Characterisation of the potential lymphoma biomarker MORC4

Naomi Lauren Ann Nabi

A thesis submitted to the University of Huddersfield in partial fulfilment of the requirements for the degree of Doctor of Philosophy

This degree is awarded by the University of Huddersfield

August 2020

Copyright statement

i. The author of this thesis (including any appendices and/ or schedules to this thesis) owns any copyright in it (the "Copyright") and s/he has given The University of Huddersfield the right to use such Copyright for any administrative, promotional, educational and/or teaching purposes.

ii. Copies of this thesis, either in full or in extracts, may be made only in accordance with the regulations of the University Library. Details of these regulations may be obtained from the Librarian. Details of these regulations may be obtained from the Librarian. This page must form part of any such copies made.

iii. The ownership of any patents, designs, trademarks and any and all other intellectual property rights except for the Copyright (the "Intellectual Property Rights") and any reproductions of copyright works, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property Rights and Reproductions cannot and must not be made available for use without permission of the owner(s) of the relevant Intellectual Property Rights and/or Reproductions.

“When someone tells me "no," it doesn't mean I can't do it, it simply means I can't do it with them.”

— Karen E. Quinones Miller

Acknowledgments

Firstly, I want to express my sincere gratitude to my PhD supervisor, Dr. Christopher Cooper. He has not only provided academic support throughout my PhD but also emotional support throughout and for that I am forever grateful. He taught me how to be a more confident and higher quality scientist over the four years that I worked in his lab. I am also exceeding grateful for his exceptional expertise and guidance, which he provided on a regular basis and for encouraging me to take care of my mental health.

I would also like to thank Dr. Richard Bingham for providing his assistance with my protein crystal mounting and data collection, especially during the final year of my lab work where we screened my crystals on a regular basis. In addition, I thank Chris' collaborators at the University of Oxford for producing and providing me with MORC4 monoclonal antibodies.

'Girl Street' was also a vital support network during my PhD. Thank you to my wonderful group of friends including Zoe, Sophie, Carla, Kim, Hollie, Megan and Courtney who willingly accepted me into their friendship group when I joined the University of Huddersfield in 2016. Without their emotional support throughout my PhD I would not have been able to make it through until the end. Thank you for the consistent support with several cups of tea and for all the brilliant nights out we experienced together to get us through the good and bad times. My family have also played a significant part in my education since day one. Especially my one-of-a-kind father who has provided emotional and financial support not only throughout my education but also my life so far. I would not be who I am today without you.

Last but not least I want to thank my Fiancé, John. He has been there to support me since my GCSE exams in High School. Thank you for consistently telling me 'you can do it' when I lack self-confidence, for allowing me to use your PC when combining my thesis chapters together and for providing tech support. Thank you for loving me endlessly and for supporting my decision to move away from home without you for three years in order to do what I loved most. I am forever grateful for your support over the last 11 years and I could not ask for a better husband-to-be.

Abstract

The microrchidia (MORC) protein family is a set of proteins comprising a conserved GHKL-type ATPase domain and a Zf-CW domain, which is associated with methylated histone binding in chromatin. This suggests an involvement in chromatin remodelling. There are 4 members of the MORC protein family in humans, MORC1, MORC2, MORC3 and MORC4. MORC2 and MORC3 are relatively well structurally and biochemically characterised with some cell biological characterisation. However, MORC4 is poorly characterised, although some evidence suggests that MORC4 could be a potential diffuse large B-cell lymphoma (DLBCL) biomarker. MORC4 mRNA is overexpressed in activated B-cell (ABC)-DLBCL cells compared with normal B cells, suggesting that MORC4 is an ABC-DLBCL biomarker. This study aims to take a broad focus to the characterisation of MORC4, both from structural and biochemical perspectives. This project also involves the investigation of RNA and protein expression analysis in several cancer cell types, in an attempt to gain an insight into the structure and function of MORC4 and further potential as a cancer biomarker.

This study involved recombinant protein design, expression and purification to produce high levels of pure MORC4 protein. This involved screening of multiple single and tandem domain MORC4 protein fragments in *Escherichia coli*, which were then purified to milligram quantities and homogeneity. Purified ATPase-Zf-CW tandem domain MORC4$^{29\text{-}480}$ protein was used for structural and biophysical analysis methods such protein crystallisation, analytical size exclusion (SEC) and SEC-multiangle laser light scattering (SEC-MALLS). Biochemical and mutagenesis approaches such as chemical crosslinking and site-directed mutagenesis (SDM), respectively, indicated that MORC4 forms ATP (AMP-PNP)-induced dimers. MORC2 and MORC3 structural analysis also indicated dimerisation, which was supported by homology modelling approaches. To investigate MORC4 protein expression in a number of cancer cell lines, novel monoclonal antibody reagent conditions were optimised in western blotting. High levels of MORC4 protein expression was identified in colorectal, breast and pancreatic cancer cell lines. This high level of MORC4 expression was supported by RNA analysis methods such as quantitative reverse transcription-polymerase chain reaction (qRT-PCR) and RT-PCR isoform analysis.

Following extensive protein crystallisation trials, optimisation and X-ray diffraction screening, the high yielding and pure MORC4$^{29\text{-}480}$ ATPase-Zf-CW tandem domain protein produced crystals diffracting to 20 Å. However, a higher resolution dataset was required for structural resolution, hence, homology modelling was instead used to predict the structure of MORC4 using MORC2 and MORC3 as templates. Homology modelling and sequence alignments suggested that the ATPase domain of MORC4 was structurally homologous to MORC3. However, the C-terminus was less conserved in comparison to other MORC proteins. This suggests that it is more difficult to predict the C-terminal structure of MORC4 and that perhaps there may be a different function between MORC paralogues. RNA and protein expression analysis identified that MORC4 RNA and protein expression was relatively high in colorectal, breast and pancreatic cancer cell lines. This indicates that MORC4 could play a role in these specific types of cancers. However, further future cellular analysis is required to investigate MORC4 function such as siRNA/CRISPR approaches to assess MORC4 knockout on apoptosis and survival. Future work is also required to determine MORC4 as an ATPase and chromatin remodeller such as ATPase assays, isothermal titration calorimetry (ITC) and chromatin association assays. Overall, this study gives preliminary insights into the structure and function of MORC4.

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Definition	Abbreviation	Definition
ABC	Activated B-cell	ILC	Infiltrating lobular carcinoma
ACLY	ATP citrate lyase	INO80	Inositol required 80
ALK	Anaplastic lymphoma kinase	IPTG	Isopropyl β-D-1-thiogalactopyranoside
AML	Acute myeloid leukaemia	IR	Ionising radiation
AMP-PNP	Adenylyl-imidodiphosphate	ISWI	Imitation switch
ARHs	ADP-ribose-protein hydrolases	JARID1	Jumonji AT-rich interactive domain-1
Arps	Actin related proteins	JDM	Juvenile dermatomyositis
ARTs	ADP-ribosyltransferases	KLD	Kinase-ligase-*Dpn*I
BAF	BRM associated factors	KM	Kaplan-Meier
BAMS	Bosma arhinia microphthalmia syndrome	LB	Luria-Bertani broth
BCA	Bicinchoninic acid assay	LCIS	Lobular carcinoma in situ
BCR	B-cell antigen receptor	LIC	Ligation independent cloning
BL	Burkitt's lymphoma	LSD1	Lysine specific demethylase 1
BM	Bone marrow	MCL	Mantle cell lymphoma
B-NHL	B-cell non-Hodgkin lymphoma	MEM-E	Minimum Essential Medium Eagle
BRCT	BRCA1 C Terminus	MGMT	O6-Methylguanine-DNA methyltransferase
BRG1	Brahma-related gene 1	Mnase	Micrococcal nuclease
BRM	Brahma	MORC	Microrchidia
BS3	Bis(sulfosuccinimidyl)suberate	MOZ	Monocytic leukemic zinc finger
BSA	Bovine serum albumin	MRC	Medical research council
CBP	CREB-binding protein	MS	Mass spectrometry
CCHS	Congenital central hypoventilation syndrome	MUSCLE	Multiple sequence comparison by Log-expectation
CDD	Conserved Domains Database	MYST	MOZ, Ybf2-Sas3, Sas2 and Tip60
CDK2	Cyclin-dependent kinase 2	MZL	Marginal zone lymphoma
CDK2AP1	Cyclin-dependent kinase 2 associated protein 1	NF	Nuclear factor
CDKs	Cyclin-dependent kinases	Ni-IMAC	Nickel-immobilised metal ion affinity chromatography
CHARGE	Coloboma, heart disease, atresia of choanae, retardation, genital anomalies, and ear anomalies	Ni-NTA	Nickel-nitrilotriacetic acid
CHD	Chromodomain-helicase-DNA binding	NLPHL	Nodular lymphocyte-predominant Hodgkin lymphoma
CHD1	Chromodomain-helicase-DNA binding 1	NuRD	Nucleosome-remodelling and deacetylase complex

Abbreviation	Definition	Abbreviation	Definition
CHRAC	chromatin accessibility complex	O-GlcNAc	β-linked N-acetylglucosamine
CML	Chronic myeloid leukaemia	ORC	Origin recognition complex
CMT	Charcot-Marie-Tooth disease	ORF	Open reading frame
CoA	Coenzyme A	PAK1	p21-activated kinase 1
Ct	Cycle threshold	PARGs	Poly (ADP-ribose) glycohydrolases
CV	Column volume	PARP1	Poly (ADP-ribose) polymerase 1
DCIS	Ductal carcinoma in situ	PBS	Phosphate buffer saline
dCTP	Deoxycytidine triphosphate	PCAF	P300/CBP-associated factor
DEPC	Diethyl pyrocarbonate	PCNA	Proliferating cell nuclear antigen
dGTP	Deoxyguanosine triphosphate	PCR	Polymerase chain reaction
DHSs	DNase I hypersensitive sites	PEG	Polyethylene glycol
DLBCL	Diffuse large B-cell lymphomas	Phyre2	Protein homology recognition engine 2
DMEM	Dulbecco's modified eagle medium	Pi	Inorganic phosphate
dNTPs	Deoxynucleotide	pI	Isoelectric point
DOC-1	Deleted in oral cancer	PI3K	Phosphoinositide 3-kinase
DSB	Double strand break	PMBL	Primary mediastinal B-cell lymphoma
dsDNA	Double stranded DNA	PML NBs	Promyelocytic leukaemia nuclear bodies
DTT	Dithiothreitol	PMSF	Phenylmethylsulfonyl fluoride
ECL	Enhanced chemiluminescence	PMTs	Post-translational histone modifications
EHMT1	Euchromatic histone-lysine N-methyltransferase 1	PR	Progesterone receptor
ELS	Early life stress	PRD	Proline-rich domain
ER	Oestrogen receptor	PSIPRED	Position specific iterative blast based secondary structure prediction
EZH1	Enhancer of zeste homolog 1	PSN	Peripheral nervous system
FAP	Familial adenomatous polyposis	PVDF	Polyvinylidene difluoride
FBS	Fetal bovine serum	qRT-PCR	Quantitative reverse transcription-polymerase chain reaction
FL	Follicular lymphoma	RPMI	Roswell Park Memorial Institute medium
FPLC	Fast protein liquid chromatography	RQ	Relative quantification
FSHD2	Facioscapulohumeral muscular dystrophy type 2	RTK	Receptor tyrosine kinase
FT	Flow through	RT-PCR	Reverse transcription-polymerase chain reaction
GC	Germinal center	RTs	Rhabdoid tumours
GCB	Germinal center B-cell	SDM	Site-directed mutagenesis

Abbreviation	Definition	Abbreviation	Definition
GENT2	Gene Expression Database of Normal and Tumour Tissues 2	SDS	Sodium dodecyl sulphate
GEPIA	Gene Expression Profiling Interactive Analysis	SDS-PAGE	Sodium dodecyl sulphate-polyacrylamide gel electrophoresis
GHKL	DNA gyrase, hsp90, histidine kinase and MutL	SEC	Size exclusion chromatography
GNAT	Gcn5-related N-acetyltransferase	SEC-MALLS	Size exclusion chromatography-Multiangle light scattering
GST	Glutathione S-transferase	Sir2	Sirtuin
HAS	helicase-SANT-associated domain	SMART	Simple Modular Architecture Research Tool
HATPase	Histidine kinase-like ATPase	SMC	Structural maintenance of chromosome
HATs	Histone acetyltransferases	SMCHD1	Structural maintenance of chromosomes flexible hinge domain containing 1
HCMV	Human cytomegalovirus	SNS	Sympathetic nervous system
HDACs	Histone deacetyltransferases	SOB	Super optimal broth
HDMs	Histone demethyltransferases	SSBs	Single strand breaks
HER2	Human epidermal growth factor receptor 2	STR	Short tandem repeat
hINO80	Human INO80	SUMO-2	Small ubiquitin-related modifier 2
HIV	Human immunodeficiency virus	SWI/SNF	The switch/sucrose non-fermenting
HKG	Housekeeper gene	TB	Terrific broth
HMT	Histone methyltransferase	TBE	Tris-borate-EDTA
HP1	Heterochromatin protein 1	TBS	Tris-buffered saline
HRP	Horseradish peroxidase	TBS-T	Tris-buffered saline-tween
HSCs	Hematopoietic stem cells	TCA	Tricarboxylic acid
HSV-1	Herpes simplex virus 1	TEV	Tobacco Etch Virus
HUSH	Human silencing hub	TFB1	Transformation buffer 1
IBD	Inflammatory bowel disease	TFB2	Transformation buffer 2
ICP0	Infected cell 0	TGS	Tris-glycine-SDS
Ies2	Ino Eighty Subunit 2	TNBC	Triple-negative breast cancer
Ies6	Ino Eighty Subunit 6	TORC1	Target of rapamycin complex 1
IEX	Ion exchange	TrxT	Thioredoxin
IgH	Immunoglobulin heavy chain	UHPLC	Ultra-high-pressure liquid chromatography
IgL	Immunoglobulin light chain	WR	Working reagent
IgM	Immunoglobulin M	WT	Wild type
ILC	Infiltrating lobular carcinoma	Zf-CW	Zinc finger-CW

CHAPTER 1

Introduction

Chapter 1 Introduction

1.1 An introduction to this study

The microrchidia (MORC) protein family is a set of proteins which have been found to have different functions. Although, one of the main functions of both MORC2 (Xie *et al.*, 2019) and MORC3 (Zhang *et al.*, 2019) is the role in chromatin remodelling and DNA repair. This chapter will give an overview of chromatin remodelling mechanisms and complexes before giving a detailed introduction to the MORC protein family. Recent evidence has also suggested that MORC4 plays a role in a specific type of lymphoma, diffuse large B-cell lymphoma (Liggins *et al.*, 2007). Therefore, this chapter will also provide an introduction to cancer and a discussion of the potential role of MORC4 in some cancer types.

1.2 An introduction to chromatin remodelling

1.2.1 DNA packaging

The nucleus in eukaryotic cells contains tightly packaged DNA, which is condensed into a nucleoprotein complex known as chromatin (van Holde and Zlatanova, 2007). The basic packaging component of chromatin is a nucleosome, which is made up of 147 base pairs of DNA that are wrapped around a histone octamer (Quina *et al.*, 2006). Each nucleosome is made up of two copies of each histone including H2A, H2B, H3 and H4 (Mariño-Ramírez *et al.*, 2005) (Figure 1.1). The following level of DNA condensation after nucleosomes is the chromatin fibre, which forms loops that are 30 nm in length. The final order of organisation is where the chromatin fibers are condensed into chromosomes (Maeshima *et al.*, 2010) (Figure 1.2).

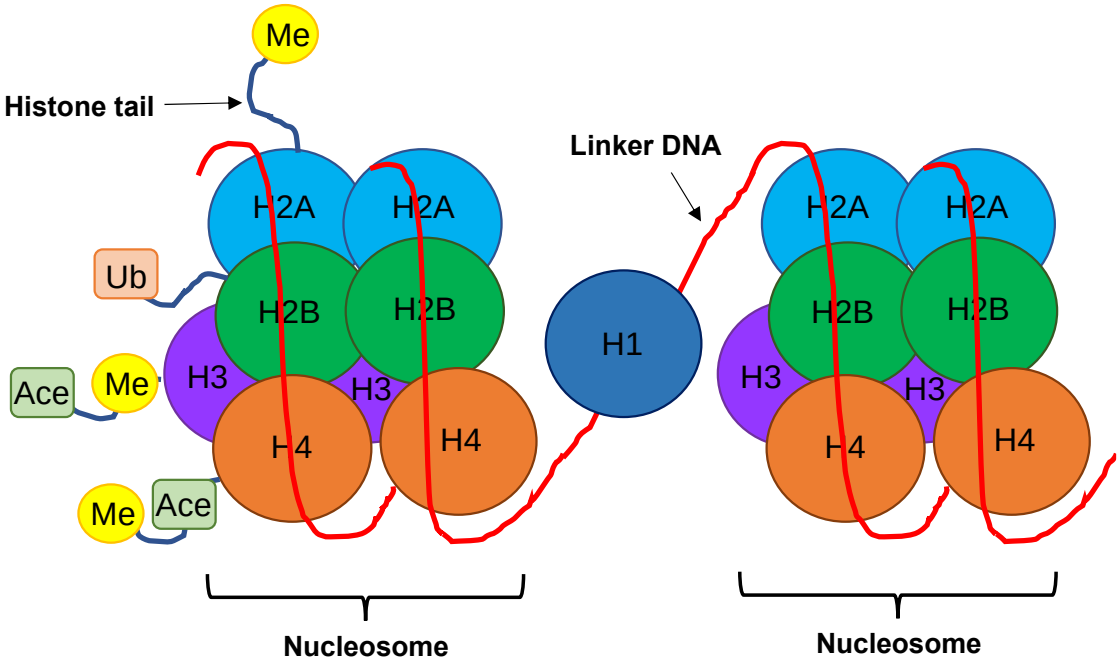


Figure 1.1: **Schematic representation of the structure of nucleosomes and some major post-translational histone modifications (PTMs).** Histone tails are only represented on one copy of histones. Yellow circles represent methylated (Me) histone tails, orange rectangles represent ubiquitination (Ub) and green rectangles represent acetylation (Ace). Figure created by author and adapted from (Dey, 2006; Starkman *et al.*, 2012; Marsh *et al.*, 2014).

There are two current structural models of the 30 nm chromatin fibres, the solenoid model (Figure 1.2) and the zig-zag model. In the solenoid model, which is also known as a one-start helix, the chromatin fiber is where nucleosomes surround one central region with approximately six nucleosomes per turn (Razin and Gavrilov, 2014). This means that each nucleosome can interact with its fifth and sixth nearby nucleosome (Luger *et al.*, 2012). The zig-zag model, which is also known as the two-start helix, proposes that each nucleosome interacts with the second nearby nucleosome (Dorigo *et al.*, 2004).

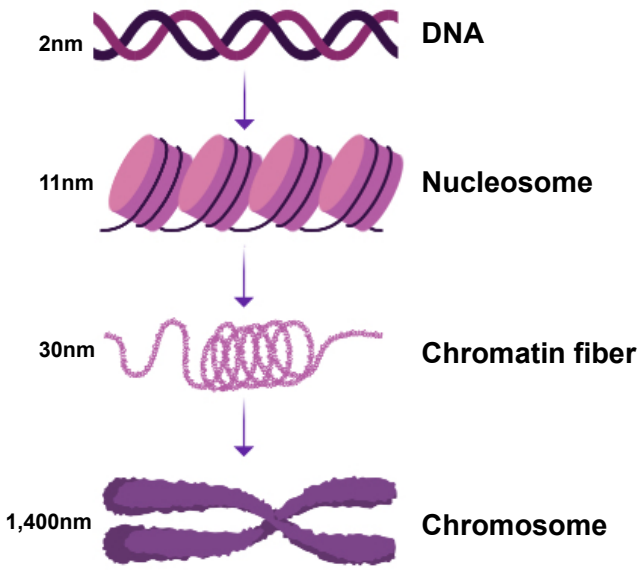


Figure 1.2: **Solenoid model of DNA condensation.** A DNA molecule is wrapped around a nucleosome which is made up of a histone octamer with a diameter of 11nm. The nucleosome is folded into chromatin fibers before subsequently being further folded into a chromosome with a dimeter of 1,400nm. Figure adapted from (Maeshima *et al.*, 2010).

In addition to packaging of DNA, chromatin is also important for other processes such as obstructing transcription when DNA is condensed into nucleosomes (Petesch and Lis, 2012) and modifications of histone tails, which is important for the regulation gene expression (Dong and Weng, 2013). Both of these functions demonstrate the importance of chromatin and it is also apparent that the chromatin structure is indeed important for the majority of DNA-processes such as transcription, replication and repair (Li *et al.*, 2007).

1.2.2 Chromatin remodelling

Chromatin remodellers are essential for all DNA-dependent pathways (Rippe *et al.*, 2007) as they are important for DNA recombination, DNA replication, DNA repair (Strahl and Allis, 2000) and gene expression (Wu, 1997). There are two classes of chromatin remodeller complexes required to complete the remodelling pathway, histone post-translational modification complexes and alteration of histone DNA through ATP hydrolysis, also known

as known as ATP-dependent chromatin remodeller complexes (Zhang *et al.*, 2016; Liu *et al.*, 2012).

1.2.2.1 Functions of chromatin remodellers

1.2.2.1.1 DNA repair

Chromatin remodelling plays an important role in DNA repair mechanisms. Chromatin remodeller complexes "loosen" the chromatin structure after detection of DNA damage has taken place, provide access to the site of DNA damage so that the chromatin remodeller complexes can reach the lesion and finally repair of the damage (Price, 2013). Mutations in chromatin remodelling complexes and associated proteins is likely to lead to lethal cancers (Chen *et al.*, 2016).

Acetylation of histone H4 creates the "loosening" of the chromatin structure so that DNA repair can progress (Price, 2013). The N-terminal of histone H4 interacts with adjacent nucleosomes, histone dimers H2A-H2B (Zhang *et al.*, 2017). There is some evidence to suggest that an increase in acetylation at histone H4 at double strand breaks (DSBs) can lead to upregulated chromatin "loosening" at the site of the DSB (Kruhlak *et al.*, 2006).

Poly (ADP-ribose) polymerase 1 (PARP1) is an important enzyme required during DNA repair and is able to aid in the repair of DNA DSBs and single strand breaks (SSBs) (Ko and Ren, 2012). Some studies indicate that PARP1 is involved in chromatin remodelling, which takes place after DNA damage and evidence also suggests that PARP1 is able to "loosen" the chromatin structure (Ji and Tulin, 2010). During this "loosening", PARP1 PARylates histones (Messner *et al.*, 2010), allowing chromatin remodellers to be recruited prior to DNA repair (Ray Chaudhuri and Nussenzweig, 2017). PARylation is a post-translational modification which is carried out by members of the PARP protein family. During PARylation, proteins are modified with the addition of ADP-ribose units (Verdone *et al.*, 2015). Chromodomain-

helicase-DNA-binding protein 1-like (CHD1L) is an ATPase remodeller (Ahel *et al.*, 2009) and a member of the switch/sucrose non-fermenting **(**SWI/SNF) family of chromatin remodellers that bind PARP1 (Tsuda *et al.*, 2017). CHD1L has also been found to be recruited to DNA lesions through PARP1 with the interaction with the CHD1L macro-domain (Jiang *et al.*, 2015).

1.2.2.1.2 Gene expression

Chromatin remodelling is also responsible for regulation of gene expression, which is performed by the alteration of the accessibility of chromatin to transcription factors (Morrison, 2020). This involves the release of DNA from the nucleosome and activation or suppression of gene expression, which is performed by either the histone modifications or ATP-dependent chromatin remodelling proteins (Zhang *et al.*, 2016). A specific type of ATP-dependent chromatin remodeller, inositol required 80 (INO80), regulates metabolic signalling of the target of rapamycin complex 1 (TORC1) pathway (Beckwith *et al.*, 2018). The TOR pathway is involved in the regulation of gene expression in response to nutrient availability (Zoncu *et al.*, 2011). For example, a lack of nutrients can influence carbon gene expression (Rohde and Cardenas, 2003), chromatin structure and expression of ribosomal genes (Tsang *et al.*, 2003).

1.2.2.1.3 DNA replication

Chromatin remodelling is also important for DNA replication, as this process requires significant rearrangement of chromatin, including nucleosomes and DNA loop domains (Demeret *et al.*, 2001). It is thought that condensed packing of DNA into chromatin can prevent DNA transcription and ultimately DNA replication (Groth *et al.*, 2007). In addition, chromatin is thought to increase the stabilisation of binding of the origin recognition complex (ORC) (Bellush and Whitehouse, 2017), which is responsible for the initiation of DNA replication (Li *et al.*, 2018). Furthermore, it is evident that mammalian cell DNA replication initiation sites have the same histone modifications found at sites of gene transcription (Lubelsky *et al.*, 2014).

Another study which investigated ORC-binding patterns, identified that the most common predicted chromatin feature associated with ORC-binding was DNase I hypersensitive sites (DHSs) (Miotto *et al.*, 2016), which are regions highly sensitive to DNase I (Lyu *et al.*, 2018). These DHSs contain chromatin histone modifications such as H3 acetylation and H3K4 methylation (Miotto *et al.*, 2016). Following successful DNA replication, DNA topoisomerase II is responsible for altering the structure of chromatin and reducing torsional stress caused by DNA replication (Roca, 2011) specifically, the unwinding of DNA (Demeret *et al.*, 2001). Another study also suggests that the chromatin accessibility complex (CHRAC) (Scacchetti *et al.*, 2018), is required for termination of DNA replication (Demeret *et al.*, 2001).

1.2.2.2 Deregulation of chromatin remodelling and diseases

Mutations affecting the chromatin remodelling processes may result in several diseases, including most human cancers (Koschmann *et al.*, 2017) and several syndromes including but not limited to, Rett syndrome and Kleefstra syndrome (Mirabella *et al.*, 2016). SMARCB1 is a member of the SWI/SNF family of chromatin remodellers in humans, which functions as a tumour suppressor. However, a biallelic inactivation mutation (a mutation that occurs in both alleles of a gene) in the *SMARCB1* gene can lead to Rhabdoid tumours (RTs) (Geller *et al.*, 2015). RTs are rare carcinomas that can develop anywhere in the body including the brain and lungs, however, they most commonly occur in the kidneys (Brennan *et al.*, 2013).

Deregulation of non-ATP dependent chromatin remodellers, specifically histone acetyltransferases (HATs), has a direct link to metastasis (Di Cerbo and Schneider, 2013). A study found that some mutations in CREB-binding protein (CBP), a transcription coactivator (Korzus *et al.*, 2004), lead to breast, colorectal and gastrointestinal cancers (Iyer *et al.*, 2004). Evidence has also suggested that abnormal expression of histone deacetyltransferases (HDACs) could also lead to cancer (Barneda-Zahonero and Parra, 2012).

Another disease that is associated with the deregulation of chromatin remodelling is CHARGE (coloboma, heart disease, atresia of choanae, retardation, genital anomalies, and ear anomalies) syndrome, which develops in 1 in every 10,000 people. CHARGE is caused by heterozygous frameshift and deletion mutations (Tyagi *et al.*, 2016) in the *CHD7* gene (Basson and van Ravenswaaij-Arts, 2015). CHD7 is part of the family of chromodomain-helicase-DNA binding (CHD) chromatin remodellers and is a transcriptional regulator in the nucleoplasm (Zentner *et al.*, 2010). CHARGE syndrome is a rare disorder that affects several areas of the human body. Patients with CHARGE can have a severe range of side effects including but not inclusive to, growth deficiency and spinal abnormalities (Pampal, 2010).

1.2.2.3 Post-translational histone modification (PMT) complexes

Histone methylation (Vidanes *et al.*, 2005) acetylation (Drazic *et al.*, 2016), ubiquitination (Xu and Jaffrey, 2013) and phosphorylation (Duan and Walther, 2015) are examples of post-translational histone modifications (PMTs). Each histone has an N-terminal tail that is important for histone modification complexes to remove or write one PMT (Liu *et al.*, 2012) or multiple PMTs, (Figure 1.1) which can either occur on the same histone tail (*cis*) or different histone tails (*trans*) (Musselman *et al.*, 2012; Rossetto *et al.*, 2012). These modified histones affect nucleosome: DNA interactions such as shielding of the positive charge of lysine on N-terminal tail, which helps to open up the nucleosome (Chang and Takada, 2016). This is important for initiating ATP-dependent chromatin remodelling (Bannister and Kouzarides, 2011).

1.2.2.3.1 Histone acetylation

Histone acetylation was the first PMT identified (Allfrey *et al.*, 1964) and has since then become one of the most characterised types of PMTs (Kurdistani and Grunstein, 2003). Acetylation of histones generally occurs on lysine residues and produces a neutral charge on histones, subsequently leading to an increase in accessibility of chromatin (Verdone *et al.*,

2006), allowing chromatin remodelling. In addition, histone lysines can also have an acetylation-deacetylation switch mechanism, which is regulated through the action of histone acetyltransferases (HATs) and histone deacetyltransferases (HDACs), which are involved in the addition of an acetyl group to a lysine residue on the N-terminus of a histone tail (Yang, 2004).

1.2.2.3.1.1 Histone acetyltransferases (HATs)

The Gcn5-related N-acetyltransferase (GNAT) and the MOZ, Ybf2-Sas3, Sas2 and Tip60 (MYST) (Avvakumov and Côté, 2007) are two main families of well characterised HATs (Wapenaar and Dekker, 2016), which are enzymes responsible for the addition of an acetyl group to histone tails (Wapenaar and Dekker, 2016). There are two known types of GNAT HATs in humans, GNC5 (Ud-Din *et al.*, 2016) and P300/CBP-associated factor (PCAF) (Suryanarayanan *et al.*, 2018). GNC5 in humans is one of the most characterised types of HATs (Xue-Franzén *et al.*, 2013) and is involved in the acetylation of histones and transcription factors (Nagy and Tora, 2007). GNC5 is involved in the acetylation of lysine residues specifically histones H3 and 2B (Imoberdorf *et al.*, 2006). PCAF is also involved in acetylation of lysine residues (Ogryzko *et al.*, 1998), specifically lysine 9 and 14 of histone H3 in addition to acetylation of transcriptional regulators (Love *et al.*, 2012) such as p53 (Xenaki *et al.*, 2008).

The highly conserved MYST family all contain a MYST domain consisting of an acetyl-CoA binding domain and a zinc finger domain (Avvakumov and Côté, 2007). Some members of this family also have an additional domain such as chromodomains (Utley and Côté, 2003). The monocytic leukemic zinc finger (MOZ) is a member of the MYST family, which acetylates histones H3, H4, H2A and H2B (Carlson and Glass, 2014) and is a fusion partner of CBP, specifically in a 8;12 translocation in acute myeloid leukaemia (Crowley *et al.*, 2005).

1.2.2.3.1.2 Histone deacetyltransferases (HDACs)

HDACs are responsible for the deacetylation of lysine residues in histones (Seto and Yoshida, 2014) and enabling of histone-DNA interactions, which subsequently leads to blocked access to DNA (Farria *et al.*, 2015). There are two families of HDACs, the histone deacetylase family and the Sirtuin (Sir2) family (Seto and Yoshida, 2014). HDACs are also separated into classes based on sequence identity. Class I and class II are part of the histone deacetylase family and class IV and class III are members of the Sir2 family (Seto and Yoshida, 2014; Haberland *et al.*, 2009).

1.2.2.3.1.3 Mechanism of the HAT and HDAC switch

Both HAT and HDAC are thought to act as a switch mechanism, which is able to shift between activation and repression of chromatin accessibility (Eberharter and Becker, 2002), allowing other mechanisms to take place such as DNA transcription, DNA repair (Gong and Miller, 2013), transcriptional regulation (Jacobson and Pillus, 2004) and chromatin remodelling (Yamada *et al.*, 2004). HATs are able to increase access to chromatin by impairing histone-DNA interactions (Bannister and Kouzarides, 2011) (Figure 1.3). Conversely, HDAC stimulates condensation of nucleosomal DNA, which can also involve methylation of lysine residues (Eberharter and Becker, 2002) (Figure 1.3). Methylated histones, which are produced by histone methyltransferase (HMT) (Stewart *et al.*, 2005) (Rice and Allis, 2001), recruit heterochromatin protein 1 (HP1), leading to silencing of heterochromatin regions (Bird, 2001) (Figure 1.3).

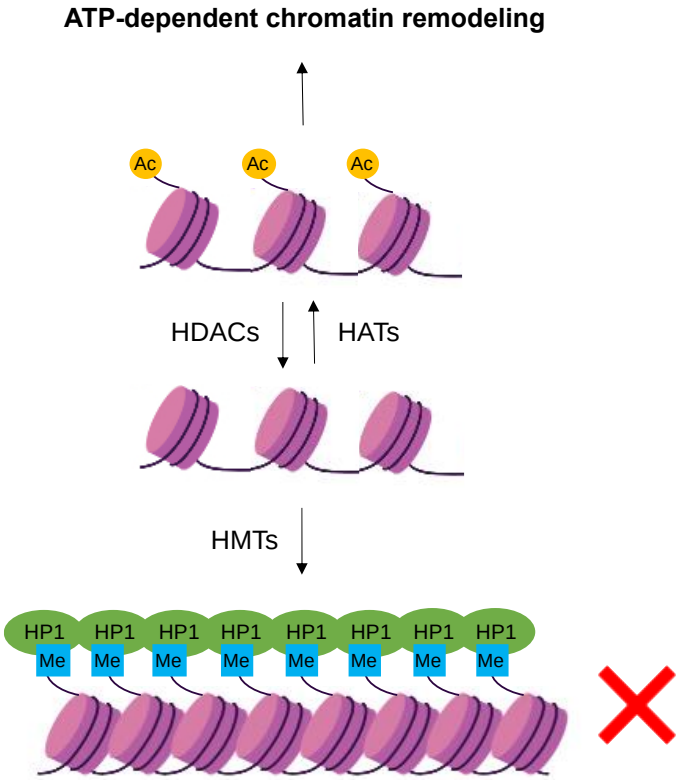


Figure 1.3: **Histone acetylation switch mechanism.** Histone acetyltransferases (HATs) increase access to chromatin with acetylation of histone tails (Acetylated histone tails are represented as yellow circles). Histone deacetyltransferases (HDACs) deacetylate histone tails leading to blocked access of DNA (highlighted by the red cross). Histone methyltransferase (HMT) catalyses the methylation of histone tails leading to blocked access to DNA (Methylated histones are represented as blue squares and heterochromatin protein 1 (HP1) is represented as green ovals). Figure created by author and adapted from (Eberharter and Becker, 2002).

1.2.2.3.2 Histone methylation

Histone methylation is a PMT, which is important for a number of cellular functions including chromatin remodelling (Geiman and Robertson, 2002) and gene expression (Greer and Shi, 2012). The most common histone methylation regions are on lysine and arginine residues and can involve either mono-, di-, or trimethylation of lysine or mono- or demethylation of arginine (Bannister *et al.*, 2002). Methylation involves the addition of a methyl group to the C5 position of the cytosine base (Geiman and Robertson, 2002), forming 5-methylcytosine (Moore *et al.*, 2013). Unlike histone acetylation, histone methylation does not disrupt the charge one histones, suggesting that methylation does not affect the histone-DNA structure, however, it interferes

with the interaction of chromatin binding-proteins and DNA (Zhang *et al.*, 2016). Histone methylation is regulated by two classes of enzymes, HMTs and histone demethyltransferases (HDMs) (Keppler and Archer, 2008). HMTs are a group of writers which catalyse the transfer of methyl group(s) (Kwon *et al.*, 2003). Each HMT is specific to histone residues e.g. enhancer of zeste homolog 1 (EZH1) is responsible for methylation of H3K27 (Shen *et al.*, 2008), however, euchromatic histone-lysine N-methyltransferase 1 (EHMT1) is responsible for methylation of H3K9 (Tachibana *et al.*, 2008). Conversely, HDMs are a group of erasers, which remove methyl group(s) (Demeret *et al.*, 2001). Similar to HMTs, each HDM is also specific to histone residues, e.g. lysine specific demethylase 1 (LSD1) is responsible for the removal of methyl groups from H3K4me1 or H3K4me2 (Hyun *et al.*, 2017). Whereas, jumonji AT-rich interactive domain-1 (JARID1) demethylates H3K9me1 and H3K9me2 (Yamane *et al.*, 2006). Readers are also important in this process as they are able to recognise and regulate the methylation of histones (Hyun *et al.*, 2017). An example of a reader is chromodomain helicase DNA binding protein 1 (CHD1), which is thought to be involved in the regulation of H3K4 methylation (Quan and Hartzog, 2010).

1.2.2.3.3 Histone phosphorylation

Histone phosphorylation is important for a number of cellular processes such as DNA damage repair (Rossetto *et al.*, 2012), transcriptional regulation (Banerjee and Chakravarti, 2011) and chromatin condensation (Sawicka and Seiser, 2014). Histone kinases are responsible for phosphorylation of serine, threonine and tyrosine residues (Rossetto *et al.*, 2012) and phosphatases are responsible for the removal of phosphate groups (Zhang *et al.*, 2016). Upon histone phosphorylation, the positive charge on the histone tail is reduced (Bannister and Kouzarides, 2011), consequently causing destabilisation of histone-DNA interactions (North *et al.*, 2014), leading to increased access to DNA and ultimately DNA repair and other processes (Fernandez-Capetillo *et al.*, 2004). The BRCA1 C Terminus (BRCT) reader protein

has been shown to recognise phosphorylation of S139 residue (Stucki *et al.*, 2005) in response to DNA damage repair (Jin *et al.*, 2005).

1.2.2.3.4 Rare histone modifications

ADP-ribosylation is one type of rare histone modification, which involves the use of ADP-ribosyltransferases (ARTs) to transfer one ADP-ribose from NAD^+ to several residues such as lysine, arginine or glutamate (Messner and Hottiger, 2011). Conversely, ADP-ribose-protein hydrolases (ARHs) (Rack *et al.*, 2018) or poly (ADP-ribose) glycohydrolases (PARGs) (Byrne *et al.*, 2016) are responsible for the removal of ADP-ribose. ADP-ribosylation of histones increases the negative charge (Laing *et al.*, 2011) on histones, resulting in "loosening" of the chromatin structure and consequently allowing DNA repair mechanisms (Beneke, 2012).

Sumoylation and ubiquitylation are similar rare histone modifications, which involve the attachment of small ubiquitin-like or ubiquitin proteins to lysine residues on histones (Zhang *et al.*, 2016). Histone ubiquitin ligases are responsible for the ubiquitylation of histones and deubiquitating enzymes are required for removal of ubiquitin proteins (Cao and Yan, 2012). It is thought that upon DNA DSBs, ubiquitylation is induced, causing recruitment of DNA repair proteins (Yu *et al.*, 2020). Sumoylation is similar to ubiquitylation, however, some evidence suggests that it is involved in transcriptional repression (Nathan *et al.*, 2003).

In addition, there are some histone modifications which have been more recently identified. Some histones are modified by β-linked N-acetylglucosamine (O-GlcNAc) sugars at serine and threonine residues (Sakabe *et al.*, 2010). O-GlcNAc transferase is responsible for the transfer of sugars to histones and O-GlcNAcase is required to remove sugars (Vocadlo, 2012). Furthermore, histone tail clipping is another rare histone modification which is the process of removing the first 21 residues in the histone N-terminal tail (Santos-rosa *et al.*, 2012), affecting processes such as transcriptional regulation (Yi and Kim, 2018).

1.2.2.4 ATP-dependent chromatin remodellers

ATP-dependent chromatin remodelling utilises energy from ATP hydrolysis (Varga-Weisz, 2001) to produce a motor motion to translocate DNA, which breaks DNA-histone interactions and consequently moves DNA along the histone surface resulting in chromatin remodelling (Clapier *et al.*, 2017; Vignali *et al.*, 2000). There are 4 families of ATP-dependent chromatin remodellers, the switch/sucrose-non-fermenting (SWI/SNF), the imitation switch (ISWI), the chromodomain-helicase-DNA binding (CHD) and the inositol requiring 80 (INO80) (Tyagi *et al.*, 2016) (Table 1.1).

Table 1.1: **ATP-dependent chromatin remodellers.** Examples of each type of chromatin remodeller including the family, and species.

Family	Chromatin remodeller	Species
The switch/sucrose non-fermenting (SWI/SNF)	BAF	Human
The imitation switch (ISWI)	Snf2	*Saccharomyces cerevisiae*
	WICH	Human
	NoRC	
	RSF	
	ACF	
	CHRAC	
	NURF	
	CERF	
The chromodomain-helicase-DNA binding (CHD)	CHD 1-9	Human
The inositol requiring 80 (INO80)	hIN08	Human

1.2.2.4.1 Domain architecture

The domain architecture of ATP-dependent chromatin remodellers consists of a main ATPase domain and several other specialised domains (Tyagi *et al.*, 2016; Längst and Manelyte, 2015). All ATP-dependent chromatin remodellers have a main ATPase domain, which includes both DExx and HELICc, that are separated by a linker and is responsible for ATP hydrolysis (Längst and Manelyte, 2015). DExx and HELICc are two RecA type folds (Hopfner *et al.*, 2012), which are related to helicases (Dürr *et al.*, 2006). Helicases are enzymes which utilise energy from ATP hydrolysis (Wiegand *et al.*, 2019) to translocate DNA (Nakano *et al.*, 2013), interrupt binding of complementary strands of DNA and unwind DNA structures (Datta and Brosh, 2018). An example of a well characterised chromatin remodeller is the *Saccharomyces cerevisiae* CHD1, which has been shown to interact with nucleosomes, leading to chromatin remodelling (Sundaramoorthy *et al.*, 2018) (Figure 1.4).

