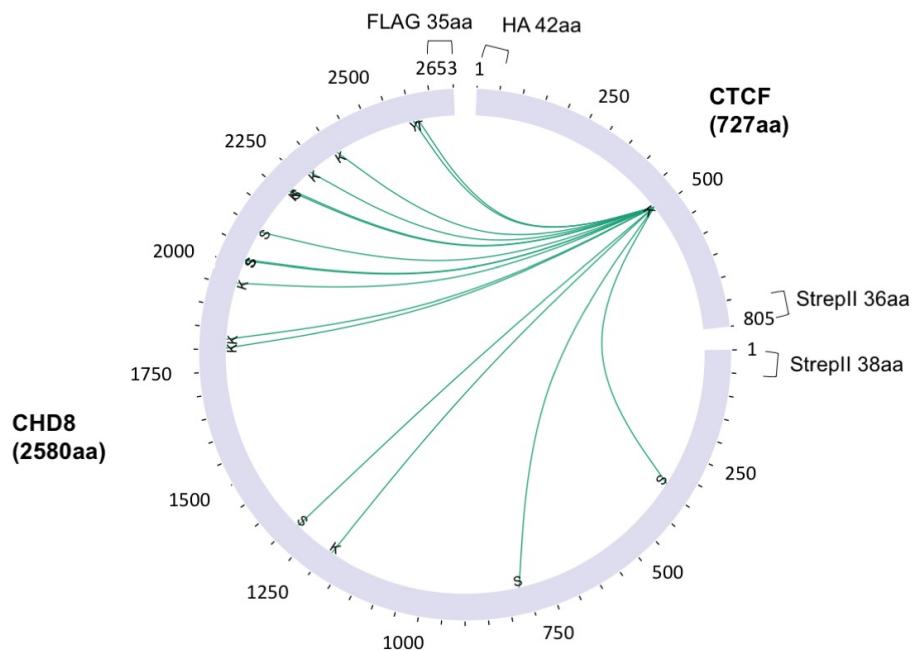
In addition to the regions where CTCF and CHD8 showed cross-linking, it was also interesting to understand the self-associating regions within these two proteins. The results suggested that CHD8 was tightly packed as seen by the extensive self-associating cross-links starting from 600 aa (Figure 6.7C). This was consistent with the characterised domain architecture of CHD8 since the first reported structured domains in CHD8 are the two N-terminal chromodomains located between residues 640 – 790 (Marfella and Imbalzano, 2007). Moreover, the results revealed that the chromodomains might be in close proximity with the C-terminal region. Extensive cross-linking was also observed within residues 1100 to 1750 which was expected as this region corresponded to the SF2 helicase domain that has been shown to be tightly packed for ATP hydrolysis (Farnung *et al.*, 2017). On the other hand, only a few self-crosslinking was observed above the score threshold for CTCF. The most abundant self-association was found at residues near 450 aa, which were also involved in the CTCF-CHD8 interaction. Lastly, a marked portion of the self-interactions within CHD8 overlapped with the regions involved in the CTCF-CHD8 interactions (Figure 6.7D). This finding corroborated that the self-

interacting regions in CHD8 were packed in close proximity and these regions might be the key for CTCF-CHD8 interactions.

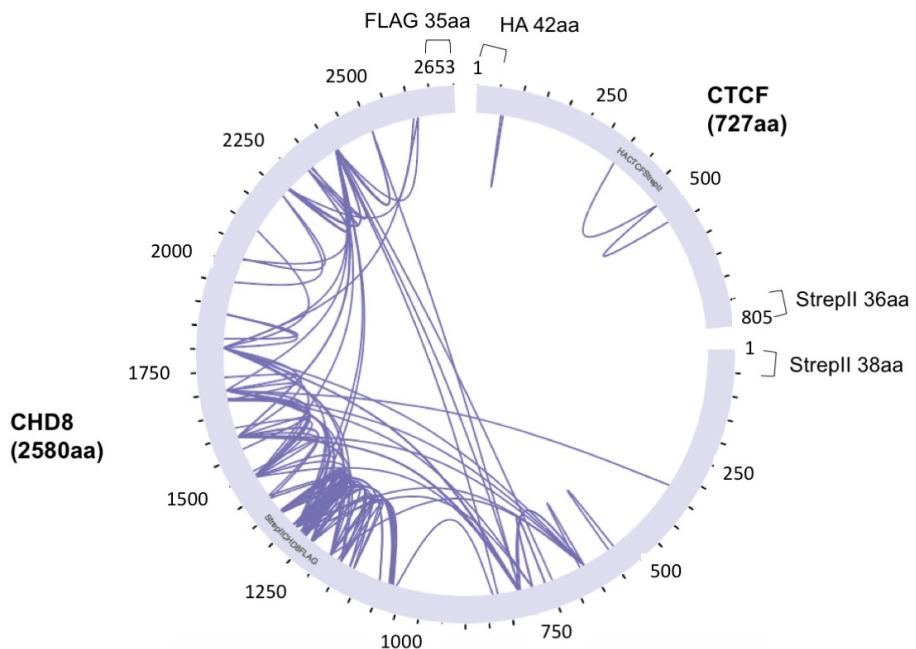
A

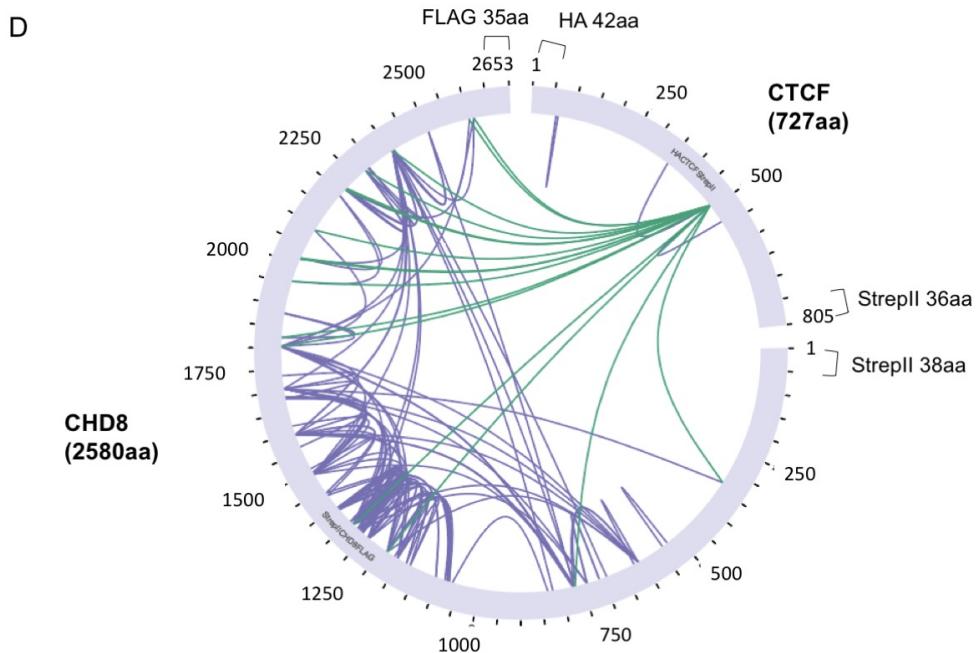


B



C





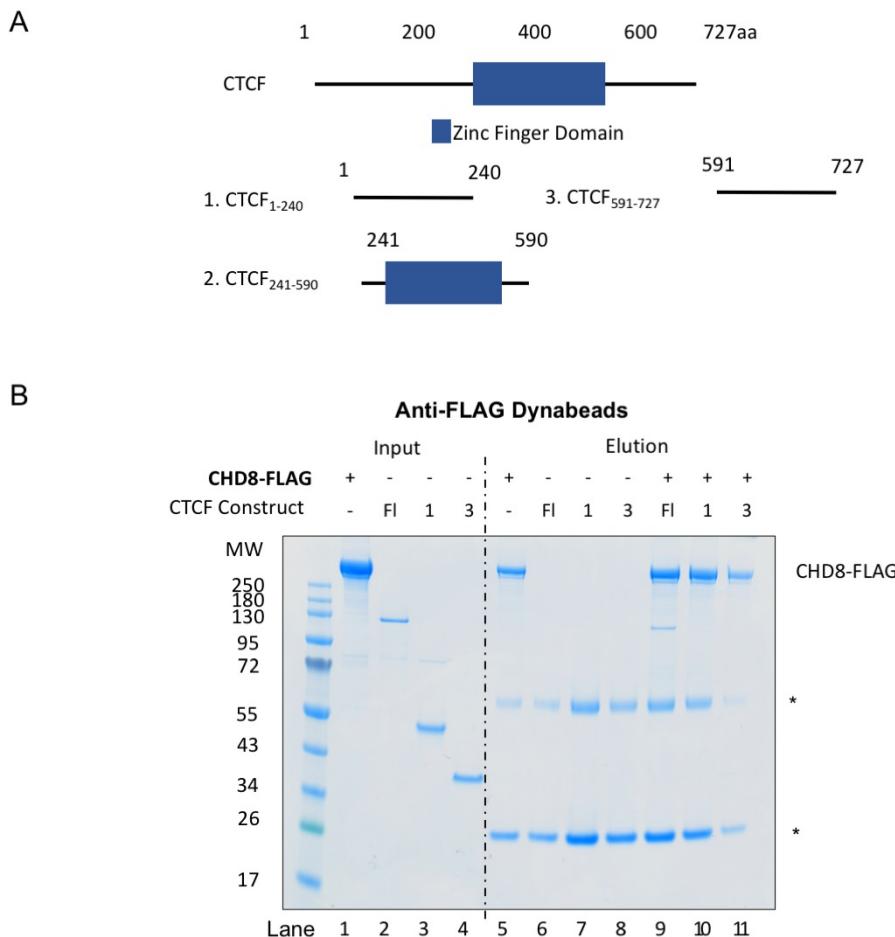
**Figure 6.7 Cross-Link Mass Spectrometry**

A: SDS-PAGE analysis of the CTCF and CHD8 protein samples pre- and post- cross-linked with DSBU. B: The interacting regions between CTCF and CHD8 plotted in a circular diagram. The lengths of protein tags were also indicated. C: Self-crosslinked regions within CTCF and CHD8 were plotted in a circular diagram. D: Overlapping of the cross-linked mass spectrometry results. Green lines corresponded to the cross-links observed between CHD8 and CTCF, whereas purple lines corresponded to the self-crosslinks observed within CTCF or CHD8.

#### 6.1.5.2 Pull-Down Assays Between CTCF and CHD8 Constructs

Various constructs of CTCF and CHD8 were designed to confirm the results obtained from the cross-link mass spectrometry experiment. To identify the region of CTCF that interacted with CHD8, three truncated versions of CTCF were designed based on its domain architecture (Figure 6.8A). These constructs were expressed in High5 cells and purified following the protocols mentioned in the methods section 2.3.2. However, due to the instability of the eleven zinc fingers, the zinc finger domain ( $\text{CTCF}_{241-590}$ ) was not well-expressed and therefore could not proceed to purification. The remaining two purified constructs ( $\text{CTCF}_{1-240}$  and  $\text{CTCF}_{591-737}$ ), along with purified full length CTCF were incubated with CHD8-FLAG that was immobilised on anti-FLAG beads. The results suggested that CHD8 was able to capture full length CTCF

(Figure 6.8B), but not the other two constructs, suggesting a possible interaction between CHD8 and the zinc finger region of CTCF, and this result was consistent with the cross-link mass spectrometry observation where amino acids around 450, corresponding to ZF7 in the zinc finger region of CTCF was involved in the interaction with CHD8.

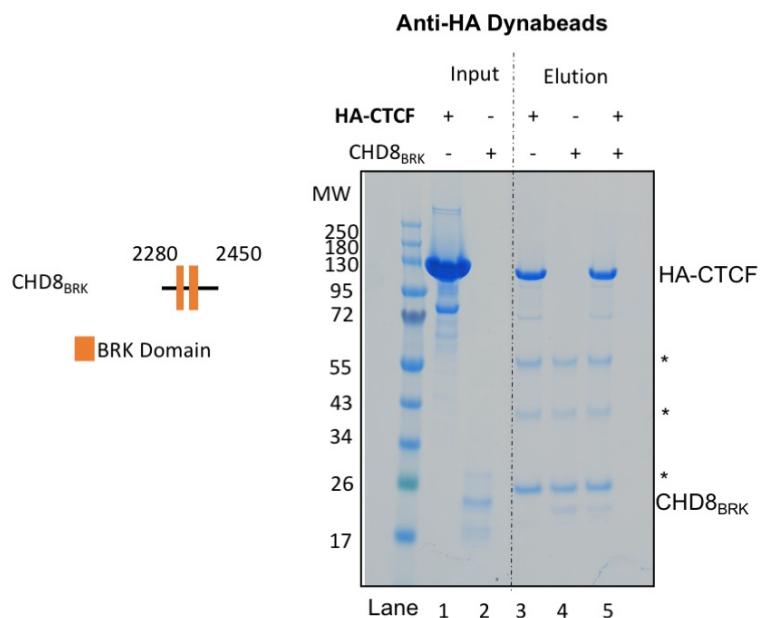


**Figure 6.8 CTCF Domain Involved the CTCF-CHD8 Interaction**

A: CTCF constructs that were designed and purified. The domain architecture was illustrated in a bar diagram. B: Pull down assay between CTCF constructs and full length CHD8 analysed by SDS-PAGE. CTCF<sub>241-590</sub> could not be expressed. Only full length CTCF was captured by CHD8. Bands marks with asterisks (\*) corresponded to the heavy and light chains of the anti-FLAG antibodies.

The interaction between CTCF and the C-terminal domain of CHD8 was further explored because one of the main features at the CHD8<sub>CTD</sub> was the BRK domain that is often found in transcription factors, however the function associated with this domain remained unknown. A CHD8 construct comprised

of the BRK domain with short flanking regions was designed and expressed in *E. coli*. The purified protein CHD8<sub>BRK</sub> was incubated with full length HA-tagged CTCF immobilised on anti-HA beads. The resulting SDS-PAGE showed that the BRK domain was not captured by CTCF (Figure 6.9), as the band observed in the CTCF lane likely corresponded to non-specific BRK-bead binding as seen in the control lane. Hence suggesting that this purified domain construct was not sufficient to mediate the interaction between CTCF and CHD8.



**Figure 6.9 CHD8 BRK Domain is not Involved in CTCF-CHD8 Interaction**

Pull down assay between recombinantly purified full length CTCF and CHD8<sub>BRK</sub>. Anti-HA dynabeads were used to immobilise HA-tagged CTCF. No interaction was observed between these two proteins. Protein bands marked with asterisks (\*) corresponded to the heavy and light chains of the anti-HA antibodies.

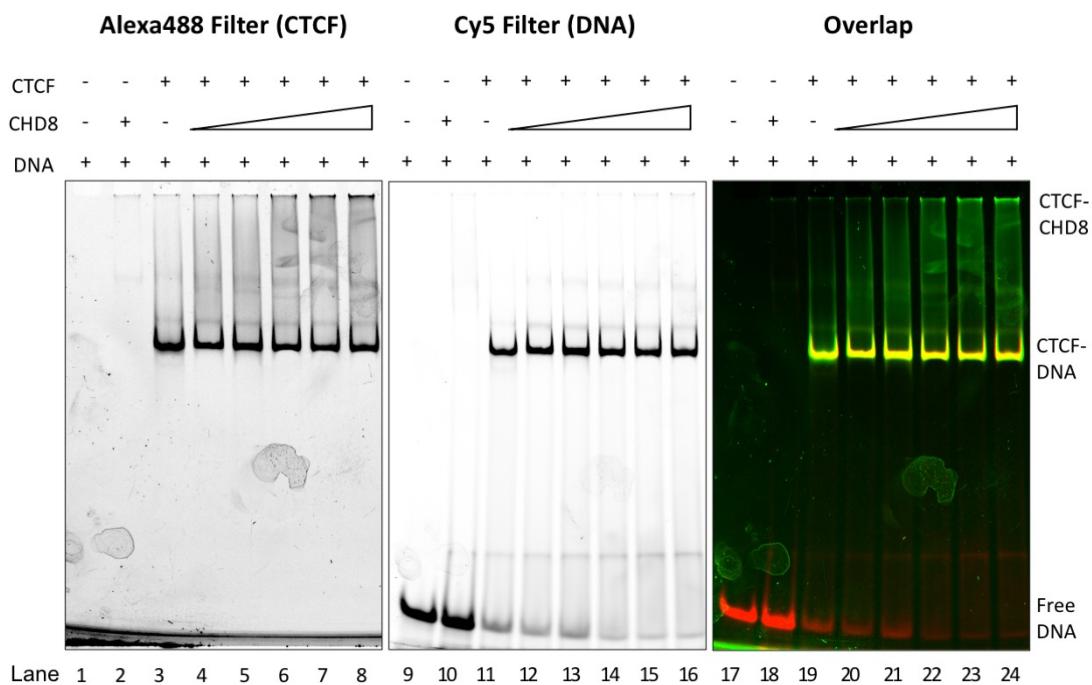
To conclude, the results from the cross-link mass spectrometry were in accordance with the results from the pull-down assays where the interaction between CTCF and CHD8 was through the zinc finger region of CTCF (ZF7). Although the BRK domain construct designed for this experiment did not show binding to CTCF, this did not exclude the possibility that a longer flanking region around the BRK domain might be needed to mediate the interaction.

## 6.1.6 CTCF Cannot Interact with CHD8 and DNA Simultaneously

### 6.1.6.1 CHD8 Does Not Interact with Pre-formed CTCF-DNA Complex

As shown in the cross-linked mass spectrometry and pull-down experiments, CHD8 interacted with CTCF through CTCF's zinc finger domain which is also its DNA binding region. Therefore, the next question was to investigate whether or not CTCF can interact with DNA and CHD8 at the same time to form CTCF-CHD8-DNA complex. EMSAs and pull-down assays were performed to investigate this hypothesis. These two techniques allowed the visualisation of how the protein and DNA substrates are involved in the interactions.

The aim of the EMSA was to test if CHD8 was able to bind to the CTCF-DNA complex to create a slower migrating species. In this assay, CTCF consensus DNA was labelled with Cy-5 and CTCF protein was conjugated with Alexa488, enabling both DNA and protein to be visualised in the EMSA. Four-fold molar excess of CTCF<sub>Alexa488</sub> was incubated with 37 bp Cy-5 labelled consensus DNA on ice for 20 minutes, followed by the addition of increasing concentrations of CHD8, to a maximum of 16-fold molar excess relative to the CTCF-DNA complex. The samples were incubated on ice for another 20 minutes before they were analysed using Native PAGE. Only one sharp band shift was observed in this EMSA experiment, corresponding to the CTCF-DNA complex (Figure 6.10). As the concentration of CHD8 increased, a band at the top of the gel was observed. The band was detected by the Alexa488 filter which might correspond to the interaction between excess CTCF and CHD8. Since the concentration of CHD8 was high, the protein complex might have aggregated and hence was not able to enter the gel matrix. No other band shifts were detected, indicating that CHD8 was not able to interact with the pre-formed CTCF-DNA complex.



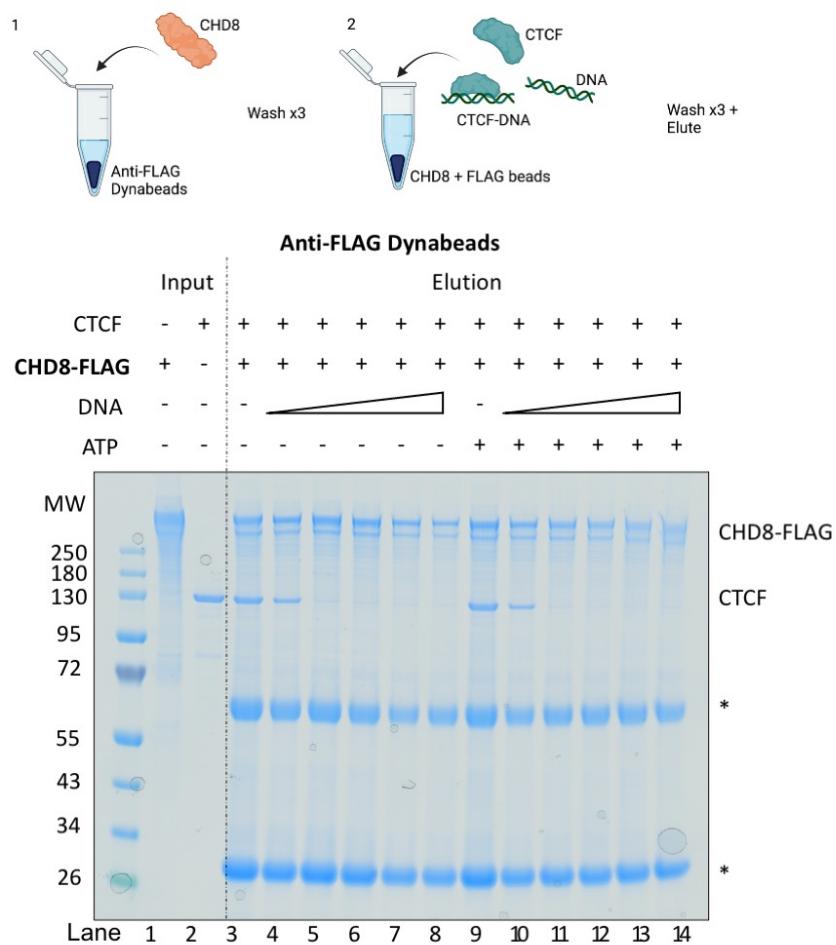
**Figure 6.10 EMSA Between CHD8 and CTCF-DNA Complex**

EMSA assay between CTCF-DNA complex incubated with different concentrations of CHD8. The individual channels for Alexa488 (lanes 1-8) and Cy-5 (lanes 9-16) were displayed in black and white. The overlay of both filters was shown in lanes 17-24, where CTCF conjugated to Alexa488 was shown in green, whereas Cy-5 labelled DNA was shown in red, therefore the CTCF-DNA band appeared to be yellow.