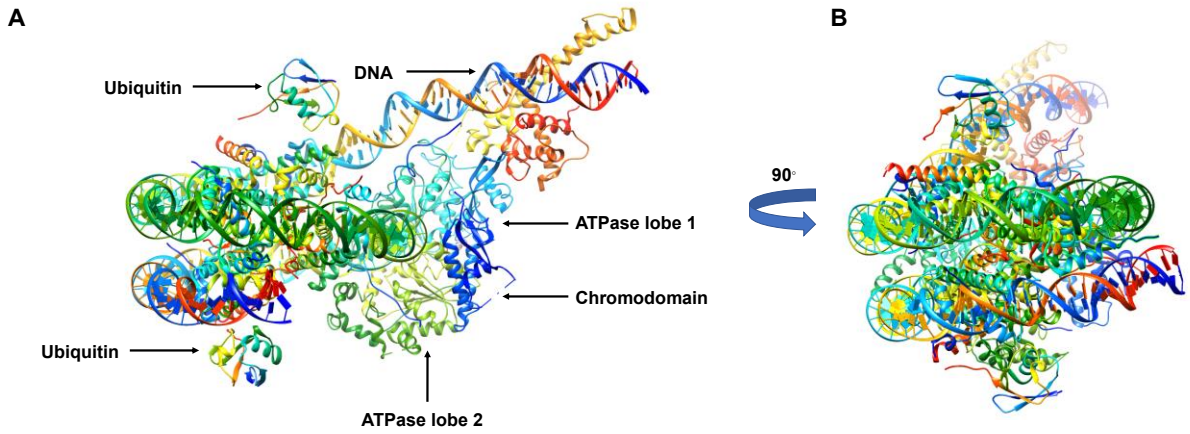


Figure 1.4: **Structure of the highly characterised *Saccharomyces cerevisiae* helicase, CHD1 bound to DNA and ubiquitin. (A)** ribbon representation of CHD1 showing ubiquitin, chromodomain, ATPase lobe 1 and 2 and DNA double helix. **(B)** Structure rotated at a 90° angle (PDB: 6FTX). Figure created by author and adapted from (Sundaramoorthy *et al.*, 2018).

SWI/SNF remodellers also contain C-terminal bromodomains, which are responsible for the binding of acetylated lysines (Tang *et al.*, 2010). The ISWI family of remodellers have C-terminal HAND, SANT and SLIDE domains, which are important for nucleosome spacing and

interaction (Bartholomew, 2014). The CHD family have two N-terminal chromatin organisation modifier (chromo) domains, which are crucial for mediating chromatin interactions by binding to methylated histones (Concetta and Imbalzano, 2007). Finally, both the SWI/SNF and INO80 chromatin remodellers have the N-terminal helicase-SANT-associated (HAS) domain, which is vital for the binding of actin and actin related proteins (Szerlong *et al.*, 2008) (figure 1.5).

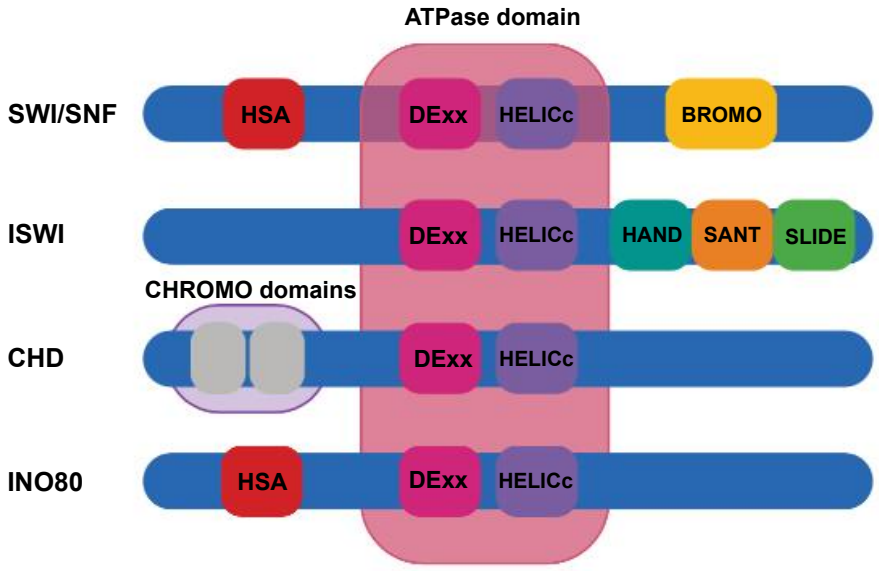


Figure 1.5: **Schematic representation of the domain architecture of the family of ATP-dependent chromatin remodellers**. The central DExx (pink) and HELICc (purple) (ATPase domain) is fully conserved in all ATP-dependent chromatin remodellers. HSA (red) is present in both SWI/SNF and INO80 remodellers. The bromodomain (yellow) is exclusive to the SWI/SNF remodellers along with the HAND, SANT, SLIDE domains (turquoise, orange, green, respectively), which is limited to ISWI. Two chromodomains are characteristic of the CHD remodellers. Figure adapted from (Tyagi *et al.*, 2016).

1.12.2.4.2 General mechanism of action

The chromatin remodelling complex binds to DNA on hydrolysis of ATP (Liu *et al.*, 2011), which is thought to occur through minor groove interactions (Euskirchen *et al.*, 2012). Upon ATP hydrolysis, the conformation of nucleosomes changes as a result of histone-DNA interactions, with some studies suggesting that translocation causes sliding of the DNA around

the octamer (Lorch *et al.*, 2010). This leads to "loosening" of the chromatin structure and consequently, chromatin remodelling (figure 1.6) (Vignali *et al.*, 2000). Chromatin remodelling may result in either sliding nucleosomes in *cis* (to a different position in the same DNA molecule) or by displacing nucleosomes in *trans* (to a different DNA segments) (Chandy *et al.*, 2006).

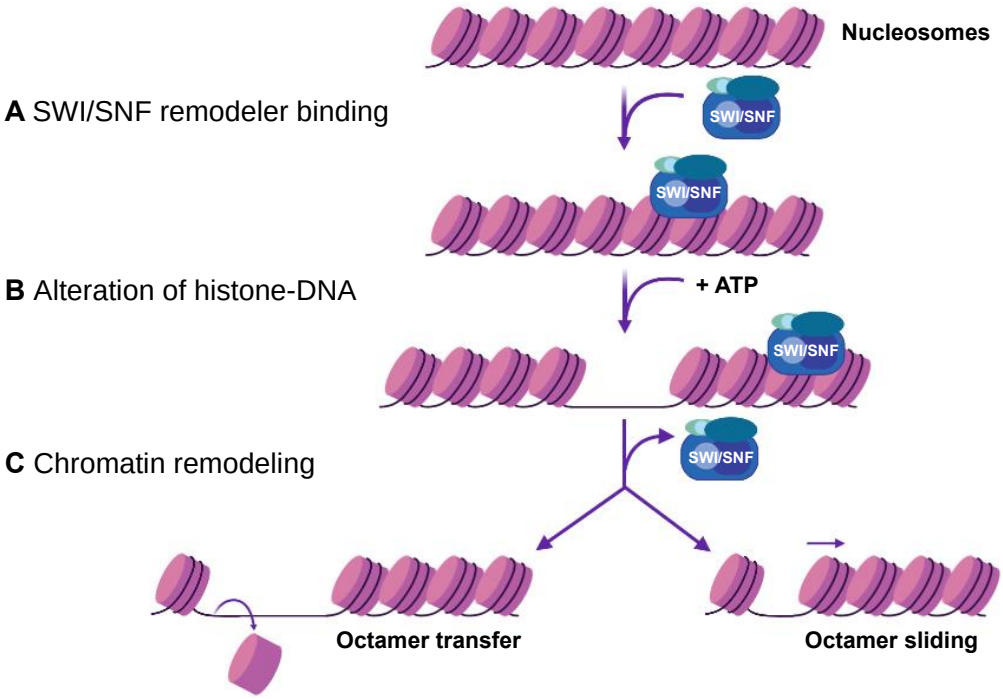


Figure 1.6: **Schematic representation of the mechanism of the ATP-dependent SWI/SNF chromatin remodeller.** The SWI/SNF chromatin remodeller binds to chromatin on the hydrolysis of ATP. **(B)** The release of ATP initiates the alteration of histone-DNA interactions and the "loosening" of the chromatin structure. **(C)** This alteration of histone-DNA results in the remodelling of chromatin and either octamer transfer or octamer sliding. Figure adapted from (Vignali *et al.*, 2000; Roberts and Orkin, 2004).

1.2.2.4.3 The switch/sucrose non-fermenting (SWI/SNF) chromatin remodeller

The switch/sucrose non-fermenting (SWI/SNF) was initially discovered in *Saccharomyces cerevisiae* (yeast) (Neigeborn and Carlson, 1984). The SWI/SNF subunits are approximately 2000 kDa, which includes Brahma-related gene 1 (BRG1)/ Brahma (BRM) associated factors (BAF) (Figure 1.7) (Sokpor *et al.*, 2017; Sena *et al.*, 2013), which is the human homologue of SWI/SNF (Alfert *et al.*, 2019) (figure 2.2) (Table 1.1). This chromatin remodelling complex

interacts with acetylated histone tails, which is regulated by the bromodomain at the C-terminal of the SWI/SNF complex (Marmorstein and Berger, 2001).

The SWI/SNF remodellers also have an N-terminal HSA domain, which is required for the binding of actin or actin related proteins (Arps) (Olave *et al.*, 2002). Some research suggests that actin/arps is required to regulate the catalytical and nucleosome targeting and the sliding mechanism of the SWI/SNF remodeller (Turegun *et al.*, 2018).

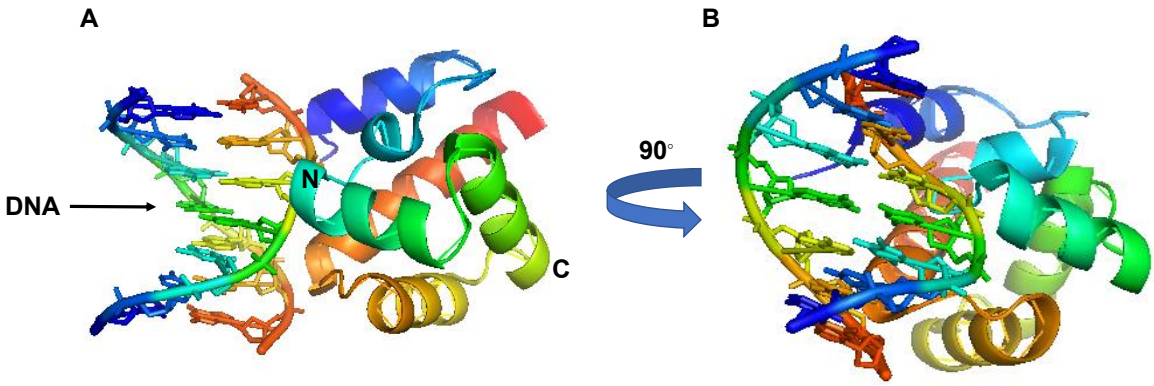


Figure 1.7: **Crystal structure of the mammalian ATP-dependent chromatin remodeller, BAF, bound to DNA. (A)** Ribbon representation of the BAF monomer showing the N-terminal HhH motif (light green), a pseudo HhH motif (blue) bound to DNA. **(B)** Structure has been turned at a 90° angle (PDB: 2BZF). Figure created by author and adapted from (Christina Marchetti Bradley *et al.*, 2005).

1.2.2.4.4 The imitation switch (ISWI) chromatin remodeller

The imitation switch (ISWI) chromatin remodeller was first discovered in *Drosophila melanogaster* in the search for the *Drosophila brahma (brm)*, which is highly analogous to the *Saccharomyces cerevisiae* chromatin remodeller, SNF2 (Elfring *et al.*, 1994). Unlike the SWI/SNF family, the ISWI family involves the use of a linker of DNA, which promotes binding to nucleosomes. Evidence also suggests that the longer the length of the DNA linker, the higher the affinity of the ISWI for the nucleosome, resulting in regulation of nucleosome sliding (Kagalwala *et al.*, 2004). A minimum of 20-23 base pairs of DNA is required for ISWI to bind to nucleosomes (Zofall *et al.*, 2004).

There are currently seven types of ISWI chromatin remodellers that are present in mammals including, WICH, NoRC, RSF, ACF, CHRAC, NURF and CERF (Table 1.1). The ISWI family all have a central ATPase domain (consisting of DExx and HELICc) (Aydin *et al.*, 2014) and a C-terminal HAND SANT SLIDE domain (Figure 1.4) (Bartholomew, 2014).

The nucleosome remodelling factor (NURF) is a chromatin remodelling complex, which comprises four subunits including NURF140 (the ATPase region), NURF55 (involved in histone metabolism), NURF38 (pyrophosphatase) and NURF301 (Badenhorst *et al.*, 2002; Alkhatib and Landry, 2011). NURF301 is vital subunit of NURF, which is important for the interaction of transcription factors, in particular, the GAGA transcription factor (Xiao *et al.*, 2001). GAGA is a multifunctional protein involved in chromatin remodelling and can adjust nucleosome locations which is needed for modification of the accessibility of promotors (Chopra *et al.*, 2008).

1.2.2.4.5 The chromodomain-helicase-DNA binding (CHD) chromatin remodeller

In mammals, there are nine distinct types of chromodomain-helicase-DNA binding (CHD) chromatin remodellers (CHD1-9), which are able to form complexes with other proteins (Hota and Bruneau, 2016) such as histone deacetylases 1 (HDAC1) (Lai and Wade, 2011). Each of the nine members of CHD contains two N-terminal chromodomains and a central ATPase domain (Murawska and Brehm, 2011).

One of the most highly characterised CHD chromatin remodellers is the nucleosome-remodelling and deacetylase (NuRD) complex (Chiu *et al.*, 2017) (Table 1.1). NuRD functions as a ATP-dependent chromatin remodeller and also has histone deacetylase enzyme activity (Basta and Rauchman, 2009). The NuRD complex consists of two histone deacetylase proteins, HDAC1 and HDAC2, which remove lysine-acetyl groups on histones and other proteins (Gonneaud *et al.*, 2019), such as p53 (Stojanovic *et al.*, 2017). NuRD also consists of either

CHD3 (Mi2-α) or CHD4 (Mi2-β), which are important in the DNA damage response (DDR) and have been shown to be recruited to areas of DNA that have been damaged with UV radiation (Smith *et al.*, 2018). Methyl-CpG-binding domain protein, MBD2 or MBD3, is another component of the NuRD complex. MBD2 is able to bind methylated DNA (Figure 1.8) (Torchy *et al.*, 2015), unlike MBD3 (Menafra and Stunnenberg, 2014). Metastasis-associated proteins, MTA1, MTA2 or MTA3 regulate binding to DNA and HDAC1 (Figure 1.8) (Kumar and Wang, 2016). Another component of the NuRD complex is the histone-binding proteins Rbbp46 and Rbbp48, which are histone chaperones (Figure 1.8) (Allen *et al.*, 2013). The nuclear zinc-finger protein Gata2a and Gata2b (p66α/β) has a coiled-coil interaction with MDB2 (Gnanapragasam *et al.*, 2011) but the function of p66α/β alone is not yet known.

Finally, deleted in oral cancer (DOC-1) is a vital part of the NuRD complex and is an interaction partner of cyclin-dependent kinase 2 (CDK2) (a fundamental regulator of the cell cycle) (Peng *et al.*, 2016), which explains its other name, cyclin-dependent kinase 2 associated protein 1 (CDK2AP1) (Mohd-Sarip *et al.*, 2017). Some of the individual structures of domains and mechanisms of the NuRD complex have been characterised (Figure 1.8), however the entire complex has not yet been characterised structurally (Torchy *et al.*, 2015).

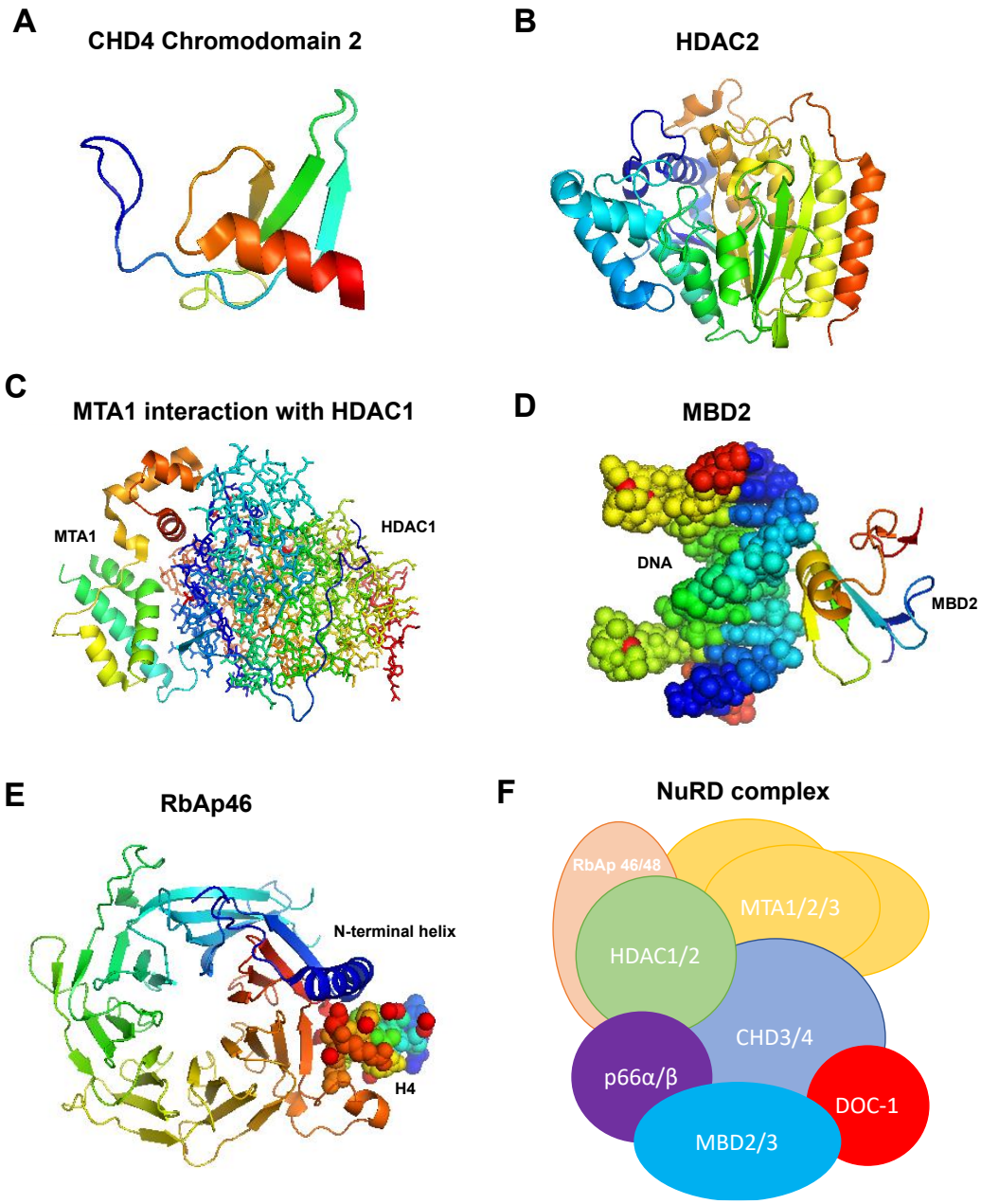


Figure 1.8: **Structures and a schematic representation of the human CHD chromatin remodelling complex, NuRD. (A)** NMR structure of a chromodomain from the CHD4 complex (PDB:2EE1). **(B)** Crystal structure of the HDAC2 domain (PDB:4LXZ). **(C)** Crystal structure of HDAC1 (cartoon representation) interacting with MTA1(stick representation) (PDB:4BKX). **(D)** NMR structure of MBD2 (cartoon representation) interacting with methylated DNA (sphere representation) (PDB:2KY8). **(E)** Crystal structure of RbAp46 (cartoon representation) showing the N-terminal helix and interaction with histone 4 (sphere representation) (PDB:3CFV). **(F)** Schematic representation of the NuRD complex. Figure created by author and adapted from (Torchy *et al.*, 2015).

1.2.2.4.6 The inositol requiring 80 (INO80) chromatin remodeller

The inositol required 80 (INO80) chromatin remodeller is a large multi-subunit complex chromatin remodeller, which was initially discovered in *Saccharomyces cerevisiae* and was purified by immunoprecipitation (Chen *et al.*, 2011). INO80 is vital for DNA repair and DNA replication (Hur *et al.*, 2010). Each complex contains a unique insertion in the middle of the ATPase domain, which causes the domain to split in two (Bao and Shen, 2009). The human INO80 (hINO80) complex (Table 1.1) consists of eight subunits (Mendiratta *et al.*, 2016). At the N-terminal of hINO80 are the actin, Arp4, Arp5 and Arp8 subunits (SC1) with some evidence to suggest that this SC1 region interacts with histones (Saravanan *et al.*, 2012). In addition to the SC1 subunit, there is the hINO80 ATPase and the AAA^+ ATPases, TBP-interaction proteins 49a and 49b (Tip49a and Tip49b) (Figure 1.9), with some evidence suggesting that they interact with nuclear factors and possess DNA helicase activity (Makino *et al.*, 2000). Ino Eighty Subunit 2 and 6 (Ies2 and Ies6) are two additional proteins that are present in the hINO80 complex. Ies2 is involved in the activation of enzyme activity of the hIN080 and Ies6 is involved in the binding of the hIN080 ATPase to nucleosomal DNA (Chen *et al.*, 2013). Evidence also suggests that Ies6 also binds to Arp5, which could be important for enzymatic pathways (Yao *et al.*, 2015) (Figure 1.9). However, both Ies2 and Ies6 are relatively uncharacterised (Yao *et al.*, 2015).

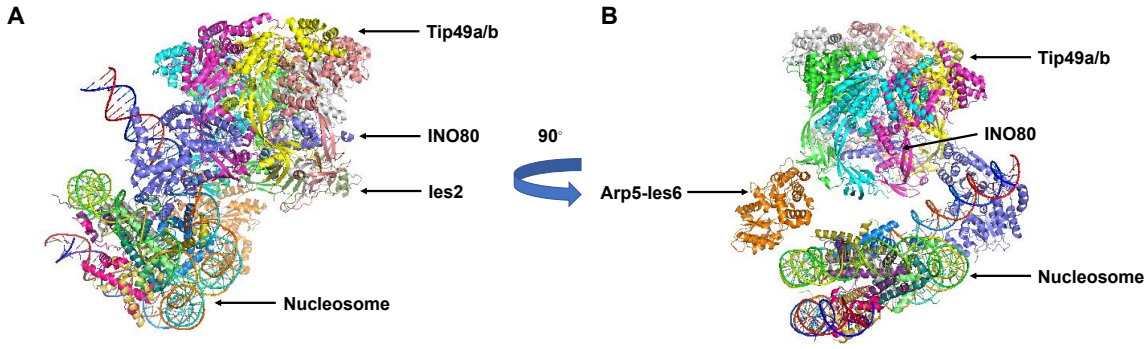


Figure 1.9: **Crystal structure of the human ATP-dependent chromatin remodeller, hINO80, bound to a nucleosome. (A)** ribbon representation of the hINO80 chromatin remodeller showing the Ies2 subunit (light green), INO80 ATPase subunit (purple), the Tip49a/b subunit (light pink & dark pink) and a nucleosome (rainbow coloured). **(B)** INO80 Structure has been turned at a 90° angle (PDB: 6HTS). Figure created by author and adapted from (Ayala *et al.*, 2018).

1.3 An introduction to the MORC family of proteins

Microrchidia (MORC) is a highly conserved family of proteins with 5 human members including MORC1, MORC2, MORC3, MORC4 and structural maintenance of chromosomes flexible hinge domain containing 1 (SMCHD1) (Li *et al.*, 2013). Each MORC has a different function and expression patterns (Table 1.2) ranging from MORC1 involvement in spermatogenesis (Pastor *et al.*, 2014) to MORC3, which is important for viral response mechanisms (Sloan *et al.*, 2016) and chromatin remodelling (Zhang *et al.*, 2019).

MORC2, MORC3 and SMCHD1 have been structurally and biochemically characterised (Li *et al.*, 2013; Pedersen *et al.*, 2019; Douse *et al.*, 2018). Conversely, only potential functions of MORC1 have been proposed (Thomas *et al.*, 2020; Pastor *et al.*, 2014), but MORC4 remains uncharacterised. Conservation of all MORC paralogues throughout mammals suggests similarities in terms of amino acid sequence and domain structure e.g. MORC3 and MORC4 (Figure 1.10). However, differences in sequence insertions and deletions suggest specific roles for each MORC subtype.

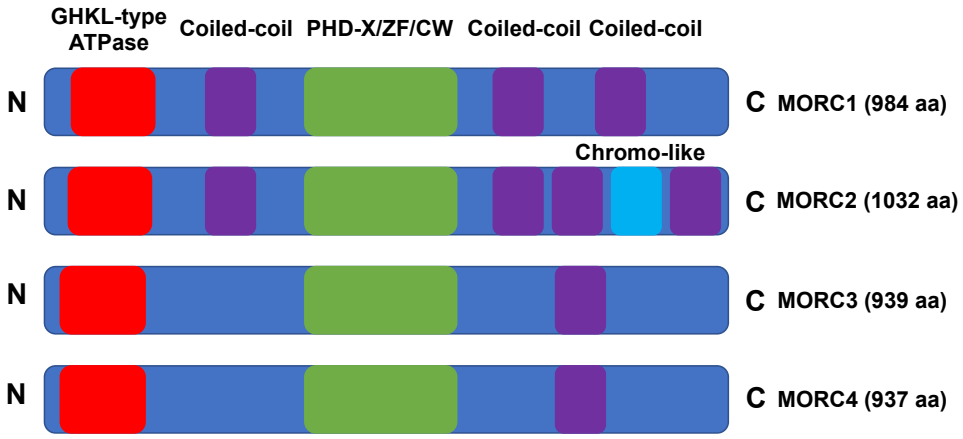


Figure 1.10: **A schematic representation of the domain architecture of the MORC family of proteins.** All MORCs have a GHKL-type ATPase domain (red) at the N-terminal, coiled-coil domains (purple) and a PHD-X/ZF-CW domain (green). MORC2 has an additional chromo-like domain (blue). Figure created by author and adapted from (Li *et al.*, 2013).

Genes encoding MORCs are also found in plants, for instance, *Arabidopsis thaliana*. There are 7 members of the MORC family in *Arabidopsis* (AtMORC1-7) (Koch *et al.*, 2017). Mutations in *Arabidopsis* AtMORC1 and AtMORC6 cause decondensation of heterochromatin and transcriptional faults, without having an effect on DNA methylation (Moissiard *et al.*, 2014) suggesting that MORCs are involved in regulating gene expression (Moissiard *et al.*, 2012). More recently, it has become apparent that that AtMORC6 interacts with AtMORC1, forming a heterodimer, leading to gene silencing (Moissiard *et al.*, 2014).

1.3.1 MORC domain architecture

All MORC members comprise a DNA gyrase, hsp90, histidine kinase and MutL (Iyer *et al.*, 2008) (GHKL)-type ATPase domain, which is one of several conserved domains in the MORC family of proteins (Hong *et al.*, 2017). The GHKL-type ATPase domain is characteristic of the MORC family at the N-terminus (Figure 1.10) (Li *et al.*, 2013), which is often combined to the C-terminus of the S5 domain, which forms an active ATPase domain when bound together (Dong *et al.*, 2018). This GHKL-type ATPase domain has a different architecture to the tandem RecA ATPase domain found in *Saccharomyces cerevisiae* CHD1 chromatin remodeller. When

active, the GHKL-type ATPase is important for DNA metabolism (Hong *et al.*, 2017), protein folding (Hu *et al.*, 2003) and signal transduction (Lorkovic, 2012).

Table 1.2: **Human MORC4 subfamily** including chromosome location, function of each MORC and expression patterns and/or mutations.

MORC	Chromosome location	Function	Expression pattern and/or mutations
MORC1	3q13.13	Spermatogenesis (Pastor *et al.*, 2014) Potential biomarker for depression (Thomas *et al.*, 2020).	Male germline cells (Pastor *et al.*, 2014), multiple myeloma (van Duin *et al.*, 2011) and breast cancer (Shah *et al.*, 2009).
MORC2	22q12.2	Lipogenesis, adipogenesis (Sanchez-Solana *et al.*, 2014) and the DNA damage response (Zhang and Li, 2019).	High expression in breast cancer and liver cancer (Pan *et al.*, 2018). Mutations can cause Charcot-Marie-Tooth disease (Douse *et al.*, 2018).
MORC3	21q22.12	Viral response mechanism (Sloan *et al.*, 2016), binding of histones & chromatin remodelling (Zhang *et al.*, 2019).	Leukocyte expression is altered after treatment with chemotherapy (Gonzalez-Fernandez *et al.*, 2012). Upregulated in Down Syndrome (H Andrews *et al.*, 2016)
MORC4	Xq22.3	Unknown	DLBCL and Hodgkin lymphoma expression (Liggins *et al.*, 2007)
SMCHD1	18p11.32	Tumour suppressor (Leong *et al.*, 2013) and chromosomal factor (Chen-Yu Wang *et al.*, 2019).	Mutations in SMCHD1 cause tumorigenesis in mice (Leong *et al.*, 2013).

Another domain essential for the MORC family of proteins, which is present at the C-terminus of MORCs, is the PHD-X/Zf-CW domain (Figure 1.10) (Perry and Zhao, 2003). The zinc finger CW (Zf-CW) domain consists of approximately 60 amino acids with 4 cysteine and 2 tryptophan residues (He *et al.*, 2010). This domain functions as a histone recognition domain, where it recognises methylated H3K4 (histone H3, lysine 4) (Sanchez and Zhou, 2011; Li *et al.*, 2019). This suggests that proteins with the Zf-CW domain are involved in epigenetic regulation (Hoppmann *et al.*, 2011).

The chromo-like domain (Figure 1.10) consists of approximately 50 amino acids and is characteristic of MORC2 (Figure 1.10) (Li *et al.*, 2013). The function of the chromo-domain in MORC2 has not yet been elucidated, however, it is thought that the domain could be involved in epigenetic regulation (Cavalli and Paro, 1998) and the binding of methylated histones (Bottomley, 2004), similarly to chromodomains which are found in the well characterised CHD chromatin remodeller (Concetta and Imbalzano, 2007).

In addition to chromo-like domain found in MORC2, all members of the MORC protein family also comprise interspersed coiled-coil domains (Figure 1.10), containing 2-5 α-helices wrapped around one another to form a supercoil (Mason and Arndt, 2004). There are several types of coiled-coils including an undistorted α-helix, a supercoiled α-helix, a left-handed coiled coil, parallel and anti-parallel coiled-coils. Each coiled-coil can have insertions of up to 6 amino acids, which produces discontinuities in the coiled-coil and potential structural deformities (Truebestein and Leonard, 2016). Coiled-coil domains can have several roles including the ability to initiate conformational changes (Carter *et al.*, 2016) recognition of DNA (Kim *et al.*, 2005) and they can help the formation of molecular spacers (regions of non-coding DNA), which can be important for helping to scaffold large protein complexes (Xi *et al.*, 2012).

MORC1 contains three coiled-coil regions and MORC2 contains four, which are dispersed throughout the proteins (Figure 1.10). However, MORC3 and MORC4 contain one coiled-coil domain near the C-terminus (Figure 1.10) (Manning and Cantley, 2003). The function of coiled-coil domains in MORCs has not yet been elucidated, however, some studies indicate that they could be involved in gene transcription (Watanabe *et al.*, 1998) and the DNA damage response (Hohl *et al.*, 2011; Truebestein and Leonard, 2016). MORCs are structurally analogous to other proteins involved in the DNA repair response and chromatin remodelling (Ryan and Owen-Hughes, 2011), for instance, the sucrose nonfermenting (SNF) 2 (Neigeborn and Carlson, 1984; Tang *et al.*, 2010). This family of proteins is analogous to MORCs in that it contains an ATPase domain and is involved in the DNA repair response (Zhihui Zhang *et al.*, 2018). Similar to SNF2 chromatin remodellers, MORC proteins also have a targeting domain such as a chromodomain or Zf-CW domain. In the case of CHD1L/ALC1, there is a macrodomain that binds to poly(ADP-ribose) (PAR) chains of poly(ADP-ribose) polymerase (PARP1), leading to ATPase-mediated chromatin remodelling (Gottschalk *et al.*, 2012; Ahel *et al.*, 2009).

1.3.2 MORC1

1.3.2.1 MORC1 function and role in disease

MORC1 is the ‘founding’ member of the MORC protein family and was initially discovered in mice. One study indicates that MORC1 knockouts results in infertility in male mice, however, female mice were not affected by this gene knockout (Watson *et al.*, 1998). This is where it was discovered that MORC1 plays a role in spermatogenesis (Table 1.2), which is where the name Microrchidia derives, Microrchidia being the medical term for abnormally small testes (Watson *et al.*, 1998). In addition to this, the *MORC1* gene is only expressed in meiotic and mitotic germ cells, which also provides further evidence that MORC1 is essential for spermatogenesis (Mundorf *et al.*, 2018).

Recent evidence has suggested a new role for MORC1, which is as a biomarker for early life stress (ELS) and depression (Thomas *et al.*, 2020) (Table 1.2). In addition to infertility in males, *MORC1* knockouts have more recently been associated with depression in mice (Schmidt *et al.*, 2016). One study suggested that ELS in mice can cause methylation of *MORC1* leading to depressive symptoms (Thomas *et al.*, 2020). MORC1 is also thought to be involved in gene silencing and compaction of chromatin, specifically in *Arabidopsis thaliana* (Moissiard *et al.*, 2012). Recent advances indicate that *Caenorhabditis elegans* MORC1 binds to DNA and compacts chromatin with the formation of DNA loops (Kim *et al.*, 2019).

1.3.3 MORC2

1.3.3.1 Structural and biochemical mechanisms

MORC2 has been shown to dimerise in the presence of adenylyl-imidodiphosphate (Figure 1.11a and b) (AMP-PNP) (Douse *et al.*, 2018), which is a non-hydrolysable analogue of ATP (Padrón and Huxley, 1984). In addition to AMP-PNP, Mg^{2+} is also present as ATP binds to the ATPase domain in the presence of Mg^{2+} (Figure 1.11b) (Kawamura *et al.*, 2001). Further evidence also suggests that the addition of Mg^{2+} to AMP-PNP creates a ternary complex, resulting in tighter binding of AMP-PNP to the ATPase (Robinson, 1980). When the MORC2 ATPase domain is mutated (N39A) the folding of the protein was not affected, however, AMP-PNP could not bind to MORC2, with no corresponding ATPase activity and the protein was in a monomeric state. This indicates that ATP binding is essential for dimerisation (Douse *et al.*, 2018).

The Zf-CW domain of MORC2 does not prevent binding of DNA when bound to the ATPase domain (Douse *et al.*, 2018), unlike MORC3 (Zhang *et al.*, 2019). There are several residues that are unconserved between MORC2 (Arg254, Arg266, and Thr496) and MORC3 (Glu184, Arg195, Lys216, Tyr217, Arg405, Arg444, and Asp454), which could contribute to the

difference in Zf-CW-ATPase binding of DNA (Douse *et al.*, 2018). In MORC2, this binding of the Zf-CW domain to the ATPase domain causes epigenetic silencing by the human silencing hub (HUSH) (Figure 1.11c) at histone H3 trimethylated at lysine 9 (H3K9me3), which is vital for the silencing of transgenes that have been introduced to the DNA (Tchasovnikarova *et al.*, 2015).

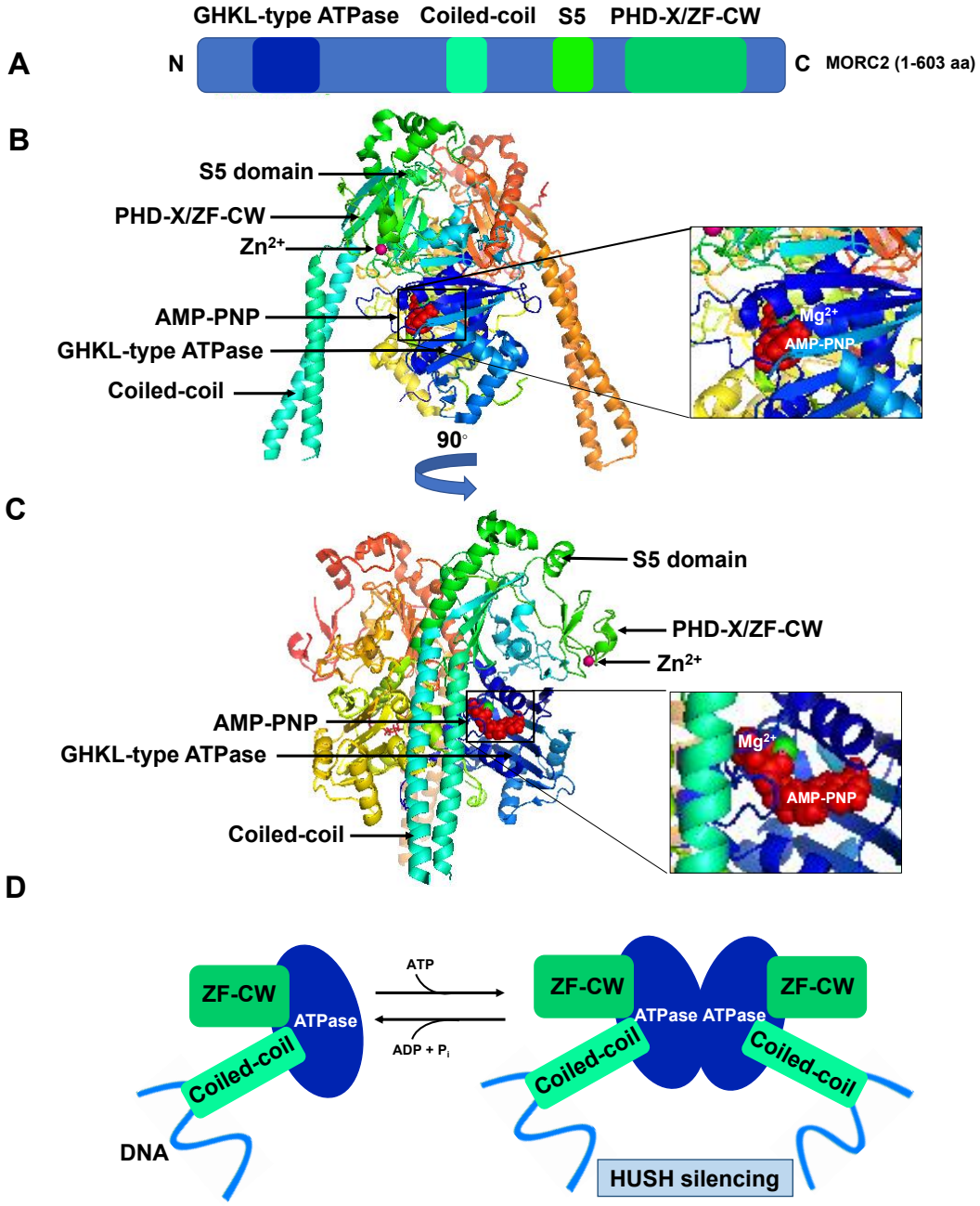


Figure 1.11: **A schematic representation and crystal structure of dimerised human MORC2$^{(1\text{-}603)}$ in the presence of AMP-PNP. (A)** Schematic representation of the domain architecture of MORC2$^{(1\text{-}603)}$. The colour of each domain corresponds to the structure below. **(B)** Ribbon representation of the structure of human dimerised MORC2. One promotor is coloured according to figure A and the other protomer is coloured in orange and yellow. GHKL-type ATPase (blue), the PHD-X/ZF-CW domain (dark green), the coiled-coil (light green), the S5 domain (bright green), Zn^{2+} (pink sphere), AMP-PNP (red spheres) and Mg^{2+} (green sphere). **(C)** MORC2$^{(1\text{-}603)}$ structure has been rotated at a 90° angle (PDB: 5OF9). **(D)** MORC2 HUSH silencing mechanism. Figure created by author and adapted from (Douse *et al.*, 2018).

1.3.3.2 MORC2 function and role in disease

MORC2 has been shown to respond to DNA damage by inhibiting gene expression and inducing chromatin remodelling (Sanchez-Solana *et al.*, 2014). MORC2 is also thought to be a substrate of p21-activated kinase 1 (PAK1) (Jagadeeshan *et al.*, 2016). PAK1 is a serine/threonine kinase, which has been shown to be upregulated in various human cancers (Kumar *et al.*, 2006; Vadlamudi and Kumar, 2003) and is involved in several cellular pathways that are hallmarks of cancer (Radu *et al.*, 2014). Upon DSB DNA damage, PAK1 phosphorylates MORC2 at Ser739, which then initiates chromatin relaxation leading to chromatin remodelling. This study also suggested that upon ionising radiation (IR) of mammalian cell lines, MORC2 was recruited to chromatin, which was confirmed by western blotting. In addition, this study also revealed that increased expression of MORC2 was associated with an increase in micrococcal nuclease (MNase), an enzyme required to digest linker areas between nucleosomes, which is also associated with a relaxed chromatin state (Ziv *et al.*, 2006). This increase in MNase in addition to MORC2 was further increased with IR, suggesting that MORC2 increases chromatin relaxation upon DNA damage, specifically, DSBs (D Q Li *et al.*, 2012).

The MORC2-PAK1 complex is important for maintaining genome integrity and the DNA damage response (D Q Li *et al.*, 2012). Evidence also shows that when the MORC2 ATPase domain was mutated (D68A), MORC2 could still dimerise, however, ATP could not bind or hydrolyse meaning that chromatin could not be relaxed and therefore, there was no chromatin remodelling and repair of DSBs (D Q Li *et al.*, 2012). The architectural similarity of MORC4 to MORC2, suggests MORC4 could play a similar role in the DNA damage response and could potentially be involved in chromatin remodelling, following some stimulus of DNA damage such as ultraviolet radiation or mutagenic chemical damage (Lans *et al.*, 2012).

MORC2 has also been found to function in lipogenesis and adipogenesis with some evidence suggesting that ATP citrate lyase (ACLY) binds to MORC2 (Sanchez-Solana *et al.,* 2014). ACLY is an enzyme involved in the conversion of coenzyme A (CoA) and citrate to produce acetyl-CoA and oxaloacetate, a vital step in fatty acid synthesis (Verschueren *et al.*, 2019) and the mevalonate pathway (cholesterogenesis) (Granchi, 2018). Some evidence suggests that MORC2 upregulates ACYL, which increases acetyl-CoA levels, ultimately increasing fatty acid biosynthesis. This indicates that the MORC2-ACLY interaction could be a novel therapeutic target for human diseases with upregulated lipogenesis (Sanchez-Solana *et al.*, 2014).

In addition to fatty acid biosynthesis and cholesterogenesis, acetyl-CoA is also important for the modification of proteins such as histone acetylation (Takahashi *et al.,* 2006) and gene expression (Martínez-Reyes and Chandel, 2018). ACLY upregulates acetylation of histones (Wellen *et al.,* 2009) and citrate produced in the tricarboxylic acid (TCA) cycle nuclear membranes, allowing histone modification (Deb *et al.*, 2017).

Several studies have shown that an arginine to tryptophan mutation at residue 252 (R252W) in the ATPase domain of MORC2 upregulates HUSH silencing in neuronal cells and therefore causes over-repression of HUSH genes in neuronal cells. This mutation causes axonal Charcot-Marie-Tooth disease (CMT) (Tchasovnikarova *et al.,* 2018), which is a genetic neuropathic muscular disorder, which causes weakness in the upper limbs and feet (Tácio Quadros Santos Monteiro Fonseca and Zanoteli, 2018).

Studies have also shown that MORC2 is highly expressed in breast cancer (Pan *et al.*, 2018). In particular, one study demonstrates how a point mutation of methionine to isoleucine at residue 276 (M276I) enhances development of triple-negative breast cancer (TNBC) and increases cell migration and invasion (Fang Lin Zhang *et al.*, 2018). TNBC is a type of breast

cancer which has little or no expression of the oestrogen receptor (ER), progesterone receptor (PR) or human epidermal growth factor receptor 2 (HER2), unlike other types of breast cancer (Bianchini *et al.*, 2016). Another study suggests that the proline-rich domain (PRD) (residues 601-734) is vital for MORC2 and the upregulation of breast cancer. PRD silencing also abolished breast cancer metastasis (Liao *et al.*, 2017).

1.3.4 MORC3

1.3.4.1 Structural and biochemical mechanisms

MORC3 is 939 residues in length (Figure 1.10) and perhaps the best characterised of the MORC protein family. Some studies suggest that the Zf-CW domain of MORC3 interacts with methylated lysine 4 of histone 3 (H3K4me) tails (Li *et al.*, 2012, 2016) . The histone H3 peptide binds to the Zf-CW domain of MORC3 with a third β-strand, which is bound to the β1 strand of MORC3 with the addition of hydrogen bonds for stabilisation. This binding of the Zf-CW to histone H3 suggests that MORC3 functions as a chromatin-associated ATPase and is therefore, involved in chromatin remodelling (H Andrews *et al.*, 2016). X-ray crystallography studies have shown that MORC3 dimerises in the presence of AMP-PNP, similar to MORC2 (Figure 1.12) (Li *et al.*, 2016).

Structural analyses show the Zf-CW domain of MORC3 also interacts with the ATPase domain (Zhang *et al.*, 2019). Upon ATP hydrolysis, binding of the MORC3 Zf-CW domain to histone H3 of the nucleosome exposes the ATPase domain from the ZF-CW domain, causing de-repression of the ATPase domain of MORC3 and increased MORC3 activity (Figure 1.12). It has also been revealed that the Zf-CW domain can obstruct DNA binding, leading to a decreased ATPase activity and an autoinhibited state (H Andrews *et al.*, 2016).

It is thought that when the ATPase domain is isolated from the CW domain, there is no ATPase activity, however, with the addition of DNA, there was a significant increase in ATPase

activity. This indicates that the activity of the ATPase domain of MORC3 is DNA dependent. Furthermore, the presence of the ATPase-CW domain with the addition of DNA did not increase ATPase activity. However, with the addition of both DNA and H3 peptide, the ATPase activity significantly increased. This suggests that the interactions of the ATPase-CW with DNA and the H3 tail are key for MORC3 activation (Zhang *et al.*, 2019). In addition, ATPase activity of MORC3 was also affected by the oligomeric status of the protein. An experiment suggested that dimerisation of MORC3 is required for ATPase activity. Mutated ATPase domain of MORC3 did not dimerise in the presence of AMP-PNP and did not show any ATPase activity with the addition of DNA and H3 peptide. This suggests that dimerisation is required for ATP hydrolysis of MORC3 (Zhang *et al.*, 2019).

MORC3 homodimerisation could play a role in this autoinhibition/activation mechanism as homodimerisation only occurs on ATP binding (Li *et al.*, 2016). ATP-dependent homodimerisation is characteristic of many GHKL-ATPases (Corbett and Berger, 2003; Corbett and Berger, 2005; Dutta and Inouye, 2000). This mechanism appears like a switch, with DNA binding and H3 exposing the active ATPase domain, stimulating the catalytic activity of MORC3. However, the autoinhibited state of MORC3 is formed when the CW domain is bound to the ATPase domain, abrogating binding of the ATPase domain to DNA (Zhang *et al.*, 2019). Similar autoinhibition mechanisms have been found in other chromatin remodellers including the well characterised CHD1 helicase, which utilises a chromodomains

to abrogate the ATPase domain, blocking binding to DNA (Hauk *et al.*, 2010). Overall, this data gives an insight into the structure and function of MORC3.

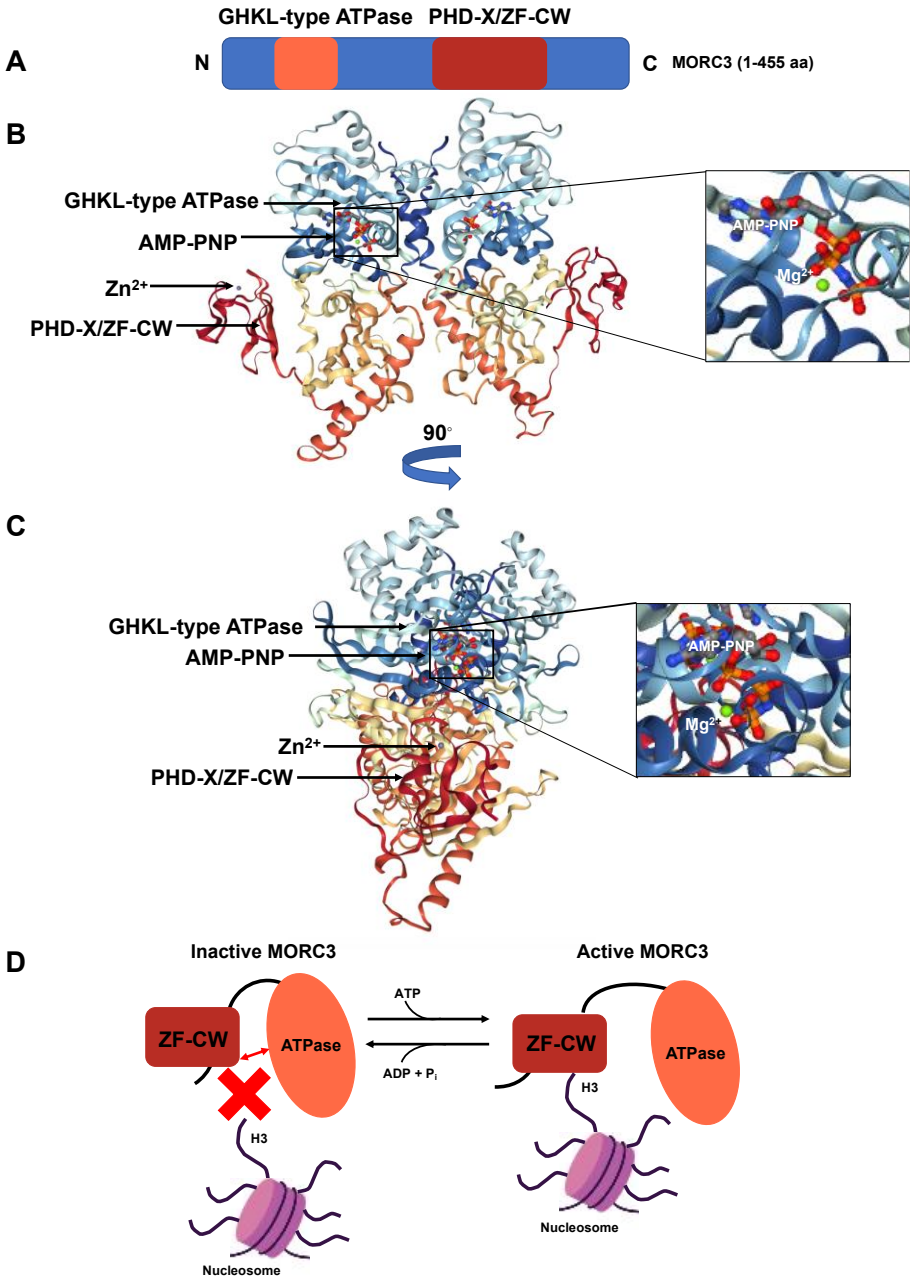


Figure 1.12: **A schematic representation and crystal structure of dimerised human MORC3$^{(1\text{-}455)}$ in the presence of AMP-PNP and activation of the MORC3-ATPase-CW cassette. (A)** Schematic representation of the domain architecture of MORC3$^{(1\text{-}455)}$. The colour of each domain corresponds to the structure below. **(B)** Ribbon representation of the structure of human dimerised MORC3. GHKL-type ATPase (orange), the PHD-X/ZF-CW domain (red), Zn^{2+}(grey sphere), AMP-PNP (ball & stick representation) and Mg^{2+} (green sphere). **(C)** MORC3$^{(1\text{-}455)}$ structure has been rotated at a 90° angle (PDB: 6O1E). **(D)** MORC3-ATPase-CW cassette mechanism showing autoinhibited MORC3 and active MORC3. Figure created by author and adapted from (Zhang *et al.*, 2019).

1.3.4.2 MORC3 function and role in disease

Leukocyte MORC3 expression is upregulated after treatment with chemotherapy reagents (Gonzalez-Fernandez *et al.,* 2012). MORC3 is also localised in promyelocytic leukaemia nuclear bodies (PML NBs), which is where it activates p53, causing arrest of cell proliferation (Mimura *et al.*, 2010). One study indicates that histone binding via the MORC3 CW domain is important for its role in PML NBs (H Andrews *et al.*, 2016). Some studies also indicate that MORC3 could be involved in transcription following its ability to promote gene silencing when bound to small ubiquitin-related modifier 2 (SUMO-2) (Rosendorff *et al.*, 2006).

MORC3 has also been found to be upregulated in humans with Down's syndrome, which could result from the *MORC3* presence in chromosome 21 (Table 1.2). Down's syndrome is a genetic disorder, which is caused by an extra copy of chromosome 21, therefore, it is often described as trisomy 21 (Diamandopoulos and Green, 2018). MORC3 is highly expressed in cells from people with Down's syndrome, which is highly conserved throughout different cell types (H Andrews *et al.*, 2016).

MORC3 is also a target of herpes simplex virus 1 (HSV-1) infected cell 0 (ICP0)-mediated degradation (Sloan *et al.*, 2015). HSV-1 is one of the world's most common viruses infecting humans, causing both recurrent oral and genital lesions (Kelly *et al.*, 2009). ICP0 is an intermediate protein that is produced by the HSV-1 in the early stage of the virus life cycle (Smith *et al.,* 2011) which is vital for infecting and manipulating DNA in human cells after infection of the virus (Smith *et al.*, 2011). MORC3 has antiviral activity following infection of HSV-1, however, this is prevented by ICP0. In addition to HSV-1, MORC3 also has antiviral activity to human cytomegalovirus (HCMV), a virus that affects the unborn foetus or a newly born child (Fletcher-Etherington *et al.,* 2019). Following infection with the HSV-1 virus, MORC3 is recruited to the nucleus following infection of HSV-1 of the host cell and then the RING finger domain of ICP0 is able to degrade MORC3, allowing ICP0 to infect cells (Sloan

et al., 2016). This evidence provides an insight into MORC3 function and its role in the immune response.

There is also evidence that MORC3 is associated with an increase in cancer incidence in individuals with juvenile dermatomyositis (JDM) (Gunawardena *et al.*, 2009; Hong *et al.*, 2017). JDM is an autoimmune disease which causes inflammation of the muscles and skin rashes in children (Martin *et al.*, 2012). Patients with JDM who have been diagnosed with cancer have antibodies to MORC3 (Fiorentino *et al.*, 2013). MORC3 is either mutated or deleted in several different types of cancer including bladder, stomach and diffuse large B-cell lymphomas (DLBCL) (H Andrews *et al.*, 2016). These findings suggest a role for MORC3 in JDM and cancer, however, MORC3 function at a molecular level has not yet been determined.

A point mutation (R420Q) is common in MORC3 Zf-CW domain and is often found in several types of diseases including prostate and stomach cancer (H Andrews *et al.*, 2016). One study found that the R420Q mutant decreased binding of the Zf-CW domain to H3K4me significantly, when compared to the wild type (WT). This indicates that when MORC3 is mutated, histone binding is affected, which can lead to cancer (H Andrews *et al.*, 2016).

1.3.5 MORC4

The structure, biochemistry or mechanism of action of MORC4 has not yet been characterised, however, a histone-tail binding assay suggests that the Zf-CW domain of MORC4 binds to H3K4me, in particular, H3K4me2 (Hoppmann *et al.*, 2011). This suggests that MORC4 could play a role in chromatin remodelling similarly to MORC3 and MORC2. This project leads to a greater understanding of the structure, function and biochemistry of MORC4, in particular to assess if it has a similar structural biochemistry to MORC2/MORC3 or plays a role in chromatin remodelling or genome integrity as with MORC2.

1.3.6 SMCHD1

Structural maintenance of chromosomes flexible hinge domain containing 1 (SMCHD1) has not been structurally characterised however, sequence analysis has determined an N-terminal GHKL-ATPase domain, similarly to other MORCs (Lemmers *et al.*, 2019). SMCHD1 also has a C-terminal SMC hinge domain, which is vital for interactions between the hinge domain and the N-terminus (Hirano, 2006).

SMCHD1 is a member of the MORC protein family and also the structural maintenance of chromosome (SMC) gene family (Li *et al.*, 2013). Some SMCHD1 mutations cause facioscapulohumeral muscular dystrophy type 2 (FSHD2) (Shaw *et al.*, 2017), one of the most common types of muscular dystrophies (Statland and Tawil, 2014). Mutations in SMCHD1 can also cause Bosma arhinia microphthalmia syndrome (BAMS) (Lemmers *et al.*, 2019) a disease, which causes patients to be born without a nose (Gordon *et al.*, 2017).

SMCHD1 is also an epigenetic regulator, which promotes inactivation of chromosome X (Blewitt *et al.*, 2008; Sakakibara *et al.*, 2018). A common characteristic of epigenetic regulators is that they are often mutated in human cancers (Kanwal and Gupta, 2012). One study demonstrates how silencing of SMCHD1 causes development of cancer, which leads to the conclusion that SMCHD1 acts as a tumour suppressor (Leong *et al.*, 2013).

1.4 An introduction to cancer

In 2018, there were 18.1 million cancer diagnoses and 9.6 million reported cancer deaths worldwide, with lung cancer being the most commonly diagnosed and female breast cancer being the second most common (Bray *et al.*, 2018). If there is a mutation in a gene which is vital for cellular function, this can lead to disease development (Sawyer *et al.*, 2007). Approximately 90% of cancers are caused by somatic mutations (Paul *et al.*, 2019) (mutations that do not occur in germline cells, which can be inherited), which are often caused by

environmental factors including, but not limited to, UV radiation (Saini *et al.*, 2016). There are several hallmarks of cancer that facilitate tumour growth (Hanahan and Weinberg, 2011) including cells resisting cell death (Fulda, 2010), activating angiogenesis to increase the blood supply to the tumour (Prager and Zielinski, 2013) and evading growth suppressor signalling (Amin *et al.*, 2015).

1.4.1 The cell cycle and cancer

The cell cycle has 4 consecutive phases, S phase, M phase, G_1 phase, G_2 and G_0 phase. S phase is where DNA replication occurs, and M phase is where the cell divides into two daughter cells (mitosis followed by cytokinesis). In between phases S and M there are phases G_1 and G_2, G_2 occurs after the S phase and involves preparation of the cell so it is ready to enter mitosis (M phase). G1 occurs after mitosis and involves cell growth (Williams and Stoeber, 2012) and G_0 is cell cycle arrest when cells are either non-proliferating or quiescent (Tomura *et al.*, 2013) (Figure 1.13).

The cell cycle can become disrupted following exposure to exogenous or endogenous sources of DNA damage and oncogene activation (Foster, 2008). p53 is a transcription factor, which is activated by cellular stress such as DNA damage. When p53 becomes activated the most likely conclusion is cell cycle arrest and apoptosis (Chen, 2016) however, it could also lead to DNA repair (Giono and Manfredi, 2006). If a mutation occurs in the *p53* gene and is not repaired it could lead to tumourigenesis and an accumulation of the p53 protein, which is a common hallmark of cancerous cells (Rivlin *et al.*, 2011).

There are several checkpoints throughout the cell cycle which are regulated by the cyclin-dependent kinases (CDKs). There are 20 different recognised types of CDKs (Malumbres and Barbacid, 2005), which are serine/threonine protein kinases, that phosphorylate specific substrates to enable DNA synthesis (Barnum and O'Connell, 2014). Mutations in CDKs can

dysregulate the cell cycle and cause CDK hyperactivation, leading to uncontrolled cell proliferation and tumour cell formation (Peyressatre *et al.*, 2015).

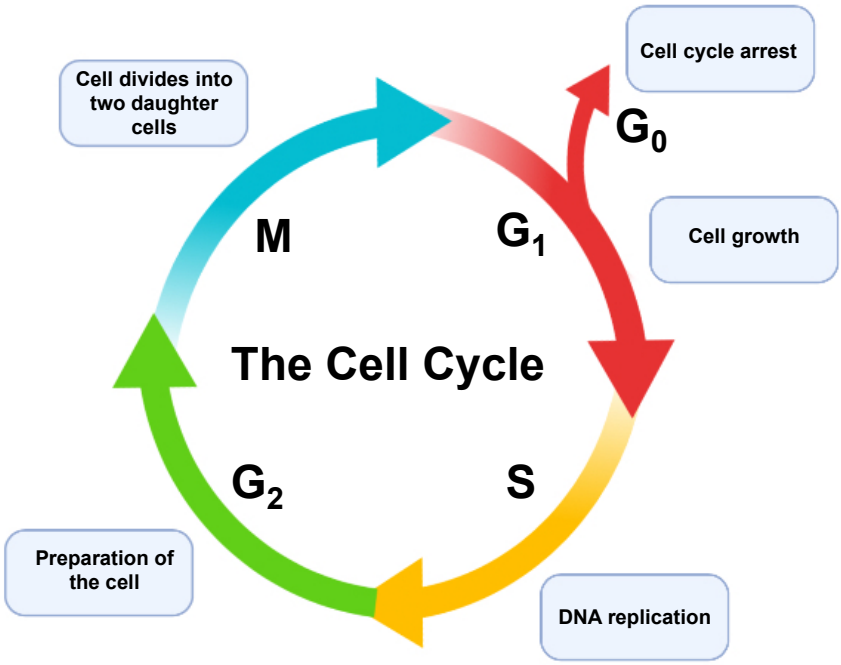


Figure 1.13: **Schematic representation of the eukaryotic cell cycle**. The cell cycle begins with G_1 phase, which is cell growth, the next stage is S phase which is important for DNA replication of each of the 46 chromosomes. G_2 phase is where the cell is prepared for mitosis. Phase M is where mitosis occurs and the cell divides into two daughter cells (cytokinesis). G_0 is where cell cycle arrest occurs when a cell is either non-proliferating or quiescent. Figure created by author and adapted from (El-Aouar Filho *et al.*, 2017).

1.4.2 Cancer and deregulated gene expression

Cancer is a disease which is caused by alterations in DNA, including somatic mutations, copy number modifications and DNA methylation. (Gao and Teschendorff, 2017). These alterations in DNA have the potential to up or down-regulate gene expression, which could drive cancers, for example loss of tumour suppressors can lead to cancer directly (Li Hui Wang *et al.*, 2019). In addition, gene expression in human cancers can be investigated to find potential biomarkers of cancer and for therapy and treatment (Kamel and Al-Amodi, 2017). There have been several reported genes which have been shown to be upregulated in human cancers, with many of the

genes also reported in cancer types with aggressive cancer progression (Mengyuan Li *et al.*, 2017).

For example, PARP1 is an important regulator in DNA repair and it is thought that upregulation of PARP1 can cause diseases such as triple-negative breast cancer (Ossovskaya *et al.*, 2010) (Gilabert *et al.*, 2014), which is negative for oestrogen receptor, progesterone receptor and human epidermal growth factor receptor (HER2) (Wahba and El-Hadaad, 2015). One study also found that silencing of PARP1 using inhibitors can lead to cancer cell death (Pazzaglia and Pioli, 2019).

Some proteins are also downregulated in cancer such as the p53-mediated DNA repair protein, O^6-Methylguanine-DNA methyltransferase (MGMT) (Bocangel *et al.*, 2009). MGMT is responsible for the repair of DNA lesions (Fan *et al.*, 2013; Soniya Sharma *et al.*, 2009) and some studies suggest that p53 binds to the SP1 transcription factor, therefore, leading to is abrogated binding of p53 to MGMT, leading to lower expression levels (Bocangel *et al.*, 2009). Overall, the low expressing genes in cancer are likely to have low expression levels in cancerous cells compared to normal cells, suggesting that perhaps gene expression is a driver of cancer progression (Mengyuan Li *et al.*, 2017).

1.4.3 Cancer types

There are several types of cancer which can exist in several areas of the human body. Lymphoma is a specific type of cancer which affects the blood and has several sub-types (Laurent *et al.*, 2015) and known cells of origin. Neuroblastoma is another type of cancer which affects children (Papaioannou and McHugh, 2005) whereas, colorectal carcinoma (Mauri *et al.*, 2019) and breast cancer (Anders *et al.*, 2009) are adult onset cancers.

1.4.3.1 Lymphoma

Lymphoma and leukaemia are two types of cancer affecting the blood, however, acute myeloid leukaemia (AML) and chronic myeloid leukaemia (CML) both affect the myeloid cells within the bone marrow (Jamieson *et al.*, 2004) whereas, lymphoma causes malignancy of the T (thymus derived)-cells or B (bone marrow derived)-cells (Xu *et al.*, 2002) (Ninkovic and Lambert, 2017). There are two main groupings of lymphomas, Non-Hodgkin lymphoma and Hodgkin lymphoma. A subtype of Non-Hodgkin lymphoma has been found to be upregulated in MORC4 and could be a potential biomarker of lymphoma (Liggins *et al.*, 2007).

B-cell Non-Hodgkin lymphomas include Diffuse-large B-cell lymphoma, Burkitt's lymphoma, Follicular lymphoma, Marginal zone lymphoma and Mantle cell lymphoma. T-cell Non-Hodgkin lymphomas include Peripheral lymphoma and Cutaneous lymphoma (Ninkovic and Lambert, 2017). There are two types of Hodgkin lymphoma including classical Hodgkin lymphoma and nodular lymphocyte-predominant Hodgkin lymphoma (NLPHL). The type of Hodgkin lymphoma is determined by the cell morphology and symptoms of the patient (Fields and Wrench, 2017).

1.4.3.1.1 B-cell development

B lymphocytes are vital white blood cells, which present antibodies (Chen and Jensen, 2008) and are required during the human immune response to fight infection. B-cells develop from hematopoietic stem cells (HSCs) in the bone marrow (BM) and follow a number of stages of maturation to differentiated plasma cells, which are ready to help fight infections (Harnett *et al.*, 2005) (Figure 1.14).

The initial step of B-cell maturation involves the rearrangement of the immunoglobulin (Ig) heavy chain (IgH) (V_H, D_H and J_H) and light chain (IgL) (V_L-J_L) (ALT *et al.*, 1986), which are also known as antibodies. These antibodies are glycosylated proteins that are situated on the

surface of B-cells (Hoffman *et al.*, 2016). The rearrangement of the IgH and IgL helps to generate different combinations of antibodies, which can recognise over 50 million different antigens (Pieper *et al.*, 2013). Following on from the rearrangement of the IgH and IgL chains, the next stage is the pre-B-cell stage. During the pre-B-cell stage, the cells undergo cell division (van Zelm *et al.*, 2007) and immunoglobulin M (IgM) is presented on the cell surface, which are then identified as immature B-cells, which can leave the BM and travel to the spleen (Pieper *et al.*, 2013).

Once the B-cell antigen receptor (BCR) is formed on the surface of the B-cell, the cell may enter the pre B-cell stage of maturation (LeBien and Tedder, 2008). B-cells are then differentiated into plasma cells or memory B-cells, which occurs in spleen and lymph nodes at the germinal center (GC) (Kurosaki, 2010; Victora and Nussenzweig, 2012). 40% of adult B-cells are memory B-cells, which develop from naive B-cells (Seifert and Küppers, 2016). Memory B-cells are a crucial part of the immune response, which have the ability to recognise antigens from pathogens from previous infections (Vikstrom and Tarlinton, 2011). Plasma cells secret antibodies (Khodadadi *et al.*, 2019), which are essential for the immune response (Hoffman *et al.*, 2016).

1.4.3.1.2 B-cell non-Hodgkin lymphoma (B-NHL)

B-cell non-Hodgkin lymphoma (B-NHL) is the most common form of lymphoma and the sixth most common cancer in the UK (Ninkovic and Lambert, 2017), which forms from GC B-cells (Ramezani-Rad and Rickert, 2017). The majority of GC B-cells will undergo apoptosis, however, some GC B-cells will upregulate phosphoinositide 3-kinase (PI3K) signalling (Pfeifer *et al.*, 2013) and expression of *Myc* (Hoffman and Liebermann, 2008), an important proto-oncogene involved in progression of B-NHL (Cai *et al.*, 2015). There are several sub-types of B-NHL including, Burkitt's lymphoma (BL), Mantle cell lymphoma (MCL), Marginal

zone lymphoma (MZL), Follicular lymphoma (FL) and Diffuse large B-cell lymphoma (DLBCL). The type of B-NHL is dependent on the cell of origin (Figure 1.14).

1.4.3.1.2.1 Diffuse large B-cell lymphoma (DLBCL)

DLBCL is responsible for approximately 31% of B-NHL in Western countries (Martelli *et al.*, 2013) and is one of the most aggressive types of B-NHL (Ninkovic and Lambert, 2017). There are several elements that can increase the risk of DLBCL including the human immunodeficiency virus (HIV) (Besson *et al.*, 2017), pesticide use (Altahan and Martin, 2018) and a high fat and high sugar diet (Mozaheb *et al.*, 2012). There are several symptoms of DLBCL, however, the majority of patients have enlarged and inflamed lymph nodes (Liu and Barta, 2019) and B symptoms, which includes night sweats, a high fever and weight loss (Rohini Sharma *et al.*, 2009). The most commonly used chemotherapy that is used to treat DLBCL is the monoclonal antibody, rituximab (Dotan *et al.*, 2010). Rituximab cures 50-60% of diagnosed patients (Liu and Barta, 2019), however, patients who relapse following treatment of rituximab have poor prognosis (Gisselbrecht *et al.*, 2010).

There are three types of DLBCL, including germinal centre B-cell (GCB)-like, activated B-cell (ABC)-like DLBCL (Figure 1.14) and the more recently discovered, primary mediastinal B-cell lymphoma (PMBL) (Lenz *et al.*, 2008). Patients with GCB-like DLBCL have an overall higher survival rate than patients with ABC-like DLBCL (Alizadeh *et al.*, 2000). Although patients with DLBCL usually respond well to treatment, less than half are cured (Fisher *et al.*, 1993; Liggins *et al.*, 2007). The morphology of DLBCL cells is often characterised by the proliferation of large sized B-cells (Iqbal *et al.*, 2016; Gouveia *et al.*, 2012).

The tumour suppressor p53 has been associated with cases of DLBCL. Patients with silenced or mutated *p53* have a more aggressive type of DLBCL (Leroy *et al.*, 2002). Mutated or silenced *p53* can have an effect on DNA repair, leading to DLBCL (Gouveia *et al.*, 2012). *Myc*

rearrangement has also been associated with a more aggressive type of DLBCL and causes a lower survival rate (Dunleavy, 2015). Another common abnormality in DLBCL patients is the *BCL6* gene at the 3q27 locus, which is vital for the formation of the GC (De Leval and Harris, 2010).

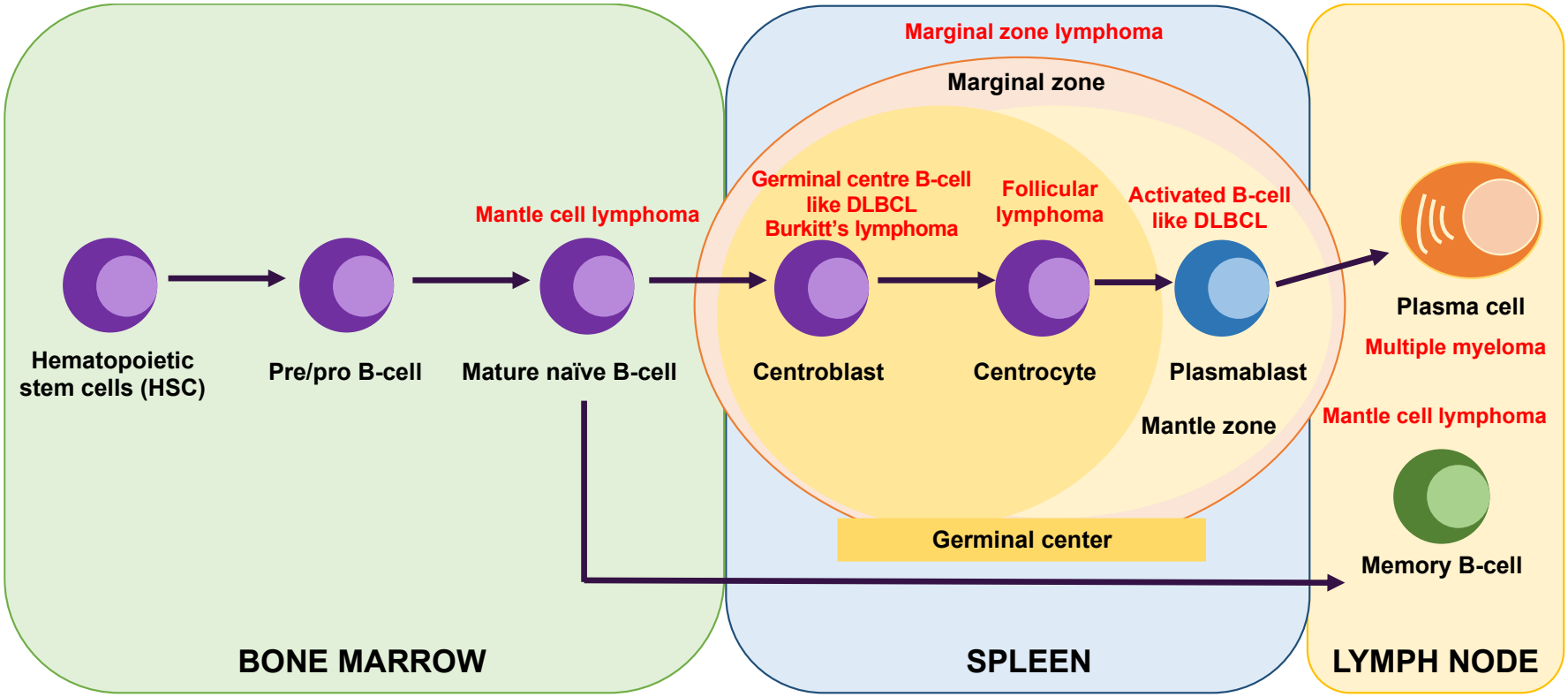


***Figure 1.14*: B-cell maturation and development of B-cell lymphomas.** B-cell maturation begins in the bone marrow as B-cells develop from HSCs. Memory B-cells and plasma cells develop from pre/pro B-cells, which are formed from the germinal centre. The B-cell lymphomas are shown in red, which indicates the cell of origin. Mantle cell lymphomas are most likely to derive from mature naïve B-cells (purple) in the bone marrow; however, they can also derive from memory B-cells (green) in the lymph nodes. Germinal center B-cell (GCB)-like diffuse large B-cell lymphoma (DLBCL), Burkitt's lymphoma (BL) and Follicular lymphoma (FL) are all derived from the germinal centre (dark yellow) in the spleen. Activated B-cell (ACB)-like DLBCL is derived from plasmablast cells in the mantle zone of the germinal center. Figure created by author and adapted from (Klein *et al.*, 2003; Weigert and Weinstock, 2012; Scott and Gascoyne, 2014; Pieper *et al.*, 2013).

1.4.3.1.2.1.1 DLBCL subtypes

GCB-like DLBCL is derived from normal B-cells in the germinal center and ABC-like DLBCL is derived from post germinal center cells, which are cells that have the potential to differentiate into plasma cells (figure 1.14) (Lenz *et al.*, 2008). More recently, there has been a third type of DLBCL identified, which is known as PMBL (Savage *et al.*, 2003; Rosenwald *et al.*, 2003). PMBL is derived from thymic B cells (Lenz *et al.*, 2008), which are B-cells that are found at the thymus (Perera and Huang, 2015) and patients with this type of DLBCL are often young in comparison to other types of DLBCL. PMBL also shows similarities to a subtype of Hodgkin lymphoma (Savage *et al.*, 2003), known as nodular sclerosis, which has similar sites of disease and pathology (Savage, 2006). Each type of DLBCL are individual diseases as they occur at different points in the B-cell maturation process (Staudt and Dave, 2005), have separate behaviours in response to chemotherapy (Lenz *et al.*, 2008) and oncogenic pathways (Dunleavy and Wilson, 2014).

Patients with GCB-like DLBCL are more likely to be cured with chemotherapy than those with ABC-like DLBCL (Wright *et al.*, 2003). Following treatment with anthracycline, a chemotherapeutic agent, patients with GCB DLBCL have a 5 year survival rate of 60% compared to ABC DLBCL and PMBL, which only has a 35% and 39% chance of survival after 5 years, respectively (Rosenwald *et al.*, 2002).

The nuclear factor- κB (NF-κB) is a transcription factor (Hayden and Ghosh, 2004), which regulates the immune system and oncogenic pathways (Moynagh, 2005). Gene expression profiling results reveal that ABC DLBCL has higher expression of NF-κB target genes in comparison to GCB DLBCL (Davis *et al.*, 2001). Silencing of the NF-κB pathway in ABC DLBCL was toxic to cells, however, it was not toxic to GCB DLBCL cells (Staudt and Dave, 2005). More recently, PMBL was also found to express NF-κB target genes (Lam *et al.*, 2005).

Overall, this evidence suggests that the NF-κB pathway is a potential therapeutic target for both ABC DLBCL and PMBL (Staudt and Dave, 2005).

1.4.3.2 Neuroblastoma

Neuroblastoma is a childhood cancer that develops from neural crest stem cells and is often diagnosed between the age of 18 to 22 months old (Newman *et al.*, 2019; London *et al.*, 2005). Neural crest cells are a group of cells found in vertebrate embryos (Achilleos and Trainor, 2012) and are able to migrate throughout the embryo to different locations where they contribute to different tissues including the peripheral nervous system (PNS) (Prendergast and Raible, 2014) and the facial skeleton (Bronner, 2012).

Neuroblastoma causes approximately 15% of childhood cancer deaths every year in the United States (Li *et al.*, 2008) and has an overall survival rate of less than 40% (Bosse and Maris, 2016). Anaplastic lymphoma kinase (ALK) is a receptor tyrosine kinase (RTK) oncogene, which is often mutated in inherited neuroblastoma (Umapathy *et al.*, 2019; Mossé *et al.*, 2008). Another known but less common mutation, that can cause neuroblastoma is the paired–like homeobox 2B (*PHOX2B*) gene (Mosse *et al.*, 2004), which is also responsible for congenital central hypoventilation syndrome (CCHS) (Trochet *et al.*, 2004), a disorder that affects breathing (Doherty *et al.*, 2007).

Neuroblastoma can develop anywhere in the sympathetic nervous system (SNS) but can progress, in particular, in the adrenal medulla in the adrenal gland of the kidney (Maris, 2010). Tumours are often present in either the adrenal gland or in the neck, chest, abdomen or pelvis (Newman *et al.*, 2019). Once metastasis begins, the tumours will often migrate to the lymph nodes and the bone marrow (Mullassery and Losty, 2016).

1.4.3.3 Colorectal carcinoma

Colorectal carcinoma is responsible for approximately 10% of cancer incidence in Europe (Kuipers *et al.*, 2015) and is one of the most common cancers in the world (Haggar and Boushey, 2009). This type of cancer often develops over several years and can cause tumours to develop on the lining of the rectum or colon (Marley and Nan, 2016). However, approximately 90% of colorectal cancers are adenocarcinomas, which is a type of cancer that develops from the epithelial cells in the colon (Fleming *et al.*, 2012).

There are several subtypes of colorectal carcinoma syndromes which are responsible for hereditary colorectal cancer with the most common syndrome known as Lynch syndrome (hereditary nonpolyposis colorectal cancer) (Kuipers *et al.*, 2015), which increases the risk of colorectal carcinoma (Al-Sukhni *et al.*, 2008). Lynch syndrome is caused by an genetic mutation in one of the DNA mismatch repairs genes, often leading to carcinomas (Cohen and Leininger, 2014). Furthermore, familial adenomatous polyposis (FAP) syndrome is another disease which causes hereditary colorectal cancer and consequently leads to thousands of polyps which can be found in the colon and rectum (Bojuwoye *et al.*, 2018). Patients with inflammatory bowel disease (IBD) are also at an increased risk of developing colorectal carcinoma (Stidham and Higgins, 2018).

The most common mechanism for the formation of colorectal carcinoma begins as a polyp, which has a 5% chance of developing into a benign tumour on a gland (Bujanda *et al.*, 2010). This tumour progresses before eventually becoming colorectal cancer. This type of cancer can take up to approximately 10 years to develop (Simon, 2016) with the addition of several consecutive mutations (Armaghany *et al.*, 2012). If the patient has a colorectal carcinoma syndrome, they may develop the cancer more quickly (Kuipers *et al.*, 2015).

1.4.3.4 Breast cancer

Breast cancer is the most common type of cancer which affects women in the USA (Toriola and Colditz, 2013) and around the world. However, the life expectancy in several developed countries is approximately 80 years, depending on the aggressiveness of the tumour (Becker, 2015). Breast cancer can also lead to tumour metastasis if not diagnosed or treated in the early stages, where it can spread to several organs throughout the human body such as bones, the lungs and the brain (Sun *et al.*, 2017). There are also several risk factors which can increase the likelihood of developing breast cancer. Age is the most common risk factor, as studies have found that the most common age is around 40 to 59 years as breast cancer is also most commonly found in women around the menopause age (Kamińska *et al.*, 2015). In addition, exposure to oestrogen has also been found to be linked to an increase risk of breast cancer (Travis and Key, 2003). Furthermore, it appears that there is a genetic link to breast cancer as women with relatives who have had the disease are also more likely to develop this cancer (Brewer *et al.*, 2017) and it is known that this is related to mutations of the breast cancer oncogenes, *BRCA1* and *BRCA2* (Mehrgou and Akouchekian, 2016).

There are several types of breast cancer such as the common lobular carcinoma in situ (LCIS), ductal carcinoma in situ (DCIS), infiltrating lobular carcinoma (ILC) and less common types such as medullary carcinoma, inflammatory breast cancer and Paget's disease of the nipple (Sharma *et al.*, 2010). The most common type of non-invasive breast cancer is DCIS (Bergholtz *et al.*, 2020), which affects the ducts of the breast (van Seijen *et al.*, 2019). It is also known that the oestrogen receptor (ER) is important for regulating breast cancer development, progression and for therapeutics of breast cancer (Rozeboom *et al.*, 2019; Williams and Lin, 2013). Furthermore, breast cancer can be either ER-positive or ER-negative with those diagnosed with a ER-negative breast cancer types often diagnosed earlier in life with a higher mortality rate (Bauer *et al.*, 2007).