Although the EMSA experiment suggested that CHD8 was not able to bind onto CTCF that was pre-occupied by CTCF-consensus DNA, it did not determine whether or not CHD8 was able to disrupt the CTCF-DNA complex and interact with the dissociated CTCF. The aim of the next pull-down assay was to determine if DNA affects the ability of CHD8 to pull down CTCF. CTCF-DNA complex was formed by mixing CTCF with increasing concentrations of consensus DNA (0.25x – 4x). This pre-formed CTCF-DNA mixture was then incubated with CHD8 and anti-FLAG resins and washed with binding buffer. The proteins were eluted from the beads and analysed using SDS-PAGE. The gel showed that as the concentration of DNA increased, the amount of CTCF captured by CHD8 decreased (Figure 6.11, lane 3-8). This might be because at lower DNA concentrations, a larger proportion of CTCF protein was unbound, allowing CHD8 to capture the excess CTCF, a population which was

reduced as the DNA concentration increased, thus suggesting that CHD8 was able to bind to free CTCF only.

CHD8 is an ATP-dependent chromatin remodelling factor, hence it was interesting to understand if the presence of ATP would modulate the behaviour of CHD8 to disrupt the interaction between CTCF and its consensus DNA. The pull-down assay was repeated in the presence of 2 mM ATP. Similar results were observed, where the amount of CTCF pulled down by CHD8 decreased as the concentration of DNA increased (Figure 6.11, lane 9-14). This demonstrated that neither the presence of ATP nor ATP hydrolysis of CHD8 affected the binding activities between CHD8, CTCF and DNA.



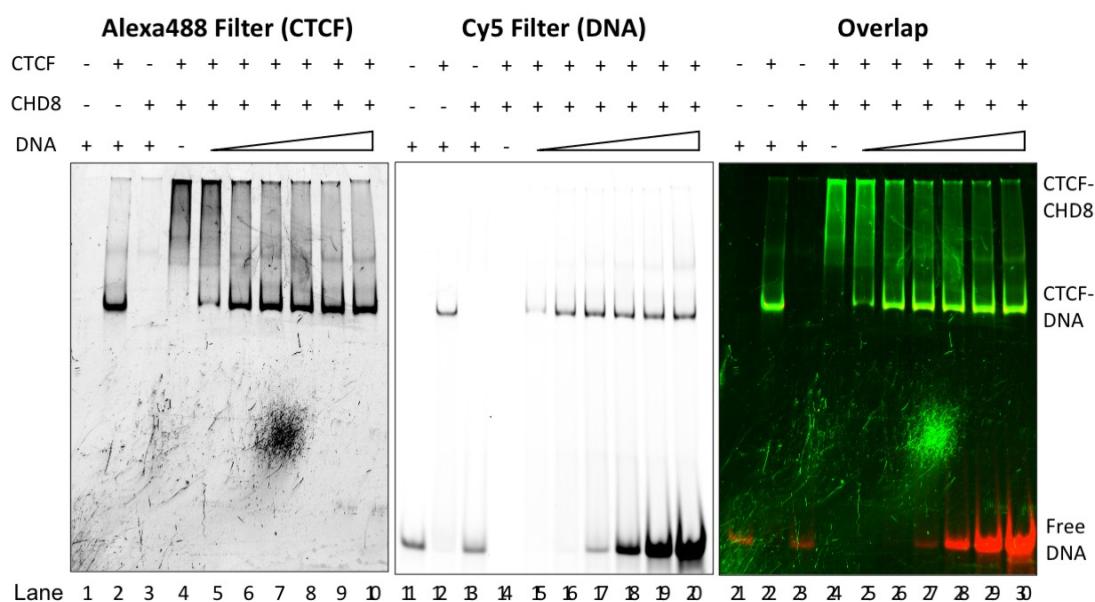
**Figure 6.11 CHD8 Does Not Interact with Pre-formed CTCF-DNA Complex**

Pull-down assay between pre-formed CTCF-DNA complex and CHD8 with and without ATP. Protein bands marked with asterisks (\*) corresponded to the heavy and light chains of the anti-FLAG antibodies.

#### 6.1.6.2 Interaction Between CHD8-CTCF is Disrupted by Consensus DNA

As confirmed in the previous section, once CTCF formed a complex with its consensus DNA, CHD8 was not able to disrupt the interaction. A second set of experiments were carried out to understand if DNA was able to break the pre-formed CTCF-CHD8 complex.

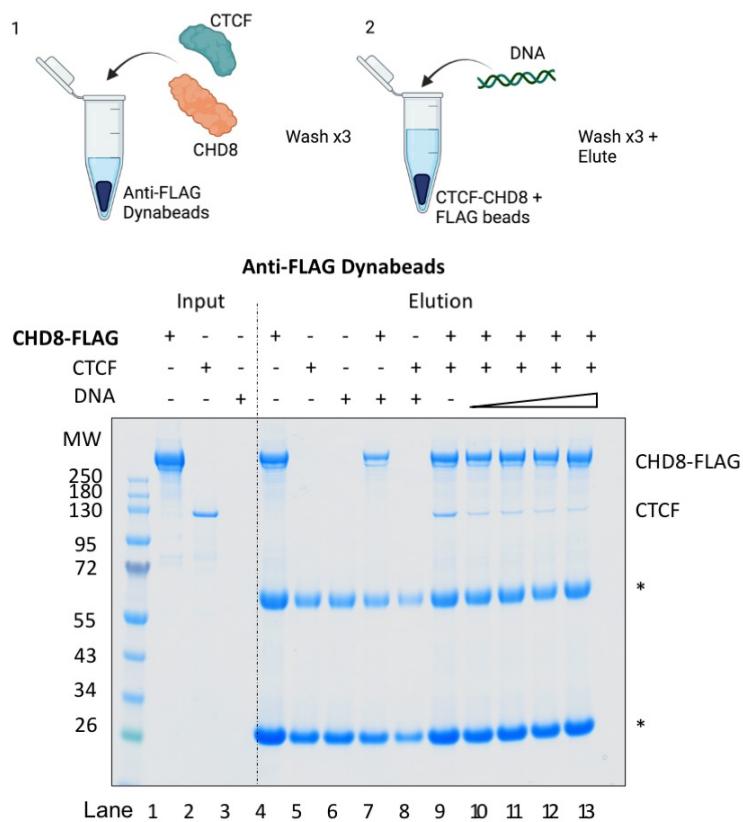
EMSA were carried out to understand if the pre-formed CTCF-CHD8 complex was able to bind to DNA to form CTCF-CHD8-DNA complex. CTCF conjugated to Alexa488 was incubated with two molar excess of CHD8, followed by incubation with 0.5x – 16x molar excess of CTCF consensus DNA. The samples were analysed using Native PAGE (Figure 6.12). Band shifts with the same level of migration were detected in lanes 25 to 30, and as the concentration of DNA increased, the intensity of the bands increased. These band shifts corresponded to the CTCF-DNA complex as it showed similar mobility as the CTCF-DNA control (lane 22). The CTCF-DNA complex band was observed starting from the lowest concentration of DNA, hinting the strong affinity between the two molecules. This result suggested that not only that CTCF-CHD8 was not able to interact with the CTCF consensus DNA, but also CTCF might have dissociated from CHD8 and to form a complex with DNA.



**Figure 6.12 EMSA Between DNA and CTCF-CHD8 Complex**

EMSA assay with pre-formed CTCF-CHD8 complex and increasing concentrations of consensus DNA. The individual channels for Alexa488 (lanes 1-10) and Cy-5 (lanes 11-20) were displayed in black and white. The overlay of both filters was shown in lanes 21-30, where CTCF conjugated to Alexa488 was shown in green, whereas Cy-5 labelled DNA was shown in red, therefore the CTCF-DNA band appeared to be yellow.

To verify that consensus DNA was able to disrupt the interaction between CHD8 and CTCF, a pull-down assay was performed. FLAG-tagged CHD8 was immobilised on anti-FLAG Dynabeads followed by the addition of CTCF. Washes were performed before increasing concentrations of DNA (0.5x – 4x) were added to the reaction. The samples were eluted through boiling and analysed using SDS-PAGE (Figure 6.13). As the concentration of DNA increased, the level of CTCF captured by CHD8 decreased, suggesting the addition of DNA disrupted the CHD8-CTCF complex. This was in accordance with the results from the previous section 6.1.6.1, where CTCF showed higher affinity towards consensus DNA and therefore the formation of CTCF-DNA complex was preferred over the CTCF-CHD8 complex.



### **Figure 6.13 Consensus DNA Disrupts the Interaction between CTCF and CHD8**

Pull-down assay between pre-formed CTCF-CHD8 complex and different concentrations of consensus DNA. Protein bands marked with asterisks (\*) corresponded to the heavy and light chains of the anti-FLAG antibodies.

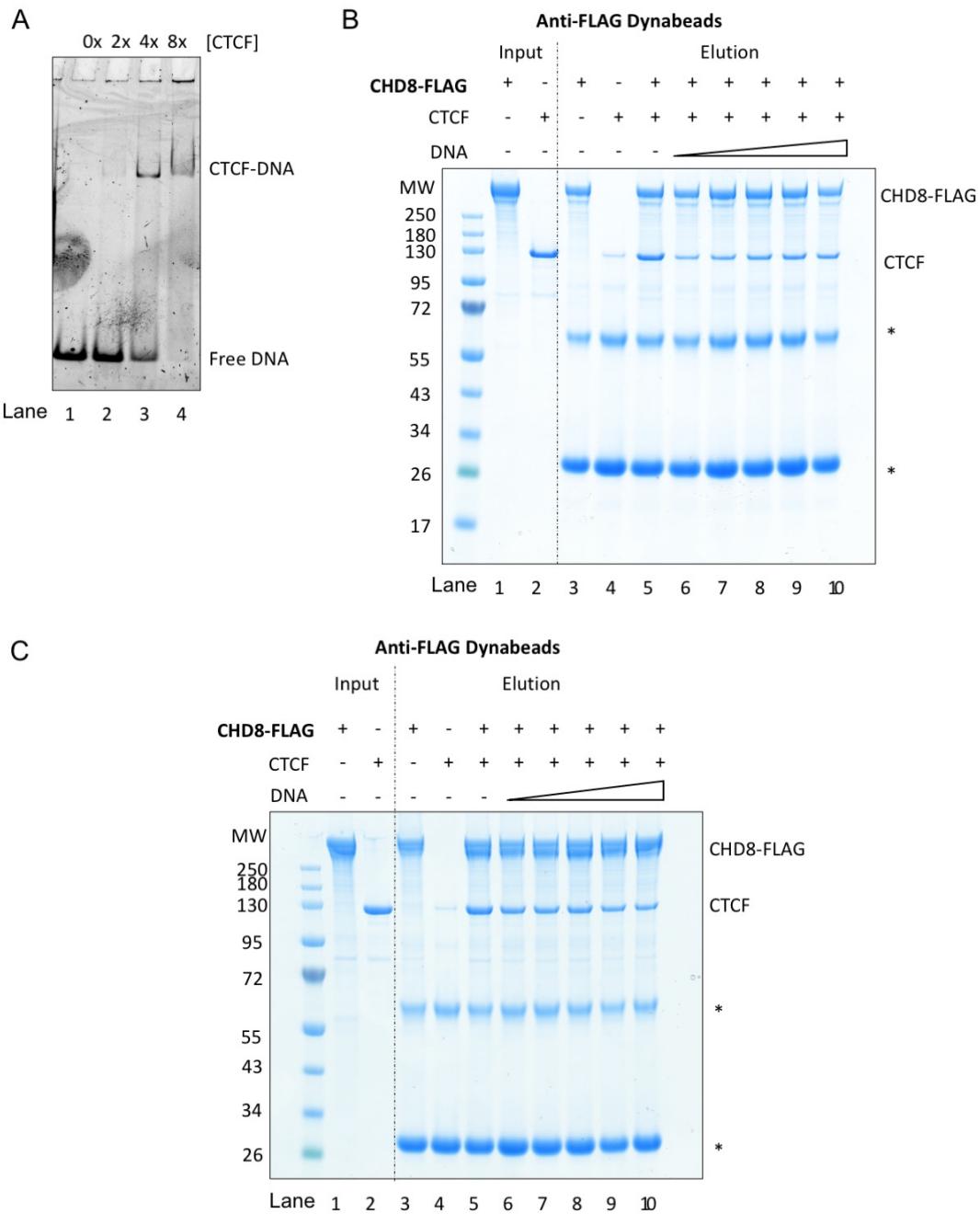
#### **6.1.6.3 CTCF Prefers CHD8 Over Non-Consensus DNA**

Following the observation that the interaction between CTCF and CHD8 could be disrupted by the addition of CTCF consensus DNA, another similar set of experiments were carried out, where the DNA was replaced with a sequence lacking the CTCF consensus motif. This was to understand the importance of the CTCF-consensus site in the competition between CHD8, CTCF and DNA.

Firstly, it was crucial to confirm that CTCF was able to bind to DNA without its consensus site. An EMSA assay was performed and the results showed a CTCF-DNA complex band shift, hence confirmed that CTCF was able to interact with DNA lacking its recognition sequence (Figure 6.14A).

Next, CTCF was incubated with increasing concentration of non-consensus DNA prior to further incubation with FLAG-tagged CHD8 and anti-FLAG resin. The experiment protocol was identical to the pull-down assay described in section 6.1.6.1 with the exception that the consensus DNA was replaced with the non-consensus DNA. As the concentration of DNA increased, the amount of CTCF captured by CHD8 was not affected (Figure 6.14B). This indicated that CHD8 was able to disrupt the interaction of the pre-formed complex between CTCF and the DNA lacking the CTCF consensus site.

The third assay was similar to the pull-down experiment described in section 6.1.6.2 where the CTCF-CHD8 complex was formed prior to the addition of DNA. Likewise, the DNA was replaced with the substrate lacking the consensus site. The results were analysed by SDS-PAGE (Figure 6.14C). The amount of CTCF captured by CHD8 remained constant, and CTCF was not dissociated from CHD8 even in the presence of four-fold molar excess of DNA. This confirmed that the dissociation of the CTCF-CHD8 complex was dependent on the DNA CTCF-consensus motif.



**Figure 6.14 Interaction Between CTCF, CHD8 and Non-consensus DNA**

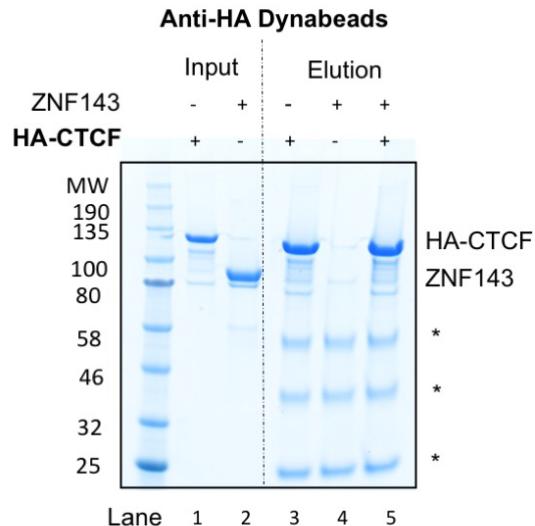
A: EMSA assay between CTCF and DNA without its consensus motif. CTCF was able to bind to DNA lacking CTCF consensus site. B: SDS-PAGE analysis of pre-formed CTCF-DNA complexes added to the immobilised CHD8-beads mixture. C: SDS-PAGE analysis of CTCF-CHD8 complex supplemented with increasing concentrations of non-consensus DNA. Protein bands marked with asterisks (\*) corresponded to the heavy and light chains of the anti-FLAG antibodies.

## 6.2 Interactions Studies of ZNF143

### 6.2.1 ZNF143 and CTCF

Through ChIP-seq studies, ZNF143 has been shown to be the only protein that is correlated to both the distal regulatory element and promoter regions at RNA polymerase II and RNA polymerase III transcription start sites (Bailey *et al.*, 2015). Since CTCF occupies distal regulatory elements (DRE) and mediates short-range intra-TAD (topologically associating domains) interactions (James Faresse *et al.*, 2012; Ngondo-Mbongo *et al.*, 2013), it has been speculated that chromatin interactions involving ZNF143 and CTCF are required to initiate transcription. This interaction will allow the enhancer (which is located next to the DRE) and the promoter to be in close proximity to initiate transcription (Bailey *et al.*, 2015). The potential role of ZNF143 acting as an enhancer facilitator suggests that there might be a direct interaction between ZNF143 and CTCF to allow crosstalk between the two gene elements (Ong and Corces, 2014; Bailey *et al.*, 2015; Ye *et al.*, 2016). In addition, it has been suggested that ZNF143 and CTCF co-localise extensively at TAD boundaries to form chromatin loops (Ye *et al.*, 2016). Hence, it is intriguing to understand if ZNF143 and CTCF interact directly to play an important role in chromatin regulation.

To identify if there is a direct binding between ZNF143 and CTCF, recombinantly purified proteins were used to perform a pull-down assay. Anti-HA dynabeads were used to capture HA-tagged CTCF and purified ZNF143 was added subsequently. Proteins that were eluted from the beads were analysed by SDS-PAGE. It revealed that CTCF was not able to capture ZNF143 (Figure 6.15). The absence of direct interaction between ZNF143 and CTCF hinted that a potential mediator might be required to facilitate the chromatin interaction and to initiate transcription at specific promoters.



**Figure 6.15 ZNF143 Does Not Interact with CTCF**

SDS-PAGE of pull-down assay between ZNF143 and CTCF. HA-tagged CTCF was used as bait and it was not able to capture ZNF143. The bands labelled with asterisks (\*) corresponded to the heavy and light chains of the anti-HA antibody.

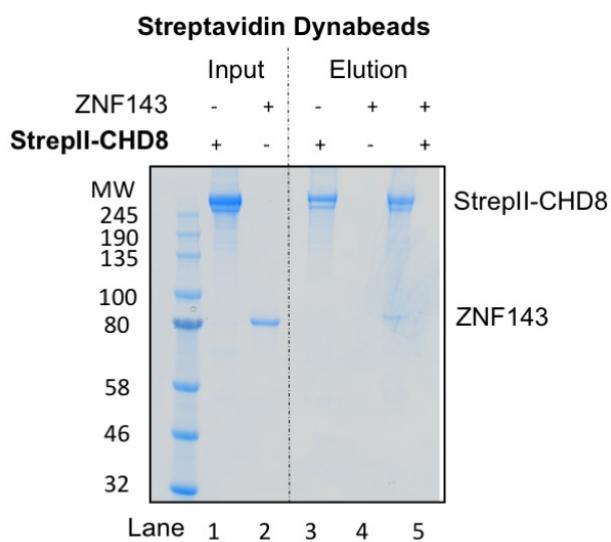
## 6.2.2 ZNF143 and CHD8

### 6.2.2.1 ZNF143 Interacts Directly with CHD8

Based on previous publications, CHD8 is the only protein that has been shown to associate with both ZNF143 and CTCF and therefore it could be the mediator that facilitates the interaction between the two transcription factors. CHD8 was identified as one of the targets that ZNF143 interacted with through immunoprecipitation (Yuan *et al.*, 2007). ZNF143 is recruited to the RNA polymerase III *U6* promoters to initiate transcription (Ramsay and Vannini, 2018). The co-localisation of ZNF143 and CHD8 together with the evidence that CHD8 contributes to efficient transcription at the *U6* promoters suggest that CHD8 and ZNF143 might interact at *U6* promoters and that transcription initiation might involve chromatin remodelling (Yuan *et al.*, 2007). However, the mechanism of how ZNF143 and CHD8 interact and activate transcription remains to be elucidated.

Recombinant protein pull-down assay was performed to determine if ZNF143 and CHD8 interact directly *in vitro*. Strep-tagged CHD8 was incubated with Streptavidin dynabeads and washed with interaction buffer. Two-fold molar

excess of ZNF143 was then added to the beads and washed before eluting through incubating at 100 °C. The ZNF143 only control confirmed that ZNF143 did not bind non-specifically to the beads (Lane 4, Figure 6.16) Although the SDS-PAGE analysis revealed ZNF143 and CHD8 interact directly, the binding between the two proteins were non-stoichiometric (Lane 5, Figure 6.16). The intensity of the ZNF143 band was weaker compared to the CHD8 band, suggesting a weak interaction which may require DNA or nucleosomes to increase avidity.