1.4.4 MORC4 and cancer

1.4.4.1 MORC4 is a potential biomarker of DLBCL

Recent evidence has found a link between MORC4 and DLBCL (Liggins *et al.*, 2007). One study analysed MORC4 mRNA and found that in normal cell lines, MORC4 was expressed at low levels, although it had the highest expression in testes and placenta tissues. Whereas in DLBCL cell lines, MORC4 mRNA was significantly upregulated in more than half of DLBCL cell lines. More specifically, they found that the DLBCL subtype, ABC DLBCL cell lines had increased MORC4 expression in comparison to GCB DLBCL, which had no measurable expression of MORC4. Similar results were found in biopsies from patients with DLBCL subtypes. ABC DLBCL has a poor prognosis, which suggests that there could be a link between MORC4 and the aggressive disease, ABC DLBCL. Overall, this indicates that MORC4 could be a novel biomarker of DLBCL subtypes (Liggins *et al.*, 2007).

1.4.4.2 MORC4 is a novel breast cancer oncogene

More recently, MORC4 has been identified as a potential breast cancer oncogene. One study identified a higher expression of *MORC4* in breast cancer tissue when compared with normal tissues, which suggests a role for MORC4 in breast cancer. In addition to this there is further evidence to suggest that MORC4 regulates apoptosis in breast cancer cells. When MORC4 was knocked down via siRNA, a decrease in cell growth of breast cancer cells was observed, along with a decreased in the antiapoptotic protein, Bcl-2. This suggests that MORC4 could be playing an important role in breast cancer cell survival by regulating apoptosis (Yang *et al.*, 2019).

1.5 Aims

The main aims of this project are to characterise MORC4 biochemically and structurally in addition to an investigation into MORC4 as a cancer biomarker with RNA and protein expression analysis methods. Both MORC2 and MORC3 are relatively well structurally and biochemically characterised. However, MORC4 remains poorly characterised with some evidence suggesting that perhaps MORC4 is a potential biomarker of a specific type of non-Hodgkin's lymphoma known as diffuse large B-cell lymphoma (DLBCL). In an attempt to solve the structure of MORC4, multiple domain boundaries will be designed and cloned into *E. coli* expression vectors with an aim to produce pure protein for X-ray crystallography trials. Biochemical assays such as chemical cross-linking will also be utilised to examine the oligomeric state of MORC4 and to gain an understanding of potential biochemical mechanisms. Finally, MORC4 expression patterns will be investigated in an attempt to identify cancer cell lines with expression of MORC4 to help determine MORC4 function. Overall, this project aims to give an insight into the structure and biochemistry of MORC4 in addition to expression of MORC4 in cancer.

CHAPTER 2

Materials and Methods

Chapter 2 Material and Methods

2.1 Materials

All chemicals and reagents were from ThermoFisher Scientific (Loughborough, UK) unless otherwise stated. ddH_2O was purified using a Barnstead™ Nanopure™ Ultrapure (Barnstead™ Nanopure™, California) water purification system, which purified ddH_2O to 18.2 MΩ.

2.2 Recombinant DNA and cloning techniques

2.2.1 Preparation of competent MACH1 and Rosetta2 *Escherichia coli* cells

2x Luria-Bertani broth (LB) was prepared (formulation per 1L: 10g peptone 140, 5g yeast extract and 170 mM NaCl) and sterilised by autoclaving. 5 mL LB was inoculated from a MACH1 or Rosetta2 glycerol stock and incubated overnight at 37°C, 200 rpm (Rosetta2 culture required 34 µg/ml chloramphenicol antibiotic). Following overnight incubation, 5 mL of the overnight culture was used to inoculate 200 mL super optimal broth (SOB) medium (20 g/L bacto peptone casein, 5 g/L bacto yeast extract, 9 mM NaCl, 2.5 mM KCl and 20 mM $Cl_2Mg \cdot 6H_2O$) and was incubated at 37°C, 200 rpm, until the OD_{600} reached 0.45-0.55. 50mL tubes were pre-chilled on ice, and cells were harvested at 1300 rpm for 5 minutes.

The supernatant was discarded, and the pellets were gently resuspended in 160 mL (per 500 mL of culture) pre-chilled transformation buffer 1 (TFB1) (TFB1 buffer: 30 mM $C_2H_3KO_2$, 100 mM RbCl, 10 mM $CaCl_2 \cdot 2H_2O$, 150 mL/L glycerol made to pH 5.85 (with 1 M acetic acid) and then addition of 50 mM $Cl_2Mn \cdot 4H_2O$). Resuspended cells were spun at 1300 rpm for 5 minutes and the supernatant was discarded. The pellets were gently resuspended on ice in 30 mL (per 500 mL culture) pre-chilled transformation buffer 2 (TFB2) (TFB2 buffer: 10 mM $C_7H_{15}NO_4S$, 11.027g/L $CaCl_2 \cdot 2H_2O$, 1 mM RbCl and 150mL/L glycerol made to pH 6.5 with KOH), snap frozen as aliquots in liquid nitrogen and stored at -80°C until required for use.

2.2.2 Polymerase Chain Reaction (PCR)

MORC DNA sequences were amplified using primers containing the following extensions: 5'-TACTTCCAATCCATG-3' (the ATG was in-frame) and 5'-TATCCACCTTTACTG*TCA*-3' (the termination codon was within the 3' LIC site). For each PCR reaction, a 25 µL (final volume) master mix was made, containing 1x Phusion® high-fidelity buffer (Phusion HF buffer) (New England BioLabs, Ipswich, MA, USA), 200 µM deoxynucleotide (dNTPs) mix and 0.25 µL of Phusion® high-fidelity DNA polymerase (New England BioLabs, Ipswich, MA, USA).

5 ng of template DNA and 2.5 µM of each primer (forward and reverse) was used for each PCR sample. The PCR was run under touch-down conditions (Hecker and Roux, 1996) consisting of the following cycle: 95°C for 10 minutes, 95°C, 30 sec; 68°C, 30 sec; 72°C, 1-3 min* x5 cycles, 95°C, 30 sec; 60°C, 30 sec; 72°C, 1-3 min* x5 cycles, 95°C, 30 sec; 55°C, 30 sec; 72°C, 1-3 min* x5 cycles, 95°C, 30 sec; 50°C, 30 sec; 72°C, 1-3 min* x10 cycles, 72°C, 10 min and 15°C on hold (* the extension time was dependent on the size of the PCR product, 1 kb ~ 1 minute).

2.2.3 Agarose gel DNA electrophoresis

DNA samples were diluted with water and 1 x gel loading dye (New England BioLabs). The DNA was resolved on a 0.8% gel (stained with 0.4 x SYBR® safe DNA gel stain) with 1 x 2-log DNA ladder (New England BioLabs) for size comparison, for approximately 60 minutes at 80 V in 1 x Tris-borate-EDTA (10 x TBE composition: 1 M tris base, 1 M boric acid and 20 mM EDTA (disodium salt)) and viewed under ultraviolet light (UV) using an InGenius gel viewer (Syngene, Bangalore, India). Images were edited and saved using GeneSnap software (Syngene).

2.2.4 Plasmid DNA extraction and purification

Glycerol stocks of *E. coli* cells containing plasmid DNA were used to inoculate 5 mL of LB. Inoculated LB was incubated overnight at 37°C at 180 rpm. For a higher concentration of resultant plasmid DNA, cells were grown in 5 mL LB for 8 hours at 37°C at 180 rpm. Following an 8-hour incubation, the 5 mL culture was used to inoculate 200 mL of LB and incubated overnight at 37°C at 180 rpm. Cells were aliquoted into 50 mL tubes and harvested by centrifugation at 2000 rpm for 20 minutes. The supernatant was removed, and DNA was extracted and purified with the use of either a miniprep or maxiprep kit (Qiagen, Hilden, Germany) (Table 2.1).

Table 2.1: **DNA purification miniprep and maxiprep kit** culture volumes, volume of plasmid eluted and expected yield of plasmid.

DNA purification kit	Culture volume (mL)	Plasmid eluted (µL)	Expected yield of plasmid (ng/µL)
Miniprep kit	5	50	50-200
Maxiprep kit	200	1000	300-700

2.2.5 DNA quantitation

DNA was quantified using a nanodrop 2000 (260 nm) (ThermoFisher Scientific). The nanodrop was blanked using 2 µL nuclease-free water and 2 µL DNA was quantified. The 260/280 nm ratio was used to assess purity of DNA (pure DNA ~1.8).

2.2.6 DNA sequencing

Purified cloned DNA plasmids were sent to SourceBioscience (Cambridge, UK) for Sanger sequencing to confirm DNA sequences. Plasmid DNA was provided at 100 ng/µL and primers at 3.2 µM.

2.2.7 Transformation of recombinant plasmid DNA into *E. coli*

30 µL of MACH1, NEB5-α or Rosetta2 competent *E. coli* cells were added to the pre-chilled ligation mix and incubated on ice for 30 minutes. Cells were transformed by incubating at 42°C for 45 seconds in a water bath and immediately chilled on ice for 2 minutes. 60 µL of 2x LB was added to the cells and incubated at 37°C for at least an hour, at 180 rpm. 100 µL of the transformed bacteria was spread onto LB agar plates with 5% (w/v) sucrose and appropriate antibiotic (50 µg/mL kanamycin, 34 µg/ml chloramphenicol or 200 µg/mL ampicillin) and incubated overnight at 37°C.

2.2.8 Glycerol stock preparation of recombinant bacteria

Single bacterial colonies were grown overnight at 37°C, 180 rpm in 5 mL LB with the appropriate antibiotic (50 µg/mL kanamycin, 34 µg/ml chloramphenicol or 200 µg/mL ampicillin). To prepare a glycerol stock, 1 mL of overnight culture was added to 333 µL 60 % (v/v) glycerol (15 % glycerol final concentration) and stored at -80°C.

2.2.9 Bacmid preparation for insect cell expression

2.2.9.1 Preparation of DH10bac competent *E. coli* cells

10 mL LB (containing 50 mg/mL kanamycin and 5 mg/mL tetracycline) was inoculated with a DH10bac glycerol stock and incubated overnight at 37°C at 180 rpm. 100 mL LB (containing 50 mg/mL kanamycin and 5 mg/mL tetracycline) was inoculated with 1 mL of the overnight culture and grown to an OD_{600} of 0.4-0.6. Cells were incubated on ice for 30 minutes and then centrifuged for 5 minutes at 4000 rpm at 4°C. The supernatant was discarded, and the pellet was resuspended in 10 mL pre-chilled 100 mM $CaCl_2$ on ice and then incubated for an additional 30 minutes. The cells were centrifuged for 5 minutes at 4000 rpm at 4°C, the supernatant was discarded, and the pellet was resuspended in 5 mL pre-chilled 50 mM $CaCl_2$

with 20% (v/v) glycerol on ice. Cells were incubated on ice for a further 30 minutes and then aliquoted into 300 µL volumes in pre-chilled 1.5 mL tubes and stored at -80°C.

2.2.9.2 Bacmid transposition

DH10bac selective plates comprised of LB agar with 50 µg/mL kanamycin, 7 µg/mL gentamycin, 10 µg/mL tetracycline, 40 µg/mL IPTG and 200 µg/mL X-gal. 20 µL DH10bac cells were added to 3 µL 10-200 ng/µl DNA and incubated on ice for 30 minutes. Cells were transformed by heating at 42°C in a water bath and immediately returned to ice. 900 µL of LB (no antibiotics) was added to the cells and incubated at 37°C for 5 hours, 180 rpm. The cultures were diluted 10-fold in LB and 50 µL was spread onto DH10bac selective agar plates. The plates were incubated at 37°C for 48 hours covered in foil, and then incubated at 4°C for 24-48 hours. White colonies were streaked to dilution on DH10bac selective plates with a blue colony used as a control.

2.2.9.3 Bacmid DNA production

White colonies from the streaked dilution plate were used to inoculate LB (containing 50 µg/mL kanamycin, 7 µg/mL gentamycin, 10 µg/mL tetracycline) and were grown overnight at 37°C, 180rpm. A glycerol stock was made of each culture (Section 2.2.8) and the remaining culture was centrifuged at 4°C at 2600 x g for 30 minutes. Using buffers from the plasmid DNA extraction kit (Qiagen), the pellet was resuspended in 250 µL buffer P1 containing RNase A. 250 µL of buffer P2 (lysis buffer) was added to each sample, gently inverted and incubated at room temperature for 10 minutes. 300 µL N3 buffer (precipitation buffer) was added, mixed gently and the samples were placed on ice for 20 minutes. The samples were centrifuged at 2600 x g for 30 minutes at 4°C and the supernatant was transferred to a new tube and centrifuged for a second time.

2.2.10 Ligation independent cloning (LIC)

2.2.10.1 Plasmid vector digestion

5 µg of plasmid (pNH-TrxT, pNIC-ZB, pGTVL2 or pFB-LIC-Bse) with 1x NEBuffer 3.1 (New England BioLabs) was mixed with 3 µL of enzyme (*Bsa*I or *Bse*RI (New England BioLabs). The samples were made up to a total volume of 100 µL with ddH_2O, which was adjustable depending on the concentration of the plasmid. The vector digest was incubated overnight at 50°C or 37°C, depending on the enzyme, and resolved on a 0.8 % gel to confirm digestion. Following vector digestion, the plasmid DNA was purified using a QIAquick PCR purification kit (Qiagen) and eluted in 50 µL elution buffer.

2.2.10.2 Preparation of vector and insert cohesive ends

Digested and purified vector was treated with 1x NEB buffer 2.1 (New England BioLabs), 2.5 µM deoxyguanosine triphosphate (dGTP), 10 mM dithiothreitol (DTT) and 2.5 µL T4 DNA polymerase (New England BioLabs). The treated vector was incubated for 30 minutes at 22°C and heat inactivated for 20 minutes at 75°C in a thermocycler. LIC insert cohesive ends were prepared in a similar way, however, 2.5 µM deoxycytidine triphosphate (dCTP) (New England BioLabs) was used.

2.2.10.3 Removal of template DNA and purification of PCR products

*Dpn*I treatment was only required if the template DNA used for PCR had the same antibiotic resistance as the cloning vector. *Dpn*I (New England BioLabs) was prepared as a 1 in 20 dilution with NEB buffer 2.1. 1 µL of the freshly prepared *Dpn*I stock was added to the PCR product and incubated at 37°C for 1 hour. Following *Dpn*I treatment, the insert DNA was purified using a QIAquick PCR purification kit (Qiagen) and eluted in 50 µL elution buffer.

2.2.10.4 Ligation of treated LIC vector and insert

To ligate the insert with the plasmid, a ratio of 1:2, 2:2 and 2:4 of vector: insert was used. The plasmid was added to the insert and incubated at room temperature (22°C) for 30 minutes.

2.2.10.5 Colony PCR recombinant clone screening

Colony PCR forward and reverse primers (pLICF/pLICR, FBAC1/FBAC2/M13bac_rev or pLICfor/pLICrev for *E. coli*, baculoviral or mammalian constructs, respectively) were mixed to give a final concentration of 10 µM. For a single colony PCR reaction, 5 x My Taq™ reaction buffer red (Bioline, London, UK) was added to 3 % dimethyl sulfoxide (DMSO), 3 % glycerol, 1 µM primer mix and 5 U/20 µL MyTaq™ DNA polymerase (Bioline). The reaction was made up to 20 µL with ddH_2O and 0.5 µL of the overnight bacterial culture containing the potential recombinant plasmid DNA, was added to the PCR tube.

The thermocycler was pre-heated to 94°C (as MyTaq™ is not a hot-start DNA polymerase) and the following PCR conditions were used: 94°C for 10 minutes 10, 94°C for 30 seconds, 50°C for 30 seconds, 72°C for 1-3 minutes* x 30 cycles, 72°C for 5 minutes and hold at 15°C (* the extension time was dependent on the size of the PCR product, 1 kb ~1 minute, an additional 1 minute was added to the colony PCR thermocycler conditions). 10 µL of the colony PCR screen was resolved on a 0.8% agarose gel to determine the size of each recombinant DNA, to determine if ligation had been successful into the respective vectors. Glycerol stocks containing recombinant DNA of the correct size were used to inoculate 5 mL LB medium and incubated overnight at 37°, 180 rpm. Cells were harvested, and the recombinant DNA was extracted and purified (Section 2.2.4). Purified recombinant DNA was quantified using a nanodrop 2000 (Section 2.2.5) and sent for DNA sequencing to confirm DNA sequences prior to being transformed into competent *E. coli* cells.

2.2.11 Site-directed mutagenesis (SDM)

The Q5 site-directed mutagenesis (SDM) kit (New England BioLabs, Ipswich, Massachusetts, United States) was used to create substitution mutations in MORC4 recombinant plasmid DNA for downstream experiments.

2.2.11.1 PCR

Primers were designed with the substitution mutation in the central region of the primers with the 5' end of the first primer situated adjacent to the 5' end of the second primer with amplification occurring outwards from each (Appendix 8.1 and 8.2). Recombinant plasmid DNA that had been previously constructed by ligation independent cloning (LIC) was used for PCR amplification. The PCR was prepared similar to section 2.2.2, however, an extension time of 10 minutes was required to amplify the MORC4 gene and vector sequences.

2.2.11.2 Phosphorylation, ligation and removal of template DNA

Kinase-ligase-*Dpn*I (KLD) treatment was required to phosphorylate, ligate and remove template DNA, respectively. Prior to KLD treatment an additional *Dpn*I treatment was essential to ensure complete removal of the template DNA (Section 2.2.10.3). Each sample contained a total of 10 µL reaction mix including, 1 µL of PCR product (*Dpn*I treated), 1 x KLD reaction buffer, 1 x KLD enzyme and 3 µL nuclease-free water. Samples were incubated for 5 minutes at room temperature.

2.2.11.3 Transformation of mutated plasmid DNA into *E. coli*

5 µL KLD mix containing the mutated plasmid DNA was added to 30 µL MACH1 competent *E. coli* cells and incubated at 4°C for 30 minutes. Samples were incubated at 42°C for 45 seconds and then at 4°C for 3 minutes. 180 µL LB was added to each sample and incubated at 37°C for 75 minutes at 180 rpm. 50 µL of culture was spread onto an agar plate containing 50

µg/mL kanamycin and incubated overnight at 37°C. Glycerol stocks were made the following day (Section 2.2.8).

2.2.11.4 Mutagenic colony PCR screen

Mutagenic colony PCR screen was performed as in section 2.2.10.5, however, only pLICF/pLICR primers were used with an extension time of 3 minutes. Glycerol stocks containing mutated plasmid DNA of the correct size were incubated overnight, cells were harvested, and the recombinant DNA was extracted and purified (Section 2.2.4). Purified recombinant DNA was quantified using a nanodrop 2000 (Section 2.2.5) and was sent to SourceBioscience (Cambridge, UK) for Sanger sequencing (Section 2.2.6) to confirm DNA sequences including the mutation.

2.3 Recombinant protein expression and purification

2.3.1 Recombinant protein expression

2.3.1.1 50mL protein expression testing and large-scale protein expression in *E. coli*

Cultures (from Rosetta *E. coli* glycerol stocks) were grown overnight in LB (with antibiotics 50 µg/mL kanamycin and 34 µg/mL chloramphenicol) or terrific broth (TB) (TB medium composition per 1L: 12 g bacto peptone, 24 g bacto yeast extract, 4 mL glycerol and 100 mL potassium phosphate solution containing 170 mM KH_2PO_4 and 710 mM K_2HPO_4) (with 50 µg/mL kanamycin and 25 µg/mL chloramphenicol) at 37°C, 180 rpm. 50 mL or 1 L LB/TB (with 50 µg/mL kanamycin) was inoculated with 500 µL or 10 mL overnight culture in 250 mL or 2 L baffled flasks (Thomson Instrument Company, CA, USA) and incubated at 37°C, 200 rpm to mid-log phase (OD_{600} of ~0.8 for LB or ~1.5-2.0 for TB cultures).

Cells were cooled to 18°C and protein expression was induced with 100 mM (final concentration) isopropyl β-D-1-thiogalactopyranoside (IPTG) and incubated overnight at 180 rpm. The overnight cultures were harvested by centrifugation at 4000 rpm for 30 minutes at

4°C. The supernatant was discarded, and the pellet was stored at -80°C until it was required for protein purification.

2.3.2 Recombinant protein purification

2.3.2.1 Preparation of nickel-nitrilotriacetic acid (Ni-NTA) beads

Nickel-nitrilotriacetic acid (Ni-NTA) beads (Qiagen) were prepared by spinning 1 volume of beads for 2 minutes at 1500 rpm and 1 volume of beads and 1 volume of ddH_2O at 1500 rpm for 2 minutes, twice. 1 volume of lysis buffer 0 (500 mM NaCl, 50 mM HEPES pH 7.5, 5 % (w/v) glycerol and 10 mM imidazole) was added to the beads and spun 3 times (1500 rpm for 2 minutes).

2.3.2.2 50 mL protein purification test

2.3.2.2.1 Nickel-immobilised metal ion affinity chromatography (Ni-IMAC)

Frozen pellets from 50 mL *E. coli* protein expression were thawed briefly at 37°C. A stock of 5 x lysis buffer was prepared (2.5 M sodium chloride, 250 mM HEPES (pH 7.5), 25 % glycerol, 5 mM phenylmethylsulfonyl fluoride (PMSF), 5 mM benzamidine and 20 mM β-mercaptoethanol) and the pellets were resuspended in 3 mL lysis buffer 1 (1 x lysis buffer, 10 mM imidazole (pH 8.0) and 1250 units (U) per 100 mL Basemuncher nuclease (Expedeon, Swavesey, UK). Samples were lysed by sonication, on ice for 2 minutes (LB) or 3 minutes (TB) (20 seconds on, 20 seconds off), split into centrifuge tubes and spun at 13500 rpm for 30 minutes at 4°C.

3 mL of clarified supernatant was transferred into a 15 mL tube with 3 mL lysis buffer 2 (1 x lysis buffer and 10 mM imidazole pH 8.0) and 200 µL of 50 % nickel-nitrilotriacetic acid (Ni-NTA) beads (Qiagen). Samples were batch bound to the beads by incubating at 8°C at 100 rpm for 30 minutes. The lysate was transferred to a gravity flow column and allowed to drip through at 4°C. 5 x 2 mL wash buffer (1 x lysis buffer and 30 mM imidazole pH 8.0) was added to the

columns and allowed to drip through. Proteins were eluted in 200 µL elution buffer (1 x lysis buffer with 500 mM imidazole pH 8.0) and stored at 4°C until analysis by SDS-PAGE.

2.3.2.3 Large-scale protein purification

2.3.2.3.1 Ni-IMAC

Frozen pellets containing overexpressed protein (from 1 L or greater culture) were prepared similarly to 50 mL test expression samples (section 2.3.2.2) however, pellets were resuspended in lysis buffer (50 mL/1 L TB pellet) with addition of 0.5 mg/ml lysozyme (50 mM HEPES pH 7.5, 500 mM NaCl, 5 % glycerol, 10 mM imidazole, 4 mM βME with the addition of phenylmethylsulfonyl fluoride (PMSF), benzamidine, 1250 units (U) per 100mL Basemuncher nuclease and 0.5 mg/ml lysozyme). Samples were sonicated on ice for 15 minutes in 50 mL batches (10 seconds on, 10 seconds off) and centrifuged at 33000 x g for 1 hour at 4°C.

The Ni-NTA IMAC column containing 2 mL bed volume of Ni-NTA beads was equilibrated using 10 column volume (CV) of ddH_2O and 10 CV lysis buffer (no lysozyme/PMSF/benzamidine/Basemuncher). The supernatant from the clarified lysate was applied to the column and the flow through was collected. The column was washed with 10 CV lysis buffer, 10 CV lysis buffer with 30 mM imidazole (W30). Protein was eluted with 4 ml lysis buffer containing 300 mM imidazole (E300) in 2 CV fractions (Figure 2.1). Small samples were collected at each wash and elution step and 20 µL of each sample was resolved by SDS-PAGE to assess yield and purity.

2.3.2.3.2 Tobacco Etch Virus (TEV) cleavage and Ni-IMAC rebind

Protein elutions from Ni-IMAC were pooled and dialysed overnight at 4°C with the addition of 0.1 mg/mL 6xHis tagged TEV protease, which was produced in house from the expression plasmid (From Opher Gileadi, The University of Oxford). Dialysis was performed using Snakeskin dialysis membrane (3500 Da MWCO, Fisher) with at least a 1:100 volume ratio of

protein: dialysis buffer (5 % glycerol, 20 mM HEPES, 500 mM NaCl made with ultra-pure MilliQ ddH_2O).

The Ni-IMAC was equilibrated as before (section 2.3.2.2.1) and dialysed protein was applied to the column. Lysis buffer containing 20 mM, 40 mM, 100 mM and 300 mM imidazole were used to collect 2 CV elution fractions. The elutions were resolved on a 10 % TGS SDS-PAGE gel (section 2.3.3).

2.3.2.3.3 Ion exchange (IEX) chromatography

Pooled protein following TEV cleavage (section 2.3.2.3.2) was diluted 5-fold in dilution buffer (20 mM HEPES pH 7.5 and 5 % glycerol) to reduce the NaCl concentration to 100 mM and was concentrated to 5 mL in either a 30, 10 or 3 kDa MWCO concentrator, depending on the size of the protein. A 1 ml HiTrap Q-IEX/SP-IEX column (GE Healthcare, UK) was washed with ddH_2O at 1 mL/min on an ÄKTA Pure fast protein liquid chromatography (FPLC) system (GE Healthcare) and equilibrated with 20 CV buffer A (20 mM Tris/HEPES pH 7.5, 100 mM NaCl, 5 % glycerol and 1 mM DTT).

The column was washed with buffer B (20 mM Tris pH 7.5, 1000 mM NaCl, 5 % glycerol and 1 mM DTT) and re-equilibrated with buffer A. 5 mL of concentrated protein was loaded onto a washed 5 mL capillary loop and the protein loaded at 1 mL/min. Protein was eluted using a 0-100 % buffer B gradient (100-1000 mM NaCl) with 2 mL fractions. The buffers used depended on the column used for IEX. 20 mM Tris pH 7.5 was used for Q-IEX column and 20 mM HEPES pH 7.5 was used for SP-IEX column. 20 µL of each protein elution was resolved on a 10% SDS-PAGE gel (section 2.3.3).

2.3.2.3.4 Size exclusion chromatography (SEC)

Pooled proteins from IEX chromatography were concentrated to 5 mL using either a 30, 10 or 3 kDa MWCO concentrator, depending on the size of the protein. A HiLoad 16/600 Superdex

200 pg column (GE Healthcare) was washed with ddH_2O at 1 mL/min and equilibrated with SEC buffer (20 mM HEPES, 500 mM NaCl, 5 % glycerol and 1 mM DTT) on an ÄKTA Pure FPLC system. 5 mL of protein was loaded into a washed 5 mL capillary loop, loaded and resolved at 2 mL/min. Proteins were eluted in 2 mL fractions and 20 µL was resolved on a 10% SDS-PAGE gel (section 2.3.3).

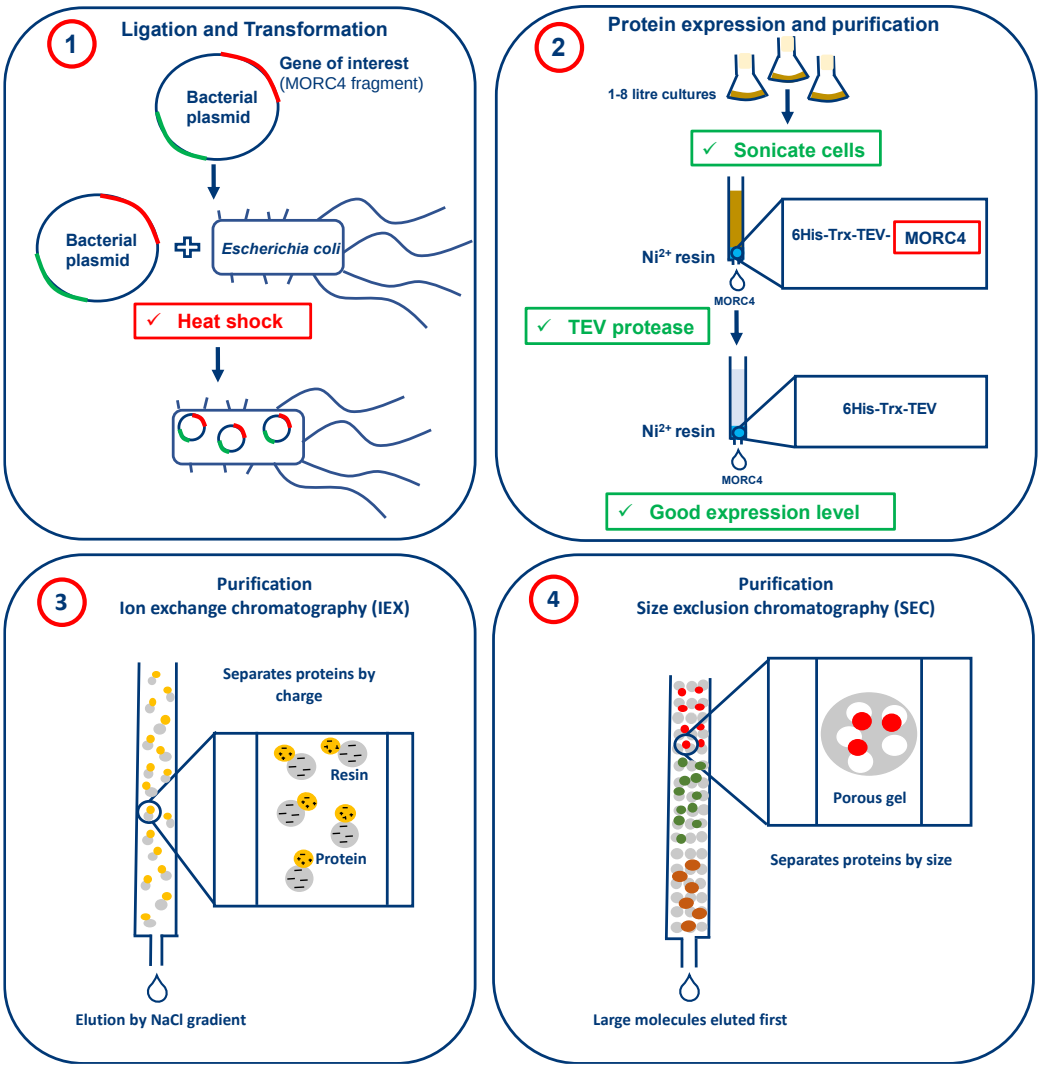


Figure 2.1: **Schematic representation of an overview of the generalised large-scale protein expression and purification procedure of MORC4.** Representation demonstrates each purification step from protein expression, Ni-IMAC affinity and details of ion exchange chromatography (IEX) and size exclusion chromatography (SEC).

2.3.3 Sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE)

Loading buffer was prepared with 4 x NuPAGE LDS sample buffer (ThermoFisher) and 200 mM dithiothreitol (DTT). Loading buffer was diluted to 1 x with sample and heated for 5 minutes at 96°C. 10 µL of Precision Plus protein dual colour standards protein ladder (Bio-Rad, Hercules, CA, USA) and 20 µL of sample was loaded and resolved on either precast 4-20 % gradient or 10% tris-glycine gels (Table 2.2) in 1 x tris-glycine-SDS (TGS) (10x TGS

composition: 250 mM Tris, 1.92 M glycine and 1 % SDS) buffer at 200 V for 50 minutes (ThermoFisher Scientific).

The gels were visualised with Instantblue protein stain (Expedeon, Heidelbery, Germany) overnight and destained with deionised water for 3 x 10-minute washes and 1 x 2-hour wash. Alternatively, gels were stained with Coomassie blue stain (0.1 % w/v Coomassie Blue R-250, 50 % v/v methanol and 10 % v/v acetic acid)/destain solutions (50 % v/v methanol and 10 % v/v acetic acid). The gels were imaged under white light using an InGenius gel viewer (Syngene).

Table 2.2: **TGS SDS-PAGE gel recipe** for 10% SDS-PAGE.

Component	10% gel	4% stacking gel
ddH_2O	19.4mL	9.5mL
1.5 M Tris pH 8.8	10mL	-
0.5 M Tris pH 6.8	-	3.78mL
10% SDS	400µL	150µL
40% acrylamide	10mL	1.5mL
TEMED	20µL	15µL
10% APS	200µL	75µL
Total Volume	**40mL**	**15mL**

2.3.4 Protein quantitation

2.3.4.1 Spectrophotometric determination

Pure protein was concentrated to ~200 µL using either a 30, 10 or 3 kDa MWCO concentrator, depending on the size of the protein. Protein was quantified using a nanodrop 2000 (ThermoFisher Scientific) and the 1 Abs = 1mg/ml default setting which measured absorbance at 280 nm.

2.3.4.2 Bicinchoninic acid assay (BCA)

Protein quantification was also determined using the bicinchoninic acid assay (BCA) protein assay kit. Several bovine serum albumin (BSA) dilutions of known concentration were prepared (Table 2.3) from a 2 mg/ml stock, which were used as controls for the BCA assay. The BCA assay working reagent (WR) was prepared by mixing 50 parts of reagent A and 1 part of reagent B. Using a flat bottomed 96-well plate, 25 µL of each standard of BCA were pipetted into the plate, in triplicate. Each sample of cell lysate or protein solution was diluted accordingly, highly concentration samples were diluted 1 in 12 with lysis buffer. 200 µL of WR was added to each sample and mixed using a KS250 IKA labortechnik orbital shaker (IKA, Staufen im Breisgau, Germany) for 30 seconds. The plate was incubated at 37°C for 30 minutes on a digital block heater. The 96-well plate was cooled the room temperature and the absorbance was measured at 562 nm using a BMG LabTech Spectrostar Nano plate reader (Offenburg, Germany).

Table 2.3: **BSA standard dilutions** required for the BCA assay protein quantification kit.

BSA standard (mg/ml)	BSA (µL)	ddH_20 (µL)
0	0	100
0.2	10	90
0.4	20	80
0.8	40	60
1.0	50	50
1.4	70	30
2.0	100	0

2.3.4.2.1 BCA assay analysis and preparation for western blotting

Following the BCA assay, the absorbance readings were analysed and were used to calculate the concentration of protein in each sample. The concentration of each sample was used to calculate the volume of each sample required to load 20 µg of protein on an SDS-PAGE gel required for western blotting. Samples were prepared with the addition of lysis buffer to dilute

the protein and 1 x loading dye (4 x NuPAGE LDS sample buffer and 200 mM dithiothreitol (DTT).

2.3.5 Protein storage

Pure protein was flash-frozen in liquid nitrogen and stored at -80°C until required for further analysis.

2.3.6 Western blotting

2.3.6.1 SDS-PAGE

SDS-PAGE gels were performed similarly to section 2.3.3 however, 1 µL fluorescent PageRuler (ThermoFisher) prestained protein ladder was used as a marker.

2.3.6.2 Transfer

The blotting cassette sponges and filter paper from the mini trans-blot cell (Bio-Rad) were fully submerged in 1 x transfer buffer (10 x transfer buffer and 10 % methanol). Immobilon™-FL polyvinylidene difluoride (PVDF) membrane (Merk Millipore, Darmstadt, Germany) was activated in methanol for 10 seconds, washed in ddH_2O for 30 seconds and then transfer buffer for 30 seconds. The blotting cassette (Bio-Rad) was loaded in the following order: sponge, filter paper, gel, activated PVDF membrane, filter paper and sponge. The cassettes were loaded into the tank with a magnetic stirrer and an ice pack. The tank was filled with transfer buffer and the voltage was set to 100 V for 90 minutes with stirring.

2.3.6.3 Fluorescent detection of western blots

2.3.6.3.1 Blocking and antibody staining

All antibodies were from Abcam, Cambridge, UK, unless otherwise stated. The membrane was blocked overnight at 4°C or for 1 hour at room temperature on a KS250 IKA labortechnik orbital shaker with 5 % skimmed milk and Tris-buffered saline (TBS). The membrane was washed 3 x 5 minutes in 0.1 % Tris-buffered saline-tween (TBS-T) and incubated in a primary

antibody (Table 2.4) with 5 % milk TBS-T for 1 hour at room temperature or overnight at 4°C, on the orbital shaker.

The membrane was washed 3 x 5 minutes with TBS-T and incubated with fluorescently labelled secondary antibody (1:20,000 of anti-mouse or anti-rabbit; Table 2.4) in 5 % milk TBS-T for 1 hour at room temperature on the orbital shaker. The membrane was washed 3 x 5 minutes with 0.01 % sodium dodecyl sulphate (SDS) and TBS/T for 5 minutes, then a final wash in TBS.

Table 2.4: **Antibodies used for fluorescent western blotting** including details of primary and secondary antibodies, dilutions and incubation procedures.

Catalogue Number	Antibodies		Dilutions		Incubation details	
	Primary	**Secondary**	**Primary**	**Secondary**	**Primary**	**Secondary**
f3165	**FLAG**	Anti-mouse 680nm	1:20,000	1:20,000	1 hr, room temp	1 hr, room temp
ab8224	**β-actin**	Anti-mouse 790nm	1:20,000			
C29F4	**HA**	Anti-rabbit 680nm	1:1,000			
ab220788	**TBP**	Anti-rabbit 790nm	1:2,000			
NB-34-0669-36	**MORC1**	Anti-rabbit 680nm	1:1,000	1:20,000	1 hr, room temp	1 hr, room temp
CSB-PA896555LA01HU	**MORC2**		1:1,000			
STJ117609	**MORC3**		1:2,000			
HPA000395	**MORC4**		1:1,000		Overnight, 4°C	
Alison Banham, The University of Oxford	**MORC4 161A/3** **MORC4 97F/9**	Anti-mouse 680nm	range (including neat)	1:20,000	Overnight, 4°C	1 hr, room temp

2.3.6.3.2 Protein detection

Following western blotting and antibody incubation, the membrane was imaged on a Li-Cor 9120 Odyssey infrared system (Li-Cor, Lincoln, NE, USA). The intensity of the fluorescent detection was optimised manually, by adjusting the exposure, to ensure protein detection without over saturation of the signal.

2.3.6.4 Enhanced chemiluminescence (ECL) detection of western blots

2.3.6.4.1 Blocking and antibody staining

Membranes were blocked and primary antibody stained similar to fluorescent detection western blotting (Section 2.3.6.3) however, either 5 % milk and BSA were used for incubation of antibodies and horseradish peroxidase (HRP) goat-anti-mouse was used as the secondary antibody (Table 2.5).

2.3.6.4.2 Protein detection

To detect proteins, a BM enhanced chemiluminescence (ECL) blotting substrate (POD) kit (Roche, Basel, Switzerland) was used. Membranes which had been previously blocked, and antibody stained, were placed on plastic film within a cassette and any excess TBS was removed. Blotting substrate and developing reagent were mixed in a radio of 1:100 and a thin coating was applied to each membrane for 1 minute. Any excess solution was removed and placed onto plastic film within a cassette.

In a dark room, membranes were exposed to autoradiography film for 1 hour (GE healthcare, Chicago, Illinois, United States) and were full immersed into multigrade developer (Ilford, Cheshire, United Kingdom) for 2 minutes until protein bands appeared. Following exposure to the developer, the film was fully immersed briefly into dH_2O, rapid fixer (Ilford) for 2 minutes and then left to dry overnight.

Table 2.5: **Antibodies used for enhanced chemiluminescence (ECL) western blotting** including details of secondary antibodies, dilutions and incubation procedures.

	Antibodies		Dilutions		Incubation details	
	Primary	**Secondary**	**Primary**	**Secondary**	**Primary**	**Secondary**
Monoclonal	**MORC4 161A/3** **MORC4 97F/9**	HRP Anti-mouse	1:500	1:20,000	Overnight, 4°C OR 2 hrs room temp	1 hr, room temp

2.4 Protein structural and biophysical analysis

2.4.1 Mass spectrometry (MS) analysis

Mass spectrometry (MS) analysis was performed on either gel bands (MSMS) or concentrated purified protein samples (intact MS). For MSMS tryptic identification, gel bands were excised from an SDS-PAGE gel and stored in 200 µL 10 % methanol. For intact MS analysis, 0.02 mg/mL of protein was prepared in 0.1 % formic acid. The samples were analysed by Dr. Rod Chalk at the Structural Genomics Consortium, University of Oxford.

2.4.2 X-ray crystallography

2.4.2.1 Protein crystallography trials and optimisation

Purified MORC4 protein was defrosted quickly in lukewarm water and centrifuged at 13500 rpm for 5 minutes in a microfuge at 4°C, to remove precipitated protein. The protein was diluted 1 in 4 in dilution buffer (20 mM HEPES pH 7.5 and 5 % glycerol and 1 mM DTT) and incubated with 1 mM adenylyl-imidodiphosphate (AMP-PNP) ligand (Bio-Techne, Minneapolis, Minnesota) and 1.25 mM $MgCl_2$. Some protein was also diluted but in the absence of AMP-PNP/$MgCl_2$. Both samples were incubated at 4°C for 2 hours and concentrated to ~3 mg/ml each. Some samples were also concentrated prior to incubation with AMP-PNP.