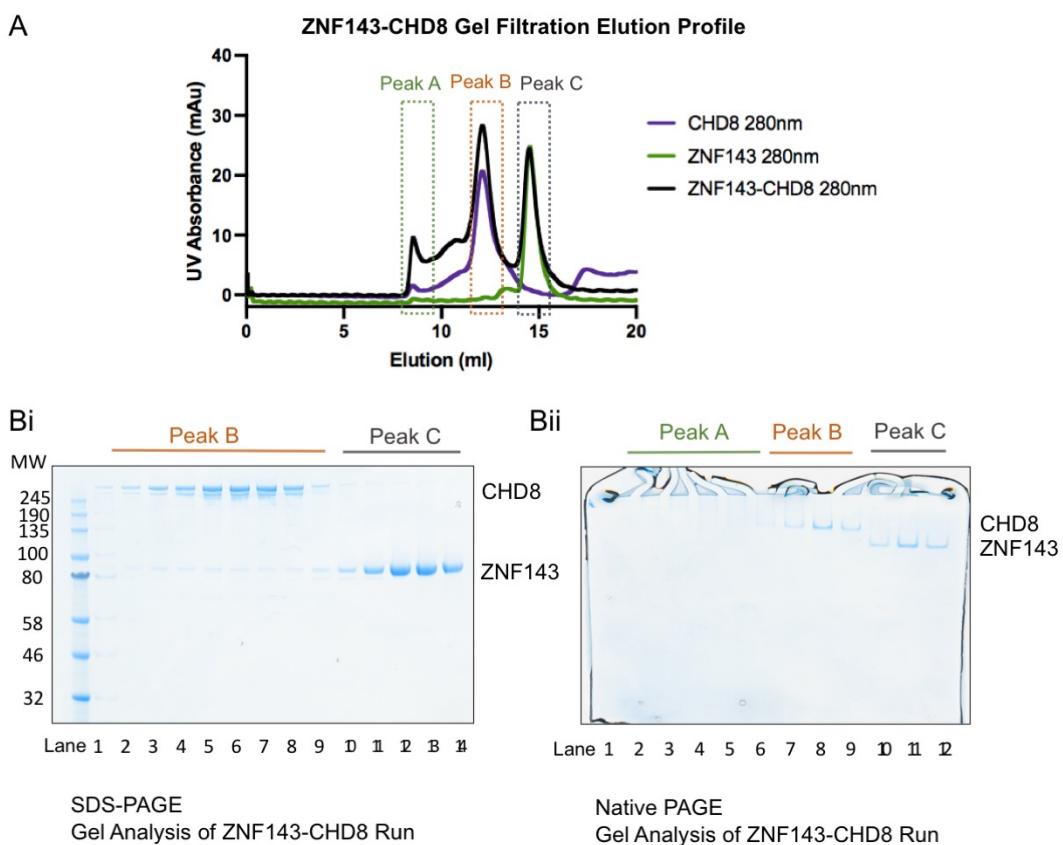


**Figure 6.16 Direct Interaction Between ZNF143 and CHD8**

SDS-PAGE of pull-down assay between ZNF143 and CHD8 using StrepII-CHD8 as bait, where StrepII-CHD8 was immobilised onto Streptavidin dynabeads. Weak interaction between CHD8 and ZNF143 was observed.

**6.2.2.2 ZNF143 and CHD8 Do Not Form a Stable Complex Over Gel Filtration**  
A stable complex would be essential for structural studies. To explore whether the interaction observed was stable, CHD8 and ZNF143 samples were passed through analytical gel filtration column. Four-fold molar excess of ZNF143 was incubated with CHD8 prior to injection onto a Superose 6 10/300 column. This ratio was chosen to allow all CHD8 to be bound to ZNF143, and therefore only two peaks will be expected, the first one corresponding to CHD8-ZNF143 complex, whereas the second peak would be ZNF143 only. However, three peaks were observed in the experiment (Figure 6.17A). Peak A was eluted close to the void volume (~8 ml), whereas peaks B and C were present at 12

ml and 14.5 ml respectively. The eluted fractions were analysed by SDS-PAGE (Figure 6.17Bi) and Native-PAGE (Figure 6.17Bii). Based on the Native PAGE, the proteins in peak A were mostly aggregated as they were precipitated in the wells, and therefore peak A might correspond to CHD8-ZNF143 complex or CHD8/ZNF143 alone which might have aggregated due to the difference in the buffer composition. Peak B was composed mostly of CHD8 with a weak ZNF143 band, and peak C was comprised of ZNF143 only. This data suggested a non-stoichiometric and weak interaction between ZNF143 and CHD8, as a result it was difficult to proceed with structural analysis. Addition of DNA, nucleosomes or crosslinking may be required for further structural analysis.

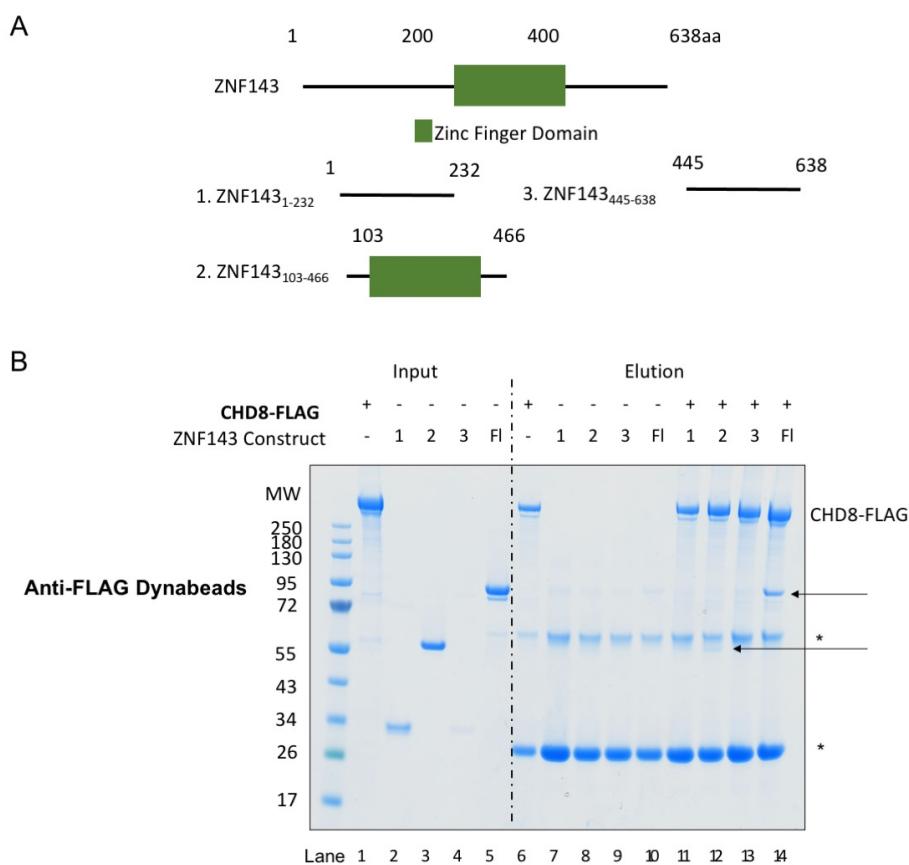


**Figure 6.17 Analytical Gel Filtration of ZNF143 and CHD8**

A: Elution profile of the analytical gel filtration between CHD8 and ZNF143. Three runs were plotted on the same graph: ZNF143 only, CHD8 only and ZNF143-CHD8 complex runs. Three peaks were observed, where peak B corresponded to CHD8 and peak C corresponded to ZNF143. Bi: SDS-PAGE analysis of peak B and peak C from the ZNF143-CHD8 complex run. Bii: Native PAGE analysis of ZNF143-CHD8 run. Peak B and peak C corresponded to CHD8 and ZNF143 respectively.

### 6.2.2.3 Domains of ZNF143 and CHD8 Involved in Their Interaction

The next step was to identify the protein domains involved in the observed direct interaction to understand the binding interface. This will provide a foundation for site specific mutations and *in vivo* experiments to unravel the function and the biological relevance of the complex. Different ZNF143 and CHD8 constructs were designed and expressed, and pull-down assays were carried out to pinpoint the region involved. Four purified ZNF143 constructs (Figure 6.18A) were incubated with purified full length CHD8-FLAG as the bait using anti-FLAG dynabeads. The result of this pull-down assay was analysed by SDS-PAGE (Figure 6.18B). As expected, CHD8 was able to capture full length ZNF143. In addition, the zinc finger domain of ZNF143 ( $ZNF143_{103-466}$ ) was also pulled down by CHD8.

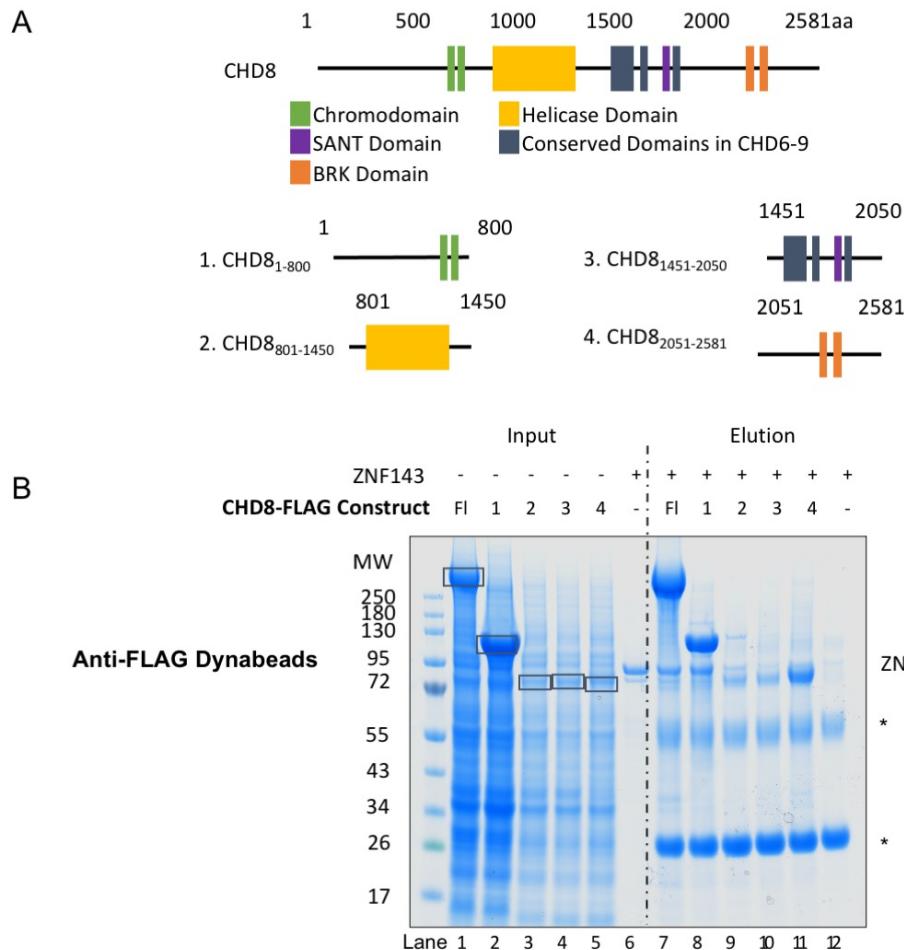


**Figure 6.18 ZNF143 Domain Involved in the ZNF143-CHD8 Interaction**

A: ZNF143 constructs that were designed and purified. The domain architecture was illustrated in a bar diagram. B: Pull down assay between ZNF143 constructs and full length CHD8 analysed by SDS-PAGE. ZNF143 constructs captured by CHD8 were indicated with an arrow. Protein bands marked with an asterisk corresponded to the heavy and light chains of the anti-FLAG antibodies. (Fl: Full length)

A preliminary experiment was performed with four distinct CHD8 constructs based on its domain architecture to identify the CHD8 domain involved in the CHD8-ZNF143 interaction (Figure 6.19A). Including full length CHD8, all of these constructs were FLAG-tagged and were expressed in High5 insect cells. 10 ml of each of the infected High5 cells were harvested, resuspended in lysis buffer, sonicated and subjected to centrifugation. The resulting lysate was applied to FLAG beads and washed before incubating with purified full length ZNF143. The results in figure 6.19B suggested that ZNF143 interacted with both the full length CHD8 and the N-terminal construct CHD8<sub>1-800</sub>, and traces of ZNF143 was also captured by CHD8<sub>801-1450</sub>. There were a few limitations in this experiment and therefore these findings would need to be corroborated through repeating the pull-down studies with purified CHD8 constructs. For instance, lysates rather than purified constructs were used in this experiment due to the difficulties in the purification processes (data not shown) and will require optimisation. In addition, the level of expression of full length CHD8, CHD8<sub>1-800</sub> and CHD8<sub>2051-2581</sub> were higher than the remaining two constructs, which might be due to the instability of the ATPase and conserved domains lacking flanking regions. The difference in CHD8 loading levels could be avoided by purifying the CHD8 constructs and hence loading the same amount of each CHD8 construct to the binding experiment, allowing for more accurate comparisons.

In summary, based on these two pull-down experiments, the zinc finger region of ZNF143 directly interacted with CHD8, however further studies will be needed to validate that ZNF143 interacts with the N-terminal domain of CHD8.



**Figure 6.19 CHD8 Domains Involved in the ZNF143-CHD8 Interaction**

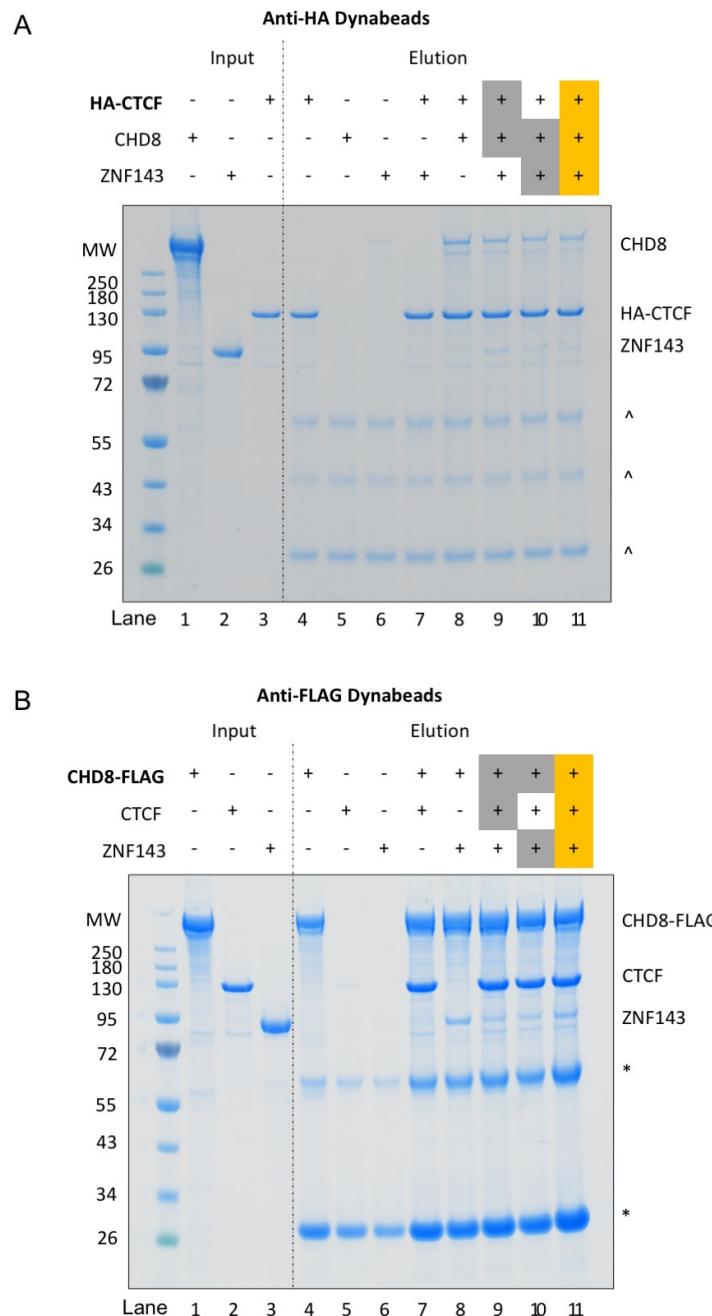
A: CHD8 constructs designed and expressed. The first and last residues of each construct were listed. The domain architecture of CHD8 was illustrated in a bar diagram. B: Pull down assay between CHD8 constructs and full length ZNF143. The black boxes in lanes 1 to 5 marked the over-expressed CHD8 construct. (Fl = full length)

### 6.3 Interaction Studies Between ZNF143, CTCF and CHD8

As shown in previous sections, CHD8 was able to interact with both ZNF143 and CTCF individually, whereas no direct interaction was observed between CTCF and ZNF143. This hinted that CHD8 might be the cofactor that mediated the chromatin loop formation between CTCF and ZNF143. In addition, the CHD8 domains involved in the interactions were different, further suggesting that CHD8, CTCF and ZNF143 might be able to form a larger complex.

HA-tagged CTCF was immobilised on anti-HA beads and incubated with ZNF143 and CHD8 in different sequential orders (Figure 6.20A). Proteins were eluted from the beads through boiling and analysed with SDS-PAGE. As expected, CTCF was able to pull CHD8 (lane 8) down in this assay but not ZNF143 (lane 7). In lane 9, CTCF was incubated with CHD8 first to allow the formation of CTCF-CHD8 complex and washed to remove excess CHD8 before the addition of ZNF143. Whereas in lane 10, CHD8-ZNF143 complex was formed prior to the incubation with CTCF. Lastly in lane 11, all proteins were added into the interaction buffer at the same time. ZNF143 and CHD8 bands were observed in all three lanes, suggesting that CTCF, CHD8 and ZNF143 form a complex and the order of protein added to the mixture did not modulate the interaction.

Another set of complementary pull-down assays was carried out to understand if the addition of a third protein (ZNF143/CTCF) would affect the pre-formed CHD8-protein (CHD8-CTCF/CHD8-ZNF143) complex. FLAG-tagged CHD8 was immobilised onto anti-FLAG resin and incubated with ZNF143 and CTCF in different sequence (Figure 6.20B). Following the incubation of immobilised CHD8 and the first protein, the beads were washed three times before the addition of the second protein. Lane 9 corresponds to the results where CHD8 was incubated with CTCF first, and further supplemented with ZNF143. In lane 10, CHD8 was mixed with ZNF143 prior to the addition of CTCF. Whereas in lane 11, all three components were mixed and incubated at the same time. All three components were detected in these three lanes and the sequence of the proteins added did not affect the amount of each of the components captured by CHD8. This suggested that the addition of a third protein did not impact the pre-formed CHD8- complexes. In addition, the assay also indicated that CTCF and ZNF143 might occupy different CHD8 regions, which was consistent with the results seen in section 6.1.5 and 6.2.2.3 where CTCF interacted with the C-terminal of CHD8 and ZNF143 interacted with the N-terminal domain of CHD8, allowing the formation of CTCF-CHD8-ZNF143 complex.



**Figure 6.20 Interaction Between CTCF, ZNF143 and CHD8**

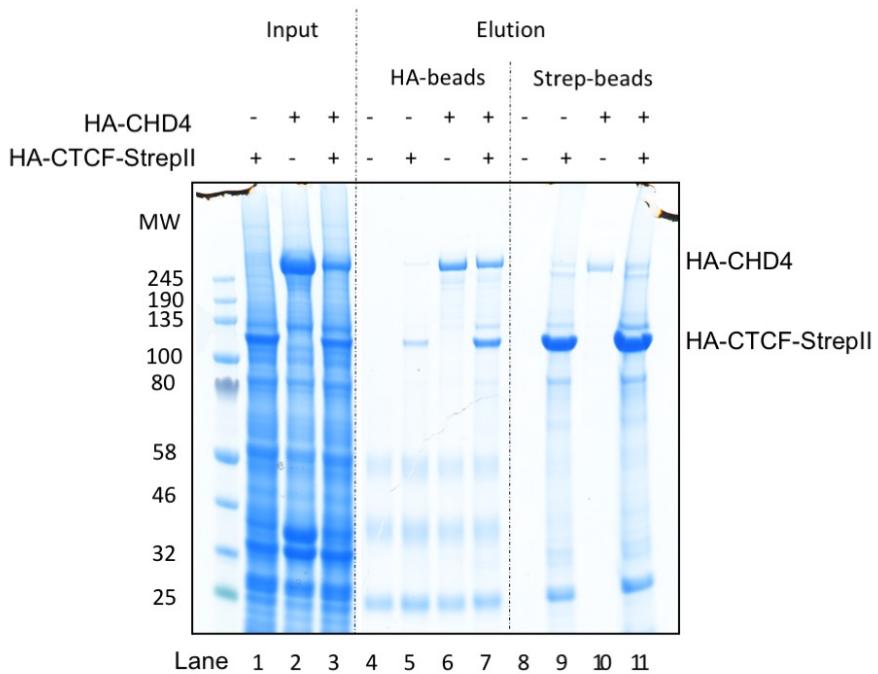
A: SDS-PAGE analysis of the pull-down assay between CTCF, ZNF143 and CHD8. HA-tagged CTCF was immobilised on anti-HA beads. The proteins were added to the reaction in different sequences, grey boxes indicated proteins that were incubated first prior to the addition of the third protein. Yellow boxes indicated proteins that were added simultaneously. Protein bands marked with carets (^) corresponded to the heavy and light antibody chains of anti-HA antibody. B: SDS-PAGE analysis of the interaction between CTCF, ZNF143 and CHD8. Anti-FLAG beads were used to immobilise FLAG-tagged CHD8. Grey boxes indicated the complex that was pre-formed prior to incubation with the third protein. Protein boxes highlighted in yellow were added to the reaction at the same time. Protein bands marked with asterisks (\*) corresponded to the heavy and light antibody chains of anti-FLAG antibody.

## 6.4 Interaction Studies with Other Chromatin Proteins

### 6.4.1 CTCF Does Not Interact with CHD4

In addition to CHD8, CTCF sites have been shown to overlap with another chromatin remodelling protein CHD4 (chromodomain helicase DNA-binding protein 4) to play a major role in chromatin modification. CHD4 has been identified to be the top CTCF-related remodeler when compared with other CHD proteins. Together with CTCF, it has been suggested that CHD4 plays an important role in establishing the asymmetry of nucleosome positioning (Clarkson *et al.*, 2019). It is therefore intriguing to understand if these two proteins interact.

Co-expression pull-down experiment was carried out to investigate the interaction between CTCF and CHD4. CHD4-HA Baculovirus was kindly provided by Dr. Erin Cutts and it was co-infected with HA-CTCF-StrepII Baculovirus in High5 cells for four days. The harvested cell pellets were incubated with Streptavidin Dynabeads and HA Dynabeads, washed three times and analysed with SDS-PAGE (Figure 6.21). The lysates were incubated with HA Dynabeads to confirm the expression of CTCF (lane 5) and CHD4 (lane 6) as they were both HA-tagged. Traces of CHD4 were bound non-specifically to the Strep beads (lane 10). However, the amount of CHD4 pulled down with CTCF (lane 11) was similar to the CHD4 control in lane 10. While pull-down experiments with CTCF and CHD8 showed a roughly stoichiometric interaction (Figure 6.1), the absence of interaction between CTCF and CHD4 indicated that the interaction between CTCF and CHD8 might be specific and CTCF do not bind to all remodelers in the CHD family.



**Figure 6.21 No Direct Interaction Between CHD4 and CTCF**

HA-CTCF-StrepII was co-expressed with HA-tagged CHD4 in High5 insect cells. Anti-HA dynabeads were used to confirm the expression of both proteins, while anti-Strep dynabeads were used to understand if CTCF was able to capture CHD4.

## 6.5 Summary

There are growing evidence suggesting the interplay between CTCF, ZNF143 and CHD8 might have an important role in genome organisation and transcription regulation (Ye *et al.*, 2020; Huning and Kunkel, 2021). In this chapter, the interactions between these three chromatin proteins were mainly characterised using *in vitro* pull-down assays. Although direct interaction between CTCF and ZNF143 was not observed, CTCF formed a strong and stable complex with CHD8, and the interacting regions were mapped through cross-link mass spectrometry. The complex was analysed with negative stain and cryo-EM to provide structural insights. However due to the low-resolution structure, the known structures of CHD8 and CTCF could not be fitted. Furthermore, since CTCF binds to CHD8 through its zinc finger domain, several assays were carried out to examine the ability of CTCF to bind to both CHD8 and DNA at the same time. Competition pull-down assays revealed that CTCF could not interact with CHD8 and DNA simultaneously. In addition,

CTCF was dissociated from the CTCF-CHD8 complex in the presence of consensus DNA and this activity was dependent on the CTCF consensus sequence, hinting a significant role of CTCF binding sites in how CTCF and CHD8 might function in the genome.

In addition to CTCF, direct interaction has been observed between CHD8 and ZNF143 where pull-down results showed that the N-terminal domain of CHD8 and the zinc finger region of ZNF143 were involved in the interaction. Since CHD8 interacted with CTCF and ZNF143 through two different domains, pull-down assays confirmed that these three chromatin proteins were able to form a complex. The interaction between CTCF, ZNF143 and CHD8 indicated that this complex might be crucial for chromatin-loop formations and further *in vivo* studies will be required to probe the importance of this complex in genome organisation.

Lastly, the ability of other chromatin interacting proteins to bind to CTCF were explored using co-expression pull down assays. CHD4, another chromatin remodelling protein that has been shown to be the top CTCF-associated remodeller (Clarkson *et al.*, 2019) was not captured by CTCF and this in turn revealed the interaction between CHD8 and CTCF might be specific. Further experiments will be needed to elucidate the importance of these interactions.

# 7 Discussion

## 7.1 Characterisation of Human CTCF

### 7.1.1 Generation of Functional Human Full-Length CTCF Protein

Although recombinantly purified CTCF has been previously reported (Chernukhin *et al.*, 2000; Yao *et al.*, 2010; Davidson *et al.*, 2016), the CTCF purified was not of sufficient quality for structural and biophysical analysis due to the level of protein and DNA contamination. Therefore in order to structurally and biophysically characterise CTCF, an optimised purification protocol was developed as part of this thesis. Human full length CTCF was expressed and purified from baculovirus infected cells, which has higher eukaryotic protein processing machinery resulting in proteins that were post-translationally modified. During the purification optimisation, recombinant CTCF protein was labelled with a N-terminal HA tag and a C-terminal StrepII tag to reduce the level of non-specific degradation from both ends which significantly reduced the quality of the protein preparations. The overall purification involved four steps, including incubation with benzonase, followed by Strep affinity column, Heparin column and gel filtration column (Figure 3.4). It was essential to incubate the lysate with benzonase at two different salt concentrations where lower salt concentration allowed efficient benzonase activity and the higher salt concentration removed non-specific interactions between DNA and CTCF. Excess DNA contamination was further removed through the heparin column. It was important to note that CTCF was eluted over a large volume from the gel filtration column which might be due to the largely unstructured N- and C-terminal regions (Martinez and Miranda, 2010) and/or the presence of CTCF oligomers which will be addressed in section 7.1.2. The quality of the protein was confirmed by SDS-PAGE analysis (Figure 3.4), as well as mass photometry where the sample was predominantly comprised of proteins at around 99 kDa that matched with the expected molecular weight of CTCF (93 kDa) (Figure 3.7). The successful purification of high quality CTCF provided a good starting point to conduct *in vitro* and structural studies to investigate the mechanism of CTCF actions.

Another CTCF construct with ybbR tag was also purified to allow protein labelling with a fluorescent dye of choice (Yin *et al.*, 2005, 2006) (Figure 3.5). This construct was used in EMSA studies for easy visualisation. This fluorescently labelled protein can also be applied in single molecule microscopy to allow efficient detection of individual molecules where these experiments would provide insights into the molecular mechanisms of CTCF alone or together with other chromatin proteins in genome organisation, in particular how CTCF mediates chromatin loop formation together with other chromatin interacting proteins would be particularly interesting.

### 7.1.2 Concentration Dependent Oligomerisation of CTCF

CTCF has a prominent role in genome organisation, acting as insulators at topologically associated domain (TAD) boundaries and as transcription activators within TAD regions (Ong and Corces, 2014; Rao *et al.*, 2014). The formation of TAD boundaries and enhancer-promoter loops are established on chromatin loops mediated by CTCF and cohesin at loop anchors (Dixon *et al.*, 2012; Rao *et al.*, 2014). Although the mechanism of loop formation remains elusive, increasing evidence is hinting that the establishment of chromatin loops are dependent on the directionality of CTCF consensus sequences where CTCF homodimers are bound. The idea of CTCF dimerization was first observed in yeast two-hybrid experiments and later confirmed in co-immunoprecipitation assays (Yusufzai *et al.*, 2004; Bonchuk *et al.*, 2020).

Previous detection of CTCF dimers were solely based on *in vivo* studies. Here, using biophysical techniques including mass photometry and dynamic light scattering (DLS), I showed that CTCF was able to self-interact and oligomerise in a concentration dependent manner. The mass photometry experiment revealed that CTCF was predominantly monomeric in solution at low concentrations (Figure 3.7), and the dynamic light scattering assay demonstrated that the hydrodynamic radius was positively correlated to the concentration of CTCF (Figure 3.8). This result was consistent with the idea that CTCF was able to self-associate to form homodimers (Yusufzai *et al.*, 2004; Bonchuk *et al.*, 2020). Although concentration-dependent CTCF

oligomerisation has not been reported, it has been shown that there are around 100,000 CTCF molecules per human cell (Cattoglio *et al.*, 2019) and these CTCF molecules are not homogeneously distributed within the nucleus, but instead forming CTCF clusters (Hansen *et al.*, 2019). These clusters might arise from CTCF oligomerisation, and the intrinsically disordered N- and C-terminal domains might also contribute to the formation of protein cluster. The functional role of CTCF clustering is still poorly understood and require further investigation.

Whilst the results presented suggested a degree of oligomerisation by CTCF *in vitro*, dynamic light scattering measured the hydrodynamic radius of the samples and therefore the measured radius did not directly reveal the number of CTCF molecules at each concentration. Hence future experiments such as size exclusion chromatography coupled to multi-angle light scattering (SEC-MALS) would be able to determine the oligomeric state of CTCF since it measures the radius of gyration and molecular weights of the samples. Understanding the oligomeric states of CTCF could reveal and validate its roles in genome organisation.

Another future experiment would be to characterise the domain involved in CTCF oligomerisation. There has been conflicting evidence on the domain that self-associates, earlier *in vivo* studies showed that the C-terminal was crucial for the dimerization (Hansen *et al.*, 2019), whereas recent *in vivo* experiments suggested that CTCF dimerises through the N-terminal domain (Bonchuk *et al.*, 2020; Nora *et al.*, 2020). Interestingly, as opposed to the results from the *in vivo* assays, SEC-MALS experiment on the purified N- and C-terminal domains (at 0.2 mg/ml) revealed that these domains alone are monomeric in solution (Martinez and Miranda, 2010). Further *in vitro* experiments include DLS measurement of different concentrations of CTCF domains and repeating the SEC-MALS experiment at higher concentrations to shed light into the domain involved in CTCF self-association. It would also be interesting to repeat the biophysical experiments in the presence of DNA or RNA since

previous studies demonstrated that CTCF self-interacts and form oligomers in a RNA-dependent manner (Hansen *et al.*, 2019; Saldaña-Meyer *et al.*, 2019).

### 7.1.3 CTCF Binds to DNA Lacking its Consensus Site

In addition to interacting with DNA containing the CTCF consensus sequence (Figure 3.9), CTCF was also able to bind to substrates lacking the CTCF motif (Figure 3.10). The 15 bp CTCF binding motif is highly versatile, with only half of the DNA bases present consistently (Nakahashi *et al.*, 2013; Hashimoto *et al.*, 2017). It has been postulated that CTCF is able to recognise sequences with high variability within the 15-bp core sequence motif due to the flexibility of its eleven zinc fingers as seen in the crystal structure of CTCF zinc fingers (Hashimoto *et al.*, 2017), however the ability to bind to substrates lacking the CTCF motif completely has not been reported. Each zinc finger binds to three adjacent DNA base pairs, but since only CTCF zinc fingers 3-7 demonstrated sequence-specific contacts, this allowed CTCF to adapt to sequence variations outside the core motif (Hashimoto *et al.*, 2017). The ability of specific residues in the zinc fingers to adopt alternative conformations might be more flexible than expected and hence CTCF was able to interact with substrates that are distinct to the CTCF core motif. Moreover, the CTCF binding motif is only present at 80% of the CTCF binding sites (Nakahashi *et al.*, 2013), thereby the ability for CTCF to bind to non-consensus sequences might enable it to localise to the remaining 20% of CTCF sites.

Although CTCF was able to bind to non-consensus sequences, future experiments such as fluorescence polarisation (FP) and microscale thermophoresis (MST) can be employed to determine the binding affinities towards different DNA constructs. Moreover, previous studies reported the presence of additional binding motifs (~10 bp upstream and downstream motifs) at around 18% of CTCF consensus sites (Nakahashi *et al.*, 2013) and therefore it would also be interesting to explore the role of these motifs on CTCF binding strength. For instance, the central CTCF core sequence can be flanked by either the upstream and/or downstream motifs to understand how these motifs might modulate CTCF binding affinity. In addition, the core

sequence be replaced with non-consensus sequences to understand if the upstream/downstream motif alone is able to modulate the binding affinity. Furthermore, since CTCF-DNA binding can be modulated by DNA methylation (Wang *et al.*, 2012), it would be intriguing to explore the effects of differential DNA methylation patterns on the binding affinities.

Lastly, optimisation of the full length CTCF-DNA complex for cryo-EM studies would shed light into the structural details of the interaction. To avoid the dissociation of DNA from the CTCF-DNA complex, buffer conditions for the CTCF-DNA complex need to be optimised. The stability of the complex in different conditions could be assessed through a variety of biophysical techniques, including dynamic light scattering, mass photometry and differential scanning calorimetry (DSF). An optimised buffer would increase the thermal stability of the protein sample and DSF is a rapid screening method that measures the temperature at which a protein unfolds through the interaction of a fluorescent dye that preferably binds to hydrophobic parts of the protein which are exposed during protein unfolding (Niesen, Berglund and Vedadi, 2007). Following buffer optimisation, the cross-linking of CTCF-DNA complex prior to gel filtration purification might increase the stability of the complex; or a gradient fixation with cross-linkers and ultracentrifugation could be used to separate the complex from excess DNA or CTCF.

#### 7.1.4 Nucleosome Binding Studies of CTCF

CTCF has been shown to possess nucleosome positioning activity in recent studies, but how the nucleosome repeat length (distance between the centre of two adjacent nucleosomes) near CTCF binding sites are modulated remains elusive (Clarkson *et al.*, 2019; Owens *et al.*, 2019). Here, I showed that CTCF was able to bind to nucleosomes with 52 bp linker at the exit site that did not have a CTCF consensus sequence (Figure 3.13). Moreover, a supershift was observed at higher CTCF concentrations which might be due to CTCF oligomerisation or CTCF binding to a few different sites along the nucleosome. Since CTCF was able to bind to non-consensus DNA, there were a few potential interacting regions, and these include the 52 bp linker, the Widom

601 sequence wrapped around the histone core, or CTCF might directly interact with histones. Further experiments will be needed to explore the regions at which CTCF was bound to, and these include biochemical studies to test the interaction between nucleosome core particle (without linker DNA) and CTCF, and structural characterisation using cryo-EM to understand the CTCF-nucleosome complex. The ability for CTCF to position nucleosomes might also rely heavily on chromatin remodelers such as CHD4, CHD8, and SNF2H (Clarkson *et al.*, 2019). The interaction between these proteins will be discussed later in section 7.4.1.

## 7.2 Characterisation of Human ZNF143

### 7.2.1 Generation of Functional Human Full-Length ZNF143 Protein

Most of the *in vitro* studies on ZNF143 were performed with recombinantly purified *Xenopus laevis* homolog Staf (Schaub *et al.*, 1997; Schaub, Krol and Carbon, 2000), with limited studies performed with purified human ZNF143 (Yuan *et al.*, 2007). The reported purification protocols did not contain sufficient experimental details to replicate the purification of these studies, where only the presence of an N-terminal glutathione S-transferase (GST) tag was reported, and the details of the subsequent purification steps were not available (Schaub *et al.*, 1997; Yuan *et al.*, 2007). Therefore the first purification trial was carried out with N-terminally His<sub>6</sub>-GST tagged human ZNF143 that was expressed in *E. coli* and purified with a three-step protocol. However, a degradation product with truncated C-terminal region was observed during purification (Figure 4.1). To improve the stability of human ZNF143, a Strep tag was added to the C-terminus, and ZNF143 was expressed in eukaryotic baculovirus-infected insect cells to allow post-translational modification. The affinity step capturing the C-terminal tag removed degradation products that lacked the C-terminal region, and the protein purified from the insect cells appeared to be more stable as less degradation products were observed overall (Figure 4.2). In summary, the switching of expression system and tagging strategy improved the quality and the yield of ZNF143 purified recombinant protein.

Since the elution volume of ZNF143 from the gel filtration column did not match its expected molecular weight (Figure 4.2D), the oligomeric state of ZNF143 was analysed using SEC-MALS. Dimerisation of ZNF143 has not been reported and the SEC-MALS results confirmed that ZNF143 was monomeric in solution (Figure 4.3). Thus, the discrepancy in the elution volumes might arise from large hydrodynamic radius due to the disordered and extended N- and C-terminal domains of ZNF143. The monomeric state of ZNF143 can be further assessed by repeating the SEC-MALS experiment with higher protein concentrations and the hydrodynamic radius of ZNF143 could be measured through DLS assays in future experiments.

A new ZNF143 construct lacking the N- and C-terminal domains (ZNF143<sub>103-466</sub>) was designed based on a limited trypsin proteolysis experiment to remove the flexible regions in ZNF143 (Figure 4.4). The removal of the disordered terminal domains not only improved the quality of ZNF143 but also increased the yield by 50% (Figure 4.5). This construct maintained its core function as a DNA binding protein (Figure 4.7) and therefore it has been used to probe the DNA binding activities of ZNF143 (details in section 7.2.2) and for crystallisation trials (details in section 7.2.4).

### 7.2.2 ZNF143 Binds to A Variety of DNA Sequences

One of the most documented functions of ZNF143 is its role as a DNA-binding protein (Myslinski, Krol and Carbon, 1998; Schaub, Krol and Carbon, 2000). Although functional studies have been carried out on the zinc finger domain of the *Xenopus laevis* homolog Staf (Schuster *et al.*, 1995; Myslinski, Krol and Carbon, 1998; Schaub, Krol and Carbon, 2000), no experimental results have shown whether or not the human ZNF143 domains bind or modulate DNA binding indirectly. Here, I showed that the DNA-binding region of human ZNF143 was restricted to within residues 233-444 that corresponded to the zinc finger region (Figure 4.7), which is consistent with the studies performed on *Xenopus* Staf (Schaub, Krol and Carbon, 2000). Although Schaub *et al.* postulated that the binding of ZNF143 to DNA might not be limited to its zinc

finger domain, fluorescence polarisation results presented in figure 4.7 suggested the contrary where the N- and C-terminal domains did not modulate the DNA binding strength of ZNF143 significantly, hinting that the interaction between human ZNF143 and DNA might be solely dependent on its zinc finger region.

As expected, ZNF143 showed strong binding affinities towards DNA constructs comprised of its ZNF143 consensus sequence (Figure 4.8). Nonetheless, ZNF143 has also demonstrated strong interaction with constructs comprised of the CTCF-consensus sequence as well as constructs lacking the Z-motif with affinities at the nanomolar range (Figure 4.10, 4.12). This has not been previously reported, the results suggested that the zinc fingers might be able to adopt alternative conformations to form new hydrogen bonds with different DNA bases. The significance of ZNF143 binding to CTCF consensus sequence will be addressed in section 7.4.2. Competition assays revealed that although ZNF143 was able to bind to non-consensus sequences, ZNF143 showed stronger binding strength towards Z-motif DNA (Figure 4.11). Together with the difference in binding affinities measured from the fluorescence polarisation experiments (Figure 4.12), where sequences with high GC content displayed higher binding affinities hinted that specific zinc fingers might recognise and prefer guanine or cytosine to perform base-specific contacts. A previous study reported that different zinc fingers were required for different promoter binding, in particular ZF3-6 represented the minimal zinc finger region required for high DNA binding affinity (Schaub, Krol and Carbon, 1999, 2000). The non-specific base contact of ZF1-2 and ZF7 might allow ZNF143 to adapt to sequence variations to interact with different promoter regions and even with sequences that are different to the ZNF143 motif, whereas ZF3-6 might recognise guanine and cytosine bases to display base-specific interactions. Moreover, the predicted ZNF143 binding motifs were present across only 74% of all binding sites (Ngondo-Mbongo *et al.*, 2013), suggesting that ZNF143 is able to recognise sequences with high variability to accommodate the remaining binding events that lack ZNF143 motifs.

In order to elucidate the mechanism of ZNF143 interacting with different DNA sequences, a high-resolution of the tandem zinc finger array of ZNF143 in complex with DNA would allow for the basis of interaction to be directly observed. This would give important insights into the role of ZNF143 as a transcription activator. Since X-ray crystallisation allows us to visualise the protein structure at near-atomic resolution, the next step will be to crystallise ZNF143 protein with and without DNA. The details of the crystallisation trials will be mentioned in section 7.2.4.

### 7.2.3 ZNF143 Shows Strong Binding to Nucleosomes

In addition to binding to DNA, ZNF143 showed strong binding to nucleosomes with 52 bp linker at the exit site that did not have a ZNF143 binding motif (Figure 4.13). Similar to the observation in CTCF-nucleosome binding study (Figure 3.13), a supershift was observed and further experiments will be needed to probe the contacts between ZNF143 and nucleosome. Since nucleosomes play a prominent role in transcription activation at RNA polymerase III type III promoters where ZNF143 binding sites are present (Ramsay and Vannini, 2018; Huning and Kunkel, 2021), the interaction between ZNF143 and nucleosome might facilitate the formation of the stable transcription initiation complex. Yet, more experiments will be needed to characterise the interaction further. Given the strong interaction observed between ZNF143 and nucleosome, this complex could be amenable to future structural investigation with cryo-EM, which has been used to obtain the structure of numerous nucleosome-bound complexes. In addition, ZNF143 has been postulated to possess nucleosome repositioning activity at snRNA gene promoters (Yuan *et al.*, 2007), hence nucleosome remodelling assays can be performed in the future to explore the biology of the ZNF143-nucleosome complex.

### 7.2.4 Crystallisation Challenges of ZNF143

Structural analysis was undertaken after the characterisation of DNA interaction behaviour of ZNF143. The first crystallisation trial was performed with apo full length ZNF143, the protein was tested against commercial

screens with conditions at which zinc finger domain crystals have been formed, however no crystals were observed. Next, DNA was added to the screens which was thought to reduce the flexibility of the zinc finger region, yet only DNA crystals were formed (Figure 4.14). Since the N- and C-terminal domains were largely unstructured posing obstacles to protein crystallisation, trials were performed with a new ZNF143 construct ZNF143<sub>103-466</sub> where the end terminal regions were removed to generate a more rigid structure. The trial with ZNF143<sub>103-466</sub>-DNA led to tiny crystals that were later confirmed to be DNA-only crystals (Figure 4.15), hence further optimisation is required to initiate nucleation and generate protein-DNA complex crystals.