50 µL mother liquor from the JCSG-*plus*™ crystal screen, PACT *premier*™ screen, The BCS screen, Morpheus® screen (Molecular Dimensions, Sheffield, UK) or various optimisation screens were aliquoted into a 2-drop SwissCi polystyrene medical research council (MRC) crystallisation 96-well plate™ (Molecular Dimensions). Plates were prepared by the sitting drop vapour diffusion method, mixing 1 µl protein with 1 µl mother liquor (1:1 ratio) and sealed using ClearVue™ sealing sheets (Molecular Dimensions) by applying pressure to the seal around the edge of the plate. Plates were incubated at 20°C until crystals were visible under a microscope.

Following crystal production from the initial screen kits, several optimisation screens were designed in an attempt to increase the number, size and quality of the crystals. The conditions of crystals which formed from commercial kits were analysed and used to design optimisation plates, where the conditions not only included the original condition that crystallised the protein but also with ranges of concentrations of the individual condition components.

2.4.2.2 Mounting crystals

Following examination of the chosen crystal, an appropriate size round cryo loop (Molecular Dimensions) was chosen to remove the crystal. If the conditions of the crystal did not contain any cryo-protectant (e.g. glycerol or polyethylene glycol (PEG)) then 30 % glycerol was added to the well. A scalpel was used to perforate the sealing sheet under a microscope and a magnetic cryo wand with the cryo loop attached was used to extract the crystal and ensure that it was contained within the loop. If the well had a thin skin-like substance, an acupuncture needle was used to remove it prior to extracting the crystal. Cryo tongs were cooled in liquid nitrogen and the crystal was fully immersed into liquid nitrogen. Once the tools had cooled, the crystal on the cryo-loop which was connected the cryo wand, was attached to the cryo tongs and was mounted onto the goniometer head.

2.4.2.3 X-ray diffraction screening

Prior to X-ray diffraction screening, the X-ray diffraction detector D8 venture (Bruker, Billerica, Massachusetts, United States) was switched on several hours prior to use, to allow the cryostream 800 (Oxford Cryosystems, Long Hanborough, United Kingdom) to cool the goniometer. A copper (Cu) anode was required to enable the production of the X-rays. Following the mounting of the crystal cryo loop onto the goniometer head, APEX3 software was used to collect data. The crystal was aligned using a goniometer key and horizontal and vertical adjustments were essential. The crystal was spun at 90° and 180° angles until all angles of the crystal were centred. These alignment steps were vital to ensure the crystal was in the X-ray beam. Crystals were screened with a distance of 77mm, Chi (χ) 30° and Phi (φ) 180°.

2.4.3 Analytical size exclusion chromatography

A HiLoad 100/300 Superdex 200 pg column (GE Healthcare life sciences) was washed with ddH_2O at 1 mL/min and equilibrated in low salt SEC buffer (20 mM HEPES, 100 mM NaCl, 5 % glycerol and 1 mM DTT) on the ÄKTA pure system. Prior to injecting protein samples, a control marker gel filtration standard of known molecular weight (Bio-Rad) were diluted 1 in 5 in low salt SEC buffer and injected onto the column. Proteins were defrosted as in section 2.4.2.1 and were either kept in high salt SEC buffer or diluted 1 in 3 in low salt SEC buffer. Some protein samples were treated with 1 mM AMP-PNP ligand and 1.25 mM $MgCl_2$ and incubated at 4°C for 1 hour. All protein samples were injected into the HiLoad 100/300 Superdex 200 pg column for analytical size exclusion chromatography and protein was determined using the ultraviolet (UV) trace at 280 nm.

2.4.4 Size exclusion chromatography-Multiangle laser light scattering (SEC-MALS)

A Superdex 200 5/150 increase column (GE Healthcare life sciences) was washed with ddH_2O at 0.2 mL/min and was equilibrated in low salt size exclusion chromatography-multiangle light scattering (SEC-MALS) buffer (20 mM HEPES, 150 mM NaCl and 5 % glycerol). Protein

samples were purified (Section 2.3.2.3), defrosted (Section 2.4.2.1) and either diluted 1 in 3 with dilution buffer (20 mM HEPES pH 7.5 and 5 % glycerol, 1 mM DTT) or remained as concentrated protein. Some protein samples were incubated with 1 mM AMP-PNP ligand and 1.25 mM $MgCl_2$. 30 µL of protein was loaded onto the ultra-high-pressure liquid chromatography (UHPLC) system (Shimadzu, Kyoto, Kyoto Prefecture, Japan) and 25 µL was injected at 4°C for SEC. Following SEC, the multiangle laser light scattering (MALLS) detector DAWN 8+ (Wyatt Technology, Santa Barbara, California, United States), refractometer with extended range (T-rEX) (Optilab, Phoenix, Arizona, United States) and a RF-20A spectrofluorometric detector (Shimadzu) were used to determine mass and structural properties of proteins. ASTRA 6 software (Wyatt Technology) was used to analyse the data and there was 5 %± error in calculations of the mass of proteins. All SEC-MALS experiments were performed with the assistance of Dr. Maren Thomsen, The University of Leeds.

2.5 Biochemical analysis techniques

2.5.1 ATPase assays

2.5.1.1 Malachite green phosphate detection assay

The malachite green phosphate assay kit (Sigma, St. Louis, Missouri, United States) colorimetric test was used to quantify inorganic phosphate (Pi) release to determine ATPase activity. To create a standard curve, several phosphate (KH_2PO_4/K_2HPO_4) stocks were prepared ranging from 0-50 µM from a 1 mM stock of KH_2PO_4/K_2HPO_4 pH 7.5 (Table 2.6). 80 µL of each phosphate standard was dispensed into a flat bottomed 96-well plate (Sarstedt, Newton, North Carolina).

Table 2.6: **Phosphate standard dilutions** required for the ATPase assay phosphate quantification colorimetric kit.

[Phosphate] (µM)	Phosphate stock 50µM (µL)	ddH_20 (µL)
50	400	0
40	320	80
32	256	144
24	192	208
16	128	272
12	96	304
8	64	336
0	0	400

For ATPase samples of an unknown phosphate concentration, 1 x base buffer was prepared consisting of 10 mM $MgCl_2$, 100 mM NaCl, 1 mM DTT and 50 mM Tris pH 7.5. 35 µL of base buffer was dispensed into the flat-bottomed 96-well plate. 500 nM control protein, RecA (New England Biolabs), was also prepared at 4°C, in 1 x base buffer ranging from 0.1-10 µM, in 10 µM increments. Each sample of RecA had the addition of 5 µL single stranded DNA (ssDNA) (10 µM of 40 bp oligonucleotide, MORC4-r006 primer), double stranded DNA (100 ng/µL plasmid DNA, pcDNA3-N-HA-LIC) (dsDNA) or no DNA. 1 µM of oligo ssDNA such as a primer, was added to the RecA protein. 10 ng/µL plasmid DNA was used for dsDNA and nuclease-free H_2O was used for samples containing no DNA.

5 µL protein prepared from serial dilutions was added to each sample along with 50 µM ATP (Sigma) from a 100 mM stock (made with 1 M Tris pH 7.5, pH adjusted with 3 M NaOH and 1 M $MgCl_2$ was added in a ratio of 1:3). The 96-well plate was incubated at room temperature for 30 minutes and the reaction was quenched with the addition of 22.5 mM EDTA.

For both protein and phosphate standards samples, working reagent was prepared. Working reagent consisted of freshly prepared reagent B and reagent A in a ratio of 1:100. 20 µL working reagent was added to each well and a pipette was used to mix. For colour development,

the 96-well plate was incubated for 30 minutes at room temperature and the absorbance was measured at 620 nm using a BMG LabTech Spectrostar Nano plate reader (Offenburg, Germany).

2.5.2 Protein chemical crosslinking

Purified protein was defrosted in lukewarm water and centrifuged for 5 minutes at 13500 rpm in a microfuge to eliminate any precipitated protein. Protein was diluted to 2 mg/ml in SEC buffer (20 mM HEPES, 500 mM NaCl, 5 % glycerol and 1 mM DTT). WT protein and 3 mutants (K124R, N60A and D88A) were used for cross-linking in the presence of no ligand, AMP-PNP (Bio-Techne) and ADP (Sigma) each with bis(sulfosuccinimidyl)suberate (BS^3) crosslinker (Fisher) ranging from 0 nM to 0.1 nM, in 0.05 nM increments. Pre-weighed 2 mg aliquots of BS^3 were used with each aliquot dissolved in 1 x phosphate buffer saline (PBS) to make a 50 mM stock of BS^3. Protein crosslinking was also examined with glutaraldehyde (Sigma), which was diluted from a 50 % stock to a 1 % stock also in 1 x PBS. Glutaraldehyde dilutions ranged from 0 % to 1 %.

Proteins were diluted 1 in 5 with dilution buffer (20 mM HEPES and 5 % glycerol only) in an attempt to allow ligands to bind to the protein. To remove precipitants, samples were centrifuged for 5 minutes at 13500 rpm in a microfuge. Each diluted protein was split into 3 aliquots to account for BS^3/glutaraldehyde dilutions and supplemented with 1 mM AMP-PNP or ADP ligand and 1.25 mM $MgCl_2$. Samples were incubated at 4°C for 1 hour.

2.5 µL of each BS^3/glutaraldehyde dilution for each sample of protein was dispensed into PCR tubes. 10 µL of protein, which had been previously diluted in low salt SEC buffer and treated with ligand/Mg^{2+}, was added to the BS3/glutaraldehyde dilution. The protein/ligand/crosslinker mixture was incubated at 4°C for 60 minutes, the reaction was quenched with 250 mM Tris pH 7.5 and incubated for 15 minutes at room temperature. 1 x

NuPAGE LDS sample buffer (containing 50 mM DTT) was added and incubated at 95°C for 2 minutes on a digital block heater. Samples were analysed using SDS-PAGE (Section 2.3.2).

2.6 Cell biology techniques

2.6.1 Cell lines

Several cells lines (Appendix 8.3) were maintained at 37°C with 5 % CO_2 in either Dulbecco's modified eagle medium (DMEM) (without glutamine), Roswell Park Memorial Institute medium (RPMI), a mixture of both DMEM and RPMI or Minimum Essential Medium Eagle (MEM-E). All media was supplemented with the addition of 100 µg/mL penicillin, 100 µg/mL streptomycin, glutamine and 10 % fetal bovine serum (FBS), prior to use. Cell lines were maintained and grown with additional support from Dr. Christopher Cooper (The University of Huddersfield).

2.6.2 Transfection

2.6.2.1 Seeding cells

To prepare cells for transfection, U2OS, HEK293FT, HeLa or COS-1 cell lines were seeded at a density of 0.3 x 10^6 in a 6 well plate or 2.2 x 10^6 in a 10 cm^3 (Table 2.7) dish and incubated at 37°C with 5 % CO_2 for ~48 hours and until cells reached 60-70 % confluency.

2.6.2.2 Chemical-based transfection

The medium was removed from cells and full fresh medium was applied. Lipofectamine 2000 (Fisher) and plasmid DNA was prepared in serum free medium (medium without any antibiotics or FBS), as shown in table 2.7. The DNA was added to the lipofectamine samples, mixed and incubated at room temperature for 20 minutes. The DNA/lipofectamine mix was added to the cells dropwise, the cells were mixed and incubated at 37°C with 5 % CO_2 for 48 hours.

Table 2.7: **Preparation of lipofectamine and DNA samples for transfections** showing seeding density, volume of media, concentration of DNA, and volume of lipofectamine.

Plate or dish	Seeding density	Medium required in plate/dish (mL)	DNA (µg)	Medium required for DNA (mL)	Lipofectamine (µL)	Medium required for lipofectamine (mL)
6 well plate	0.3×10^6	3	1.5 or 2.5	0.125	3 or 1.5	0.125
10 cm^3 dish	2.2×10^6	10	12	1	20	1

2.6.2.3 Harvesting adherent cells

To harvest transfected HEK293FT cells, the medium was removed, and the cells were washed with enough phosphate buffer saline (PBS) to cover the bottom of the 6 well plate or 10 cm^3 dish (1 mL or 2 mL PBS, respectively). Cells were removed using PBS and centrifuged at 1000 x g for 5 minutes at 4°C. U2OS, HeLa and COS-1 cells were harvested similarly to HEK293FT cells however, 0.25 % trypsin-EDTA was used to remove the cells. Pellets were stored at -80°C until they were required for cell lysis.

2.6.3 Cell lysis

Transfected cell pellets harvested from a 6 well plate were resuspended in 100 µL lysis buffer or cell pellets at $1x10^7$ cells per mL, were resuspended in 500 µL lysis buffer unless, the cell pellet had less cells e.g. $0.5x10^7$, then 400 µL of lysis buffer was used for resuspending. (50 mM Tris pH 7.5, 150 mM NaCl, 1 % Triton, 1 mM DTT and 1 mini protease inhibitor tablet (ThermoFisher) per 10mL lysis buffer and any remaining lysis buffer was stored at -20°C). 250 units (U) BaseMuncher nuclease per 1 mL of lysis buffer was added on the day of use). Transfected cell pellets harvested from a 10 cm^3 dish were resuspended in 500 µL lysis buffer. Cell resuspensions were incubated at 4°C at 15 rpm for 15 minutes. Resuspensions were sonicated on ice for 20 seconds, 5 seconds on, 15 seconds off twice (4 times if the cells were

harvested from a 10 cm^3 dish) for transfected cells and 1 minute 20 seconds (5 seconds on, 15 seconds off x 4) for cell pellets at $1x10^7$ cells per mL. Cells were incubated for a further 15 minutes at 4°C at 15 rpm and then centrifuged at 13500 rpm at 4°C for 15 minutes in a microfuge. The supernatant was stored at -80°C until they were required for western blotting. Lysates required for protein and RNA analysis were quantitated using the BCA assay.

2.6.4 Preparation of cell lines for protein and RNA analysis

2.6.4.1 Harvesting suspension cell lines

Several suspension cell lines (Appendix 8.3) were maintained and grown until they reached approximately $1x10^6$ in a T175 cm^3 flask. If there were multiple T175 cm^3 flasks, they were pooled together, and a sample was taken, which was used to count the number of cells using a haemocytometer. Cells were harvested by gently centrifuging in 50 mL tubes at 200 x g for 5 minutes. Excess media was removed with a 1 mL pipette and resuspended in enough room temperature 1x PBS to give $1x10^7$ cells per mL (e.g. 200 mL of cells at $1x10^6$ gives $2x10^8$ cells in total. 20 mL of 1x PBS is required for $1x10^7$ cells per mL). Resuspended cells were transferred to 1 mL tubes and gently centrifuged at 200 x g for 5 minutes. The supernatant was removed, and the pellets were put on ice (4°C) immediately. Cell pellets were stored at -80°C until required for cell lysis or for downstream RNA analysis.

2.6.5 RNA analysis

2.6.5.1 Maintaining RNase free conditions

Prior to RNA extraction, the specified RNA-only bench area and pipettes, were cleaned with 1 M NaOH. This helped to eliminate any contaminants such as RNase or DNA. Filter pipette tips were RNase free and water used was diethyl pyrocarbonate (DEPC)-treated water. Gloves were also discarded regularly throughout experiments to reduce contamination.

2.6.5.2 Total RNA extraction and purification

Harvested suspension cell line (Section 2.6.4.1) pellets containing 0.5-1 x 10^7 cells were used for extracting RNA. RNA was extracted and purified using a RNeasy Mini Kit (Qiagen, Hilden, Germany). Total RNA samples were in the range of 500-1000 ng/µL and were stored at -80°C until required for use, to prevent degradation.

2.6.5.3 RNA quantitation and quality analysis

RNA was quantified using a nanodrop 2000 (260 nm) (ThermoFisher Scientific). The nanodrop was blanked using 2 µL nuclease-free water and 2 µL RNA was used for quantitation. The 260/280 nm ratio was used to assess purity of DNA (pure RNA is ~2.0).

2.6.5.4 First-strand cDNA synthesis

First-strand cDNA synthesis was performed under RNase free conditions (Section 2.6.5.1). For each sample, 0.5 µL of oligo$(dT)_{20}$, 5 µg total RNA (which had been previously diluted in RNAse-free water) and 1 µL 10 mM dNTPs (dATP, dGTP, dCTP, dTTP) mix was dispensed into RNAse-free PCR tubes. 6.5µL RNAse-free water was added to each tube. Samples were heated to 65°C for 5 minutes (in a digital block heater) and incubated at 4°C for 1 minute.

Following incubation, 4 µL 5 x first-strand buffer, 1 µL 0.1 M DTT, 0.5 µL RNaseIn and 0.9 µL SuperScript III reverse transcriptase (ThermoFisher) were added to each tube. The samples were incubated at 50°C for 60 minutes and the reaction was terminated by heating the samples to 70°C for 15 minutes. cDNA was transferred to a 1 mL tube and stored at -20°C until required for subsequent experiments.

2.6.5.5 Quantitative reverse transcription (qRT)-PCR

In a 96-well PCR plate, each well required 7.5 µL 2 x mix, 0.75 µL probe (MORC4, TBP or both probes) and 1.75 µL water. The probe mixture was mixed gently with a pipette and 5 µL cDNA (diluted 1 in 25) was added to each well in quadruplicate. The plate and well were

covered with a plastic seal and the plate was put into a Step-One Plus real-time PCR system (Applied Biosystems, Foster City, California, United States) for qPCR. Each cycle had a holding stage at 95°C for 2 minutes followed by a cycling stage which was 95°C for 10 seconds and then 60°C for 50 seconds. Each cycle was completed 40 times.

2.6.5.5.1 qRT-PCR analysis

Following completion of the qPCR cycles, Step-One Plus real-time PCR software was required for qPCR analysis. Using the software, any outliers were eliminated, and the data was exported into Microsoft excel. Each cycle threshold (Ct) value for the housekeeper gene (HKG) was averaged to give an average HKG Ct value. The cell line with the lowest expression (highest Ct value) was used as a normalisation control and to normalise other cell line RNA expression. The control Ct value was mean averaged and the ΔCt value was calculated by subtracting the HKG Ct value from Ct value from each cell line. The ΔΔCt was calculated by subtracting the ΔCt value of the control cell line from the ΔCt value of other cell lines. Finally, the relative quantification (RQ) value was calculated by $2^{-\Delta\Delta Ct}$ of each sample. The RQ value was important for determining the fold change in relation to the control sample i.e. the control cell sample RQ value was 1.0.

2.6.5.6 Reverse transcription (RT)-PCR alternative splice variants

To determine potential MORC4 splice variants in cell lines (Appendix 8.3), reverse transcription (RT)-PCR was used to amplify MORC4 DNA found in different cell lines using cDNA (Section 2.6.5.4). Each sample was prepared with 0.2 µL MyTaq DNA polymerase, 4 µL 5 x red buffer mix, 0.5 µM of each forward and reverse primer (MORC4_f1/r1, MORC4_f2/r2 or MORC4_f2/r3), 15.1 µL nuclease-free water and 0.5 µL cDNA. PCR products were resolved on an agarose gel similar to section 2.2.3, however, a 2% gel was required due to the size of PCR products. Each cycle consisted of 98°C for 5 minutes, 95°C for

30 seconds, 55°C 30 seconds and 72°C x min (1kb/min). Each cycle was completed either 35 or 40 times.

2.7 Bioinformatics analysis methods

2.7.1 Structural analysis

The MORC4 protein sequence was analysed via various bioinformatics analysis software methods to determine some structural characteristics of MORC4. Simple Modular Architecture Research Tool (SMART) and the Conserved Domains Database (CDD) were used to predict the domain architecture of MORC4. The bioinformatic analysis web software, GlobPlot was used to determine regions of MORC4 which were disordered or folded. Position specific iterative blast based secondary structure prediction (PSIPRED) analysis identified secondary structure details.

2.7.1.1 Homology modelling

The MORC4 protein sequence was utilised for homology modelling to predict the structure of MORC4. SWISS-MODEL (http://swissmodel.expasy.org) (Waterhouse *et al.*, 2018; Guex *et al.*, 2009; Studer *et al.*, 2020; Benkert *et al.*, 2008), the protein homology recognition engine 2 (Phyre2) (http://www.sbg.bio.ic.ac.uk/~phyre2) (Kelley *et al.*, 2015) and Robetta (https://robetta.bakerlab.org) (Song *et al.*, 2013; Raman *et al.*, 2009) bioinformatics web-servers were all used for protein homology modelling of MORC4. Predicted MORC4 models were analysed in UCSF Chimera (Goddard *et al.*, 2018).

2.7.2 RNA expression analysis

The RNA Expression Atlas by the European Bioinformatics Institute (EBI) (https://www.ebi.ac.uk) was used to analyse MORC4 expression levels in normal non-cancerous tissues and specific cell lines. The Gene Expression Database of Normal and Tumour Tissues 2 (GENT2) (http://gent2.appex.kr/gent2/) and the Gene Expression Profiling

Interactive Analysis (GEPIA) (http://gepia.cancer-pku.cn) databases were required to investigate expression of MORC4 in both normal non-cancerous and cancerous tissues. GENT2, PROGgeneV2 (http://genomics.jefferson.edu/proggene/) and Kaplan-Meier (KM) (https://kmplot.com/analysis/) databases were used for survival and prognosis analysis.

2.7.3 Protein sequence alignments

MORC4 protein sequences were identified using NCBI Protein BLAST (BLASTp) (Altschul *et al.*, 1990) in various vertebrate species. Protein sequences were uploaded to Jalview (Waterhouse *et al.*, 2009) for sequence alignment analysis and multiple sequence comparison by Log-expectation (MUSCLE) (Edgar, 2004) was utilised for the alignment. In addition, MORC (MORC1-3) human protein sequences were also aligned using MUSCLE in Jalview software.

CHAPTER 3

Results

Cloning, recombinant expression and purification of MORC4

Chapter 3 Cloning, recombinant expression and purification of MORC4

3.1 Introduction

Human microrchidia (MORC) protein MORC4 is the least characterised protein in the MORC family of proteins, with some evidence to suggest that it could be a diffuse large B cell lymphoma (DLBCL) biomarker (Liggins *et al.*, 2007) and a breast cancer oncogene (Yang *et al.*, 2019). To characterise MORC4, the protein required recombinant expression, as it is not economical or efficient to purify from a cell line culture. It is also not possible to purify from animal tissues as this project focuses on the human MORC4 protein. Therefore, *Escherichia coli* was used for recombinant protein expression in this project as it is well-established (Rosano and Ceccarelli, 2014), economical and often utilised for high-throughput recombinant protein expression (Jia and Jeon, 2016). Cloning, expression and purification are vital steps in helping to characterise the functions of MORC4, both structurally and biochemically.

Cloning multiple fragments of *MORC4* requires high-throughput methods, enabling a large array of domain boundaries to be considered, increasing the likelihood of finding one or more suitable, soluble truncated protein fragments. Several fragments of *MORC4* were designed in an attempt to increase the probability of soluble protein expression and the possibility of crystallisation, as defined protein boundaries are more likely to crystallise (Savitsky *et al.*, 2010). The cloning of multiple fragments is important as full length MORC4 may either express at low levels or it may not be soluble. Multiple soluble protein expression fragments are more likely to crystallise in X-ray crystallography trials and protein fragments can also be manipulated with multiple solubility tags to enhance expression (Savitsky *et al.*, 2010) (Cooper and Marsden, 2017).

Following cloning, small-scale test expressions identify fragments of MORC, which yield a good quantity of protein, and can be scaled up to multiple litres for expression and purification,

with the intention of producing milligram quantities of protein required for biochemical and structural analysis. The next step is identifying a fragment of MORC4 where the protein affinity tag, can be successfully cleaved, leaving a sufficient yield of protein. Overall, designing truncated fragments of *MORC4* is crucial in downstream characterisation of MORC4, for both biochemical and structural analysis.

3.2 High-throughput ligation independent cloning (LIC) creates fragments of MORC4

Ligation independent cloning (LIC) is a common method used for high-throughput cloning as it allows the use of the same PCR products for a range of expression vectors and it does not require the specific restriction site pairs used in conventional cloning (Savitsky *et al.*, 2010). LIC was utilised for creating fragments or full length *MORC4*, which were then used for protein expression and purification (Figure 3.1).

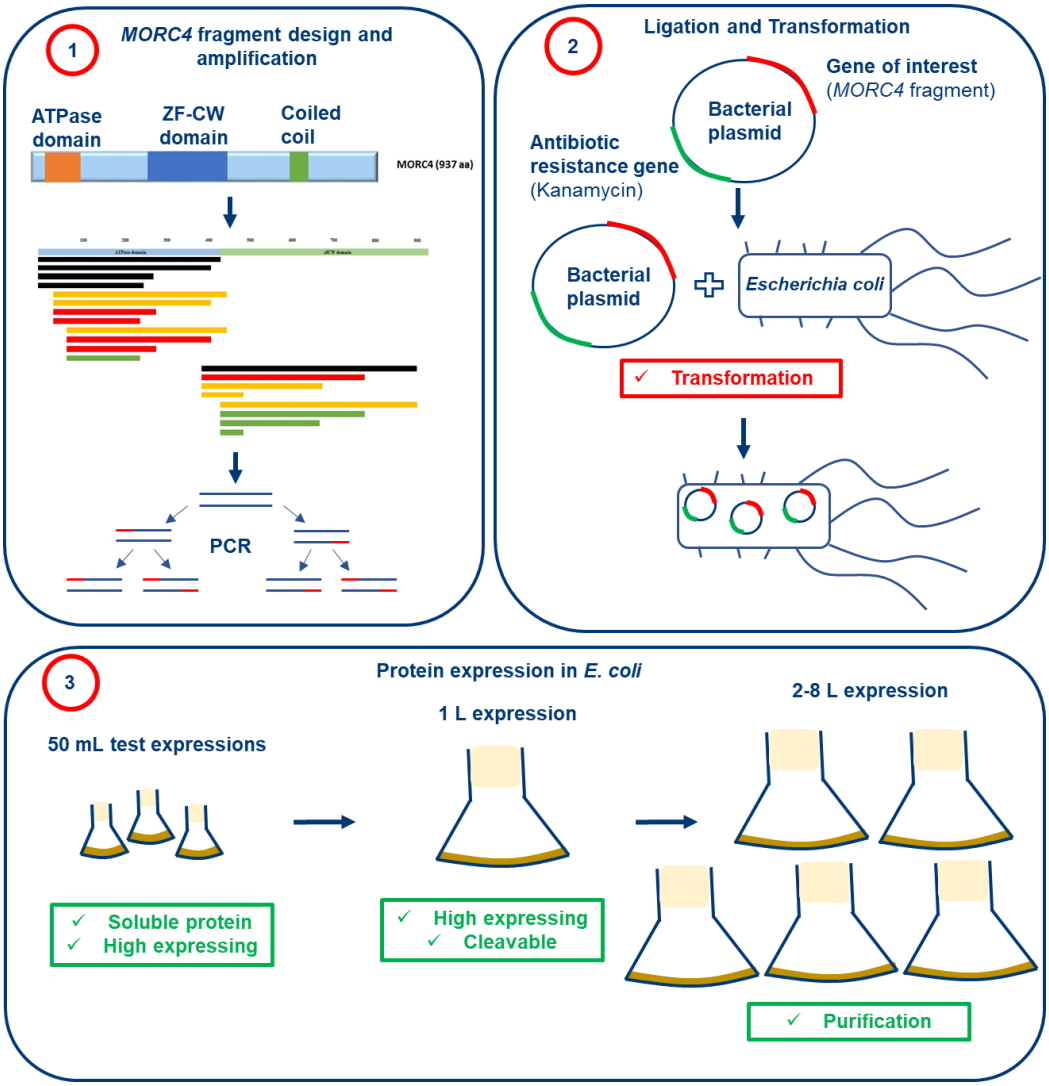


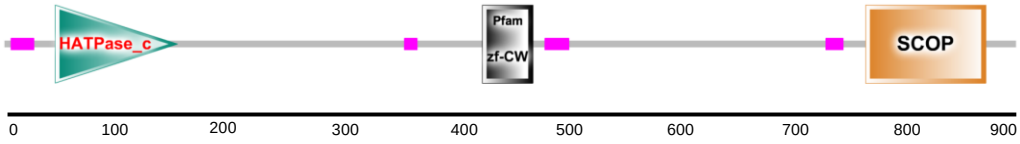
Figure 3.1: **Schematic representation of LIC and protein expression process.** *MORC4* fragments were designed and amplified using polymerase chain reaction (PCR). LIC was required to clone the gene of interest into *E. coli* expression vector with an antibiotic resistant gene. Plasmids were transformed into *E. coli* and were utilised for initial 50 mL test expressions. Fragments of MORC4 which were soluble and high expressing were used for 1 L protein expression to determine tag cleavage. Successfully tag-cleaved fragments were utilised for large-scale protein expression subsequent to purification of MORC4.

3.2.1 Bioinformatic analysis of MORC4 determines suitable domain boundaries

Truncated fragments of *MORC4* were designed by predicting soluble domain boundaries (Cooper and Marsden 2017) for high-throughput LIC cloning. Initially, a Simple Modular Architecture Research Tool (SMART) (Letunic *et al.*, 2015; Letunic and Bork, 2018) was used to predict the domains found in MORC4. The SMART analysis revealed that MORC4 has an

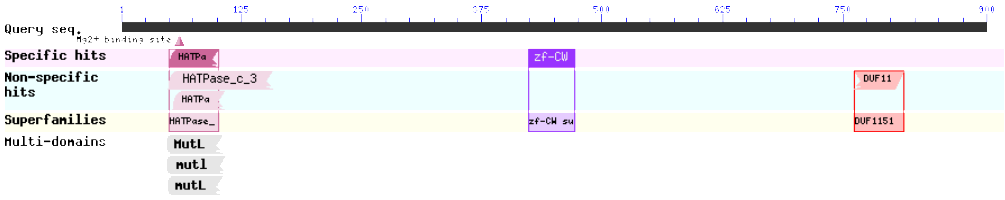
N-terminal ATPase domain known as the histidine kinase-like ATPase (HATPase) (Dutta and Inouye, 2000), which is found in several ATP-binding proteins such as histidine kinase (Song *et al.*, 2004). The SMART analysis also revealed that MORC4 contains a central Zf-CW domain (Hoppmann *et al.*, 2011) and a C-terminal coiled-coil domain (Truebestein and Leonard, 2016) (SCOP) (Figure 3.2a). In addition, the Conserved Domains Database (CDD) (Lu *et al.*, 2020) analysis also suggested that MORC4 has an N-terminal ATPase domain and a central Zf-CW domain. The CDD analysis also revealed that MORC4 has an N-terminal Mg^{2+} binding site, which could be important for ATP hydrolysis (Frick *et al.*, 2007) (Figure 3.2b).

A



Domain	Start (aa)	End (aa)
HATPase_C	45	154
Pfam zf-CW	425	470
SCOP	760	880

B

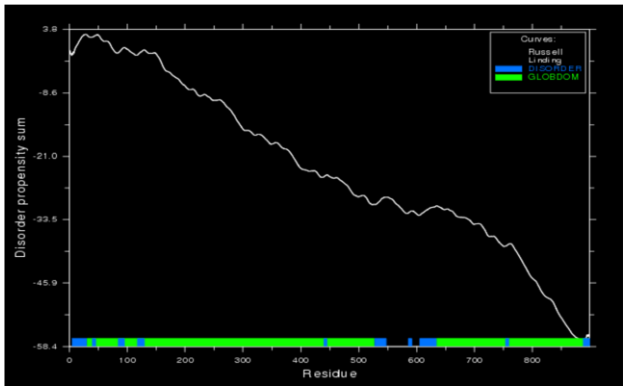


Domain	Start (aa)	End (aa)
HATPase_C	50	100
Pfam zf-CW	425	475

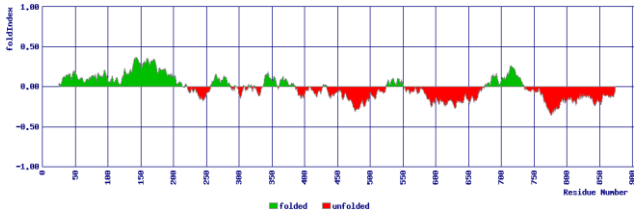
Figure 3.2: **Schematic representation of the predicted domain architecture of MORC4 following bioinformatic analysis. (A)** Predicted domain architecture of MORC4 following SMART analysis. Each domain prediction includes the associated domain boundary residue prediction. SMART analysis predicts an N-terminal ATPase domain, a central Zf-CW domain and a C-terminal coiled-coil domain. **(B)** Predicted domain architecture of MORC4 following CCD analysis. Each domain prediction includes the associated domain boundary residue prediction. The CCD analysis predicts an N-terminal ATPase domain and a central Zf-CW domain.

When designing fragments of a protein which are required for cloning and subsequent biochemical and structural experiments, it is important for the proteins to be designed in a well-folded and compact manner as disordered proteins are more difficult to crystallise (Cooper and Marsden 2017). GlobPlot is a protein analysis webserver which identifies regions of a protein that are potentially disordered or folded (Linding *et al.*, 2003). GlobPlot analysis on MORC4 revealed that there are several regions of disorder (Figure 3.3a). In particular, residues 540-640 were identified as a large region of disorder (Figure 3.3a) and smaller regions of disorder were predicted at residues 90, 120, 440 and 760 (Figure 3.3a). FoldIndex (Prilusky *et al.*, 2005) analysis also revealed areas of MORC4 which are disordered and therefore, not suitable for fragment design. More specifically, large regions of disorder included residues 552-674, residues 735-900 and residues 434-524. Smaller regions of disorder included residues 221-256, residues 324-334 and residues 386-427 (Figure 3.3b). However, residues 1-220 of MORC4 were well-folded suggesting that this region could be ideal when designing MORC4 fragments (Figure 3.3b). Overall, these results suggest that the N-terminal region of MORC4 is more structured than the C-terminal region. It also suggests that the C-terminus of MORC4 is more disordered than the N-terminal ATPase region. However, some regions of the C-terminus were predicated to be well-folded.

A



B




```
  1 MLLYRGAPAG PGAPGCGLAR PGGGPQAFGI RLSTMSPRYL QSNSSSHTRP
 51 FSAIAELLDN AVDPDVSART VFIDVEEVKN KSCLTFTDDG CGMTPHKLHR
101 MLSFGFTDKV IKKSQCPIGV FGNGFKSGSM RLGKDALVFT KNGGTLTVGL
151 LSQTYLECVQ AQAVIVPIVP FNQQNKKMII TEDSLPSLEA ILNYSIFNRE
201 NDLLAQFDAI PGKKGTRVLI WNIRRNKNGK SELDFDTDQY DILVSDFDTE
251 EKMTGGVTSE LPETEYSLRA FCGILYMKPR MKIFLRQKKV TTQMIAKSLA
301 NVEYDTYKPT FTNKQVRITF GFSCKNSNQF GIMMYHNNRL IKSFEKVGCQ
351 VKPTRGEGVG VIGVIECNFL KPAYNKQDFE YTKEYRLTIN ALAQKLNAYW
401 KEKTSQDNFE TSTVARPIPK VPDQTWVQCD ECLKWRKLPG KIDPSMLPAR
451 WFCYYNSHPK YRRCSVPEEQ ELTDEDLCLS KAKKQEQTVE EKKKMPMENE
501 NHQVFSNPPK ILTVQEMAEL NNKTIGYEGI HSPSVLPSGG EESRSPSLQL
551 KPLDSSVLQF SSKYKWILGE EPVEKRRRLQ NEMTTPSLDY SMPAPYRRVE
601 APVAYPEGEN SHDKSSSERS TPPYLFPEYP EASKNTGQNR EVSILYPGAK
651 DQRQGSLLPE ELEDQMPRLV AEESNRGSTT INKEEVNKGP FVAVVGVAKG
701 VRDSGAPIQL IPFNREELAE RRKAVESWNP VPYSVASAAI PAAAIGEKAR
751 GYEESEGHNT PKLKNQRELE ELKRTTEKLE RVLAERNLFQ QKVEELEQER
801 NHWQSEFKKV QHELVIYSTQ EAEGLYWSKK HMGYRQAEFQ ILKAELERTK
851 EEKQELKEKL KETETHLEML QKAQGFGKAY AATYPRQLSP YFCPPSLGAS
```

Figure 3.3: **MORC4 predicted regions of disorder following bioinformatic analysis. (A)** Predicted MORC4 disorder using GlobPlot analysis. Residues highlighted in green correspond to regions of putative domains, also known as GlobDom regions. Residues highlighted in blue indicate areas of protein disorder. **(B)** Predicted MORC4 disorder using FoldIndex analysis. Residues highlighted in red represent unfolded regions. Residues in green represent areas of folded MORC4.

Position specific iterative blast based secondary structure prediction (PSIPRED) (Buchan and Jones, 2019; Jones, 1999) bioinformatics analysis revealed details about the secondary structure of MORC4. PSIPRED indicated that there were several α-helices and β-strands present throughout MORC4, which should not be disrupted during cloning (Mooij *et al.*, 2009; Müller, 2017) (Figure 3.4). Overall, all of these bioinformatic analysis methods of MORC4 provide a good basis for the design of truncated fragments of *MORC4*, which will be used for subsequent cloning and protein expression.

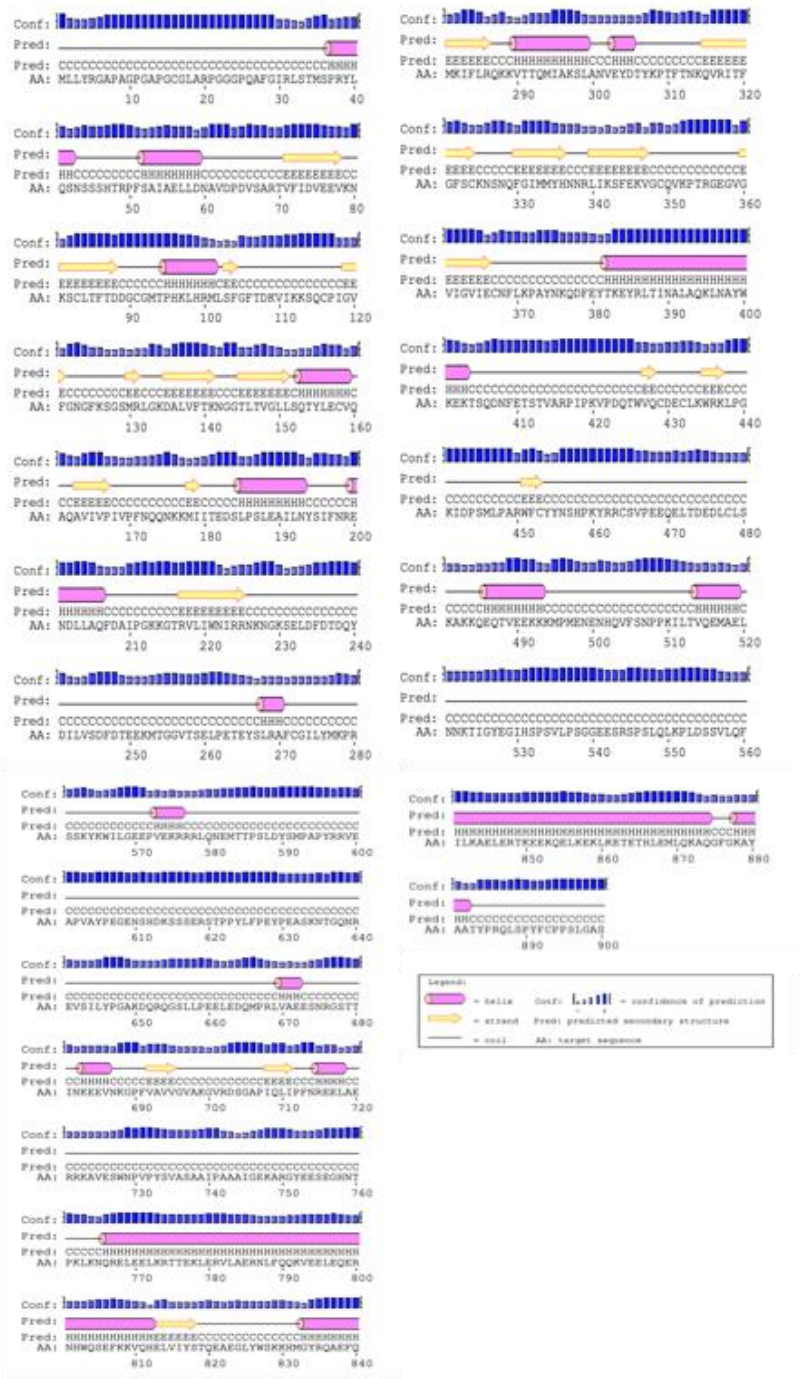


Figure 3.4: **MORC4 secondary structure prediction using bioinformatic analysis.** Predicted MORC4 secondary structure using PSIPRED. Residues highlighted in blue indicate the level of confidence of the prediction. 'Pred' represeents the predicted secondary structure of MORC4, which can either be represented as a α-helix, β-strand or coil.