The nucleation and growth of diffracting crystals has always been the rate-limiting step in structure determination with X-ray crystallography (Smyth and Martin, 2000). There are several methods that can be employed to enhance the stability of the complex and increase the chances of crystal formation. Firstly, the protein buffer can be optimised through a buffer screening assay with the use of differential scanning fluorimetry (DSF) to identify the buffer condition for the highest protein stability. Moreover, the presence of glycerol decreases nucleation and therefore screening with buffers with reduced glycerol content might aid nucleation. Secondly, proteases such as chymotrypsin or trypsin can be added to the protein samples during crystallisation screenings to increase the success rate in obtaining crystals (Dong, Xu and Edwards, 2007; Huang *et al.*, 2012). The *in-situ* proteolysis allows the digestion of the N- and/or the C-terminal regions which might improve crystal packing. However, the main caveats are that the ratio between the protein and protease requires optimisation and the crystals might be difficult to reproduce (Dong, Xu and Edwards, 2007). Thirdly, the DNA constructs used for ZNF143-DNA crystallisation screening can be further optimised. Three DNA constructs of varied lengths were used in my trials, it might be beneficial to screen more DNA constructs to increase the chances of crystallisation, including DNA with sticky ends, where single-stranded bases are present on nucleic acid ends can be introduced for screening (Krauss *et al.*, 2013). Biophysical experiments such as fluorescence polarisation can be

performed to identify the DNA construct with the highest affinity prior to setting up crystallisation screens. Lastly, through analysing successfully crystallised zinc finger proteins, the number of zinc fingers solved by X-ray crystallography was usually limited to under five zinc fingers. Thereby new ZNF143 constructs with reduced number of zinc fingers might increase the chances of crystallisation. For instance, the tandem array of seven zinc fingers can be designed into three new constructs, where each of them is comprised of four zinc fingers, ZF1-4, ZF3-6 (minimal DNA binding region), and ZF4-7, respectively. Preliminary expression and purification tests suggested these constructs are soluble (data not shown) and could proceed to crystallisation following the identification of a suitable DNA substrate.

### 7.2.5 Potential Interacting Partners of ZNF143

Most of the ZNF143 interacting proteins discovered by co-immunoprecipitation assay (co-IP) were involved in the mRNA splicing via the spliceosome and microtubule assembly hinting potential roles of ZNF143 in these pathways (Table 4.11). Although no previous publications have suggested that ZNF143 is directly associated with specific proteins within these pathways, it has been shown that ZNF143 acts as regulators at gene promoters of RNA binding proteins involved in mRNA splicing (Fish *et al.*, 2019). Co-IP is a powerful technique used to analyse protein-protein interactions *in vivo*. It is important to note that although the co-IP experiment was performed on the nuclear extract and DNase was added, traces of DNA might still be present and hence the chromatin and/or other cofactors might act as mediators between the proteins captured by endogenous ZNF143. In addition, there might be a small degree of mixing of cellular compartments during cell lysis which might lead to false positives. Therefore, several experiments can be carried out to validate the interactions, these include adding cross-linkers prior to the cell lysis step to capture transient interactions or repeating the co-IP experiment where the bait is replaced with the captured protein of interest rather than ZNF143. Another experiment would be to observe the interaction *in vitro* through carrying out pull-down assays with recombinantly purified proteins.

On the other hand, it is interesting to note that two proteins (CTCF and CHD8) that have been shown to associate with ZNF143 in previous publications (Ishihara, Oshimura and Nakao, 2006; Yuan *et al.*, 2007) scored very low in the co-IP assay, hinting that the direct interaction between ZNF143 and CTCF / CHD8 might not exist (Table 4.11). There are a few possibilities to account for the conflicting results. The association between ZNF143 and CTCF / CHD8 might be relatively transient or weak and therefore the interaction was not captured by co-IP; or these proteins do not interact directly, but rather require a cofactor to facilitate the interaction. Whether or not ZNF143 interacts with CTCF and/or CHD8 are analysed through *in vitro* pull-down assays which will be discussed in section 7.4.2 and 7.4.3, respectively.

### 7.3 Characterisation of Human CHD8

#### 7.3.1 Generation of Functional Human Full-Length CHD8 Protein

Recombinant CHD8 has been purified in previous studies for *in vitro* assays (Thompson, Lin and Bochar, 2008; Manning and Yusufzai, 2017). To ensure the quality of purified CHD8 is suitable for structural characterisation, the purification protocol has been adapted where a Q column ion exchange chromatography was included as an extra step to remove DNA and protein contaminants. Human CHD8 (with N-terminal StrepII and C-terminal FLAG tag) was expressed and purified from baculovirus infected insect cells following a three-step purification protocol (Figure 5.2). Since CHD8 was expressed in insect cells, the purified protein might be heavily phosphorylated (Figure 5.3). Although the function of different CHD8 phosphorylation states have not been documented, CHD8 phosphorylation might be vital in acting as a molecular switch in the regulation of specific pathways which will require further investigation.

A functional protein is the foundation for biochemical and structural characterisation. CHD8 was confirmed to be functional and active through ATPase hydrolysis assays in the presence of nucleosomes, including nucleosome EMSA and nucleosome remodelling assays. Wild type CHD8

exhibited nucleosome-stimulated ATPase activity and the mutation in the ATPase domain (mutant CHD8<sub>K842R</sub>) abolished the activity (Figure 5.5). These results were in accordance with previous studies (Thompson, Lin and Bochar, 2008; Manning and Yusufzai, 2017), confirming that the recombinantly purified wild type CHD8 was active.

Although the analysis from mass photometry suggested CHD8 was monomeric in solution (Figure 5.6), the negative stain EM micrographs revealed the contrary where the particles were seen to be heavily aggregated (Figure 5.7). The discrepancy might be due to the difference in the concentration of protein used in the two experiments and CHD8 might aggregate due to the deposition onto carbon surface grids or uranyl acetate staining during EM screening. While the concentration must be kept low in mass photometry to allow the detection of individual molecules, negative stain EM required higher concentrations to generate enough particles for particle picking and model building in the downstream processes. To obtain higher quality homogenous protein sample, buffer optimisation could be performed using DSF as previously mentioned. It might also be beneficial to avoid freeze-thaw cycles by performing another round of size-exclusion chromatography straight after the gel filtration purification step followed by negative stain screening. Furthermore, substrates such as nucleosomes might be required to improve the stability of CHD8 to allow further structural analysis.

### 7.3.2 CHD8 Performs Nucleosome Sliding through ATP Hydrolysis

CHD8 belongs to the chromodomain helicase DNA-binding family that utilises ATP hydrolysis to catalyse a broad range of chromatin interactions including nucleosome sliding. The SF2 helicase present in all CHD proteins is the key to ATP hydrolysis and nucleosome remodelling (Tyagi *et al.*, 2016). Most of our current understanding on the mechanistic role of CHD proteins have been based on yeast Chd1 and human CHD4 as seen in the recent structural and functional studies (Farnung *et al.*, 2017; Farnung, Ochmann and Cramer, 2020; Zhong *et al.*, 2020). Firstly, I showed that CHD8 was able to bind nucleosomes with 63 bp linker at the exit site (Figure 5.9). Supershifts were observed as the

concentration of CHD8 increased to four-fold molar excess ratio of nucleosomes, indicating that two or more CHD8 molecules might be binding to one nucleosome. This observation has also been characterised previously in Chd1 and CHD4 where the second band shift was later confirmed to be a 2:1 complex in which each superhelical location 2 (SHL2) site was occupied by a CHD molecule (Farnung *et al.*, 2017; Zhong *et al.*, 2020). Comparing the EMSA band shifts, nucleosomes acted as a better substrate for CHD8 than DNA, which was likely due to the binding of CHD8 to histone H3 tails which has been suggested to increase the affinity between the two molecules (Yuan *et al.*, 2007).

Next, I performed nucleosome binding studies to show that CHD8 repositioned nucleosomes along the nucleosomal DNA in the presence of ATP. A mutation in the ATPase site of CHD8 ( $\text{CHD8}_{\text{K842R}}$ ) that was confirmed to be deficient in ATP hydrolysis significantly reduced nucleosome sliding activity, indicating that the ATPase domain was essential for the activity (Figure 5.11). It was intriguing to observe that the binding of CHD8 to nucleosome did not require ATP (Figure 5.10), suggesting that ATP hydrolysis might be correlated to nucleosome sliding to a higher extent than to nucleosome binding.

A preliminary 3D map of crosslinked CHD8-nucleosome complex from negative stain EM was presented (Figure 5.15). AMP-PNP was added to the reaction to capture the state of CHD8 poised for catalytic activity. Although the complex was crosslinked, there were still a significant portion of particles that were nucleosome only as seen by its characteristic disc shape. These particles were removed in the subsequent steps and the final resolution of the map was 31.5 Å. Despite the low final resolution, the 3D map appeared to be comprised of two components, where a disc-shaped structure was observed indicating the presence of nucleosome, and CHD8 might be positioned in the remaining density. The approximation of molecular weight using Chimera also suggested the presence of both CHD8 and nucleosome in the 3D map. Proceeding to cryo-EM did not result in data collection due to the dissociation of the DNA from the complex (Figure 5.16). To enhance the stability and the

homogeneity of the CHD8-nucleosome complex, a few variables can be optimised. Firstly, the choice of crosslinker and its concentration can affect the stability of the complex. Glutaraldehyde was used in this experiment due to its high reactivity. It bridges amine groups with a spacer arm length of 5 Å (Migneault *et al.*, 2004). Other crosslinkers including BS3 can be adopted. BS3 acts as an amine-to-amine crosslinker with a longer spacer arm length of 11 Å, which might result in a complex that is more stable. Next, different nucleosomal DNA lengths might affect the affinity between the nucleosome and CHD8. It has been shown that extranucleosomal DNA is required for stronger binding between CHD remodelers and nucleosomes, yet the length of the linker required varies (McKnight *et al.*, 2011; Farnung *et al.*, 2017; Farnung, Ochmann and Cramer, 2020). Lastly, crosslinking the CHD8-nucleosome complex prior to glycerol gradient and dialysis might avoid the dissociation of the complex due to the difference in buffer conditions.

It will also be intriguing to understand the role of CHD8 in nucleosome sliding using single molecule fluorescence resonance energy (smFRET). smFRET monitors the change in distance between the DNA and histone as the nucleosome is repositioned through a pair of donor and acceptor fluorophores. The overall nucleosome sliding mechanism is expected to be similar to other CHD proteins where they all support the twist diffusion model (Winger *et al.*, 2018; Sabantsev *et al.*, 2019; Zhong *et al.*, 2020). On the other hand, there might be slight differences in the mode of unwrapping DNA from the histones due to the difference in the auxiliary domains across the CHD families. Together with structural characterisation, these results will provide valuable insights into the how CHD8 is engaged with nucleosomes and the mechanism for CHD8 nucleosome sliding.

## 7.4 The Interaction between Chromatin Proteins CTCF, ZNF143 and CHD8

### 7.4.1 The Interplay Between CTCF and CHD8

The association between human CTCF and CHD8 was first observed through immunoprecipitation assays (Ishihara, Oshimura and Nakao, 2006). ChIP-seq analysis further revealed that CHD8 and CTCF colocalise extensively at CTCF insulator binding sites to mediate insulator activities (Ishihara, Oshimura and Nakao, 2006). The direct interaction between CTCF and CHD8 has been demonstrated with mouse CTCF and CHD8 in yeast two-hybrid screening and pull-down assay (Ishihara, Oshimura and Nakao, 2006). Here, I showed that human CTCF and CHD8 interacted directly (Figure 6.1) and these two proteins formed a stable complex over gel filtration (Figure 6.2), indicating a strong association between them. The domains involved in the interaction were pinpointed with cross-link mass spectrometry (XL-MS) (Figure 6.7) and confirmed with pull-down assays (Figure 6.8, 6.9).

The CTCF-CHD8 complex was analysed with electron microscopy. Using negative stain EM, a preliminary 3D map at a resolution of 28.9 Å was reconstructed using *ab initio* model (Figure 6.5). Due to the low resolution, it was unable to fit the known/homology structures of CHD8 and CTCF into the map, but the approximation of the molecular weight based on the map indicated the presence of a CTCF-CHD8 complex at 1:1 stoichiometric ratio. High signal to noise in the background of the 2D classes from the cryo-EM data collection limited the resolution (Figure 6.6) and indicated that further optimisation will be needed to achieve a higher resolution structure. This includes collecting more data to increase the number of particles for 2D classification to enhance the protein features, introducing a second gel filtration protocol or glycerol gradient to remove potential heterogeneity, moderating the cryogenic conditions to optimise ice thickness, and substituting carbon-supported grids with copper or gold grids to improve imaging.

One particular region of CTCF within the zinc finger domain interacted with CHD8 across the C-terminal domain, spanning between residues 1750 – 2550

aa (Figure 6.7). The multiple cross-linked sites in CHD8 hinted that these residues may be in close proximity in 3D space even if they were not adjacent to each other in the amino acid sequence. CHD8 might interact with zinc finger 7 in CTCF, suggesting that the zinc finger domain of CTCF was able to interact with proteins in addition to its crucial DNA-binding role. The CTCF zinc finger domain is comprised of eleven tandemly repeated zinc finger DNA binding motifs, and in particular, ZF3-7 has been shown to direct base-specific interaction with DNA base pairs (Hashimoto *et al.*, 2017). The interaction between CHD8 and CTCF might also be modulated by the terminal zinc fingers (ZF8-11) that lacked DNA specificity. Although the BRK domain was one of the most prominent features within the C-terminal domain of CHD8, direct interaction was not observed between the purified CHD8<sub>BRK</sub> construct and full length CTCF (Figure 6.9). However, this did not exclude the possibility that a longer flanking region around the BRK domain might be required to mediate the interaction. XL-MS provided low-resolution structural information, and the results from XL-MS might be biased depending on the distance between the cross-linked regions which can be affected by the type of cross-linkers and the chemical and physical properties of the protein complex (Merkley, Cort and Adkins, 2013). Hence in addition to validating the results from XL-MS through pull-down assays with purified protein constructs, further structural analysis will be needed to reveal the details in the protein interaction.

Following the identification of the CTCF zinc finger domain interacting with CHD8, it was intriguing to understand if CTCF was able to bind to CHD8 and DNA simultaneously. Our findings suggested that CTCF cannot interact with CHD8 and DNA at the same time (Figure 6.10-13). This indicated that the CTCF zinc finger domain might only be able to accommodate either CHD8 or DNA. ZF3-7 in CTCF has been shown to direct base-specific contacts with DNA (Hashimoto *et al.*, 2017), since CHD8 interacted with ZF7, it might be likely that the extended CHD8 terminal domains inhibited the binding of DNA to the remaining zinc fingers. DNA was able to disrupt the interaction between the pre-formed CTCF-CHD8 complex, but this was dependent on the presence of the CTCF-consensus motif (Figure 6.10-14). Moreover, CHD8

was only able to interact with free CTCF (Figure 6.11) and CTCF showed stronger binding affinity towards DNA with CTCF-motif over CHD8 (Figure 6.11-13). Together, these results revealed a significant role of CTCF-consensus sites in the interaction between CTCF and CHD8. Lastly, since DNA methylation modulates CTCF-DNA binding affinities (Wang *et al.*, 2012) and CTCF-CHD8 complex has been shown to participate in CpG methylation (Ishihara, Oshimura and Nakao, 2006), it would be intriguing to repeat these experiments with DNA constructs with different methylation patterns.

The colocalization of CHD8 with CTCF at CTCF-binding sites has been reported to modulate the average distance between the centres of neighbouring nucleosomes and perform insulator functions (Ishihara, Oshimura and Nakao, 2006; Clarkson *et al.*, 2019). However, it is unclear how CHD8 colocalises with CTCF across the genome, in particular how these two proteins are recruited to the CTCF-binding sites remain elusive. As CHD8 does not bind to or recognise a specific consensus sequence, it might be possible that the CTCF-CHD8 complex is recruited to CTCF-binding sites recognised by CTCF. The complex dissociates as it comes into close proximity with CTCF-binding sites and this allows the free CHD8 to induce nucleosome remodelling activities on nearby chromatin. On the other hand, it might also be possible that as CHD8 is recruited to the chromatin to remodel nucleosomes, where a previously occluded CTCF site might then be exposed. Hence the association between CHD8 and CTCF would enable rapid deposition of CTCF at the newly exposed CTCF-binding site. Further *in vivo* experiments will be needed to confirm these speculations. Furthermore, CTCF has been shown to modulate nucleosome positions (Clarkson *et al.*, 2019), it will also be intriguing to explore the nucleosome remodelling activities of CHD8 in the presence of CTCF.

Interaction was not observed between CTCF and CHD4 (another chromatin remodeller) through a co-expression assay (Figure 6.21). In addition to CHD4 and CHD8, other chromatin remodelling proteins such as EP400 (INO80 family) and SNF2H (ISWI family) have been shown to colocalise at CTCF

binding sites (Wiechens *et al.*, 2016; Clarkson *et al.*, 2019). In particular, CHD4 has been identified to be the top CTCF-related remodeler through colocalisation when compared with other CHD proteins, and CHD8 ranked second (Clarkson *et al.*, 2019). The lack of interaction between CTCF and its top related remodeler CHD4 might suggest CTCF colocalise with CHD4 through different mechanisms, which might require mediators or cofactors. Importantly, this experiment indicated that the interaction between CTCF and CHD8 might be specific and significant.

#### 7.4.2 The Interplay Between CTCF and ZNF143

Increasing evidence has been suggesting that the interaction between ZNF143 and CTCF might be associated with chromatin interactions (Xie *et al.*, 2013; Bailey *et al.*, 2015; Ye *et al.*, 2016). Through ChIP-seq and ChIA-PET studies, ZNF143 has been shown to be required at promoters to mediate chromatin loop formation with distal regulatory elements (Bailey *et al.*, 2015). The colocalisation of ZNF143 at CTCF binding sites were not limited to within enhancer-promoter anchors but the enrichment of ZNF143 was also detected at chromatin TAD boundaries (Ye *et al.*, 2016). Here, I demonstrated that the colocalization of ZNF143 at CTCF binding sites was not due to the direct interaction between the two proteins. Although direct interaction was not observed between recombinantly purified ZNF143 and CTCF in pull-down assay (Figure 6.15), ZNF143 has been shown to bind to DNA comprised of CTCF-consensus sites (Figure 4.10). In addition, despite the fact that ZNF143 interacted with CTCF-sites with lower binding affinity when compared with ZNF143-motifs, the affinity between ZNF143 and CTCF-site was 2-fold stronger than sites lacking ZNF143 or CTCF-consensus sequences (Figure 4.13). This observation was not surprising since ZNF143- and CTCF-consensus sequences share a high similarity where both of them are comprised of rich GC content (Nakahashi *et al.*, 2013; Ngondo-Mbongo *et al.*, 2013). The strong association between ZNF143 and CTCF-consensus sequence coincides with a previous study where it showed that ZNF143 consensus motifs were not always present at ZNF143 enriched anchors, and the colocalization of ZNF143 and CTCF involved weaker ZNF143 binding

strength (Bailey *et al.*, 2015; Wen *et al.*, 2018). Since the knockdown of ZNF143 has also been shown to destabilise chromatin loops (Wen *et al.*, 2018), it will also be intriguing to further explore the mechanism behind ZNF143 in chromatin loop formation, in particular if ZNF143 shares similar functions as CTCF in chromatin interaction. Furthermore, it is unclear if a potential cofactor is required to stabilise the interaction between CTCF and ZNF143, and this will be discussed in section 7.4.4.

#### 7.4.3 The Interplay Between CHD8 and ZNF143

The association between ZNF143 and CHD8 was first suggested by Yuan *et al.* through immunoprecipitation. In addition, CHD8 has been detected at ZNF143 binding sites near RNA polymerase III type III promoters to contribute to efficient transcription initiation (Yuan *et al.*, 2007). Hence the interaction between these two proteins has been postulated to play an important role at these promoters. Although CHD8 was not detected in our endogenous ZNF143 co-immunoprecipitation experiment (Table 4.11), direct interaction was validated through *in vitro* pull-down assay (Figure 6.16). High concentration of purified proteins were used in the pull-down assay, and this might shift the interaction equilibrium to allow detection of interaction using this method. The negative result from the co-IP assay together with the dissociation of CHD8 and ZNF143 over analytical gel filtration (Figure 6.17) suggested that the direct interaction observed might be weak and transient. Next, I demonstrated that the direct interaction to CHD8 might be through the zinc finger region of ZNF143 (Figure 6.18). This was an intriguing result because ZNF143 performs contacts with DNA through its zinc finger domains. Even though no proteins have been shown to interact with ZNF143 zinc finger domain, increasing evidence suggested that protein-protein interactions mediated by zinc fingers are more abundant than expected (Brayer and Segal, 2008). Moreover, since ZF3-6 represented the minimal zinc finger region for DNA binding (Schaub, Krol and Carbon, 2000), a subset of the remaining available zinc fingers might be involved in the CHD8 interaction, but this will require further validation. Preliminary results indicated that the N-terminal domain of CHD8 might be involved in binding to ZNF143 (Figure 6.19). This

region is largely unstructured with the presence of conserved double chromodomains that are involved in binding methylated histones and DNA (Tyagi *et al.*, 2016). Nevertheless, this observation will need to be repeated with purified CHD8 constructs to validate the experiment.