3.2.2 Cloning of *MORC4* fragments in *Escherichia coli*

Following protein sequence analysis of MORC4, several domain boundaries were chosen on the basis of the two domains (ATPase and CW), with multiple start/stop positions chosen for each, generally demarcated by the presence of secondary structural elements (Figure 3.5). The full length MORC4 protein used to design MORC4 fragments had the less common C-terminus ending in 'SLGAS' residues instead of the more common MORC4 ending in 'HILD' residues. The full length *MORC4* DNA was provided by Professor A. Banham, The University of Oxford.

```
  1 MLLYRGAPAG PGAPGCGLAR PGGGPQAFGI RLSTMSPRYL QSNSSSHTRP
 51 FSAIAELLDN AVDPDVSART VFIDVEEVKN KSCLTFTDDG CGMTPHKLHR
101 MLSFGFTDKV IKKSQCPIGV FGNGFKSGSM RLGKDALVFT KNGGTLTVGL
151 LSQTYLECVQ AQAVIVPIVP FNQQNKKMII TEDSLPSLEA ILNYSIFNRE
201 NDLLAQFDAI PGKKGTRVLI WNIRRNKNGK SELDFDTDQY DILVSDFDTE
251 EKMTGGVTSE LPETEYSLRA FCGILYMKPR MKIFLRQKKV TTQMIAKSLA
301 NVEYDTYKPT FTNKQVRITF GFSCKNSNQF GIMMYHNNRL IKSFEKVGCQ
351 VKPTRGEGVG VIGVIECNFL KPAYNKQDFE YTKEYRLTIN ALAQKLNAYW
401 KEKTSQDNFE TSTVARPIPK VPDQTWVQCD ECLKWRKLPG KIDPSMLPAR
451 WFCYYNSHPK YRRCSVPEEQ ELTDEDLCLS KAKKQEQTVE EKKKMPMENE
501 NHQVFSNPPK ILTVQEMAEL NNKTIGYEGI HSPSVLPSGG EESRSPSLQL
551 KPLDSSVLQF SSKYKWILGE EPVEKRRRLQ NEMTTPSLDY SMPAPYRRVE
601 APVAYPEGEN SHDKSSSERS TPPYLFPEYP EASKNTGQNR EVSILYPGAK
651 DQRQGSLLPE ELEDQMPRLV AEESNRGSTT INKEEVNKGP FVAVVGVAKG
701 VRDSGAPIQL IPFNREELAE RRKAVESWNP VPYSVASAAI PAAAIGEKAR
751 GYEESEGHNT PKLKNQRELE ELKRTTEKLE RVLAERNLFQ QKVEELEQER
801 NHWQSEFKKV QHELVIYSTQ EAEGLYWSKK HMGYRQAEFQ ILKAELERTK
851 EEKQELKEKL KETETHLEML QKAQGFGKAY AATYPRQLSP YFCPPSLGAS
```

Figure 3.5: **MORC4 fragment boundaries. Chosen MORC4 fragments were determined following bioinformatic analysis.** The amino acid sequence indicates the N-terminal (red) positions of each fragment of MORC4 and the C-terminal (blue) positions of each fragment of MORC4 used in cloning (Accession number: EF125209.1). Full-length MORC4 was provided by Professor A. Banham, The University of Oxford.

Truncated fragments of *MORC4* were designed in three separate categories depending on the domain: ATPase domain only, CW zinc finger domain only and both domains in tandem (Table 3.1 and figure 3.6). Pairwise arrangements allowed multiple combinations of start/stop positions to be screened during subsequent cloning and expression. This project focuses on the MORC4 tandem domains as MORC3 (Li *et al.*, 2016) and MORC2 (Douse *et al.*, 2018) tandem domains were crystallised elsewhere, during the course of this project.

***Table 3.1: E. coli* plasmid constructs containing regions of the *MORC4* gene** including the size of each clone, PCR products (bp) and protein size after the addition of the protein tag (kDa).

Residue boundary (aa)	Forward primer	Reverse Primer	Length (bp)	Vector	PCR screen expected size (bp)	Protein size (kDa)
1-430	f000	r004	1290	pNH-TrxT	1902	60
1-404	f000	r005	1212	pNH-TrxT	1824	57
1-255	f000	r006	765	pNH-TrxT	1377	40
1-229	f000	r007	687	pNH-TrxT	1299	37
29-430	f001	r004	1206	pNH-TrxT	1818	57
29-40	f001	r005	1128	pNH-TrxT	1740	54
29-255	f001	r006	681	pNH-TrxT	1293	37
29-229	f001	r007	603	pNH-TrxT	1215	34
51-430	f002	r004	1140	pNH-TrxT	1752	55
51-404	f002	r005	1062	pNH-TrxT	1674	52
51-255	f002	r006	615	pNH-TrxT	1227	35
51-229	f002	r007	537	pNH-TrxT	1149	32
381-900	f003	r000	1560	pNH-TrxT	2172	72
381-757	f003	r001	1131	pNH-TrxT	1743	55
381-677	f003	r002	891	pNH-TrxT	1503	46
381-480	f003	r003	300	pNH-TrxT	912	24
420-900	f004	r000	1443	pNH-TrxT	2055	67
420-757	f004	r001	1014	pNH-TrxT	1626	50
420-677	f004	r002	774	pNH-TrxT	1386	42
420-480	f004	r003	183	pNH-TrxT	795	19
1-900	f000	r000	2700	pNH-TrxT	3312	114
1-757	f000	r001	2271	pNH-TrxT	2883	97
1-677	f000	r002	2031	pNH-TrxT	2643	88
1-480	f000	r003	1440	pNH-TrxT	2052	66
29-900	f001	r000	2616	pNH-TrxT	3228	111
29-757	f001	r001	2187	pNH-TrxT	2799	94
29-677	f001	r002	1947	pNH-TrxT	2559	86
29-480	f001	r003	1356	pNH-TrxT	1968	63
51-900	f002	r000	2550	pNH-TrxT	3162	109
51-757	f002	r001	2121	pNH-TrxT	2733	92
51-677	f002	r002	1881	pNH-TrxT	2493	83
51-480	f002	r003	1290	pNH-TrxT	1902	61
29-874	f001	r008	2538	pNH-TrxT	3150	109
29-469	f001	r009	1323	pNH-TrxT	1935	62
51-874	f002	r008	2472	pNH-TrxT	3084	106
51-469	f002	r009	1257	pNH-TrxT	3438	122
51-900	f002	r000	2550	pNIC28-Bsa4	2835	99
51-757	f002	r001	2121	pNIC28-Bsa4	2406	83
51-677	f002	r002	1881	pNIC28-Bsa4	2166	74
51-480	f002	r003	1290	pNIC28-Bsa4	1575	52
29-900	f001	r000	2616	pGTVL2	3582	127
29-900	f001	r000	2616	pNIC-Zb	3081	108
29-757	f001	r001	2187	pGTVL2	3153	110
29-757	f001	r001	2187	pNIC-Zb	2652	91
29-480	f001	r003	1356	pGTVL2	2322	79
29-480	f001	r003	1356	pNIC-Zb	1821	60

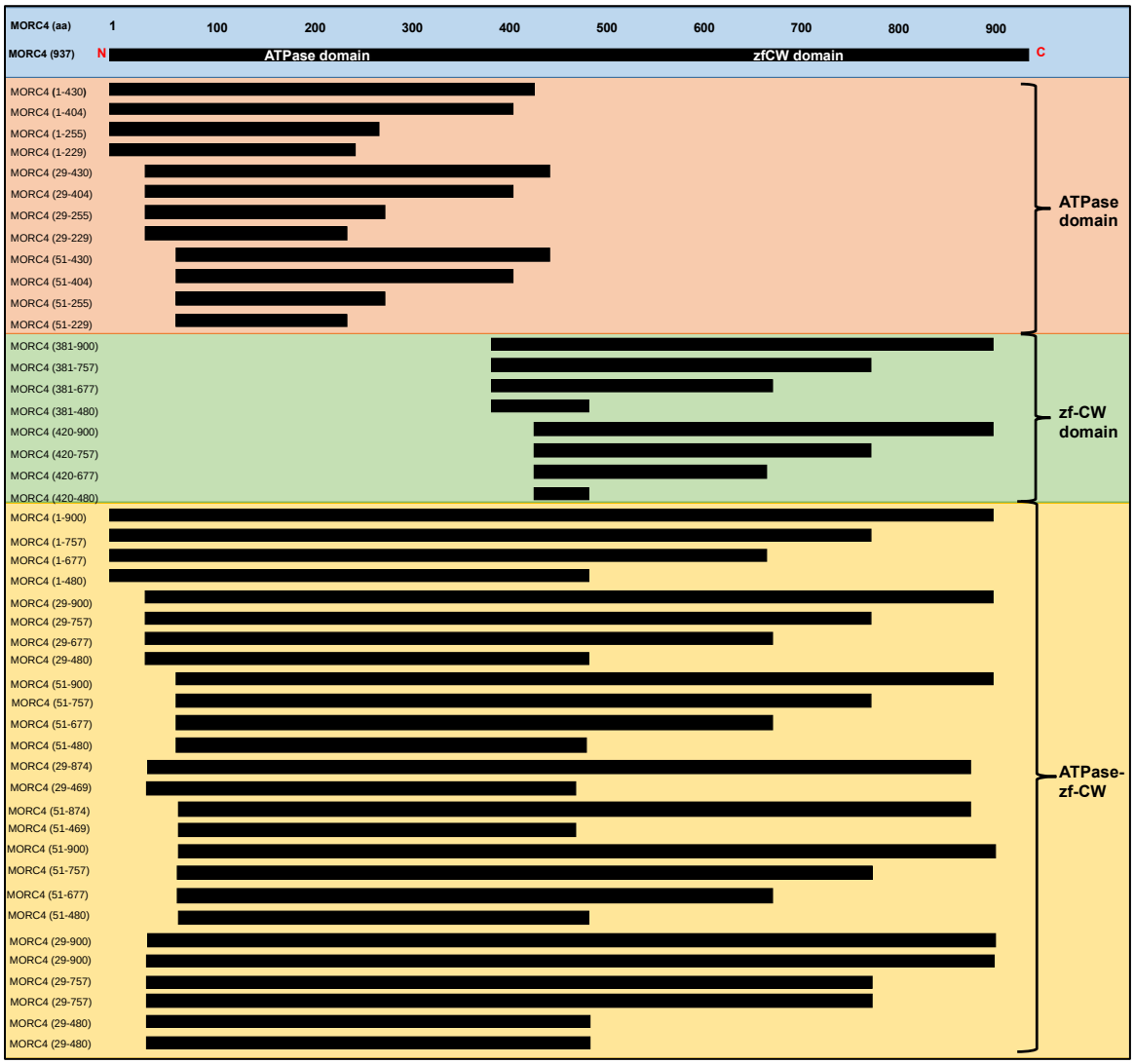


Figure 3.6: **Schematic representation architecture of each truncated MORC4 fragment following domain boundary prediction bioinformatics analysis.** The region highlighted in blue represents both the ATPase and Zf-CW domains of MORC4 (937 aa). The region highlighted in orange represents the ATPase domain only fragments of MORC4. The region in yellow represents the ATPase and Zf-CW domains fragments of MORC4. ATPase only and Zf-CW fragments were cloned by undergraduate students, Baktawar Nawaz and Sa'eeda Akbor, respectively. MORC4$^{29\text{-}874}$, MORC4$^{29\text{-}469}$, MORC4$^{51\text{-}874}$ and MORC4$^{51\text{-}469}$ fragments were cloned by undergraduate student, Will Meddick-Dyson.

3.2.2.1 Domain boundary design of MORC4

Following design of truncated *MORC4* fragments, primers were designed from *MORC4* domain boundaries (Table 3.2) (designed by Dr. Christopher Cooper, The University of Huddersfield) to produce truncated fragments of *MORC4* (Table 3.1 and figure 3.6) with the appropriate end sequences for ligation-independent cloning (LIC) into multiple pET *E. coli*

expression vectors (Appendix 8.4) including pNH-TrxT, pGTVL2, pNIC-ZB and pNIC28-Bsa4. Combinations of primers were used to create different truncated fragments boundaries of *MORC4* (Table 3.1). The use of various N-terminal thioredoxin (TrxT) (Corsini *et al.*, 2008), glutathione S-transferase (GST) (Walls and Loughran, 2011) or Zbasic (Hedhammar *et al.*, 2014) tagged vectors increases the likelihood of construct solubility and production yield (Costa *et al.*, 2014), including the pFB-LIC-Bse vector, which can be used for insect cell protein expression. In addition, a polyhistidine tag (6xHistag) was also added to the N-terminus to allow the proteins to be used for immobilised metal affinity chromatography (IMAC) purification step (Booth *et al.*, 2018) and to also increase protein expression and solubility (Graslund *et al.*, 2008). Furthermore, all vectors also contained an N-terminal tobacco etch virus (TEV) protease cleavage site, which was later required to remove the tag from the proteins (Kapust *et al.*, 2001). It is important to have a variety of truncated fragments of protein as it is not possible to guarantee soluble protein (Graslund *et al.*, 2008). Cloned fragments of *MORC4* in multiple vectors were used for protein test expressions in *E. coli*, prior to large-scale protein expression and purification for biochemical and structural analyses.

Table 3.2: MORC4* primers designed to created truncated fragments of *MORC4 following LIC. The LIC tag sequences have been highlighted in text with an underline.

Primer name	Primer sequence (5'-3')
MORC4-f000	TACTTCCAATCCATGCTCCTGTACCGAGGGGCC
MORC4-f001	TACTTCCAATCCATGGGGATCCGCCTGAGCACG
MORC4-f002	TACTTCCAATCCATGTTCAGTGCCATCGCGGAG
MORC4-f003	TACTTCCAATCCATGTATACCAAGGAGTACCG
MORC4-f004	TACTTCCAATCCATGAAGGTTCCTGACCAGACATGG
MORC4-r000	TATCCACCTTTACTGTCACGAAGCTCCAAGTGAGG
MORC4-r001	TATCCACCTTTACTGTCAACCTTCGCTCTCCTCATAGC
MORC4-r002	TATCCACCTTTACTGTCAACCTCTGTTAGATTCTTCTG
MORC4-r003	TATCCACCTTTACTGTCAGCTCAAGCACAGGTCTTC
MORC4-r004	TATCCACCTTTACTGTCAATCACACTGAACCCATGTCTG
MORC4-r005	TATCCACCTTTACTGTCATGTTTTTTCCTTCCAGTAAGC
MORC4-r006	TATCCACCTTTACTGTCAGCCAGTCATTTTTTCTTCTGTG
MORC4-r007	TATCCACCTTTACTGTCATCCATTTTTATTTCTGC

3.2.2.2 PCR amplification of *MORC4* fragments

Specific regions of *MORC4* template DNA, which had been chosen following sequence analysis of *MORC4* (Figure 3.6 and table 3.1) were amplified using polymerase chain reaction (PCR). Inserts were of the expected size with a good yield, (Figure 3.7) and could be used for subsequent ligation independent cloning (LIC) and expression in *E. coli.* However, the full length 5' of *MORC4* did not amplify following PCR, therefore, the full length 5' of *MORC4* could not be cloned (data not shown). Proliferating cell nuclear antigen (PCNA) template DNA and respective primers (Appendix 8.5) were used as a positive amplification control throughout PCR experiments.

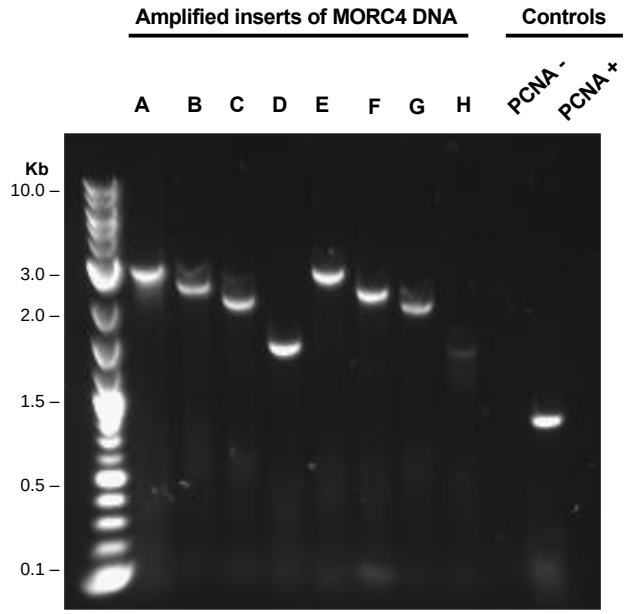


Figure 3.7: Amplified regions of *MORC4*. (A) Insert used to clone fragment MORC4$^{29-900}$. **(B)** insert used to clone fragment MORC4$^{29-757}$. **(C)** insert used to clone fragment MORC4$^{29-677}$. **(D)** insert used to clone fragment MORC4$^{29-480}$. **(E)** insert used to clone fragment MORC4$^{51-900}$. **(F)** insert used to clone fragment MORC4$^{51-757}$. **(G)** insert used to clone fragment MORC4$^{51-677}$. **(H)** insert used to clone fragment MORC4$^{51-480}$. PCNA was used as a positive control.

3.2.2.3 PCR screen of LIC truncated fragments of *MORC4*

Following amplification of *MORC4*, *MORC4* PCR inserts were purified, and cohesive ends were produced by T4 DNA polymerase treatment. Insert cohesive ends were ligated with vector (pNH-TrxT, pGTVL2, pNIC-Zb or pFB-LIC-Bse) cohesive ends. Ligations were transformed into MACH1 *E. coli* cells and a colony PCR screen using common primers binding to sites (Table 3.3) surrounding the cloning site, determined sizes of constructs (Figure 3.8). All fragments of *MORC4* were the correct size (Table 3.1), suggesting that truncated fragments of *MORC4* were successfully cloned into MACH1 *E. coli* cells. If multiple samples of the same fragment were amplified, only one sample was required for further processing. To verify cloning boundaries, a number of cloned truncated *MORC4* fragments were sequenced by Source Bioscience (Cambridge, UK).

Table 3.3: *MORC4* primers designed for screening of truncated fragments of *MORC4* following LIC. Including the primer name and primer sequence.

Primer name	Primer sequence
pcDNA3_N-Flag-fwd	TCCAAAATGTCGTAACAACTCC
pcDNA3-rev	TTTTATTAGGAAAGGACAGTGG
FBAC1	TATTCATACCGTCCCACCA
FBAC2	GGGAGGTTTTTTAAAGCAAGTAAA
pLICfor	TGTGAGCGGATAACAATTCC
pLICrev	AGCAGCCAACTCAGCTTCC

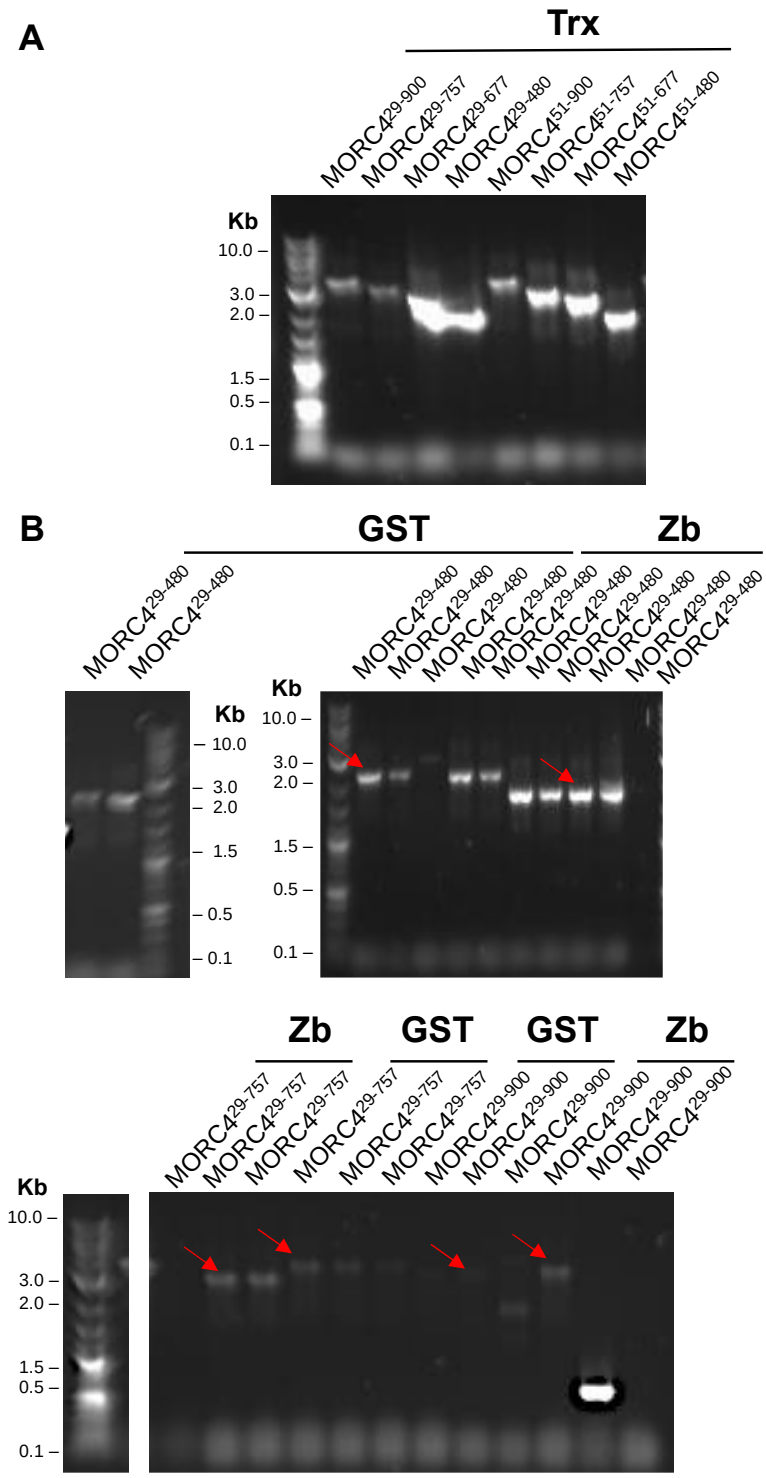


Figure 3.8: **Colony PCR screen of truncated fragments of *MORC4*.** Following a colony PCR amplification screen, cloned fragments of MORC4 were screened on a 0.8% agarose gel to confirm sizes. **(A)** *MORC4* fragments cloned into vector pNH-TrxT (Trx tag). **(B)** *MORC4* fragments cloned into vectors pGTVL2 (GST tag) and pNIC-Zb (Zb tag). For multiple samples of the same fragment, the red arrow indicates samples chosen for further processing.

3.3 Heterologous expression of fragments of MORC4 in *E. coli* and purification

Cloned truncated MORC4 plasmids were initially used for 50 mL test expression in *E. coli*. Fragments with a high yield of protein after a 50 mL test expression were subsequently used for used for 1 L expression, which helped to determine i) if there was a high enough yield for subsequent large-scale experiments and ii) if the 6xHis-(Trx/GST/Zb)-TEV affinity tag could be successfully removed using TEV protease. Proteins that expressed well and were fully TEV-cleavable were used for large-scale expression, towards biochemical and structural analysis.

3.3.1 50 mL test expressions determined high yielding truncated fragments of MORC4

Prior to 50 mL test expressions, plasmids containing cloned fragments of *MORC4* were transformed into BL21 (DE3) Rosetta2 *E. coli,* as *MORC4* fragments require a both T7 promoter for transcription and the pRARE2 plasmid to counteract potential codon usage issues, by providing additional rare tRNAs. *E. coli* expression plasmids containing truncated *MORC4* fragments, which had been transformed into Rosetta2 *E. coli,* were used for an initial 50 mL protein expression test in both LB (Figure 3.9) and TB media (Figure 3.10), to assess protein yield and solubility. Both LB and TB were used as some evidence suggests that TB produces a higher plasmid yield (Wood *et al.*, 2017). However, some evidence also suggests that TB can cause the production of additional proteins as it is used to create a large number of cells, decreasing the purity of target proteins (Kram and Finkel, 2015). Whereas, LB is most commonly used for the growth of *E. coli* as it allows a good yield of protein with a relatively quick growth time (Sezonov *et al.*, 2007). 50 mL test expressions were important as preparation for downstream large-scale protein expression at multiple litre level. In addition, an initial nickel-nitrilotriacetic acid (Ni-NTA) immobilised metal affinity chromatography (IMAC) step was performed as this helped to identify low-mid level expressors against the background of other *E. coli* proteins, which may not be distinguishable when resolving whole cell lysates.

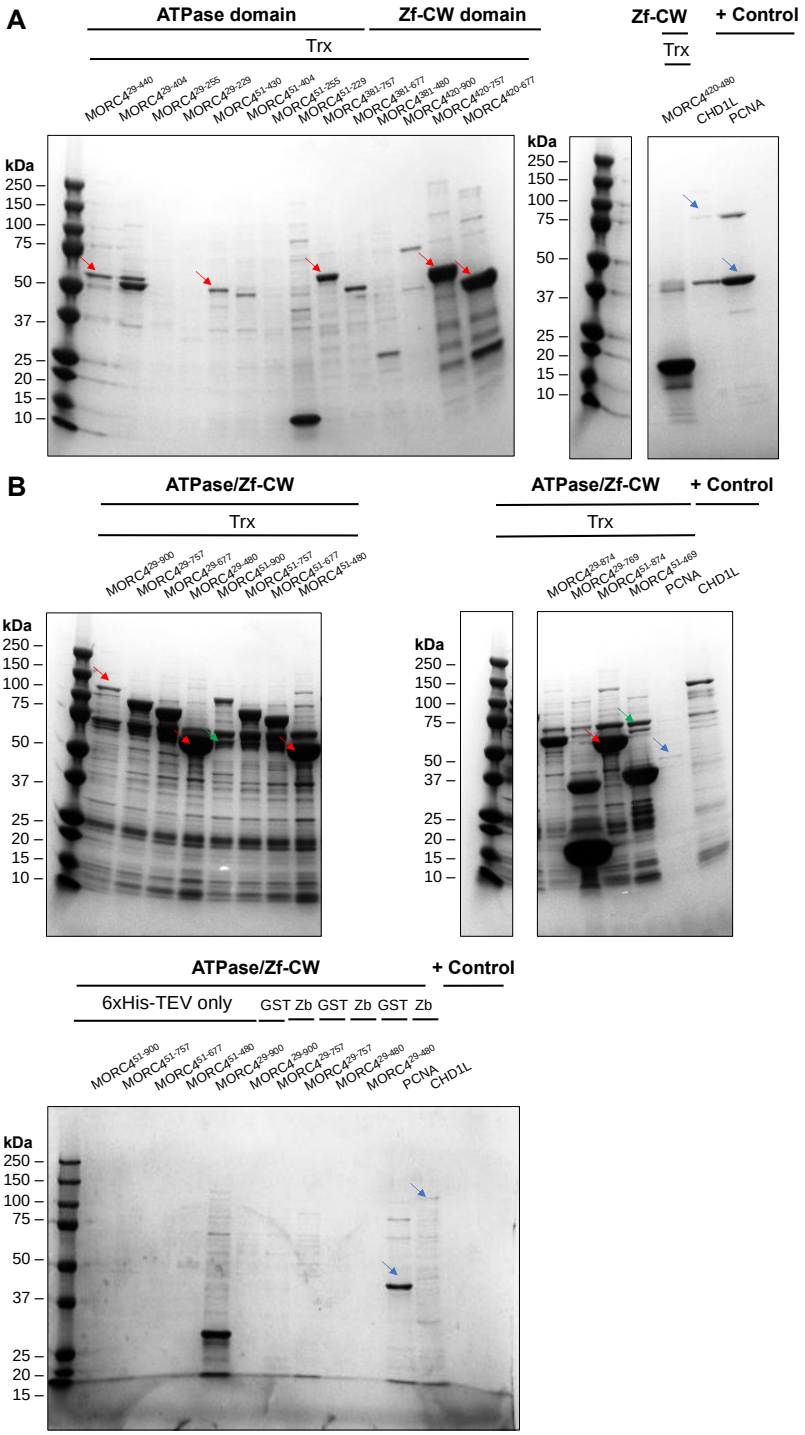


Figure 3.9: **50 mL test expressions of truncated fragments of MORC4 in LB.** Proteins were expressed in 50 mL LB, purified using a Ni-IMAC column and resolved on a 10% tris-glycine SDS(TGS)-PAGE gel. **(A)** MORC4 protein fragments with individual ATPase domain and Zf-CW domain only with 6xHis-TEV with the addition of Trx tags and CHD1L and PCNA positive controls. **(B)** ATPase/Zf-CW tandem domains with 6xHis-TEV and the addition of Trx, GST and Zb tags. Some fragments had the 6xHis-TEV tag only. Red arrows indicate examples of expressed proteins. Green arrows are examples of the potential DnaK chaperone contaminant and blue arrows indicate the positive controls (PCNA and CHD1L).

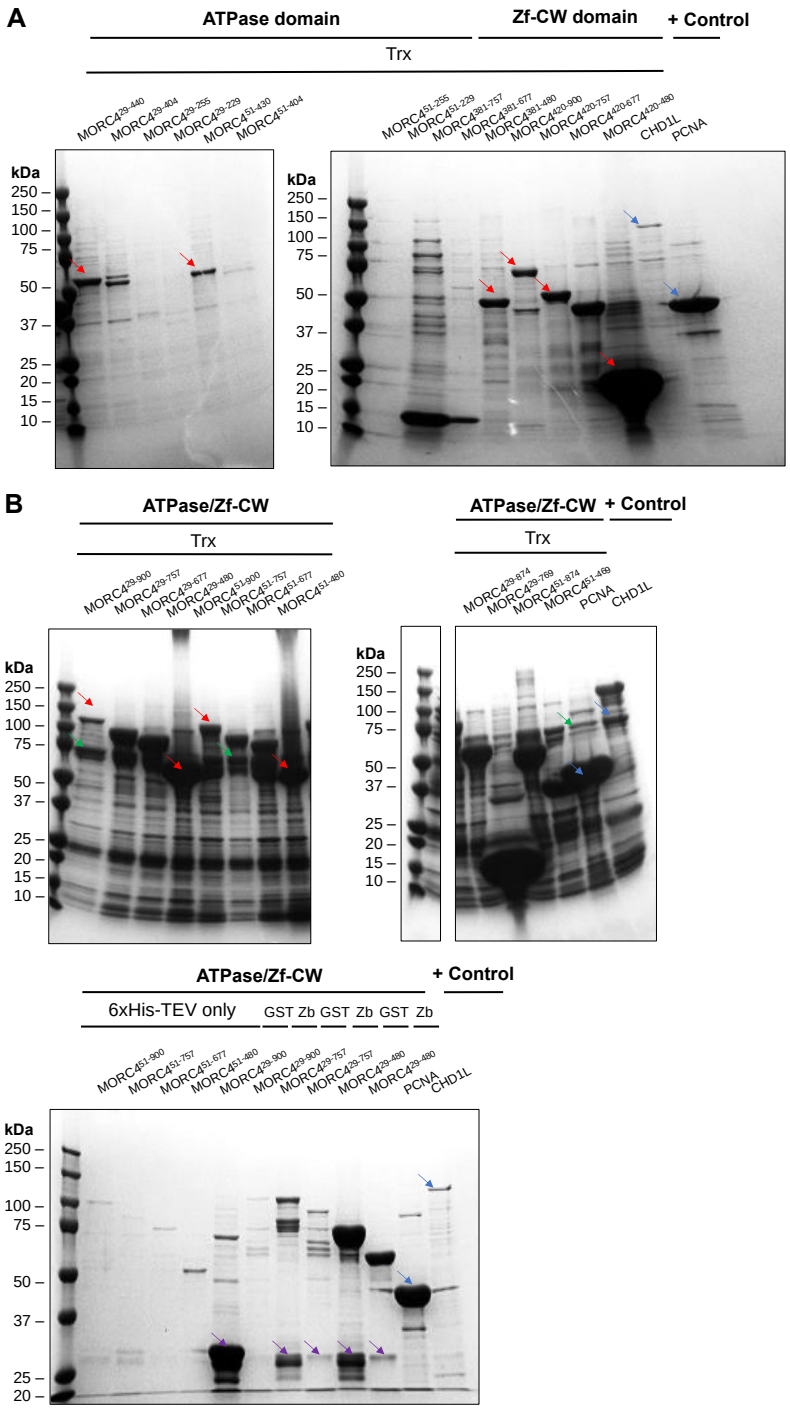


Figure 3.10: **50 mL test expressions of truncated fragments of MORC4 in TB**. Proteins were expressed in 50 mL TB, purified using a Ni-IMAC column and resolved on a 10% tris-glycine SDS(TGS)-PAGE gel. **(A)** MORC4 protein fragments with individual ATPase domain and Zf-CW domain only with 6xHis-TEV with the addition of Trx tags and CHD1L and PCNA positive controls. **(B)** ATPase/Zf-CW tandem domains with 6xHis-TEV and the addition of Trx, GST and Zb tags. Some fragments had the 6xHis-TEV tag only. Red arrows indicate examples of expressed proteins. Green arrows are examples of the potential DnaK contaminant. Blue arrows indicate the positive controls (PCNA and CHD1L), and purple arrows are examples of where the tag has been aberrantly proteolytically cleaved.

Truncated MORC4 proteins expressed in TB media appeared to have a higher yield of protein overall (Figure 3.10) (Table 3.4) when compared to those expressed in LB (Figure 3.9). More specifically, the ATPase-Zf-CW fragment MORC4$^{29\text{-}480}$ and the Zf-CW domain only fragment MORC4$^{420\text{-}480}$ both expressed at higher quantities in TB compared to LB. However, fragments of MORC4 expressed in LB media appeared to have fewer contaminants (presumably following the lower cell masses from the less rich media), but still had a good level of expression. This supports the evidence that TB medium has a higher nutrient content, and it can achieve higher *E. coli* cell densities, leading to a higher yield of protein. Truncated MORC4 proteins and positive controls (PCNA and CHD1L) expressed in both LB and TB media matched the expected sizes and there was no difference in yield between equivalent fragments in different vectors.

Table 3.4: **Expression levels of each cloned fragment of MORC4 in both LB and TB including amino acid boundary of each fragment and protein size.** MORC4$^{1\text{-}430}$, MORC4$^{1\text{-}404}$, MORC4$^{1\text{-}255}$, MORC4$^{1\text{-}229}$, MORC4$^{381\text{-}900}$, MORC4$^{1\text{-}900}$, MORC4$^{1\text{-}757}$, MORC4$^{1\text{-}677}$ and MORC4$^{1\text{-}480}$ could not be amplified by PCR, therefore, they could not be cloned.

Truncated fragment of MORC4 (aa)	Protein size (kDa)	Domain(s)	Vector	Test expression level in LB	Test expression level in TB	1 litre protein expression
MORC4 (1-430)	60	ATPase	pNH-TrxT	N/A	N/A	
MORC4 (1-404)	57	ATPase	pNH-TrxT	N/A	N/A	
MORC4 (1-255)	40	ATPase	pNH-TrxT	N/A	N/A	
MORC4 (1-229)	37	ATPase	pNH-TrxT	N/A	N/A	
MORC4 (29-430)	57	ATPase	pNH-TrxT	LOW	MEDIUM	✔
MORC4 (29-404)	54	ATPase	pNH-TrxT	LOW	MEDIUM	
MORC4 (29-255)	37	ATPase	pNH-TrxT	NONE	LOW	
MORC4 (29-229)	34	ATPase	pNH-TrxT	NONE	LOW	
MORC4 (51-430)	55	ATPase	pNH-TrxT	LOW	MEDIUM	
MORC4 (51-404)	52	ATPase	pNH-TrxT	LOW	LOW	✔
MORC4 (51-255)	35	ATPase	pNH-TrxT	NONE	NONE	
MORC4 (51-229)	32	ATPase	pNH-TrxT	MEDIUM	HIGH	
MORC4 (381-900)	72	Zf-CW	pNH-TrxT	N/A	N/A	
MORC4 (381-757)	55	Zf-CW	pNH-TrxT	MEDIUM	LOW	
MORC4 (381-677)	46	Zf-CW	pNH-TrxT	LOW	MEDIUM	
MORC4 (381-480)	24	Zf-CW	pNH-TrxT	LOW	MEDIUM	
MORC4 (420-900)	67	Zf-CW	pNH-TrxT	LOW	MEDIUM	
MORC4 (420-757)	50	Zf-CW	pNH-TrxT	HIGH	MEDIUM	✔
MORC4 (420-677)	42	Zf-CW	pNH-TrxT	HIGH	HIGH	✔

Truncated fragment of MORC4 (aa)	Protein size (kDa)	Domain(s)	Vector	Test expression level in LB	Test expression level in TB	1 litre protein expression
MORC4 (420-480)	19	Zf-CW	pNH-TrxT	HIGH	HIGH	✔
MORC4 (1-900)	114	ATPase-Zf-CW	pNH-TrxT	N/A	N/A	
MORC4 (1-757)	97	ATPase-Zf-CW	pNH-TrxT	N/A	N/A	
MORC4 (1-677)	88	ATPase-Zf-CW	pNH-TrxT	N/A	N/A	
MORC4 (1-480)	66	ATPase-Zf-CW	pNH-TrxT	N/A	N/A	
MORC4 (29-900)	111	ATPase-Zf-CW	pNH-TrxT	LOW	LOW	✔
MORC4 (29-757)	94	ATPase-Zf-CW	pNH-TrxT	MEDIUM	HIGH	✔
MORC4 (29-677)	86	ATPase-Zf-CW	pNH-TrxT	MEDIUM	HIGH	✔
MORC4 (29-480)	63	ATPase-Zf-CW	pNH-TrxT	HIGH	HIGH	✔
MORC4 (51-900)	109	ATPase-Zf-CW	pNH-TrxT	LOW	MEDIUM	✔
MORC4 (51-757)	92	ATPase-Zf-CW	pNH-TrxT	MEDIUM	MEDIUM	✔
MORC4 (51-677)	83	ATPase-Zf-CW	pNH-TrxT	MEDIUM	MEDIUM	✔
MORC4 (51-480)	61	ATPase-Zf-CW	pNH-TrxT	HIGH	HIGH	✔
MORC4 (29-874)	109	ATPase-Zf-CW	pNH-TrxT	LOW	LOW	
MORC4 (29-469)	62	ATPase-Zf-CW	pNH-TrxT	LOW	LOW	
MORC4 (51-874)	106	ATPase-Zf-CW	pNH-TrxT	LOW	LOW	
MORC4 (51-469)	122	ATPase-Zf-CW	pNH-TrxT	LOW	LOW	
MORC4 (51-900)	99	ATPase-Zf-CW	pNIC28-Bsa4	LOW	LOW	
MORC4 (51-757)	83	ATPase-Zf-CW	pNIC28-Bsa4	LOW	LOW	
MORC4 (51-677)	74	ATPase-Zf-CW	pNIC28-Bsa4	LOW	LOW	
MORC4 (51-480)	52	ATPase-Zf-CW	pNIC28-Bsa4	LOW	MEDIUM	
MORC4 (29-900)	127	ATPase-Zf-CW	pGTVL2	LOW	MEDIUM	
MORC4 (29-900)	108	ATPase-Zf-CW	pNIC-Zb	LOW	MEDIUM	
MORC4 (29-757)	110	ATPase-Zf-CW	pGTVL2	LOW	MEDIUM	
MORC4 (29-757)	91	ATPase-Zf-CW	pNIC-Zb	LOW	MEDIUM	
MORC4 (29-480)	79	ATPase-Zf-CW	pGTVL2	LOW	HIGH	
MORC4 (29-480)	60	ATPase-Zf-CW	pNIC-Zb	LOW	HIGH	

PCNA and CHD1L were chosen as control proteins during expression because they are both involved in genome transactions (Strzalka and Ziemienowicz, 2011; Lehmann *et al.*, 2017). PCNA is 28 kDa and highly expressed whereas, CHD1L is 100 kDa and expressed at lower levels (C. Cooper, *pers. comm.*), hence the use of both protein positive allowed comparative estimations of expected MORC4 protein yields. The majority of MORC4 fragment proteins were expressed within PCNA and CHD1L size range (28 kDa-100 kDa), providing further evidence that they were suitable controls. There were also contaminants co-purifying in both expressions including chaperone proteins at approximately 25 kDa and 15 kDa. Another major contaminant was expressed at approximately 75 kDa, which was likely to be the chaperone

DnaK (Rial and Ceccarelli, 2002). This 75 kDa protein was also identified co-eluting in other tests, using tryptic tandem mass spectrometry (MSMS) (C. Cooper, *pers. comm*.). This contaminant would need to be removed by further purification downstream as DnaK is an ATPase (Russell *et al.*, 1998), which means it could otherwise interfere with MORC4 ATPase assays. Following 50 mL protein test expressions of several fragments of MORC4, a number of ATPase only, Zf-CW only and ATPase-Zf-CW tandem domains were utilised for 1 L protein expression (Table 3.4). These fragments of MORC4 were chosen based on a good level of expression and purity, which was assessed following 50 mL test expressions.

3.3.2 Large scale protein expression and purification of truncated MORC4 fragments

3.3.2.1 Expression and purification of Tobacco Etch Virus (TEV) protease

Tobacco Etch virus (TEV) protease is a 27 kDa enzyme required to cleave the 6xHis-(Trx/GST/Zb)-TEV affinity tag (Wu *et al.*, 2009) from the truncated MORC4 proteins during large-scale protein purification. Removing the protein affinity tags was an important stage in the protein purification process as the tags are large and bulky with a flexible linker region. Both of which can hinder protein crystal growth (Smyth *et al.*, 2003). Some evidence also suggests that protein tags such as polyhistidine tags can also affect protein activity and/or function (Booth *et al.*, 2018). Therefore, an additional rebinding NI-IMAC following TEV protease incubation was necessary to remove the cleaved tag from the protein.

Prior to large-scale protein purification, TEV protease was expressed in *E. coli*, purified using Ni-IMAC (Figure 3.11) and eluted in fractions E300 1-3. Fractions E300 1-3 were of sufficient purity, pooled and TEV protease was concentrated, which was then used for subsequent purifications of MORC4 to remove the tags from the MORC4 protein.