Next steps to understanding potential binding between ZNF143 and CHD8 would include determining the kinetics of the interaction using techniques including isothermal titration calorimetry (ITC), microscale thermophoresis (MST) and surface plasmon resonance (SPR). Although all techniques measure binding affinities, they present different advantages. For instance, ITC provides the direct determination of the dissociation constant, however it requires high sample volume and concentration, and it can be time consuming (Freyer and Lewis, 2008). On the other hand, SPR is a label-free technique that requires low sample consumption, however it is susceptible to high background noise and non-specific interactions with the sensor surface (Douzi, 2017). It would also be important to corroborate the findings on the regions involved in the interaction. This can be achieved through cross-linking mass spectrometry where CHD8 and ZNF143 are crosslinked to provide a snapshot of the amino acid residues that are in close proximity (O'Reilly and Rappaport, 2018). Competition pull-down assays should also be performed to understand if the zinc finger domain in ZNF143 can be occupied by CHD8 and DNA at the same time, and it will also be interesting to probe whether CHD8 is able to bind to nucleosomes and ZNF143 simultaneously.

Furthermore, structural analysis of the ZNF143-CHD8 complex would shed light into the mechanism behind the interaction. To overcome the complex dissociation over gel filtration, cross-linkers can be added to the reaction prior to loading onto the gel filtration column to stabilise the complex. Optimisation on the type of crosslinkers and the concentration used might be crucial to ensure the tight association between CHD8 and ZNF143. The purified stable complex can then be applied to negative stain grids and subsequently cryo-EM grids for structural characterisation.

While I demonstrated that CHD8 and ZNF143 interact directly, the mechanism by which these two proteins activate transcription from the RNA polymerase III *U6* promoter remains elusive, it is unclear if CHD8 undergoes chromatin remodelling and facilitates the recruitment of ZNF143 to its DSE binding site or if the interaction between CHD8 and ZNF143 is crucial to chromatin loop formations. High resolution structure of CHD8-ZNF143 complex in the presence of *U6* promoter could identify residues involved in mediating the interaction. Point mutations can be then be introduced to specifically modulate the binding interface and *in vivo* studies will also be essential to validate their roles involved in transcription.

#### 7.4.4 The Interplay Between CTCF, ZNF143 and CHD8

CTCF and cohesin are the two main architectural proteins that mediate chromatin looping, and the established chromatin looping model suggested that the translocation of cohesin along the DNA is stalled by a pair of CTCF dimers (Sanborn *et al.*, 2015; Fudenberg *et al.*, 2016). However, the recent evidence of ZNF143 involving in stabilisation of chromatin loops and its presence at CTCF-binding sites indicated potential new mechanisms of chromatin loop formation (Bailey *et al.*, 2015; Wen *et al.*, 2018). It is unclear if cohesin translocation can be stalled by one CTCF and one ZNF143 protein instead of CTCF homodimers. Since CTCF and ZNF143 did not interact directly (Figure 6.15) and no studies had reported the interaction between ZNF143 and cohesin, it was interesting to explore the potential cofactor that mediates the interaction. In particular, CHD8 is the only protein that has been reported to colocalise at CTCF- and ZNF143-binding sites (Ishihara, Oshimura and Nakao, 2006; Yuan *et al.*, 2007). Here, I demonstrated that CHD8 can interact with CTCF and ZNF143 simultaneously through an *in vitro* pull-down study, suggesting the formation of a CTCF-CHD8-ZNF143 complex (Figure 6.20). The interactions might be formed between the N-terminal CHD8 with ZNF143, and C-terminal CHD8 with CTCF. A recent review postulated that this complex might be involved at chromatin loop anchors (Huning and Kunkel, 2021). How these proteins facilitate chromatin interactions and if this complex is also involved in nucleosome remodelling remain elusive. Therefore, in

addition to chromatin looping, it would also be intriguing to understand if this complex might also play other significant roles in genome organisation that are yet to be determined and validated.

It is known that CTCF and cohesin interacts directly to facilitate the formation of chromatin loops (Li *et al.*, 2020), it will be intriguing to investigate if ZNF143 and/or CHD8 interact with cohesin to reveal how ZNF143 and CHD8 engage in chromatin interactions, and this can be validated through pull-down assays with recombinantly purified proteins. Moreover, cryo-EM experiments of the CTCF-CHD8-ZNF143 complex will uncover its structural and mechanistic information, whereas *in vivo* experiments will reveal the significance of the complex in regulating chromatin environment and gene transcription. The function of architecture proteins, including CTCF and cohesin have been largely explored via molecular crosslinking-based assays (Rao *et al.*, 2014; Li *et al.*, 2020), in particular the effect of CTCF and cohesin depletions have been extensively studied (Nora *et al.*, 2017; Rao *et al.*, 2017). It would be intriguing to reveal the importance of CHD8 and ZNF143 in chromatin loop formations using high-throughput 3C-based (Hi-C) and auxin-inducible degron (AID) techniques. The AID system allows the depletion of protein to below detectable levels in an acute and reversible manner (Nishimura *et al.*, 2009), hence together with Hi-C techniques, they reveal the effect of the depleted proteins on chromatin organisation in the genome-wide scale. Recent study has developed an improved AID system for ZNF143 and thereby this technique can be incorporated with Hi-C to address the fundamental questions in 3D genome organisation. Lastly, single-molecule imaging assay has been used to reveal how CTCF and cohesin co-occupy the same binding sites in chromatin loop formation (Hansen *et al.*, 2017), and hence it can also be applied to investigate the mechanism of chromatin looping involving the CTCF-CHD8-ZNF143 complex. Single-molecule fluorescence microscopy coupled with optical tweezers would provide insights into the individual molecular states and dynamics of the proteins of interest through monitoring the positions of the fluorescently labelled proteins over time. The colocalisation of proteins, rapid binding and unbinding events, and the subtle changes in the

distances and forces between molecules can be recorded and analysed (Tinoco and Gonzalez, 2011). The results from this experiment will reveal unprecedented information on the behaviour of CTCF-CHD8-ZNF143 on DNA, and how this complex might mediate chromatin loop formations.

## 7.5 Future Directions

While the findings presented in this thesis revealed interesting properties of CTCF, ZNF143 and CHD8 alone and together, they also opened up more questions that need to be answered, and these include:

- The ability for ZNF143 to bind to CTCF-consensus sequence at high affinity, interact with a diverse range of DNA sequences through different subset of zinc fingers and mediate chromatin loop formation are comparable to the functions of CTCF. Hence, it would be intriguing to investigate the significance of CTCF and ZNF143 displaying similar properties, and the functional difference between chromatin loops mediated by the presence of ZNF143/CTCF and CTCF homodimers.
- It remains unclear how CTCF cannot bind to CHD8 and DNA simultaneously and the mechanism behind CHD8 recruitment to CTCF binding sites requires validation. In addition, it would be interesting to understand the effects of the CTCF-CHD8 complex on nearby nucleosome positions.
- The formation of the CTCF-CHD8-ZNF143 complex hinted a possible role in chromatin loop formation, where CHD8 acted as a cofactor. Yet, further investigation will be needed to confirm the significance of the complex at chromatin looping and it will be intriguing to explore if this complex displays other roles, potentially at specific gene promoter regions.

In each of these cases, further structural analysis using cryo-EM might allow a hypothesis of a model for their roles in chromatin interactions with detailed dissection of the interactions between the factors allowing for a greater understanding of the complexes formed. However, to interpret such characterisation, both *in vivo* and *in situ* assays will be important to complement and unravel the functional relevance of any identified complex formation.

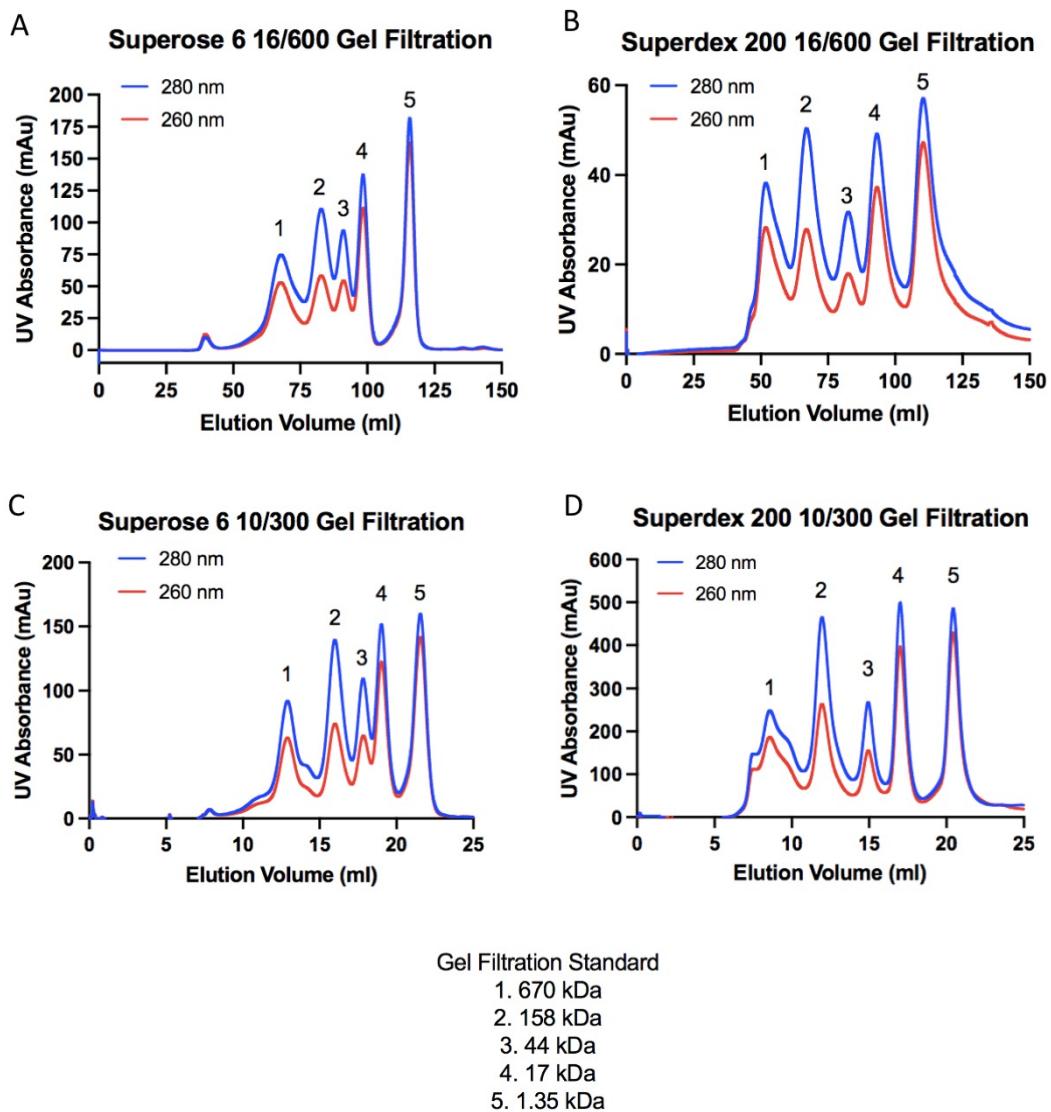
## 7.6 Conclusions

Genome organisation and gene transcription are interdependent and are fundamental to all life forms. 3D genome organisation is hierarchical through various levels of folding, including nucleosome wrapping, chromatin loops, TADs, compartments and chromosome territories (Bonev and Cavalli, 2016). These levels span from a few bases to over megabases long and are tightly regulated by an extensive list of chromatin proteins, in which CTCF and cohesin have been identified to be the two key architecture proteins underlying the formation of chromatin loops and TAD boundaries (Merkenschlager and Nora, 2016). TAD regions and chromatin loops are crucial to genome-wide expression profiles, in addition, other events including nucleosome remodelling to modulate chromatin accessibility for transcription factors by chromatin remodelling proteins are also important in regulating eukaryotic transcription (Schoenfelder and Fraser, 2019). Increasing studies demonstrated that in addition to CTCF and cohesin, ZNF143 and CHD8 play significant roles in genome organisation (Clarkson *et al.*, 2019; Huning and Kunkel, 2021). Although the interplay between CTCF, ZNF143 and CHD8 is essential for gene expression regulation, limited interaction studies have been conducted on these proteins. Hence, this thesis focused on the exploration of the interaction between CTCF, ZNF143 and CHD8, in particular the biochemical and structural features of these proteins and protein complexes.

Since the understanding of the genome architecture has been mostly explored with chromatin capture techniques, I employed biochemical pull-down assays

and structural analysis through electron microscopy to shed light into the mechanistic details. To investigate the interactions between CTCF, ZNF143 and CHD8, detailed functional assays were first carried out on individual proteins to provide insights on how they might act on the genome. Purification protocols of human CTCF, ZNF143 and CHD8 were established. Their oligomeric state, DNA- and nucleosome-binding properties were investigated with a range of biochemical and biophysical experiments. Next, direct interactions between these proteins were probed with pull-down assays using recombinantly purified proteins, where CHD8 was able to interact with CTCF and ZNF143 simultaneously. However direct binding between CTCF and ZNF143 was not observed, indicating that CHD8 might be the cofactor that mediated the association of CTCF and ZNF143 at chromatin loop sites. Since the interaction between CHD8 and CTCF was more stable, further structural analysis was performed on the complex using negative stain and cryo-EM. Together, the biochemical and biophysical studies presented in this thesis provided a fitting contribution to our current knowledge about the molecular mechanisms in genome organisation.

# Appendix



**Figure Appendix.1 Gel Filtration Standards on Different Gel Filtration Columns**

Gel filtration standards (Biorad, #1511901) were ran on four different gel filtration columns used in this thesis. Five proteins were present in the gel calibration standard, and all elution peaks were labelled.

# Bibliography

1. Aasland, R., Stewart, A. F. and Gibson, T. (1996) 'The SANT domain: A putative DNA-binding domain in the SWI-SNF and ADA complexes, the transcriptional co-repressor N-CoR and TFIIB', *Trends in Biochemical Sciences*, 21(3), pp. 87–88. doi: 10.1016/S0968-0004(96)30009-1.
2. Adrian, M. et al. (1984) 'Cryo-electron Microscopy of Viruses', *Nature*, 308, pp. 32–36. doi: 10.1038/308032a0.
3. Akdemir, K. C. et al. (2020) 'Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer', *Nature Genetics*, 52(3), pp. 294–305. doi: 10.1038/s41588-019-0564-y.
4. Alberts, B. et al. (2003) 'Molecular biology of the cell', *Biochemistry and Molecular Biology Education*, 31(4), pp. 212–214. doi: 10.1002/bmb.2003.494031049999.
5. Alpsoy, A., Sood, S. and Dykhuizen, E. C. (2021) 'At the crossroad of gene regulation and genome organization: Possible roles of ATP-dependent chromatin remodelers in regulation of 3D genome architecture', *Biology*.
6. Andersson, R. and Sandelin, A. (2020) 'Determinants of enhancer and promoter activities of regulatory elements', *Nature Reviews Genetics*, 21(2), pp. 71–87. doi: 10.1038/s41576-019-0173-8.
7. Anno, Y.-N. et al. (2011) 'Genome-wide evidence for an essential role of the human Staf/ZNF143 transcription factor in bidirectional transcription', *Nucleic Acids Research*, 39(8), pp. 3116–3127. doi: 10.1093/nar/gkq1301.
8. van Arensbergen, J., van Steensel, B. and Bussemaker, H. J. (2014) 'In search of the determinants of enhancer-promoter interaction specificity', *Trends in Cell Biology*. Elsevier Ltd, 24(11), pp. 695–702. doi: 10.1016/j.tcb.2014.07.004.
9. Bailey, S. D. et al. (2015) 'ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters', *Nature Communications*, 2, p. 6186. doi: 10.1038/ncomms7186.
10. Barnard, R. A., Pomaville, M. B. and O'Roak, B. J. (2015) 'Mutations and Modeling of the Chromatin Remodeler CHD8 Define an Emerging Autism Etiology', *Frontiers in Neuroscience*, 9, p. 477. doi: 10.3389/fnins.2015.00477.
11. Bell, A. C. and Felsenfeld, G. (2000) 'Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene', *Nature*, 405(6785), pp. 482–485. doi: 10.1038/35013100.
12. Bell, A. C., West, A. G. and Felsenfeld, G. (1999) 'The protein CTCF is required for the enhancer blocking activity of vertebrate insulators', *Cell*, 98(3), pp. 387–396. doi: 10.1016/S0092-8674(00)81967-4.
13. Bonchuk, A. et al. (2020) 'N-terminal domain of the architectural protein CTCF has similar structural organization and ability to self-association in bilaterian organisms', *Scientific Reports*, 10(1), p. 2677. doi: 10.1038/s41598-020-59459-5.
14. Bonev, B. and Cavalli, G. (2016) 'Organization and function of the 3D genome', *Nature Reviews Genetics*, 17(11), pp. 661–678. doi: 10.1038/nrg.2016.112.
15. Bornelöv, S. et al. (2018) 'The Nucleosome Remodeling and Deacetylation Complex Modulates Chromatin Structure at Sites of Active Transcription to Fine-Tune Gene Expression', *Molecular Cell*, 71(1), pp. 56-72.e4. doi: 10.1016/j.molcel.2018.06.003.
16. Bowman, G. D. (2019) 'Uncovering a New Step in Sliding Nucleosomes HHS Public Access', *Trends Biochem Sci*, 44(8), pp. 643–645. doi: 10.1016/j.tibs.2019.05.001.

17. Brayer, K. J. and Segal, D. J. (2008) 'Keep your fingers off my DNA: Protein-protein interactions mediated by C2H2 zinc finger domains', *Cell Biochemistry and Biophysics*, 50(3), pp. 111–131. doi: 10.1007/s12013-008-9008-5.
18. Brenner, S. and Horne, R. W. (1959) 'A negative staining method for high resolution electron microscopy of viruses', *BBA - Biochimica et Biophysica Acta*. Elsevier, 34(C), pp. 103–110. doi: 10.1016/0006-3002(59)90237-9.
19. Brilot, A. F. *et al.* (2012) 'Beam-Induced Motion of Vitrified Specimen on Holey Carbon Film', *J Struct Biol*, 177(3), pp. 630–637. doi: 10.1016/j.jsb.2012.02.003.
20. Bulger, M. and Groudine, M. (2011) 'Functional and mechanistic diversity of distal transcription enhancers', *Cell*. Elsevier Inc., 144(3), pp. 327–339. doi: 10.1016/j.cell.2011.01.024.
21. Campbell, M. G. *et al.* (2012) 'Movies of ice-embedded particles enhance resolution in electron cryo-microscopy', *Structure*. Elsevier Ltd, 20(11), pp. 1823–1828. doi: 10.1016/j.str.2012.08.026.
22. Cattoglio, C. *et al.* (2019) 'Determining cellular CTCF and cohesin abundances to constrain 3D genome models', *eLife*, 8. doi: 10.7554/eLife.40164.
23. de Chadarevian, S. (2018) 'John Kendrew and myoglobin: Protein structure determination in the 1950s', *Protein Science*, 27(6), pp. 1136–1143. doi: 10.1002/pro.3417.
24. Chernukhin, I. V. *et al.* (2000) 'Physical and functional interaction between two pluripotent proteins, the Y-box DNA/RNA-binding factor, YB-1, and the multivalent zinc finger factor, CTCF', *Journal of Biological Chemistry*. © 2000 ASBMB. Currently published by Elsevier Inc; originally published by American Society for Biochemistry and Molecular Biology., 275(38), pp. 29915–29921. doi: 10.1074/jbc.M001538200.
25. Choo, Y. and Klug, A. (1997) 'Physical basis of a protein-DNA recognition code', *Current Opinion in Structural Biology*, 7(1), pp. 117–125. doi: 10.1016/S0959-440X(97)80015-2.
26. Clapier, C. R. *et al.* (2017) 'Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 18(7), pp. 407–422. doi: 10.1038/nrm.2017.26.
27. Clarkson, C. T. *et al.* (2019) 'CTCF-dependent chromatin boundaries formed by asymmetric nucleosome arrays with decreased linker length', *Nucleic acids research*. Oxford University Press, 47(21), pp. 11181–11196. doi: 10.1093/nar/gkz908.
28. Cordes, R. M., Sims, W. B. and Glatz, C. E. (1990) 'Precipitation of Nucleic Acids with Poly (ethyleneimine)', *Biotechnology Progress*, 6(4), pp. 283–285. doi: 10.1021/bp00004a009.
29. Cramer, P. (2019) 'Organization and regulation of gene transcription', *Nature*, 573(7772), pp. 45–54. doi: 10.1038/s41586-019-1517-4.
30. Cuddapah, S. *et al.* (2008) 'Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains', *Genome Research*, 19(1), pp. 24–32. doi: 10.1101/gr.082800.108.
31. Davidson, I. F. *et al.* (2016) 'Rapid movement and transcriptional re-localization of human cohesin on DNA', *The EMBO Journal*, 35(24), pp. 2671–2685. doi: 10.15252/embj.201695402.
32. Dekker, J. (2002) 'Capturing Chromosome Conformation', *Science*, 295(5558), pp. 1306–1311. doi: 10.1126/science.1067799.
33. Desjarlais, J. R. and Berg, J. M. (1992) 'Toward rules relating zinc finger protein sequences and DNA binding site preferences', *Proceedings of the National Academy of Sciences of the United States of America*, 89(16), pp. 7345–7349. doi: 10.1073/pnas.89.16.7345.

34. de Dieuleveult, M. et al. (2016) 'Genome-wide nucleosome specificity and function of chromatin remodelers in ES cells', *Nature*, 530(7588), pp. 113–116. doi: 10.1038/nature16505.
35. Dixon, J. R. et al. (2012) 'Topological domains in mammalian genomes identified by analysis of chromatin interactions', *Nature*. Nature Publishing Group, 485(7398), pp. 376–380. doi: 10.1038/nature11082.
36. Dixon, J. R. et al. (2015) 'Chromatin architecture reorganization during stem cell differentiation', *Nature*. Nature Publishing Group, 518(7539), pp. 331–336. doi: 10.1038/nature14222.
37. Dong, A., Xu, X. and Edwards, A. M. (2007) 'In situ proteolysis for protein crystallization and structure determination'. doi: 10.1038/NMETH1118.
38. Douzi, B. (2017) 'Chapter 21: Protein-Protein interactions: Surface plasmon resonance', *Bacterial Protein Secretion Systems: Methods and Protocols*, 1615, pp. 257–275. doi: 10.1007/978-1-4939-7033-9.
39. Dyer, P. N. et al. (2003) 'Reconstitution of Nucleosome Core Particles from Recombinant Histones and DNA', *Methods in Enzymology*, 375, pp. 23–44. doi: 10.1016/S0076-6879(03)75002-2.
40. Eissenberg, J. C. (2012) 'Structural biology of the chromodomain: Form and function', *Gene*, 496(2), pp. 69–78. doi: 10.1016/j.gene.2012.01.003.
41. Elrod-Erickson, M. et al. (1996) 'Zif268 protein-DNA complex refined at 1.6 Å: A model system for understanding zinc finger-DNA interactions', *Structure*, 4(10), pp. 1171–1180. doi: 10.1016/S0969-2126(96)00125-6.
42. Fang, C. et al. (2020) 'Cancer-specific CTCF binding facilitates oncogenic transcriptional dysregulation', *Genome Biology*. Genome Biology, 21(1), pp. 1–30. doi: 10.1186/s13059-020-02152-7.
43. Farnung, L. et al. (2017) 'Nucleosome-Chd1 structure and implications for chromatin remodelling', *Nature*. Nature Publishing Group, 550(7677), pp. 539–542. doi: 10.1038/nature24046.
44. Farnung, L., Ochmann, M. and Cramer, P. (2020) 'Nucleosome-CHD4 chromatin remodeler structure maps human disease mutations', *eLife*, 9. doi: 10.7554/eLife.56178.
45. Faruqi, A. R. and McMullan, G. (2018) 'Direct imaging detectors for electron microscopy', *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*. Elsevier Ltd, 878(July 2017), pp. 180–190. doi: 10.1016/j.nima.2017.07.037.
46. Felsenfeld, G. et al. (2004) 'Chromatin boundaries and chromatin domains', *Cold Spring Harbor Symposia on Quantitative Biology*, 69, pp. 245–250. doi: 10.1101/sqb.2004.69.245.
47. Fischle, W., Wang, Y. and Allis, C. D. (2003) 'Histone and chromatin cross-talk', *Current Opinion in Cell Biology*, 15(2), pp. 172–183. doi: 10.1016/S0955-0674(03)00013-9.
48. Fish, L. et al. (2019) 'Nuclear TARBP2 Drives Oncogenic Dysregulation of RNA Splicing and Decay', *Molecular Cell*, 75(5), pp. 967–981.e9. doi: 10.1016/j.molcel.2019.06.001.
49. Folta-Stogniew, E. (2006) 'Oligomeric states of proteins determined by size-exclusion chromatography coupled with light scattering, absorbance, and refractive index detectors.', *Methods in molecular biology (Clifton, N.J.)*, 328, pp. 97–112. doi: 10.1385/1-59745-026-x:97.
50. Foster, H. A. and Bridger, J. M. (2005) 'The genome and the nucleus: A marriage made by evolution. Genome organisation and nuclear architecture', *Chromosoma*,

- 114(4), pp. 212–229. doi: 10.1007/s00412-005-0016-6.
51. Fraser, J. et al. (2015) ‘An Overview of Genome Organization and How We Got There: from FISH to Hi-C’, *Microbiology and Molecular Biology Reviews*, 79(3), pp. 347–372. doi: 10.1128/mmbr.00006-15.
52. Freyer, M. W. and Lewis, E. A. (2008) ‘Isothermal Titration Calorimetry: Experimental Design, Data Analysis, and Probing Macromolecule/Ligand Binding and Kinetic Interactions’, *Methods in Cell Biology*, 84(07), pp. 79–113. doi: 10.1016/S0091-679X(07)84004-0.
53. Fudenberg, G. et al. (2016) ‘Formation of Chromosomal Domains by Loop Extrusion’, *Cell Reports*. The Author(s), 15(9), pp. 2038–2049. doi: 10.1016/j.celrep.2016.04.085.
54. Gasteiger, E. et al. (2003) ‘ExPASy: The proteomics server for in-depth protein knowledge and analysis’, *Nucleic Acids Research*, 31(13), pp. 3784–3788. doi: 10.1093/nar/gkg563.
55. Gómez-Marín, C. et al. (2015) ‘Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders’, *Proceedings of the National Academy of Sciences*, 112(24), pp. 7542–7547. doi: 10.1073/pnas.1505463112.
56. Guo, Y. et al. (2015) ‘CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function’, *Cell*. Elsevier Inc., 162(4), pp. 900–910. doi: 10.1016/j.cell.2015.07.038.
57. Guo, Y. A. et al. (2018) ‘Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers’, *Nature Communications*, 9(1), p. 1520. doi: 10.1038/s41467-018-03828-2.
58. Handoko, L. et al. (2011) ‘CTCF-mediated functional chromatin interactome in pluripotent cells’, *Nature GeNetics*, 43. doi: 10.1038/ng.857.
59. Hansen, A. S. et al. (2017) ‘CTCF and cohesin regulate chromatin loop stability with distinct dynamics’, *eLife*, 6, pp. 1–33. doi: 10.7554/eLife.25776.
60. Hansen, A. S. et al. (2019) ‘Distinct Classes of Chromatin Loops Revealed by Deletion of an RNA-Binding Region in CTCF’, *Molecular Cell*, 76(3), pp. 395–411.e13. doi: 10.1016/j.molcel.2019.07.039.
61. Hashimoto, H. et al. (2017) ‘Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA’, *Molecular Cell*. Elsevier Inc., 66(5), pp. 711–720.e3. doi: 10.1016/j.molcel.2017.05.004.
62. Hauk, G. and Bowman, G. D. (2011) ‘Structural Insights into Regulation and Action of SWI2/SNF2 ATPases’. doi: 10.1016/j.sbi.2011.09.003.
63. Van Heel, M. et al. (1996) ‘A new generation of the IMAGIC image processing system’, *Journal of Structural Biology*, 116(1), pp. 17–24. doi: 10.1006/jsb.1996.0004.
64. Heidari, N. et al. (2014) ‘Genome-wide map of regulatory interactions in the human genome’, *Genome Research*, 24(12), pp. 1905–1917. doi: 10.1101/gr.176586.114.
65. Hergeth, S. P. and Schneider, R. (2015) ‘The H1 linker histones: multifunctional proteins beyond the nucleosomal core particle’, *EMBO reports*, 16(11), pp. 1439–1453. doi: 10.15252/embr.201540749.
66. Herold, M., Bartkuhn, M. and Renkawitz, R. (2012) ‘CTCF: Insights into insulator function during development’, *Development*, 139(6), pp. 1045–1057. doi: 10.1242/dev.065268.
67. Hornbeck, P. V et al. (2015) ‘PhosphoSitePlus, 2014: mutations, PTMs and recalibrations’, *Nucleic Acids Research*, 43. doi: 10.1093/nar/gku1267.
68. Huang, J. et al. (2012) ‘Seeding’ with protease to optimize protein crystallization conditions in in situ proteolysis’, *Acta Crystallographica Section F Structural Biology and Crystallization Communications*, 68(5), pp. 606–609. doi:

- 10.1107/S174430911201161X.
69. Huning, L. and Kunkel, G. R. (2021) 'The ubiquitous transcriptional protein ZNF143 activates a diversity of genes while assisting to organize chromatin structure', *Gene*. Elsevier B.V., 769(June 2020), p. 145205. doi: 10.1016/j.gene.2020.145205.
  70. Ishihara, K., Oshimura, M. and Nakao, M. (2006) 'CTCF-Dependent Chromatin Insulator Is Linked to Epigenetic Remodeling', *Molecular Cell*. Elsevier Inc., 23(5), pp. 733–742. doi: 10.1016/j.molcel.2006.08.008.
  71. Izumi, H. et al. (2010) 'Role of ZNF143 in tumor growth through transcriptional regulation of DNA replication and cell-cycle-associated genes', *Cancer Science*, 101(12), pp. 2538–2545. doi: 10.1111/j.1349-7006.2010.01725.x.
  72. James Faresse, N. et al. (2012) 'Genomic Study of RNA Polymerase II and III SNAPc-Bound Promoters Reveals a Gene Transcribed by Both Enzymes and a Broad Use of Common Activators', *PLoS Genetics*, 8(11), pp. 1–14. doi: 10.1371/journal.pgen.1003028.
  73. Jauch, A. et al. (1990) 'Chromosomal in situ suppression hybridization of human gonosomes and autosomes and its use in clinical cytogenetics', *Human Genetics*, 85(2), pp. 145–150. doi: 10.1007/BF00193186.
  74. Jung, C. et al. (2018) 'True equilibrium measurement of transcription factor-DNA binding affinities using automated polarization microscopy', *Nature Communications*. Springer US, 9(1), pp. 1–11. doi: 10.1038/s41467-018-03977-4.
  75. Kawatsu, Y. et al. (2014) 'The combination of strong expression of ZNF143 and high MIB-1 labelling index independently predicts shorter disease-specific survival in lung adenocarcinoma', *British Journal of Cancer*. Nature Publishing Group, 110(10), pp. 2583–2592. doi: 10.1038/bjc.2014.202.
  76. Kelley, L. A. et al. (2015) 'The Phyre2 web portal for protein modeling, prediction and analysis', *Nature Protocols*. Nature Publishing Group, 10(6), pp. 845–858. doi: 10.1038/nprot.2015.053.
  77. Kendrew, J. C. et al. (1960) 'Structure of Myoglobin: A Three-Dimensional Fourier Synthesis at 2 Å. Resolution', *Nature*, 185, pp. 422–427. doi: 10.1038.
  78. Kitchen, N. S. and Schoenherr, C. J. (2010) 'Sumoylation modulates a domain in CTCF that activates transcription and decondenses chromatin', *Journal of Cellular Biochemistry*, 111(3), pp. 665–675. doi: 10.1002/jcb.22751.
  79. Knezetic, J. A. and Luse, D. S. (1986) 'The presence of nucleosomes on a DNA template prevents initiation by RNA polymerase II in vitro', *Cell*, 45(1), pp. 95–104. doi: 10.1016/0092-8674(86)90541-6.
  80. Kornberg, R. D. and Lorch, Y. (1999) 'Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome', *Cell*, 98(3), pp. 285–294. doi: 10.1016/S0092-8674(00)81958-3.
  81. Krauss, I. R. et al. (2013) 'An overview of biological macromolecule crystallization', *International Journal of Molecular Sciences*, 14(6), pp. 11643–11691. doi: 10.3390/ijms140611643.
  82. Krivega, I. and Dean, A. (2017) 'CTCF fences make good neighbours', *Nature Cell Biology*, 19(8), pp. 883–885. doi: 10.1038/ncb3584.
  83. Kubo, N. et al. (2021) 'Promoter-proximal CTCF binding promotes distal enhancer-dependent gene activation', *Nature Structural & Molecular Biology*, 28(2), pp. 152–161. doi: 10.1038/s41594-020-00539-5.
  84. Kucukelbir, A., Sigworth, F. J. and Tagare, H. D. (2014) 'Quantifying the local resolution of cryo-em density maps', 11(1). doi: 10.1038/nMeth.2727.
  85. Kühlbrandt, W. (2014) 'The Resolution Revolution BIOCHEMISTRY'. doi: 10.1126/science.1251652.

86. Kujirai, T. and Kurumizaka, H. (2020) 'Transcription through the nucleosome', *Current Opinion in Structural Biology*. Elsevier Ltd, 61, pp. 42–49. doi: 10.1016/j.sbi.2019.10.007.
87. Lai, W. K. M. and Pugh, B. F. (2017) 'Understanding nucleosome dynamics and their links to gene expression and DNA replication', *Nature Reviews Molecular Cell Biology*, 18(9), pp. 548–562. doi: 10.1038/nrm.2017.47.
88. Li, B., Carey, M. and Workman, J. L. (2007) 'The Role of Chromatin during Transcription', *Cell*, 128(4), pp. 707–719. doi: 10.1016/j.cell.2007.01.015.
89. Li, W. and Mills, A. A. (2014) 'Architects of the genome: CHD dysfunction in cancer, developmental disorders and neurological syndromes'. doi: 10.2217/epi.14.31.
90. Li, Y. et al. (2020) 'The structural basis for cohesin–CTCF-anchored loops', *Nature*. Springer US, 578(7795), pp. 472–476. doi: 10.1038/s41586-019-1910-z.
91. Lichter, P. et al. (1988) 'Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant DNA libraries', *Human Genetics*, 80(3), pp. 224–234.
92. Lieberman-aiden, E. et al. (2009) 'Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome', 33292(October), pp. 289–294.
93. Lieleg, C. et al. (2015) 'Nucleosome Spacing Generated by ISWI and CHD1 Remodelers Is Constant Regardless of Nucleosome Density', *Molecular and Cellular Biology*, 35(9), pp. 1588–1605. doi: 10.1128/mcb.01070-14.
94. Liu, F., Wu, D. and Wang, X. (2018) 'Roles of CTCF in conformation and functions of chromosome', *Seminars in Cell and Developmental Biology*. Elsevier Ltd. doi: 10.1016/j.semcd.2018.07.021.
95. Liu, Y. et al. (2012) 'An atomic model of Zfp57 recognition of CpG methylation within a specific DNA sequence', *Genes and Development*, 26(21), pp. 2374–2379. doi: 10.1101/gad.202200.112.
96. Liu, Z. et al. (2014) 'Enhancer activation requires trans-recruitment of a mega transcription factor complex', *Cell*, 159(2), pp. 358–373. doi: 10.1016/j.cell.2014.08.027.
97. Lowary, P. T. and Widom, J. (1998) 'New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning', *Journal of Molecular Biology*, 276(1), pp. 19–42. doi: 10.1006/jmbi.1997.1494.
98. Ludtke, S. J. (2016) 'Single-Particle Refinement and Variability Analysis in EMAN2.1', in, pp. 159–189. doi: 10.1016/bs.mie.2016.05.001.
99. Luger, K. et al. (1997) 'Crystal structure of the nucleosome core particle at 2.8 Å resolution', *Nature*, 389(6648), pp. 251–260. doi: 10.1038/38444.
100. Luger, K., Rechsteiner, T. J. and Richmond, T. J. (1999) 'Expression and purification of recombinant histones and nucleosome reconstitution.', *Methods in molecular biology (Clifton, N.J.)*, 119(4), pp. 1–16. doi: 10.1385/1-59259-681-9:1.
101. Lupiáñez, D. G., Spielmann, M. and Mundlos, S. (2016) 'Breaking TADs: How Alterations of Chromatin Domains Result in Disease', *Trends in Genetics*. Elsevier Ltd, 32(4), pp. 225–237. doi: 10.1016/j.tig.2016.01.003.
102. Manning, B. J. and Yusufzai, T. (2017) 'The ATP-dependent chromatin remodeling enzymes CHD6 , CHD7 , and CHD8 exhibit distinct nucleosome binding and remodeling activities', 292, pp. 11927–11936. doi: 10.1074/jbc.M117.779470.
103. Marfella, C. G. A. and Imbalzano, A. N. (2007) 'The Chd family of chromatin remodelers', *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 618(1–2), pp. 30–40. doi: 10.1016/j.mrfmmm.2006.07.012.
104. Martinez, S. R. and Miranda, J. L. (2010) 'CTCF terminal segments are

- unstructured.', *Protein science : a publication of the Protein Society*, 19(5), pp. 1110–1116. doi: 10.1002/pro.367.
105. Matharu, N. and Ahituv, N. (2015) 'Minor Loops in Major Folds: Enhancer–Promoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease', *PLoS Genetics*, 11(12), pp. 1–14. doi: 10.1371/journal.pgen.1005640.
106. McKnight, J. N. et al. (2011) 'Extranucleosomal DNA Binding Directs Nucleosome Sliding by Chd1', *Molecular and Cellular Biology*, 31(23), pp. 4746–4759. doi: 10.1128/MCB.05735-11.
107. McMullan, G., Faruqi, A. R. and Henderson, R. (2016) *Direct Electron Detectors*. 1st edn, *Methods in Enzymology*. 1st edn. Elsevier Inc. doi: 10.1016/bs.mie.2016.05.056.
108. McPherson, A. and Gavira, J. A. (2014) 'Introduction to protein crystallization', *Acta Crystallographica Section F: Structural Biology Communications*. International Union of Crystallography, 70(1), pp. 2–20. doi: 10.1107/S2053230X13033141.
109. Mellacheruvu, D. et al. (2013) 'the CRaPome: a contaminant repository for affinity purification-mass spectrometry data'. doi: 10.1038/nmeth.2557.
110. Merkenschlager, M. and Nora, E. P. (2016) 'CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation', *Annual Review of Genomics and Human Genetics*, 17(1), pp. 17–43. doi: 10.1146/annurev-genom-083115-022339.
111. Merkley, E. D., Cort, J. R. and Adkins, J. N. (2013) 'Cross-linking and mass spectrometry methodologies to facilitate structural biology: finding a path through the maze', *Journal of Structural and Functional Genomics*, 14(3), pp. 77–90. doi: 10.1007/s10969-013-9160-z.
112. Migneault, I. et al. (2004) 'Glutaraldehyde: behavior in aqueous solution, reaction with proteins, and application to enzyme crosslinking', *BioTechniques*, 37(5), pp. 790–802. doi: 10.2144/04375RV01.
113. Mills, A. A. (2017) 'The Chromodomain Helicase DNA-Binding Chromatin Remodelers: Family Traits that Protect from and Promote Cancer', *Cold Spring Harbor Perspectives in Medicine*, 7(4), p. a026450. doi: 10.1101/cshperspect.a026450.
114. Mishra, A. and Hawkins, R. D. (2017) 'Three-dimensional genome architecture and emerging technologies: Looping in disease', *Genome Medicine. Genome Medicine*, 9(1), pp. 1–14. doi: 10.1186/s13073-017-0477-2.
115. Moerke, N. J. (2009) 'Fluorescence Polarization (FP) Assays for Monitoring Peptide-Protein or Nucleic Acid-Protein Binding', *Current Protocols in Chemical Biology*, 1(1), pp. 1–15. doi: 10.1002/9780470559277.ch090102.
116. El Mourad, R. and Cuvier, O. (2017) 'TAD-free analysis of architectural proteins and insulators', *Nucleic Acids Research*, 46(5). doi: 10.1093/nar/gkx1246.
117. Myslinski, E. et al. (2006) 'A genome scale location analysis of human Staf/ZNF143-binding sites suggests a widespread role for human Staf/ZNF143 in mammalian promoters', *The Journal of Biological Chemistry*, 281(52), pp. 39953–39962. doi: 10.1074/jbc.M608507200.
118. Myslinski, E. et al. (2007) 'Transcription of the human cell cycle regulated BUB1B gene requires hStaf/ZNF143', *Nucleic Acids Research*, 35(10), pp. 3453–3464. doi: 10.1093/nar/gkm239.
119. Myslinski, E., Krol, A. and Carbon, P. (1998) 'ZNF76 and ZNF143 are two human homologs of the transcriptional activator Staf', *Journal of Biological Chemistry*, 273(34), pp. 21998–22006. doi: 10.1074/jbc.273.34.21998.
120. Nakahashi, H. et al. (2013) 'A Genome-wide Map of CTCF Multivalency