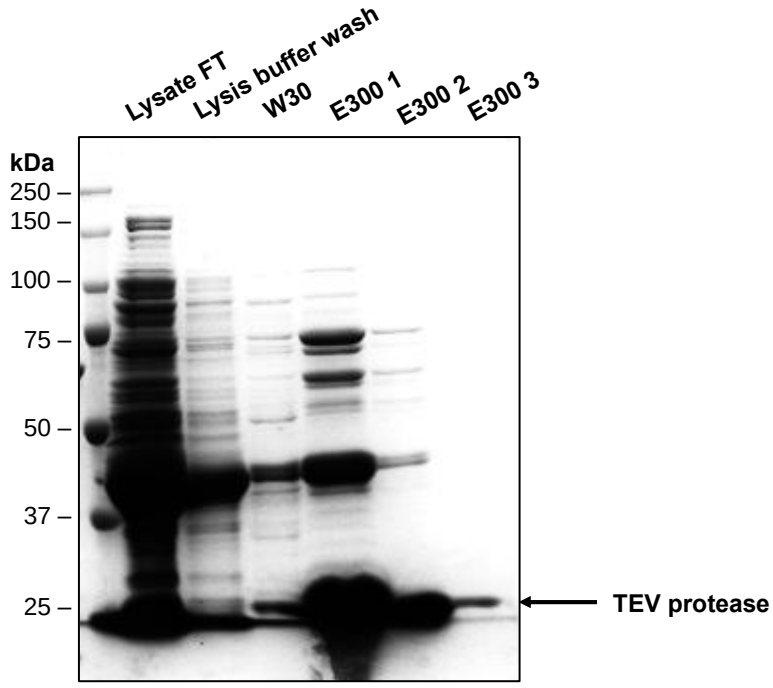


Figure 3.11: **Purification of TEV protease enzyme using Ni-IMAC.** TEV protease was expressed in 2 L LB, purified using a Ni-IMAC gravity flow column and resolved on a 10% tris-glycine SDS(TGS)-PAGE gel. The gel image shows the lysate FT (flow through), lysis buffer wash, W30 (wash 30 mM imidazole), E300 (elution 300 mM imidazole) 1, 2 and 3. The arrow indicates the bands for TEV protease.

3.3.2.1.1 Optimisation of TEV protease concentration required to cleave proteins

To identify an optimum concentration of TEV protease required to cleave the tag from partially purified MORC4 proteins, a TEV protease titration experiment was required. Both CHD1L control protein and MORC4$^{51-480}$ were used to test TEV cleavage at varying concentrations of TEV protease (0-0.6 mg/ml final concentration).

Following the addition of TEV protease and incubation at 4°C overnight, CHD1L was found to cleave in the range of 0.03-0.6mg/ml (Figure 3.12a). This suggests that only a low concentration is required to cleave CHD1L (~0.03 mg/ml). TEV protease was stored at -80°C at 3 mg/ml, which means that only 100 µl of TEV protease is required for 10 mL of protein to cleave. MORC4$^{51-480}$ was also used in this experiment as a negative control as previous experiments revealed this truncated fragment of MORC4 did not cleave. No samples of

MORC4$^{51-480}$ were cleaved with the addition of any concentration of TEV protease (Figure 3.12b), suggesting that it is not the concentration of TEV which prevents cleaving but the N-terminal structure of the protein which affects the accessibility for the TEV protease. Following this experiment, MORC4$^{51-480}$ was later abandoned for a MORC4 protein with a different N-terminus (MORC4$^{29-480}$), in an attempt to achieve TEV-cleavable MORC4.

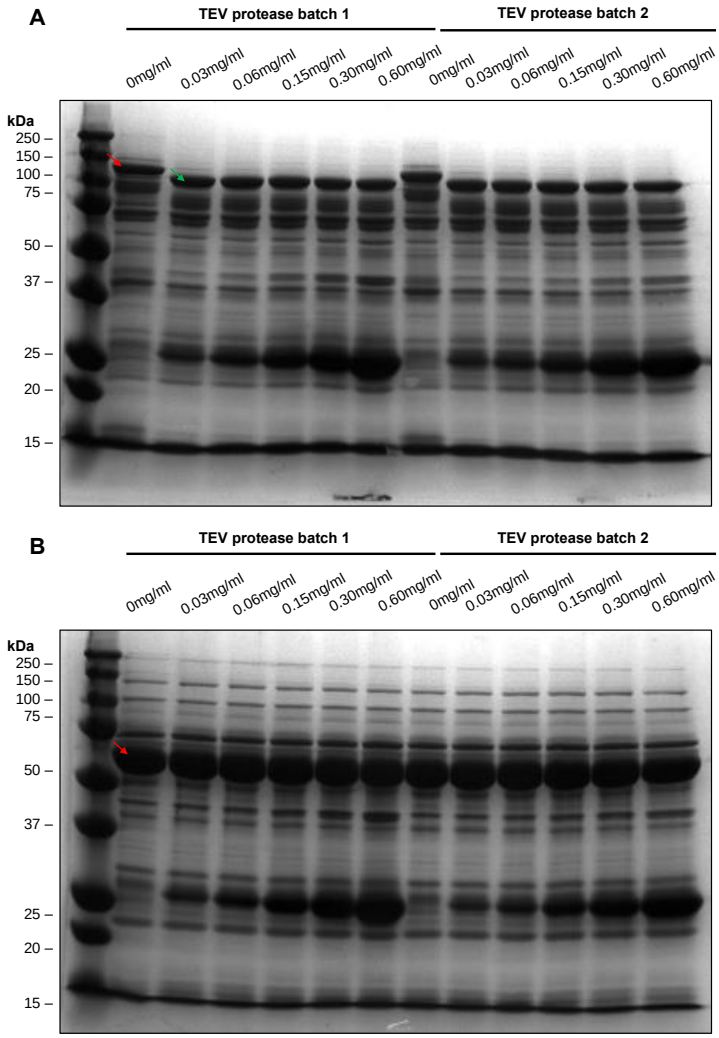


Figure 3.12: **Optimisation of TEV protease concentration required for cleaving proteins using CHD1L and MORC4$^{51-480}$.** Each Both MORC4$^{51-480}$ and CHD1L were expressed in LB, purified using a Ni-IMAC gravity flow column and TEV protease (0-0.6mg/ml) was added to each sample and left to incubate overnight at 4°C. Samples were resolved on 10% tris-glycine SDS(TGS)-PAGE homemade gels. **(A)** Positive CHD1L control showing the protein prior to cleaving (red arrow) and after TEV cleavage (green arrow). **(B)** Negative control sample MORC4$^{51-480}$ showing the uncleaved protein (red arrow).

3.3.2.2 1 litre protein expression identified TEV-cleavable ATPase-Zf-CW MORC4

Proteins that expressed a high yield of protein in the 50 mL test expression were used for 1 L large-scale expression and purification experiments. MORC4$^{29-480}$ is an ATPase-Zf-CW tandem domain MORC4 protein with a high yield of protein following 1 L expression in LB and was also successfully cleaved with the addition of TEV protease (Figure 3.13a). MORC4$^{29-480}$ was 63 kDa prior to TEV cleavage. Following 6xHis-Thioredoxin-TEV cleavage the protein size was 51 kDa (Figure 3.13b) (blue arrows represent the affinity tag removal in the 100 mM and 300 mM imidazole washes), indicating the protein had been successfully cleaved.

MORC4$^{29-900}$, MORC4$^{29-757}$ and MORC4$^{29-677}$ were also ATPase-Zf-CW MORC4 proteins which also cleaved with the addition of TEV protease (Figure 3.13b). However, they had a lower yield of protein compared to MORC4$^{29-480}$. This suggests that MORC4$^{29-480}$ is more suitable for large scale multiple litre (8 L) protein expression compared to the other ATPase-Zf-CW MORC4 proteins.

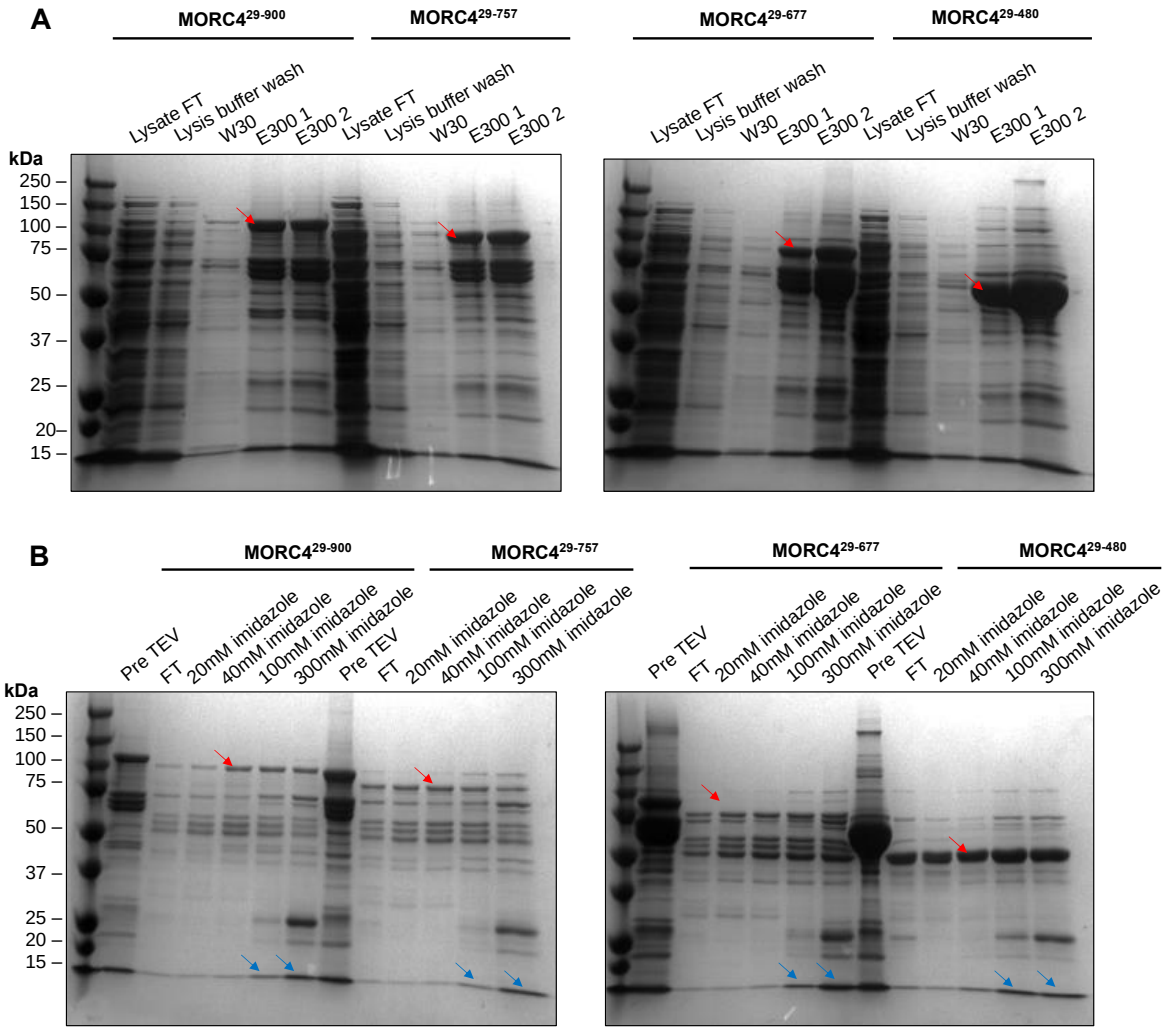


Figure 3.13: **1 L purification of cleavable MORC4$^{29\text{-}900}$, MORC4$^{29\text{-}757}$, MORC4$^{29\text{-}677}$ and MORC4$^{29\text{-}480}$ using Ni-IMAC and TEV protease.** Each fragment of MORC4 was expressed in 1 L LB, purified using a Ni-IMAC gravity flow column and cleaved using TEV protease. Samples were resolved on a 10% tris-glycine SDS(TGS)-PAGE homemade gels. **(A)** Initial Ni-IMAC purification gel showing W30 (wash 30mM imidazole) and 2 E300 (elution 300mM imidazole) washes. **(B)** Ni-IMAC gel image following addition of TEV protease at 4°C overnight. Gel images show a pre TEV sample and several imidazole washes. Red arrows indicate target protein bands. Blue arrows indicate the 6xHis-Trx tag.

MORC4$^{51\text{-}900}$, MORC4$^{51\text{-}757}$, MORC4$^{51\text{-}677}$ and MORC4$^{51\text{-}480}$ are ATPase-Zf-CW MORC4 proteins with a lower yield of protein following 1 L expression in LB, which did not cleave after the addition of TEV protease (Figure 3.14a and b). MORC4$^{51\text{-}900}$ was 109 kDa throughout the experiment, however, successful 6xHis-Thioredoxin-TEV removal would have resulted in a smaller protein of 97 kDa following overnight incubation of TEV protease. MORC4$^{51\text{-}757}$ also

remained at 92 kDa throughout the purification. Successful removal of the 6xHis-Thioredoxin-TEV affinity tag would have resulted in a protein of 80 kDa, indicating that MORC4$^{51-757}$ did not cleave.

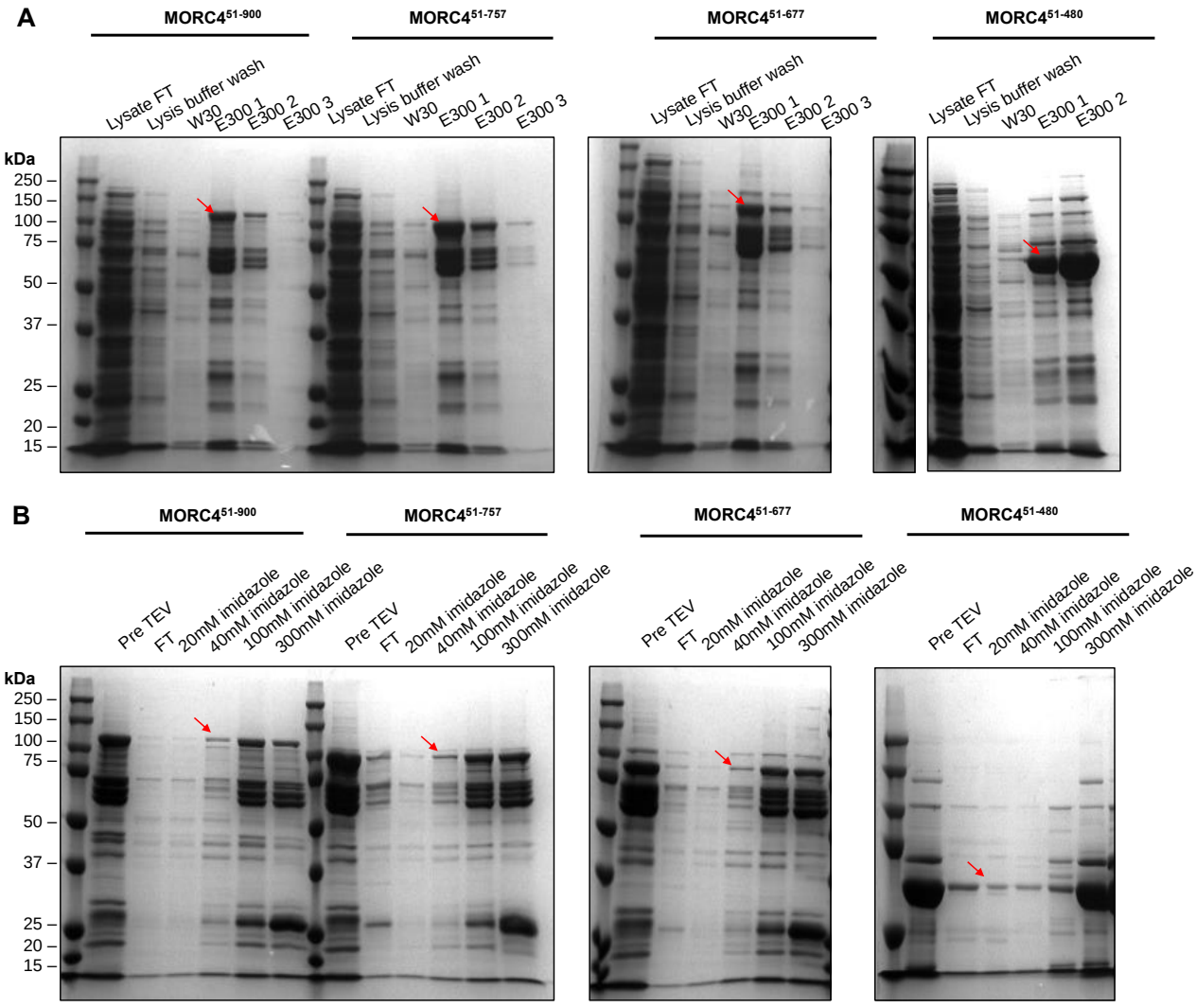


Figure 3.14: **1 L purification of non-cleavable MORC4$^{51-900}$, MORC4$^{51-757}$, MORC4$^{51-677}$ and MORC4$^{51-480}$ using Ni-IMAC and TEV protease.** Each fragment of MORC4 was expressed in 1L LB, purified using a Ni-IMAC gravity flow column and cleaved using TEV protease. Samples were resolved on a 10% tris-glycine SDS(TGS)-PAGE homemade gels. **(A)** Initial Ni-IMAC purification gel showing W30 (wash 30mM imidazole) and 2 E300 (elution 300mM imidazole) washes. **(B)** Ni-IMAC gel image following addition of TEV protease at 4°C overnight. Gel images show a pre TEV sample and several imidazole washes. Red arrows indicate target protein bands.

When analysing the protein sequences of cleavable (MORC4$^{29-480}$, MORC4$^{29-900}$, MORC4$^{29-757}$ and MORC4$^{29-677}$) and non-cleavable MORC4 (MORC4$^{51-900}$, MORC4$^{51-757}$, MORC4$^{51-677}$ and MORC4$^{51-480}$), it is interesting to note that there are differences which could contribute to if the proteins can be cleaved by TEV protease. Cleavable fragments of MORC4 begin with the residues "GIRLS…" whereas, non-cleavable fragments of MORC4 begin with "FSAI…" residues. However, both proteins have the same C-terminus ('LCLS') (Figure 3.15). This suggests that it is the N-terminal position of residues that affects whether or not MORC4$^{29-480}$ and MORC4$^{51-480}$ are cleavable by TEV protease, suggesting that these are more accessible to cleavage by TEV protease.

Figure 3.15: **Amino acid sequence of MORC4$^{51-480}$ and MORC4$^{29-480}$.** The N-terminal of MORC4$^{51-480}$ residues are highlighted in purple. The N-terminal of MORC4$^{29-480}$ residues are highlighted in red. The C-terminal of both MORC4$^{51-480}$ and MORC4$^{29-480}$ are highlighted in blue.

Overall, the 1 L protein expression and purification tests gave an indication as to potential contaminants that may be expected in large-scale purifications. In addition, it also revealed that some of the cleaved MORC4 proteins remained bound to the Ni^{2+} resin even at high

concentrations of imidazole (300 mM) where the tag has been eluted (Figure 3.13b). This could be due to patches of histidine residues in the MORC4 proteins, causing non-specific binding to the Ni^{2+} resin (Bornhorst and Falke, 2000).

3.3.2.2.1 50 mL test expression for MORC4 fragments that were uncleavable

$MORC4^{51-900}$, $MORC4^{51-757}$, $MORC4^{51-677}$ and $MORC4^{51-480}$ had a good yield of protein when expressed as 1 L in TB, which was determined by the protein band intensity on the SDS-PAGE gel (Figure 3.10b) however, they were not TEV-cleavable (Figure 3.14b). In an attempt to increase the number of cleavable MORC4 fragments, some MORC4 clones were re-cloned into pNIC28-Bsa4 vector (6xHis tag only) (Figure 3.9b and 3.10b). Cloning MORC4 fragments without Trx tag could cause the TEV cleavage site to become accessible by TEV protease, allowing cleavage of the tag. Additional MORC4 fragments, which were cleavable but had a low yield, were also re-cloned to increase overall protein expression. Unfortunately, these re-cloned MORC4 fragments did not yield a sufficient amount of protein for large-scale protein expression and purification experiments (Figure 3.9b and 3.10b), suggesting that the Trx tag was important for enhancing MORC4 protein yields.

3.3.2.3 Large-scale protein expression and purification of cleavable MORC4

High yielding proteins which were successful in 6xHis-Trx-TEV affinity tag removal, following 1 L protein expression and purification, were utilised for 8 L protein expression (green bars, figure 3.16). 8 L large-scale purifications of MORC4 fragments involved additional steps, including ion exchange chromatography (IEX) and size exclusion chromatography (SEC) to ensure purity for subsequent experiments such as biochemical assays and protein crystallography. Several MORC4 proteins were utilised for large-scale protein expression and purification including ATPase-Zf-CW, ATPase only and Zf-CW C-terminal only proteins.

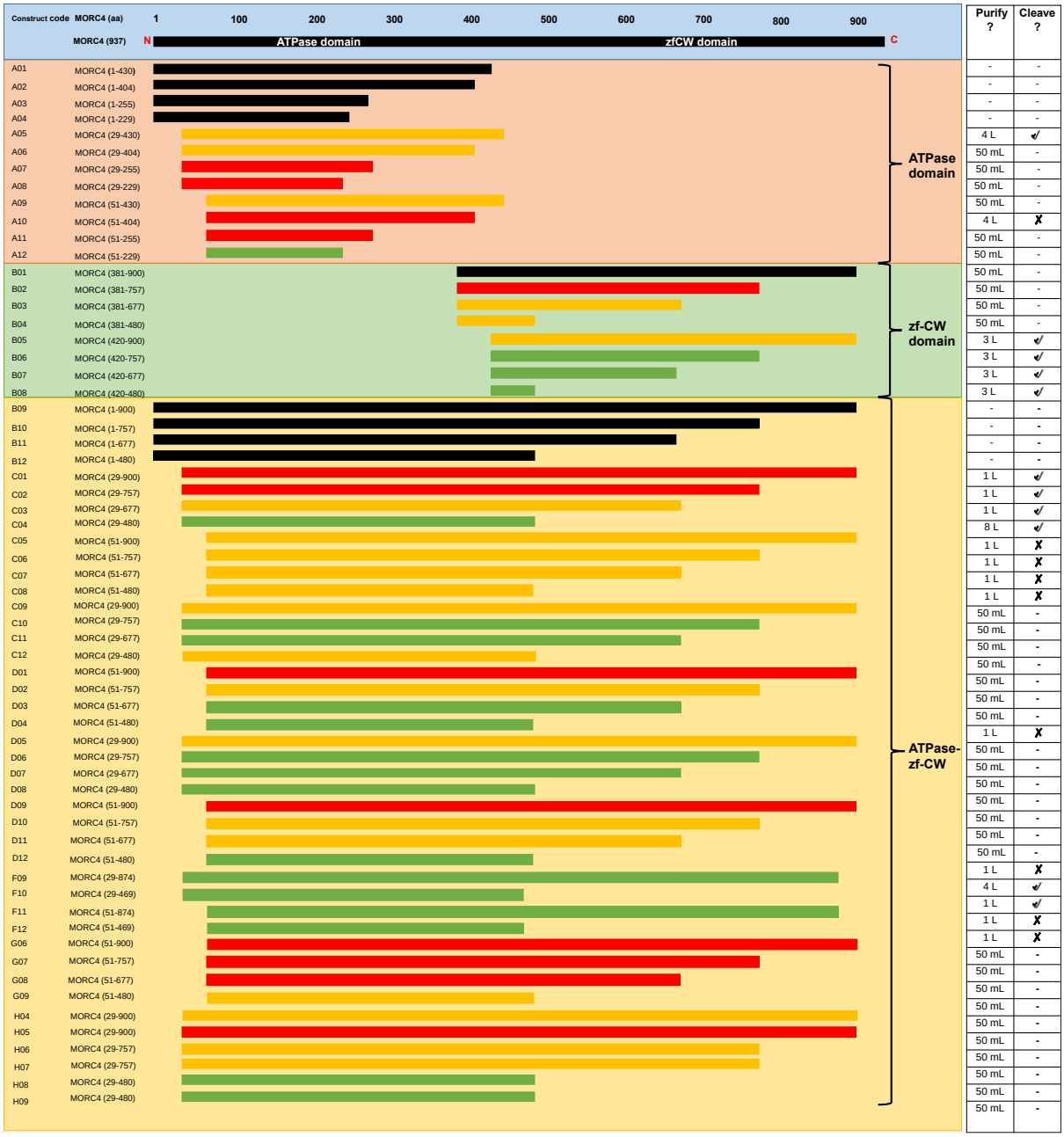


Figure 3.16: **Truncated fragments of MORC4 and protein expression levels. MORC4 expressed in 50 mL for testing, 1 L, 3 L, 4 L or 8L.** Black bars are fragments of MORC4 where cloning was unsuccessful due to failed PCRs. Red bars are low expressing fragments of MORC4, yellow bars are medium expressing fragments of MORC4, and green bars are high expressing fragments of MORC4. Each protein also has the level of expression and information regarding TEV cleavage.

3.3.2.3.1 ATPase-Zf-CW MORC4 fragments

3.3.2.3.1.1 Ni-IMAC and TEV rebind ensures affinity tag removal

The 6xHis-Trx-TEV affinity tag must be removed before utilising the protein for structural or biochemical studies, to prevent any interference with subsequent experiments. MORC4$^{29\text{-}480}$ (ATPase-Zf-CW protein) is a MORC4 protein with a high yield (Figure 3.17a), which was successfully cleaved following affinity tag removal, which was evident in the size reduction between the pre-TEV sample and post-TEV sample (Figure 3.17b). However, some of this protein appeared to remain bound to the Ni-IMAC column following TEV cleavage and rebinding to Ni-IMAC (Figure 3.17b). This suggests that either not all of all the affinity tag was removed allowing the protein to still bind to the resin or that there was a strong interaction between the protein and the resin, either from patches of histidine residues in the protein or ion exchange effects. Following the initial Ni-IMAC, elutions E300 1-3 were pooled (Figure 3.17a) for the next stage in the purification process.

Following rebinding of the TEV-cleaved sample to the Ni-IMAC resin, the flow through (FT) and 20 mM imidazole wash fractions were pooled (Figure 3.17b) for the next step of the purification process. These two samples were pooled as they had fewer contaminants in comparison to the amount of protein (Figure 3.17b). There was also no apparent TEV protease or affinity tag present in these samples.

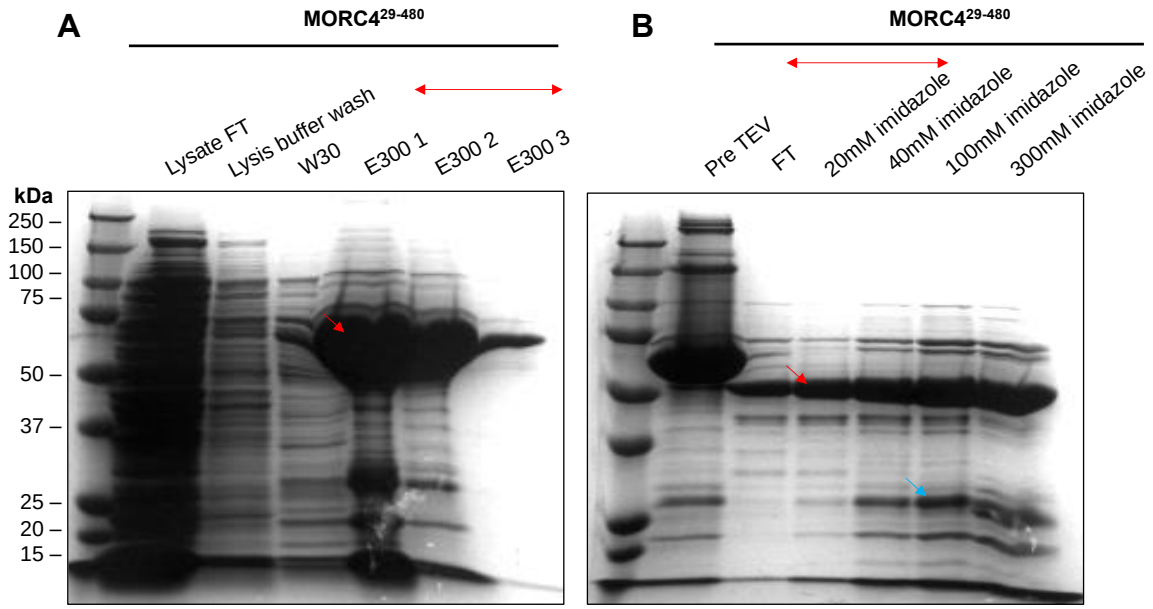


Figure 3.17: **Initial large-scale purification of MORC4$^{29-480}$ using Ni-IMAC and TEV rebind.** MORC4$^{29-480}$ was expressed in 5 L TB, purified using a Ni-IMAC gravity flow column, cleaved using TEV protease and the TEV affinity tag was removed with an additional Ni-IMAC. Samples were resolved on a 10% tris-glycine SDS (TGS)-PAGE gels. **(A)** First Ni-IMAC purification gel showing W30 (wash 30 mM imidazole) and 3 E300 (elution 300 mM imidazole) washes. **(B)** Second Ni-IMAC gel image following addition of TEV protease at 4°C overnight. Gel images show a pre-TEV sample and several imidazole washes. Red arrows indicate target protein bands and the blue arrow indicate the removed TEV protease. Red double headed arrows indicate fractions which were pooled.

3.3.2.3.1.2 Cation exchange chromatography (IEX) of MORC4$^{29-480}$

Ion exchange chromatography (IEX) separates proteins based on charge. The type of IEX column used during the purification process depends on the isoelectric point (pI) of the protein. MORC4$^{29-480}$ has a predicted pI of 8.79, which means that when in a buffer of pH 7.5 the protein will be positivity charged, indicating that this protein will bind to a negatively charged IEX column (SP-IEX) (cation exchange chromatography).

Following TEV rebind Ni-IMAC and combining of the FT and 20 mM imidazole wash, MORC4$^{29-480}$ was applied to a SP-IEX column. MORC4$^{29-480}$ was eluted from the SP-IEX column between 0.18-0.26 M NaCl (Figure 3.18a), suggesting that this protein is not bound very strongly to the column, as it did not require a high concentration of NaCl to elute. Several

fractions eluted from the SP-IEX column were resolved on an SDS-PAGE gel and fractions 11, 12 and 13 were pooled and concentrated for subsequent purification steps (Figure 3.18b).

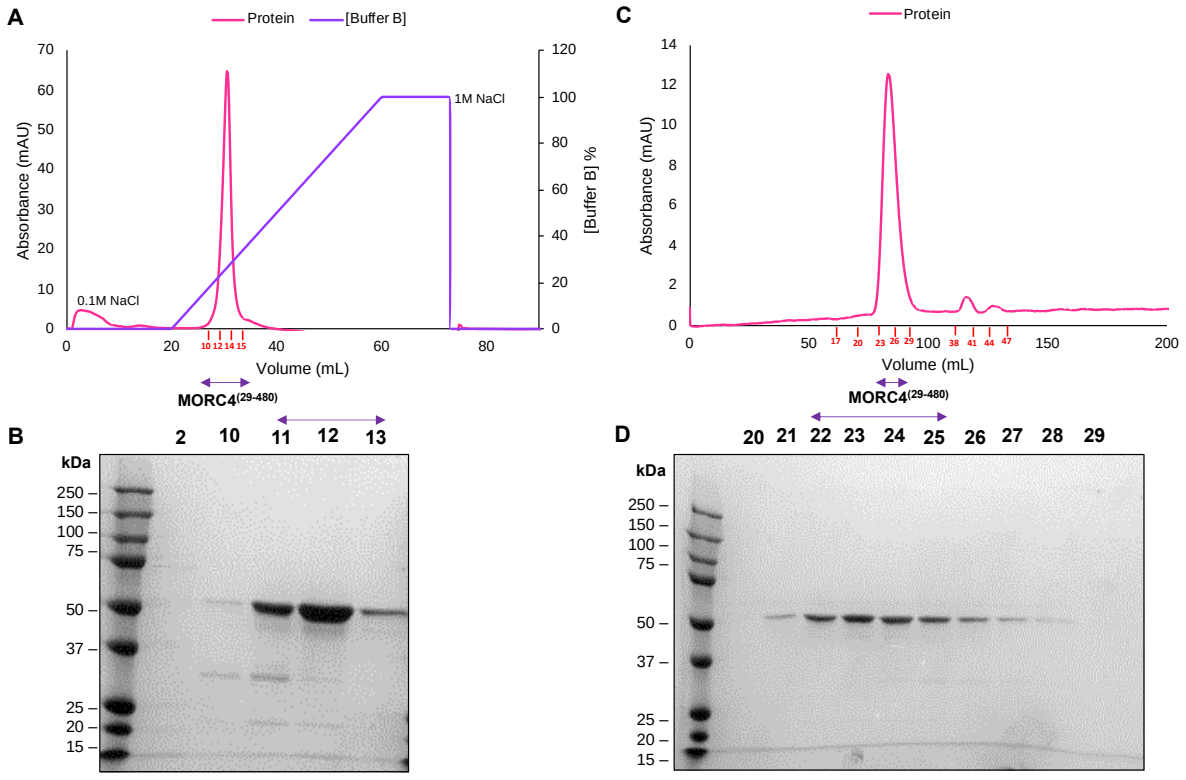


Figure 3.18: **Purification of MORC4$^{29\text{-}480}$ FT and 20mM imidazole, using SP-IEX and SEC.** Following both SP-IEX and SEC, selected samples were resolved on a 10% tris-glycine SDS (TGS)-PAGE gel. **(A)** Chromatogram of MORC4$^{29\text{-}480}$ (FT and 20 mM imidazole wash), which was purified on the A 1 ml HiTrap SP-IEX column using the ÄKTA pure. **(B)** Selected samples were resolved on an SDS-PAGE for analysis. Samples 11, 12 and 13 were pooled for further purification on a SEC column. **(C)** A chromatogram of MORC4$^{29\text{-}480}$, which had been pooled for SEC from SP-IEX and purified on the A HiLoad 16/600 Superdex 200 pg column using the ÄKTA Pure. **(D)** selected samples were resolved on an SDS-PAGE for analysis. Samples 22-25 were pooled, concentrated, flash-frozen in LN_2 and stored at -80°. Purple doubled headed arrows indicate fractions which were pooled.

3.3.2.3.1.3 Size exclusion chromatography (SEC) of MORC4$^{29\text{-}480}$

Size exclusion chromatography (SEC) separates proteins according to size and is often the final stage of protein purification, as it can separate protein oligomers (Berek, 2010), remove aggregates (Ó'Fágáin *et al.*, 2017) and also allow exchange of eluted protein into a defined buffer (Kurnik *et al.*, 1995). Following SEC of MORC4$^{29\text{-}480}$, fractions 22-25 were pooled (Figure 3.18d) (containing the 50 kDa MORC4$^{29\text{-}480}$ protein), which had a peak of 83.3 mL

(Figure 3.18b). 5 L of pure MORC4$^{29\text{-}480}$ expressed in TB resulted in 200 µL of pure and homogeneous protein at 2.1 mg/mL, however, at least 5 mg/mL is generally required for protein crystallography trials (Dessau and Modis, 2011).

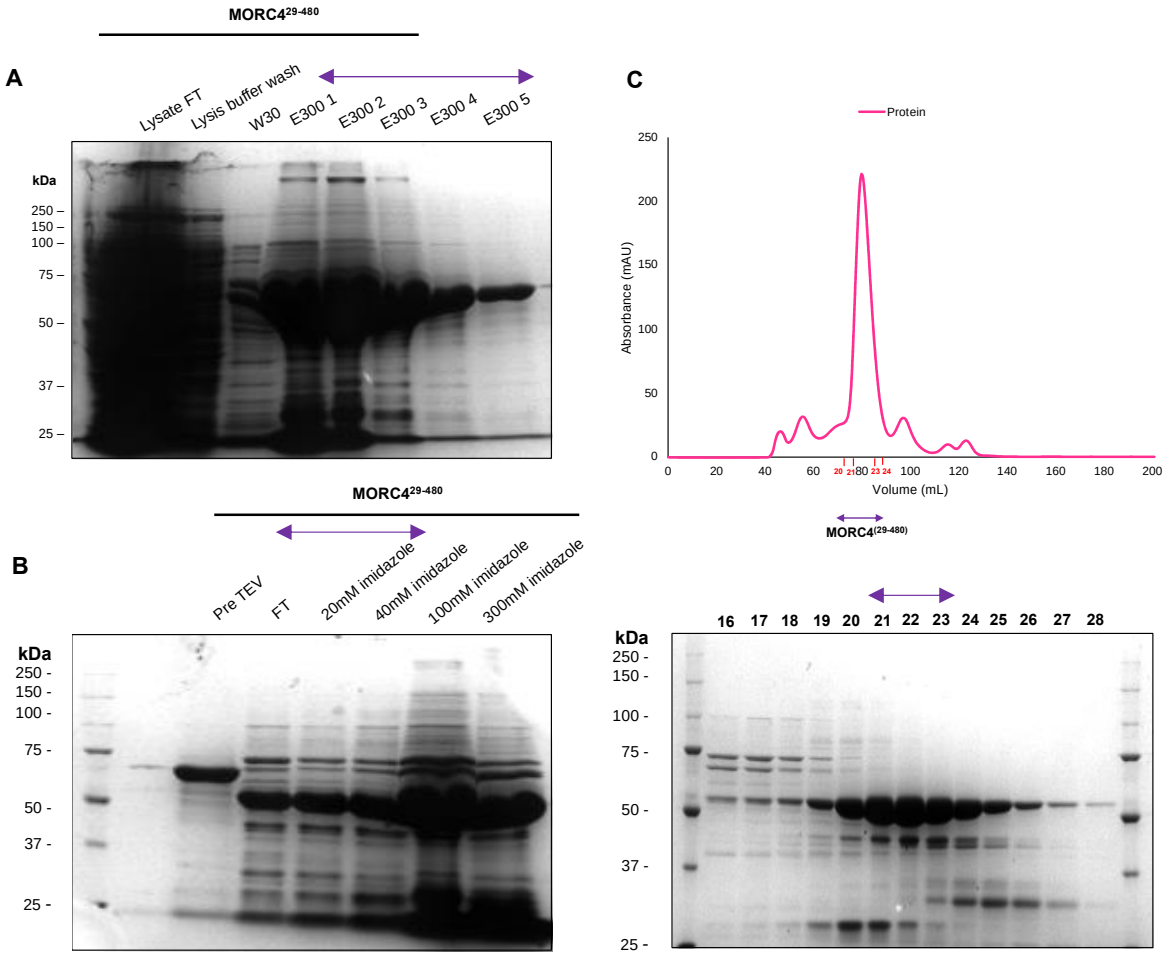


Figure 3.19: **8L Purification of MORC4$^{29\text{-}480}$ using Ni-IMAC, TEV rebind, and SEC.** MORC4$^{29\text{-}480}$ was expressed in 8 L TB, purified using a Ni-IMAC gravity flow column, cleaved using TEV protease and the TEV affinity tag was removed with an additional Ni-IMAC prior to SEC. Samples were resolved on a 10% tris-glycine SDS(TGS)-PAGE gel. **(A)** First Ni-IMAC purification gel showing W30 (wash 30 mM imidazole) and 5 E300 (elution 300 mM imidazole) washes. **(B)** Second Ni-IMAC gel image following addition of TEV protease at 4°C overnight. Gel image shows a pre-TEV sample and several imidazole washes. **(C)** A chromatogram of MORC4$^{29\text{-}480}$ following SEC, which was purified on the A HiLoad 16/600 Superdex 200 pg column using the ÄKTA Pure. Selected samples were resolved on an SDS-PAGE for analysis. Samples 21-23 were pooled, concentrated, flash-frozen in LN_2 and stored at -80°C and/or used for protein crystallography trials. Purple doubled headed arrows indicate fractions which were pooled.

To increase the yield of MORC4$^{29\text{-}480}$, the protein was purified without the use of SP-IEX (Figure 3.19). There were several more contaminants present for SEC only (Figure 3.19c)

compared to purifying MORC4$^{29-480}$ with the addition of SP-IEX. However, the SDS-PAGE gel was overloaded with protein, suggesting that if there was a smaller concentration of protein loaded, the contaminants would be minimal. SEC Purified MORC4$^{29-480}$ had a yield of 200 µl at 16.7 mg/mL (pooled samples 21-23, peak height 80.82 mL) (Figure 3.19c) and was used for subsequent protein crystallography trials. Protein required for more sensitive biochemical assays such as chemical crosslinking experiments, was also purified using SP-IEX to ensure the protein was exceptionally pure.

3.3.2.3.1.4 Optimisation of MORC4$^{29-480}$ purification

3.3.2.3.1.4.1 Optimisation of imidazole wash conditions on Ni-IMAC rebind following TEV cleavage

In an attempt to increase the overall protein yield and to determine if it was possible to purify MORC4$^{29-480}$, which was eluting with a higher concentration of imidazole (following the TEV rebind Ni-IMAC) and with other contaminants, the 100-300 mM imidazole washes containing MORC4$^{29-480}$, were pooled and purified separately from the FT and 20 mM imidazole fractions pool previously discussed (Section 3.3.2.3.1.1).