- Redefines the CTCF Code', *Cell Reports*. The Authors, 3(5), pp. 1678–1689. doi: 10.1016/j.celrep.2013.04.024.
121. Nakane, T. et al. (2018) 'Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION', *eLife*, 7. doi: 10.7554/eLife.36861.
  122. Nakane, T. et al. (2020) 'Single-particle cryo-EM at atomic resolution', | *Nature* |, 587. doi: 10.1038/s41586-020-2829-0.
  123. Nanni, L., Ceri, S. and Logie, C. (2020) 'Spatial patterns of CTCF sites define the anatomy of TADs and their boundaries', *Genome Biology*. *Genome Biology*, 21(1), pp. 1–25. doi: 10.1186/s13059-020-02108-x.
  124. Narendra, V. et al. (2015) 'CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation', *Science*, 347(6225), pp. 1017–1021. doi: 10.1126/science.1262088.
  125. Naydenova, K., Jia, P. and Russo, C. J. (2020) 'Cryo-EM with sub-1 Å specimen movement', *Science*, 370(6513), pp. 223–226. doi: 10.1126/science.abb7927.
  126. Ngondo-Mbongo, R. P. et al. (2013) 'Modulation of gene expression via overlapping binding sites exerted by ZNF143, Notch1 and THAP11', *Nucleic Acids Research*, 41(7), pp. 4000–4014. doi: 10.1093/nar/gkt088.
  127. Niesen, F. H., Berglund, H. and Vedadi, M. (2007) 'The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability', *Nature Protocols*, 2(9), pp. 2212–2221. doi: 10.1038/nprot.2007.321.
  128. Nikitina, T. et al. (2017) *DNA topology in chromatin is defined by nucleosome spacing*. Available at: <http://advances.sciencemag.org/>.
  129. Nishimura, K. et al. (2009) 'An auxin-based degron system for the rapid depletion of proteins in nonplant cells'. doi: 10.1038/nmeth.1401.
  130. Nogales, E. and Scheres, S. H. W. (2015) 'Cryo-EM: A Unique Tool for the Visualization of Macromolecular Complexity', *Molecular Cell*. Elsevier Inc., 58(4), pp. 677–689. doi: 10.1016/j.molcel.2015.02.019.
  131. Nora, E. P. et al. (2012) 'Spatial partitioning of the regulatory landscape of the X-inactivation centre', *Nature*, 485(7398), pp. 381–385. doi: 10.1038/nature11049.
  132. Nora, E. P. et al. (2017) 'Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization', *Cell*, 169(5), pp. 930-944.e22. doi: 10.1016/j.cell.2017.05.004.
  133. Nora, E. P. et al. (2020) 'Molecular basis of CTCF binding polarity in genome folding', *Nature Communications*, 11(1), p. 5612. doi: 10.1038/s41467-020-19283-x.
  134. O'Reilly, F. J. and Rappsilber, J. (2018) 'Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology', *Nature Structural & Molecular Biology*, 25(11), pp. 1000–1008. doi: 10.1038/s41594-018-0147-0.
  135. Ong, C.-T. and Corces, V. G. (2011) 'Enhancer function: new insights into the regulation of tissue-specific gene expression', *Nature Publishing Group*, 12. doi: 10.1038/nrg2957.
  136. Ong, C.-T. and Corces, V. G. (2014) 'CTCF: an architectural protein bridging genome topology and function', *Nature Reviews Genetics*. Nature Publishing Group, 15(4), pp. 234–246. doi: 10.1038/nrg3663.
  137. Orlova, E. V. and Saibil, H. R. (2011) 'Structural Analysis of Macromolecular Assemblies by Electron Microscopy', *Chemical Reviews*, 111(1), pp. 7710–7748. doi: 10.1021/cr100211k.
  138. Owens, N. et al. (2019) 'CTCF confers local nucleosome resiliency after dna replication and during mitosis', *eLife*, 8, pp. 1–26. doi: 10.7554/eLife.47898.
  139. Pant, V. et al. (2004) 'Mutation of a Single CTCF Target Site within the H19

- Imprinting Control Region Leads to Loss of Igf2 Imprinting and Complex Patterns of De Novo Methylation upon Maternal Inheritance', *Molecular and Cellular Biology*, 24(8), pp. 3497–3504. doi: 10.1128/mcb.24.8.3497-3504.2004.
140. Parelho, V. et al. (2008) 'Cohesins Functionally Associate with CTCF on Mammalian Chromosome Arms', *Cell*, 132(3), pp. 422–433. doi: 10.1016/j.cell.2008.01.011.
  141. Persikov, A. V. and Singh, M. (2014) 'De novo prediction of DNA-binding specificities for Cys2His2zinc finger proteins', *Nucleic Acids Research*, 42(1), pp. 97–108. doi: 10.1093/nar/gkt890.
  142. Perutz, M. F. et al. (1960) 'Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis', *Nature*, 185, pp. 416–422. doi: 10.1038/185.416.
  143. Phillips, J. E. and Corces, V. G. (2009) 'CTCF: Master Weaver of the Genome', *Cell*, 137(7), pp. 1194–1211. doi: 10.1016/j.cell.2009.06.001.
  144. Piazza, I. et al. (2014) 'Association of condensin with chromosomes depends on DNA binding by its HEAT-repeat subunits', *Nature Structural and Molecular Biology*. Nature Publishing Group, 21(6), pp. 560–568. doi: 10.1038/nsmb.2831.
  145. Pontén, F., Jirström, K. and Uhlen, M. (2008) 'The Human Protein Atlas—a tool for pathology', *The Journal of Pathology*, 216(4), pp. 387–393. doi: 10.1002/path.2440.
  146. Ptashne, M. (1986) 'Gene regulation by proteins acting nearby and at a distance', *Nature*, 322(6081), pp. 697–701. doi: 10.1038/322697a0.
  147. Pugacheva, E. M. et al. (2020) 'CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention', *Proceedings of the National Academy of Sciences of the United States of America*, 117(4), pp. 2020–2031. doi: 10.1073/pnas.1911708117.
  148. Punjani, A. et al. (2017) 'cryosPArc: algorithms for rapid unsupervised cryo-em structure determination'. doi: 10.1038/nmeth.4169.
  149. Radnai, L. et al. (2019) 'A Semi-High-Throughput Adaptation of the NADH-Coupled ATPase Assay for Screening Small Molecule Inhibitors', *Journal of Visualized Experiments*, (150). doi: 10.3791/60017.
  150. Ramsay, E. P. and Vannini, A. (2018) 'Structural rearrangements of the RNA polymerase III machinery during tRNA transcription initiation', *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*. Elsevier, 1861(4), pp. 285–294. doi: 10.1016/j.bbagr.2017.11.005.
  151. Rao, S. S. P. et al. (2014) 'A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping', *Cell*. Elsevier Inc., 159(7), pp. 1665–1680. doi: 10.1016/j.cell.2014.11.021.
  152. Rao, S. S. P. et al. (2017) 'Cohesin Loss Eliminates All Loop Domains', *Cell*. Elsevier Inc., 171(2), pp. 305–320.e24. doi: 10.1016/j.cell.2017.09.026.
  153. Ren, G. et al. (2017) 'CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression', *Molecular Cell*. Elsevier Inc., 67(6), pp. 1049–1058.e6. doi: 10.1016/j.molcel.2017.08.026.
  154. Ren, R. et al. (2019) 'Structural basis of specific DNA binding by the transcription factor ZBTB24', *Nucleic Acids Research*, 47(16), pp. 8388–8398. doi: 10.1093/nar/gkz557.
  155. Rodríguez-Paredes, M. et al. (2009) 'The chromatin remodeling factor CHD8 interacts with elongating RNA polymerase II and controls expression of the cyclin E2 gene', *Nucleic Acids Research*, 37(8), pp. 2449–2460. doi: 10.1093/nar/gkp101.
  156. Rosenthal, P. B. and Henderson, R. (2003) 'Optimal determination of particle

- orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy', *Journal of Molecular Biology*, 333(4), pp. 721–745. doi: 10.1016/j.jmb.2003.07.013.
157. Rowley, M. J. and Corces, V. G. (2018) 'Organizational principles of 3D genome architecture', *Nature Reviews Genetics*. Springer US, 13, p. 1. doi: 10.1038/s41576-018-0060-8.
  158. Rubio, E. D. et al. (2008) 'CTCF physically links cohesin to chromatin.', *Proceedings of the National Academy of Sciences of the United States of America*, 105(24), pp. 8309–14. doi: 10.1073/pnas.0801273105.
  159. Ruigrok, V. J. B. et al. (2012) 'Characterization of aptamer-protein complexes by x-ray crystallography and alternative approaches', *International Journal of Molecular Sciences*, 13(8), pp. 10537–10552. doi: 10.3390/ijms130810537.
  160. Rupp, B. (2009) *Biomolecular Crystallography*. 1st edn. Garland Science.
  161. Russo, C. J. and Passmore, L. A. (2014) 'Ultrastable gold substrates for electron cryomicroscopy Europe PMC Funders Group', *Science*, 346(6215), pp. 1377–1380. doi: 10.1126/science.1259530.
  162. Ryan, D. P. and Owen-hughes, T. (2014) 'Europe PMC Funders Group Snf2-family proteins : chromatin remodelers for any occasion', 15(5), pp. 649–656. doi: 10.1016/j.cbpa.2011.07.022.Snf2-family.
  163. Sabantsev, A. et al. (2019) 'Direct observation of coordinated DNA movements on the nucleosome during chromatin remodelling', *Nature Communications*. Springer US, 10(1), pp. 1–12. doi: 10.1038/s41467-019-09657-1.
  164. Saha, A., Wittmeyer, J. and Cairns, B. R. (2006) 'Chromatin remodelling: the industrial revolution of DNA around histones', *Nature Reviews Molecular Cell Biology*, 7(6), pp. 437–447. doi: 10.1038/nrm1945.
  165. Saldaña-Meyer, R. et al. (2019) 'RNA Interactions Are Essential for CTCF-Mediated Genome Organization', *Molecular Cell*, 76(3), pp. 412-422.e5. doi: 10.1016/j.molcel.2019.08.015.
  166. Sanborn, A. L. et al. (2015) 'Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 112(47), pp. E6456–E6465. doi: 10.1073/PNAS.1518552112.
  167. Sanyal, A. et al. (2012) 'The long-range interaction landscape of gene promoters ENCODE Encyclopedia of DNA Elements', *Nature*, 489. doi: 10.1038/nature11279.
  168. Schaub, M. et al. (1997) 'Staf, a promiscuous activator for enhanced transcription by RNA polymerases II and III', *EMBO Journal*, 16(1), pp. 173–181. doi: 10.1093/emboj/16.1.173.
  169. Schaub, M., Krol, A. and Carbon, P. (1999) 'Flexible Zinc Finger Requirement for Binding of the Transcriptional Activator Staf to U6 Small Nuclear RNA and tRNAsSec Promoters', *Journal of Biological Chemistry*, 274(34), pp. 24241–24249. doi: 10.1074/jbc.274.34.24241.
  170. Schaub, M., Krol, A. and Carbon, P. (2000) 'Structural organization of Staf – DNA complexes', 28(10), pp. 2114–2121.
  171. Scheres, S. H. w. (2014) 'Beam-induced motion correction for sub-megadalton cryo-EM particles', *eLife*, 3, p. e03665. doi: 10.7554/eLife.03665.
  172. Scheres, S. H. W. (2012a) 'A bayesian view on cryo-EM structure determination', *Journal of Molecular Biology*. Elsevier Ltd, 415(2), pp. 406–418. doi: 10.1016/j.jmb.2011.11.010.
  173. Scheres, S. H. W. (2012b) 'RELION: Implementation of a Bayesian approach to cryo-EM structure determination', *Journal of Structural Biology*. Elsevier Inc., 180(3),

- pp. 519–530. doi: 10.1016/j.jsb.2012.09.006.
174. Schoenfelder, S. and Fraser, P. (2019) ‘Long-range enhancer–promoter contacts in gene expression control’, *Nature Reviews Genetics*, 20(8), pp. 437–455. doi: 10.1038/s41576-019-0128-0.
175. Schuster, C. *et al.* (1995) ‘Staf, a novel zinc finger protein that activates the RNA polymerase III promoter of the selenocysteine tRNA gene’, *Embo J*, 14(15), pp. 3777–3787. Available at: [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list\\_uids=7641696](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=7641696).
176. Schuster, C., Krol, A. and Carbon, P. (1998) ‘Two distinct domains in Staf to selectively activate small nuclear RNA-type and mRNA promoters’, *Molecular and cellular biology*, 18(5), pp. 2650–2658. Available at: [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list\\_uids=9566884%5Cnhttp://mcb.asm.org/cgi/content/full/18/5/2650?view=long&pmid=9566884](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=9566884%5Cnhttp://mcb.asm.org/cgi/content/full/18/5/2650?view=long&pmid=9566884).
177. Sigworth, F. J. (2015) ‘Principles of cryo-EM single-particle image processing’. doi: 10.1093/jmicro/dfv370.
178. Silva, A. P. G. *et al.* (2016) ‘The N-terminal region of chromodomain helicase DNA-binding protein 4 (CHD4) is essential for activity and contains a high mobility group (HMG) box-like-domain that can bind poly(ADP-ribose)’, *Journal of Biological Chemistry*. © 2016 ASBMB. Currently published by Elsevier Inc; originally published by American Society for Biochemistry and Molecular Biology., 291(2), pp. 924–938. doi: 10.1074/jbc.M115.683227.
179. Smyth, M. S. and Martin, J. H. J. (2000) *Review x Ray crystallography, J Clin Pathol: Mol Pathol*.
180. Soltermann, F. *et al.* (2020) ‘Quantifying Protein–Protein Interactions by Molecular Counting with Mass Photometry’, *Angewandte Chemie - International Edition*, 59(27), pp. 10774–10779. doi: 10.1002/anie.202001578.
181. Song, S.-H. and Kim, T.-Y. (2017) ‘CTCF, Cohesin, and Chromatin in Human Cancer’, *Genomics & Informatics*, 15(4), pp. 114–122. doi: 10.5808/GI.2017.15.4.114.
182. Sonn-Segev, A. *et al.* (2020) ‘Quantifying the heterogeneity of macromolecular machines by mass photometry’, *Nature Communications*. Springer US, 11(1), pp. 1–10. doi: 10.1038/s41467-020-15642-w.
183. Splinter, E. *et al.* (2006) ‘CTCF mediates long-range chromatin looping and local histone modification in the β-globin locus’, *Genes and Development*, 20(17), pp. 2349–2354. doi: 10.1101/gad.399506.
184. van Steensel, B. and Belmont, A. S. (2017) ‘Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression’, *Cell*. Elsevier Inc., 169(5), pp. 780–791. doi: 10.1016/j.cell.2017.04.022.
185. Strahl, B. D. and David Allis, C. (2000) *The language of covalent histone modifications*, *NATURE*. Available at: [www.nature.com](http://www.nature.com).
186. Sugathan, A. *et al.* (2014) ‘CHD8 regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors’, *Proceedings of the National Academy of Sciences of the United States of America*, 111(42), pp. E4468–E4477. doi: 10.1073/pnas.1405266111.
187. Tallant, C. *et al.* (2015) ‘Molecular basis of histone tail recognition by human TIP5 PHD finger and bromodomain of the chromatin remodeling complex NoRC’, *Structure*. The Authors, 23(1), pp. 80–92. doi: 10.1016/j.str.2014.10.017.
188. Tang, Z. *et al.* (2015) ‘CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription’, *Cell*. Elsevier Inc., 163(7), pp. 1611–1627. doi:

- 10.1016/j.cell.2015.11.024.
189. Thompson, B. A., Lin, G. and Bochar, D. A. (2008) 'CHD8 Is an ATP-Dependent Chromatin Remodeling Factor That Regulates  $\beta$ -Catenin Target Genes', 28(12), pp. 3894–3904. doi: 10.1128/MCB.00322-08.
  190. Tinoco, I. and Gonzalez, R. L. (2011) 'Biological mechanisms, one molecule at a time', *Genes & Development*, 25(12), pp. 1205–1231. doi: 10.1101/gad.2050011.
  191. Tsunaka, Y. et al. (2005) 'Alteration of the nucleosomal DNA path in the crystal structure of a human nucleosome core particle', *Nucleic Acids Research*, 33(10), pp. 3424–3434. doi: 10.1093/nar/gki663.
  192. Tyagi, M. et al. (2016) 'Chromatin remodelers: We are the drivers!!', *Nucleus*. Taylor & Francis, 7(4), pp. 388–404. doi: 10.1080/19491034.2016.1211217.
  193. Venter, B. J. and Pugh, B. F. (2009) 'How eukaryotic genes are transcribed', *Critical Reviews in Biochemistry and Molecular Biology*, 44(2–3), pp. 117–141. doi: 10.1080/10409230902858785.
  194. Vietri Rudan, M. et al. (2015) 'Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture', *Cell Reports*. The Authors, 10(8), pp. 1297–1309. doi: 10.1016/j.celrep.2015.02.004.
  195. Vietri Rudan, M. and Hadjur, S. (2015) 'Genetic Tailors: CTCF and Cohesin Shape the Genome During Evolution', *Trends in Genetics*. Elsevier Ltd, 31(11), pp. 651–660. doi: 10.1016/j.tig.2015.09.004.
  196. Wagner, T. et al. (2019) 'SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM', *Communications Biology*. Springer US, 2(1), pp. 1–13. doi: 10.1038/s42003-019-0437-z.
  197. Wang, D. C. et al. (2018) 'A tour of 3D genome with a focus on CTCF', *Seminars in Cell and Developmental Biology*. Elsevier Ltd. doi: 10.1016/j.semcdb.2018.07.020.
  198. Wang, H. et al. (2012) 'Widespread plasticity in CTCF occupancy linked to DNA methylation', *Genome Research*, 22(9), pp. 1680–1688. doi: 10.1101/gr.136101.111.
  199. Wen, Z. et al. (2018) 'ZNF143 is a regulator of chromatin loop', *Cell Biology and Toxicology*. Cell Biology and Toxicology, (2014), pp. 1–8. doi: 10.1007/s10565-018-9443-z.
  200. Wiechens, N. et al. (2016) 'The Chromatin Remodelling Enzymes SNF2H and SNF2L Position Nucleosomes adjacent to CTCF and Other Transcription Factors', *PLoS Genetics*, 12(3), pp. 1–25. doi: 10.1371/journal.pgen.1005940.
  201. Winger, J. et al. (2018) 'A twist defect mechanism for ATP-dependent translocation of nucleosomal DNA'. doi: 10.7554/eLife.34100.001.
  202. de Wit, E. et al. (2015) 'CTCF Binding Polarity Determines Chromatin Looping', *Molecular Cell*. Elsevier Inc., 60(4), pp. 676–684. doi: 10.1016/j.molcel.2015.09.023.
  203. Xie, D. et al. (2013) 'Dynamic trans-Acting Factor Colocalization in Human Cells', *Cell*. Elsevier Inc., 155(3), pp. 713–724. doi: 10.1016/j.cell.2013.09.043.
  204. Xie, X. et al. (2007) 'Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites', *Proceedings of the National Academy of Sciences of the United States of America*, 104(17), pp. 7145–7150. doi: 10.1073/pnas.0701811104.
  205. Yao, H. et al. (2010) 'Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA', *Genes & Development*, 24(22), pp. 2543–2555. doi: 10.1101/gad.1967810.
  206. Ye, B.-Y. et al. (2016) 'ZNF143 is involved in CTCF-mediated chromatin interactions by cooperation with cohesin and other partners', *Molecular Biology*. doi: 10.1134/S0026893316030031.

207. Ye, B. *et al.* (2020) ‘ZNF143 in Chromatin Looping and Gene Regulation’, *Frontiers in Genetics*, 11(April), pp. 1–10. doi: 10.3389/fgene.2020.00338.
208. Yin, J. *et al.* (2005) *Genetically encoded short peptide tag for versatile protein labeling by Sfp phosphopantetheinyl transferase*. Available at: [www.pnas.org/cgi/doi/10.1073/pnas.0507705102](http://www.pnas.org/cgi/doi/10.1073/pnas.0507705102).
209. Yin, J. *et al.* (2006) ‘Site-specific protein labeling by Sfp phosphopantetheinyl transferase’, *Nature Protocols*, 1(1), pp. 280–285. doi: 10.1038/nprot.2006.43.
210. Yodh, J. (2013) ‘ATP-dependent chromatin remodeling’, *Advances in Experimental Medicine and Biology*, 767, pp. 263–295. doi: 10.1007/978-1-4614-5037-5\_13.
211. Yuan, C.-C. *et al.* (2007a) ‘CHD8 Associates with Human Staf and Contributes to Efficient U6 RNA Polymerase III Transcription’, *Molecular and Cellular Biology*, 27(24), pp. 8729–8738. doi: 10.1128/MCB.00846-07.
212. Yuan, C.-C. *et al.* (2007b) ‘CHD8 Associates with Human Staf and Contributes to Efficient U6 RNA Polymerase III Transcription’, *Molecular and Cellular Biology*, 27(24), pp. 8729–8738. doi: 10.1128/MCB.00846-07.
213. Yusufzai, T. M. *et al.* (2004) ‘CTCF Tethers an Insulator to Subnuclear Sites, Suggesting Shared Insulator Mechanisms across Species’, *Molecular Cell*, 13(2), pp. 291–298. doi: 10.1016/S1097-2765(04)00029-2.
214. Zhang, H., Roberts, D. N. and Cairns, B. R. (2005) ‘Genome-wide dynamics of Htz1, a histone H2A variant that poises repressed/basal promoters for activation through histone loss’, *Cell*, 123(2), pp. 219–231. doi: 10.1016/j.cell.2005.08.036.
215. Zheng, S. Q. *et al.* (2017) ‘MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy’. doi: 10.1038/nmeth.4193.
216. Zhong, Y. *et al.* (2020) ‘CHD4 slides nucleosomes by decoupling entry- and exit-side DNA translocation’, *Nature Communications*. Springer US, 11(1), pp. 1–14. doi: 10.1038/s41467-020-15183-2.
217. Zhou, Q. *et al.* (2021) ‘ZNF143 mediates CTCF-bound promoter–enhancer loops required for murine hematopoietic stem and progenitor cell function’, *Nature Communications*. Springer US, 12(1). doi: 10.1038/s41467-020-20282-1.
218. Zivanov, J. *et al.* (2018) ‘New tools for automated high-resolution cryo-EM structure determination in RELION-3’, *eLife*, 7, pp. 1–22. doi: 10.7554/eLife.42166.
219. Zuin, J. *et al.* (2014) ‘Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells’, *Proceedings of the National Academy of Sciences*, 111(3), pp. 996–1001. doi: 10.1073/pnas.1317788111.