Following the initial Ni-IMAC and TEV rebind Ni-IMAC, MORC4$^{29-480}$ 100-300 mM imidazole fractions were pooled for purifying via SP-IEX (Figure 3.20a and b). The SDS-PAGE gel exhibited a more intense band for the target protein (Figure 3.20b) compared to when samples FT and 20 mM imidazole wash were pooled, suggesting a higher yield of protein. However, there was a small contaminant at present approximately 23 kDa. In addition, there was a shoulder peak present in the FT sample following SP-IEX, which can be seen on the SP-IEX chromatogram (Figure 3.20a). This indicates that there were several contaminants present in the FT sample, which is confirmed by the SDS-PAGE gel image (Figure 3.20b). The gel shows several contaminants in the FT sample along with some MORC4$^{29-480}$ protein, indicating

that it did not fully bind to the SP-IEX column. Overall, the SDS-PAGE suggests that there were several more contaminants present in the FT sample of SP-IEX with 100-300 mM imidazole MORC4$^{29\text{-}480}$ than there was for the FT and 20 mM imidazole MORC4$^{29\text{-}480}$. Regardless, 100-300 mM imidazole fractions were pooled for SEC, to attempt to remove the additional contaminants.

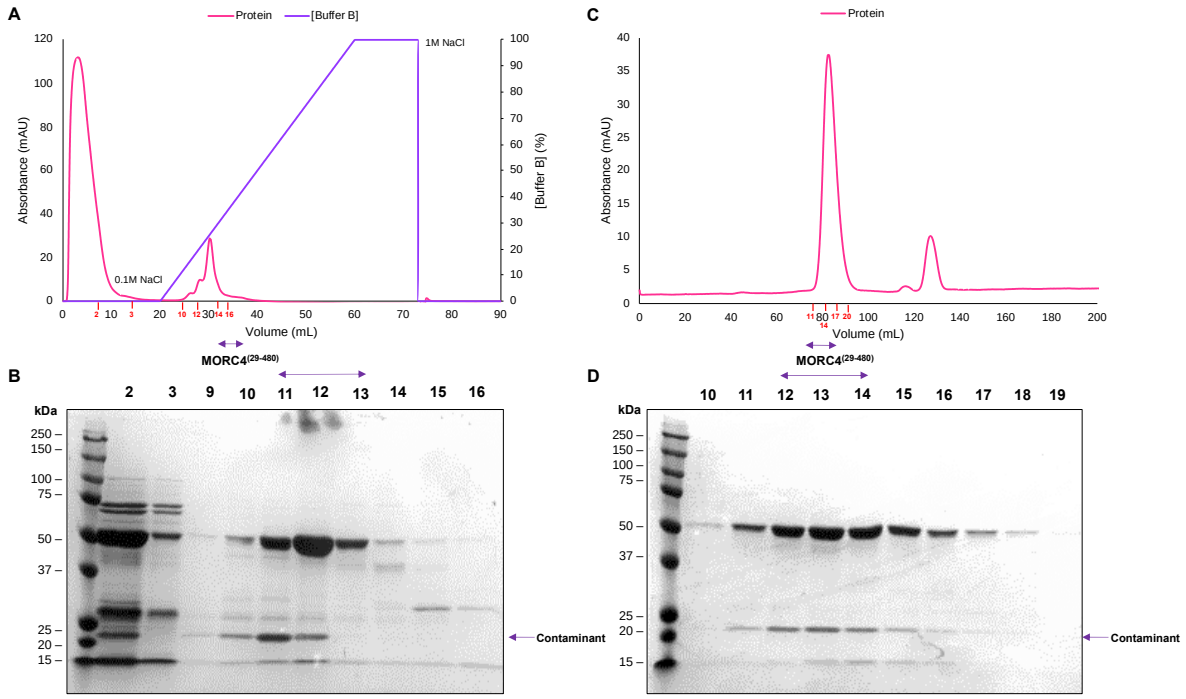


Figure 3.20: **Purification of MORC4$^{29\text{-}480}$ 100-300 mM imidazole, using SP-IEX and SEC.** Following both SP-IEX and SEC, selected samples were resolved on a 10% tris-glycine SDS (TGS)-PAGE gel. **(A)** Chromatogram of MORC4$^{29\text{-}480}$ (100-300 mM imidazole washes), which was purified on the A 1ml HiTrap SP-IEX column using the ÄKTA pure. **(B)** Selected samples were resolved on an SDS-PAGE for analysis. Samples 11, 12 and 13 were pooled for further purification on a SEC column. **(C)** A chromatogram of MORC4$^{29\text{-}480}$, which had been pooled for SEC and purified on the A HiLoad 16/600 Superdex 200 pg column using the ÄKTA Pure. D selected samples were resolved on an SDS-PAGE for analysis. Samples 12-14 were pooled, concentrated, flash-frozen in LN_2 and stored at -80°C. Purple doubled headed arrows indicate fractions which were pooled.

From the SEC chromatogram (Figure 3.20c), it is evident that there was a second UV peak, smaller than from when the protein was purified after pooling the FT and 20 mM wash. This indicates that there were more contaminants present in this sample. Following SEC, samples 12-14 were pooled (Figure 3.20d). In summary, it was clear that this protein could not be used

for subsequent biochemical assays as there was a small contaminant (~23 kDa) that co-purified after SEC (Figure 3.20d), which could interfere with experiments. This contaminant could potentially be a small chaperone from *E. coli* which was binding to and co-eluting with MORC4, as SEC separates molecules according to size.

3.3.2.3.1.4.2 1 M NaCl and 500 mM KCl has no effect on protein yield

Some MORC4$^{29\text{-}480}$ remained bound to the Ni-IMAC gravity flow column following the TEV rebind stage of the purification (Figure 3.17b). In an attempt to elute this protein from the column, 500 mM KCl and 1 M NaCl were used in the buffers for the TEV rebind step, instead of the standard 500 mM NaCl. KCl has different properties to NaCl according to the Hofmeister series, which can affect protein solubility and protein interactions (Neagu *et al.*, 2001). The reason for MORC4$^{29\text{-}480}$ remaining bound to the column could be that the protein and the Ni^{2+} were forming a strong interaction, which could not be removed with the addition of imidazole or 500 mM NaCl (Figure 3.21a and b). Therefore, increasing the concentration of NaCl could interfere with any interactions and release the protein from the column. However, 1 M NaCl had no effect on the yield of protein (Figure 3.21c), similar to 500 mM KCl, in comparison to the standard 500 mM NaCl buffer (Figure 3.21d).

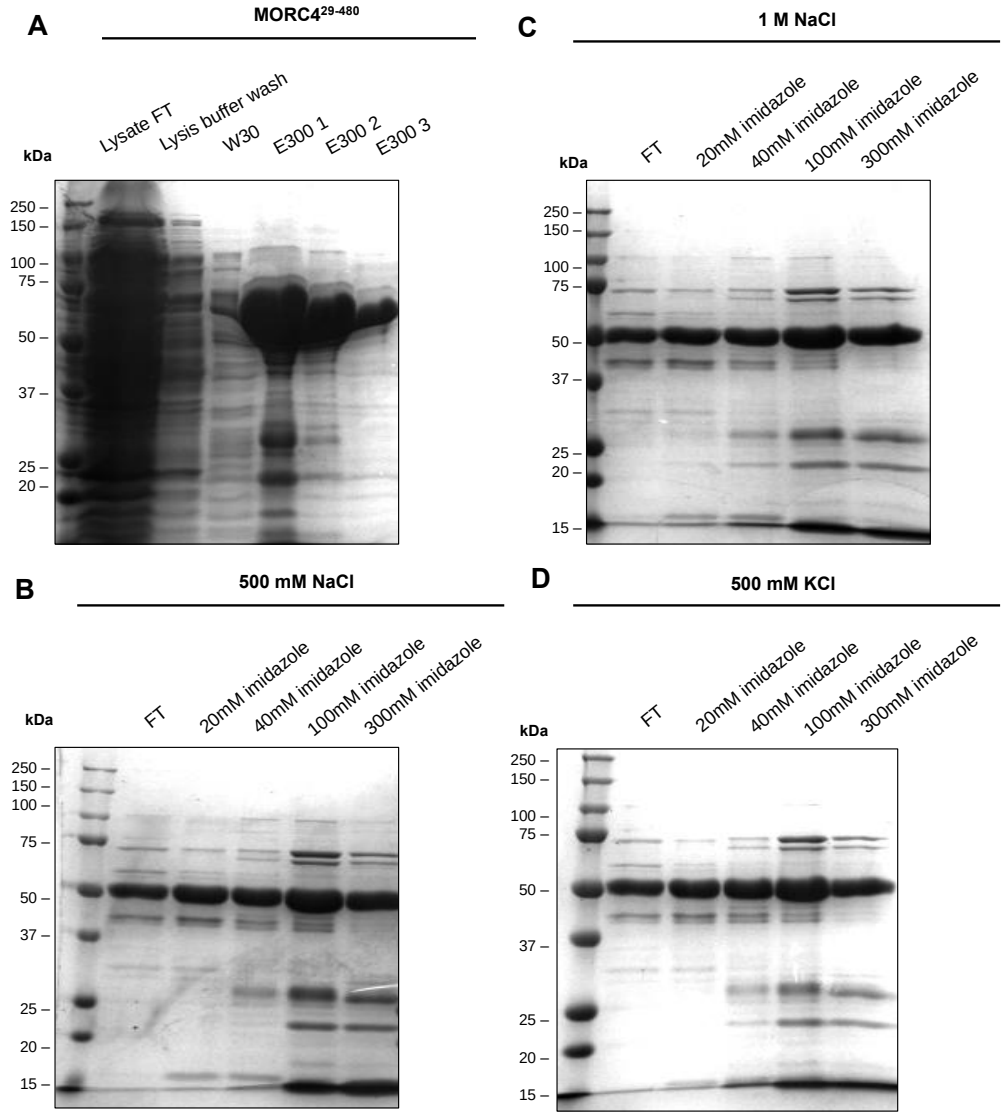


Figure 3.21: **Ni-IMAC and TEV rebind Ni-IMAC purification of MORC4$^{29\text{-}480}$ with TEV rebind buffers of 500 mM NaCl, 1 M NaCl and 500 mM KCl.** MORC4$^{29\text{-}480}$ was expressed in 2 L TB, purified using a Ni-IMAC gravity flow column, cleaved using TEV protease and the TEV affinity tag was removed with an additional Ni-IMAC. The TEV rebind Ni-IMAC was performed in triplicate using 500 mM NaCl, 1 M NaCl or 500 mM KCl in the buffers. Samples were resolved on a 10% tris-glycine SDS(TGS)-PAGE gel. **(A)** Initial Ni-IMAC. **(B)** TEV rebind Ni-IMAC with 500 mM NaCl in the buffers, as normal. **(C)** TEV rebind Ni-IMAC with 1 M NaCl in the buffers. **(D)** TEV rebind Ni-IMAC with 500 mM KCl in the buffers.

3.3.2.3.2 Purification of C-terminal domain MORC4 fragments MORC4$^{420\text{-}757}$ and MORC4$^{420\text{-}677}$

3.3.2.3.2.1 Ni-IMAC and TEV rebind ensures affinity tag removal

MORC4$^{420\text{-}757}$ (Zf-CW C-terminal fragment) was cloned by Baktawar Nawaz (undergraduate project student, 2016) and is a MORC4 C-terminal fragment, which successfully cleaved with the addition of TEV protease (Figure 3.22a). Prior to treatment with TEV protease, MORC4$^{420\text{-}757}$ was 50 kDa and after treatment it reduced in size to 38 kDa (Figure 3.22a), suggesting successful cleavage of the 6xHis-(Trx)-TEV affinity tag.

MORC4$^{420\text{-}677}$ is another MORC4 C-terminal fragment, which was also successfully cleaved on treatment with TEV protease (Figure 3.23a). Before treatment with TEV protease, MORC4$^{420\text{-}677}$ was 42 kDa and after treatment it was 30 kDa (Figure 3.23a). This is evident on the SDS-PAGE gel in the flow through (FT sample) (Figure 3.23a). Both MORC4$^{420\text{-}757}$ and MORC4$^{420\text{-}677}$ FT and 20 mM imidazole washes were pooled for purifying.

As with the ATPase-Zf-CW MORC4 fragments, some of the TEV-cleaved MORC4 CW/C-terminal domain proteins also appeared to bind to the Ni-IMAC gravity flow column, even after imidazole buffer washes. This suggests that it could be the Zf-CW domain or more specifically, the Zn^{2+} finger motif of MORC4, which is responsible for the strong binding of the protein to the Ni^{2+} resin. It could also suggest that the affinity tag was not completely removed, although this was less likely as the protein sizes were the expected sizes with the tag removed. However, failure to remove the tag completely could also be due to the Zf-CW domain, as all proteins used in this study containing the Zf-CW domain (ATPase-Zf-CW and Zf-CW proteins) showed protein binding to the Ni^{2+} resin following the imidazole buffer washes.

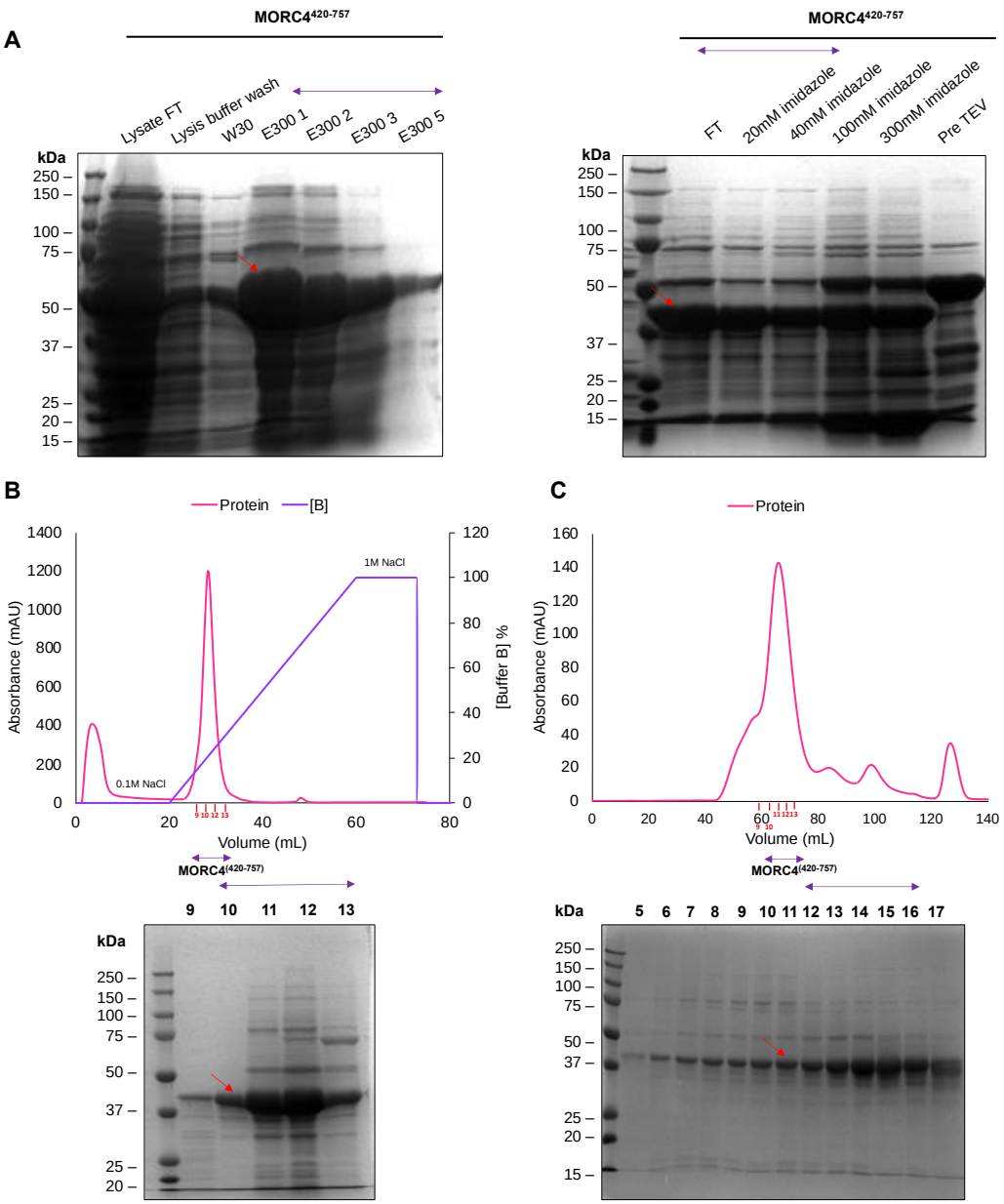


Figure 3.22: **3 L purification of MORC4$^{420\text{-}757}$ using Ni-IMAC, TEV rebind Ni-IMAC, Q-IEX and SEC.** MORC4$^{420\text{-}757}$ was expressed in 3 L TB, purified using a Ni-IMAC gravity flow column, cleaved using TEV protease and the TEV affinity tag was removed with an additional Ni-IMAC prior to SEC. Samples were resolved on a 10% tris-glycine SDS(TGS)-PAGE gel. **(A)** First Ni-IMAC purification gel showing W30 (wash 30 mM imidazole) and 5 E300 (elution 300 mM imidazole) washes. Second Ni-IMAC gel image following addition of TEV protease at 4°C overnight. Gel image shows a pre TEV sample and several imidazole washes. **(B)** Chromatogram of MORC4$^{420\text{-}757}$ and the corresponding gel image, following Q-IEX. Q-IEX was performed using a 1 ml HiTrap SP-IEX column and the ÄKTA pure. Fractions 10-13 were pooled for SEC. **(C)** Chromatogram of MORC4$^{420\text{-}757}$ and the corresponding gel image, following SEC. SEC was performed using A HiLoad 16/600 Superdex 200 pg column and the ÄKTA pure. Samples 12-16 were pooled, concentrated, flash-frozen in LN_2 and stored at -80°. Purple arrows indicate fractions which were pooled, and red arrows indicate target protein.

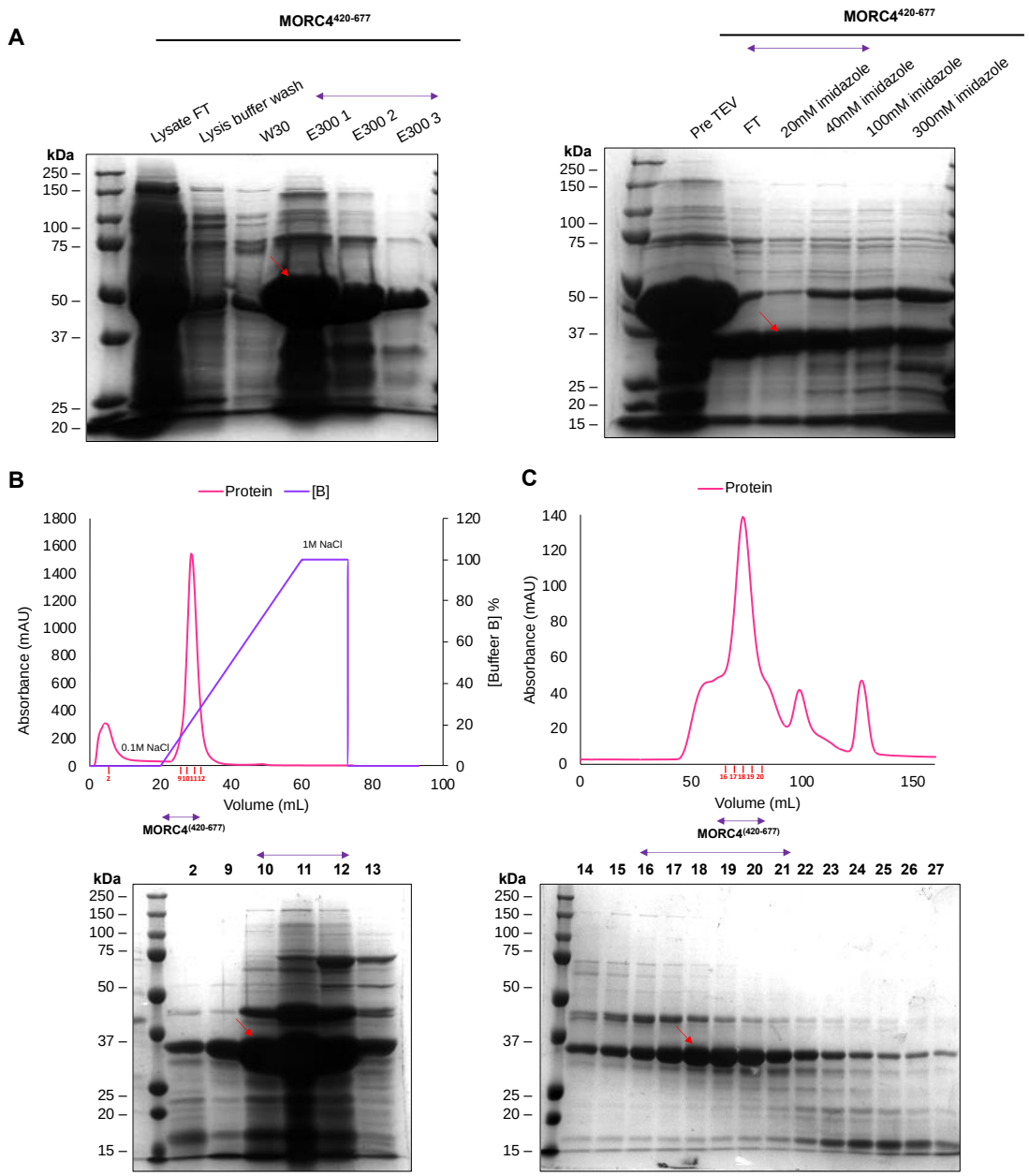


Figure 3.23: **4 L purification of MORC4$^{420-677}$ using Ni-IMAC, TEV rebind Ni-IMAC, Q-IEX and SEC.** MORC4$^{420-677}$ was expressed in 4 L TB, purified using a Ni-IMAC gravity flow column, cleaved using TEV protease and the TEV affinity tag was removed with an additional Ni-IMAC prior to SEC. Samples were resolved on a 10% tris-glycine SDS(TGS)-PAGE gel. **(A)** First Ni-IMAC purification gel showing W30 (wash 30 mM imidazole) and 5 E300 (elution 300 mM imidazole) washes. Second Ni-IMAC gel image following addition of TEV protease at 4°C overnight. Gel image shows a pre TEV sample and several imidazole washes. **(B)** Chromatogram of MORC4$^{420-677}$ and the corresponding gel image, following Q-IEX. Q-IEX was performed using a 1 ml HiTrap SP-IEX column and the ÄKTA pure. Fractions 10-12 were pooled for SEC. **(C)** Chromatogram of MORC4$^{420-677}$ and the corresponding gel image, following SEC. SEC was performed using A HiLoad 16/600 Superdex 200 pg column and the ÄKTA pure. Samples 16-21 were pooled, concentrated, flash-frozen in LN_2 and stored at -80°. Purple arrows indicate fractions which were pooled, and red arrows indicate target protein.

3.3.2.3.2.2 Anion exchange (IEX) chromatography

Following Ni-IMAC and TEV rebind, both MORC4$^{420-757}$ and MORC4$^{420-677}$ were purified using Q-IEX (anion exchange chromatography) (Figure 3.22b and 3.23b), which separated the proteins based on charge. The pI of both MORC4$^{420-757}$ and MORC4$^{420-677}$ was approximately 5.4, which means that with a buffer with a pH of 7.5, both of the proteins were negatively charged, suggesting that these proteins bound to a positively charged ion exchange column (Q-IEX).

Following Q-IEX, both MORC4$^{420-757}$ and MORC4$^{420-677}$ proteins were eluted between 0.13 M-0.24 M NaCl (Figure 3.22b and 3.23b). This suggests that the proteins bound weakly to the Q-IEX column, similar to the ATPase-Zf-CW protein, MORC4$^{29-480}$ (Figure 3.19b) (Section 3.3.2.3.1.1). Several fractions eluted from the Q-IEX column were resolved on an SDS-PAGE gel for both CW/C-terminal fragments. For both MORC4$^{420-757}$ and MORC4$^{420-677}$, fractions 10-12 were pooled for subsequent purification steps (Figure 3.22b and 3.23b).

3.3.2.3.2.3 Size exclusion chromatography (SEC)

Following Q-IEX, both MORC4$^{420-757}$ and MORC4$^{420-677}$ were purified using SEC as a final purification stage (Figure 3.22c and 3.23c), which separated any remaining contaminants based on size. Following SEC on MORC4$^{420-757}$, fractions 12-16 were pooled, which had a peak of 65.90mL (Figure 3.22c). Following 3 L expression and purification of MORC4$^{420-757}$, there was a yield of 300 µl of 14.8 mg/mL. Following SEC on MORC4$^{420-677}$, fractions 16-21 were pooled, which had a peak of 73.3mL (Figure 3.23c). Following 4 L expression and purification of MORC4$^{420-677}$, there was yield of 200 µl of 22.3 mg/ml. Overall, the final stage of SEC revealed that MORC4 was not completely pure even after 4 stages of the purification process, especially MORC4$^{420-677}$. However, the smaller contaminants could be some degradation of MORC4. Nevertheless, proteins were still deemed sufficient for crystallography trials.

3.3.2.3.3 Purification optimisation of ATPase-Zf-CW and C-terminal fragments

In an attempt to increase the yield of protein and to decrease the amount of protein which bound to the Ni-IMAC gravity flow column following TEV rebind, 0.1 mM exogenous Zn^{2+} was added before induction, at the protein expression stage prior to purification (Langelier *et al.*, 2008). All MORC4 fragment proteins with the Zf-CW domains seemed to bind to the Ni^{2+} resin following imidazole washes. This suggests that the Ni^{2+} resin could displace the Zn^{2+} in the Zn^{2+} finger motif, allowing the protein to bind to the Ni^{2+} resin despite the imidazole washes. Expressing protein in the presence of Zn^{2+} could fully saturate the protein with Zn^{2+}, therefore reducing potential aberrant interactions between the protein Zn^{2+} finger motif and the nickel resin, only allowing an interaction between the Ni^{2+} resin and the histag.

3.3.2.3.3.1 Zn^{2+} increases MORC4$^{29-480}$ protein expression yield

MORC4$^{29-480}$ was expressed in 1 L of TB in the presence and absence of exogenous Zn^{2+}. Following protein expression, MORC4$^{29-480}$ was purified using Ni-IMAC and TEV rebind Ni-IMAC in an equivalent manner for both samples. The initial Ni-IMAC showed an increased quantity of MORC4$^{29-480}$ expressed in the presence of Zn^{2+} (Figure 3.24a), this suggests that there was more functional protein present and that Zn^{2+} could be important for structural stability or folding of the CW domain, similar to other zinc finger proteins (Langelier *et al.*, 2008). However, the protein still remained bound to the Ni-NTA resin following 20 mM imidazole wash during the TEV rebind Ni-IMAC (Figure 3.24b). This indicates that Zn^{2+} does not disrupt any interaction between the Ni-NTA resin and the protein. It suggests that the use of Zn^{2+} is only required to increase protein expression yield for MORC4$^{29-480}$ rather than to reduce the amount of protein which binds to the column following affinity tag cleavage.

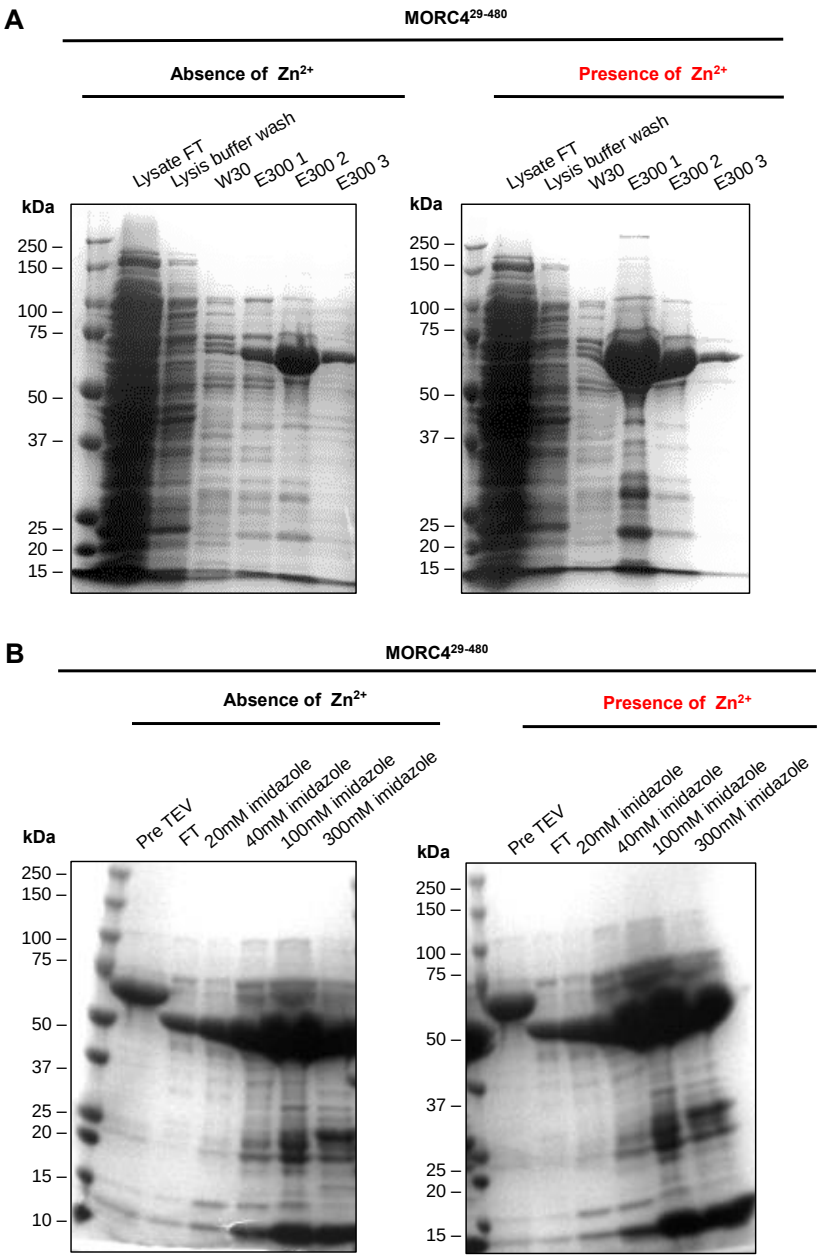


Figure 3.24: **MORC4$^{29-480}$ Ni-IMAC and TEV rebind Ni-IMAC in the absence and presence of 0.1 mM exogenous Zn^{2+}.** MORC4$^{29-480}$ was expressed in 1 L TB in the absence and presence of 0.1 mM Zn^{2+}, purified using a Ni-IMAC gravity flow column, cleaved using TEV protease and the TEV affinity tag was removed with an additional Ni-IMAC. Samples were resolved on a 10% tris-glycine SDS(TGS)-PAGE gels. **(A)** Ni-IMAC purification gel showing W30 (wash 30 mM imidazole) and 3 E300 (elution 300 mM imidazole) washes in the absence and presence of Zn^{2+}. **(B)** Ni-IMAC gel image following addition of TEV protease at 4°C overnight. Gel images show a pre-TEV sample and several imidazole washes in the absence and presence of Zn^{2+}.

3.3.2.3.3.2 Zn^{2+} has no effect on $MORC4^{420-480}$ protein expression yield

As with MORC4 ATPase-Zf-CW fragments, C-terminal fragments of MORC4 also remain bound to the Ni-IMAC column following affinity tag removal and imidazole washes. To determine if there could be an interaction between the C-terminal and the Ni-NTA resin, a duplicate experiment involving a C-terminal MORC4 fragment, $MORC4^{420-480}$, was performed (Figure 3.25a and b). Unlike $MORC4^{29-480}$, which includes the ATPase-Zf-CW domains, Zn^{2+} had no effect on the protein expression of the $MORC4^{420-480}$ CW/C-terminal fragment (Figure 3.25a). As with $MORC4^{29-480}$, Zn^{2+} had no effect on the release of protein following affinity tag removal (Figure 3.25b). Overall, these results indicate that excess Zn^{2+} assisted the tandem domain (ATPase-Zf-CW) protein to fold more efficiently, increasing the protein expression yield.

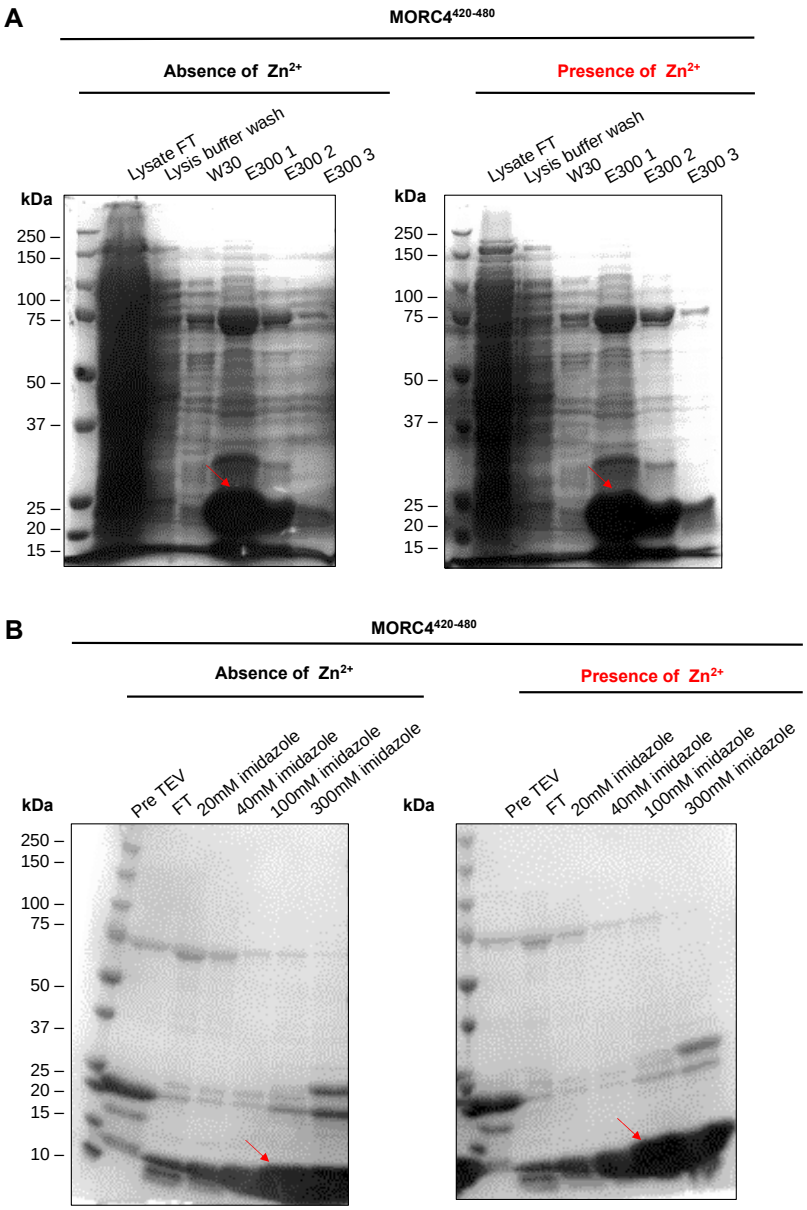


Figure 3.25: **MORC4$^{420-480}$ Ni-IMAC and TEV rebind Ni-IMAC in the absence and presence of 0.1 mM exogenous Zn^{2+}.** MORC4$^{420-480}$ (19 kDa) was expressed in 1 L TB in the absence and presence of 0.1 mM Zn^{2+}, purified using a Ni-IMAC gravity flow column, cleaved using TEV protease and the TEV affinity tag was removed (7 kDa) with an additional Ni-IMAC. Samples were resolved on a 10% tris-glycine SDS(TGS)-PAGE gels. **(A)** Ni-IMAC purification gel showing W30 (wash 30 mM imidazole) and 3 E300 (elution 300 mM imidazole) washes in the absence and presence of Zn^{2+}. **(B)** Ni-IMAC gel image following addition of TEV protease at 4°C overnight. Gel images show a pre-TEV sample and several imidazole washes in the absence and presence of Zn^{2+}. Red arrows indicate target protein.

3.3.2.3.4 Purification of ATPase domain MORC4 fragments

For downstream biochemical analysis including ATPase assays, ATPase domain-only fragments of MORC4 were expressed and purified using Ni-IMAC, IEX and SEC. In addition to ATPase-Zf-CW MORC4 proteins, the ATPase-only fragments were also important for subsequent ATPase assays as some evidence suggests that the CW domain of MORC3 can inhibit ATPase activity (Zhang *et al.*, 2019). Both MORC4$^{29-430}$ and MORC4$^{51-404}$ were cloned by Sa'eeda Akbor (undergraduate project student, 2016) and were chosen from the ATPase-only fragments as each protein had a different N- and C-termini and expressed well (Figure 3.9a and 3.10a).

3.3.2.3.4.1 MORC4$^{51-404}$ is poorly expressing and uncleavable

Following protein expression of MORC4$^{51-404}$, the initial Ni-IMAC SDS-PAGE result indicated poor expression of this ATPase domain only protein (Figure 3.26a). Prior to treatment with TEV protease, MORC4$^{51-404}$ was 52 kDa. Successful removal of the protein affinity tag should have resulted in a protein of 40 kDa. Subsequent TEV protease treatment and cleavage of the affinity tag SDS-PAGE gel revealed that MORC4$^{51-404}$ remained at 52 kDa (Figure 3.26b), suggesting that TEV protease failed to remove the 6xHis-(Trx)-TEV affinity tag.

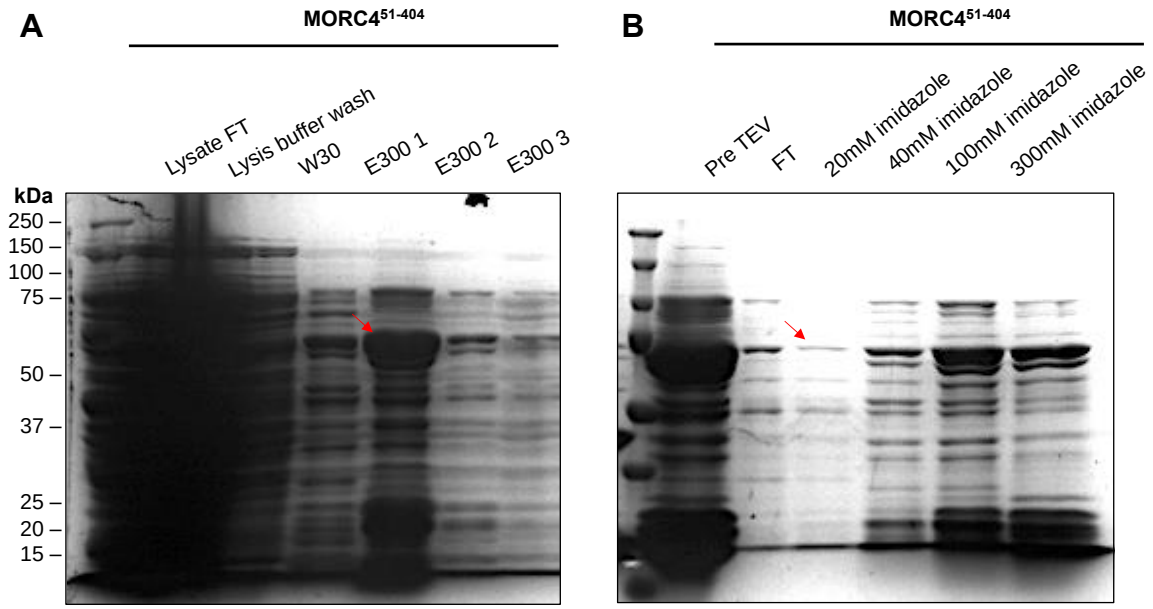


Figure 3.26: **4 L purification of ATPase only fragment, MORC4$^{51-404}$, using Ni-IMAC and TEV rebind.** MORC4$^{51-404}$ was expressed in 4 L TB, purified using a Ni-IMAC gravity flow column, cleaved using TEV protease and the TEV affinity tag was removed with an additional Ni-IMAC. Samples were resolved on a 10% tris-glycine SDS(TGS)-PAGE gels. **(A)** First Ni-IMAC purification gel showing W30 (wash 30 mM imidazole) and 3 E300 (elution 300 mM imidazole) washes. **(B)** Second Ni-IMAC gel image following addition of TEV protease at 4°C overnight. Gel images show a pre TEV sample and several imidazole washes. Red arrows indicate target protein.

3.3.2.3.4.2 MORC4$^{29-430}$ is highly expressing and TEV cleavable

MORC4$^{29-430}$ is an example of a high expressing ATPase domain only MORC4 fragment (Figure 3.27a). Following treatment of TEV protease, the Ni-IMAC SDS-PAGE gel showed a decrease in size of MORC4$^{29-430}$ to 45 kDa (Figure 3.27a). Prior to TEV protease treatment, MORC4$^{29-430}$ was 57 kDa, indicating cleavage of the 6xHis-(Trx)-TEV affinity tag.

3.3.2.3.4.2.1 Cation exchange chromatography (IEX) of MORC4$^{29-430}$

Following Ni-IMAC and TEV rebind purification, MORC4$^{29-430}$ was purified using SP-IEX. MORC4$^{29-430}$ was eluted from the SP-IEX column at approximately 0.3M NaCl (Figure 3.27b). Several fractions eluted from the SP-IEX column were resolved on an SDS-PAGE gel and fractions 15 and 16 were pooled for subsequent purification steps.

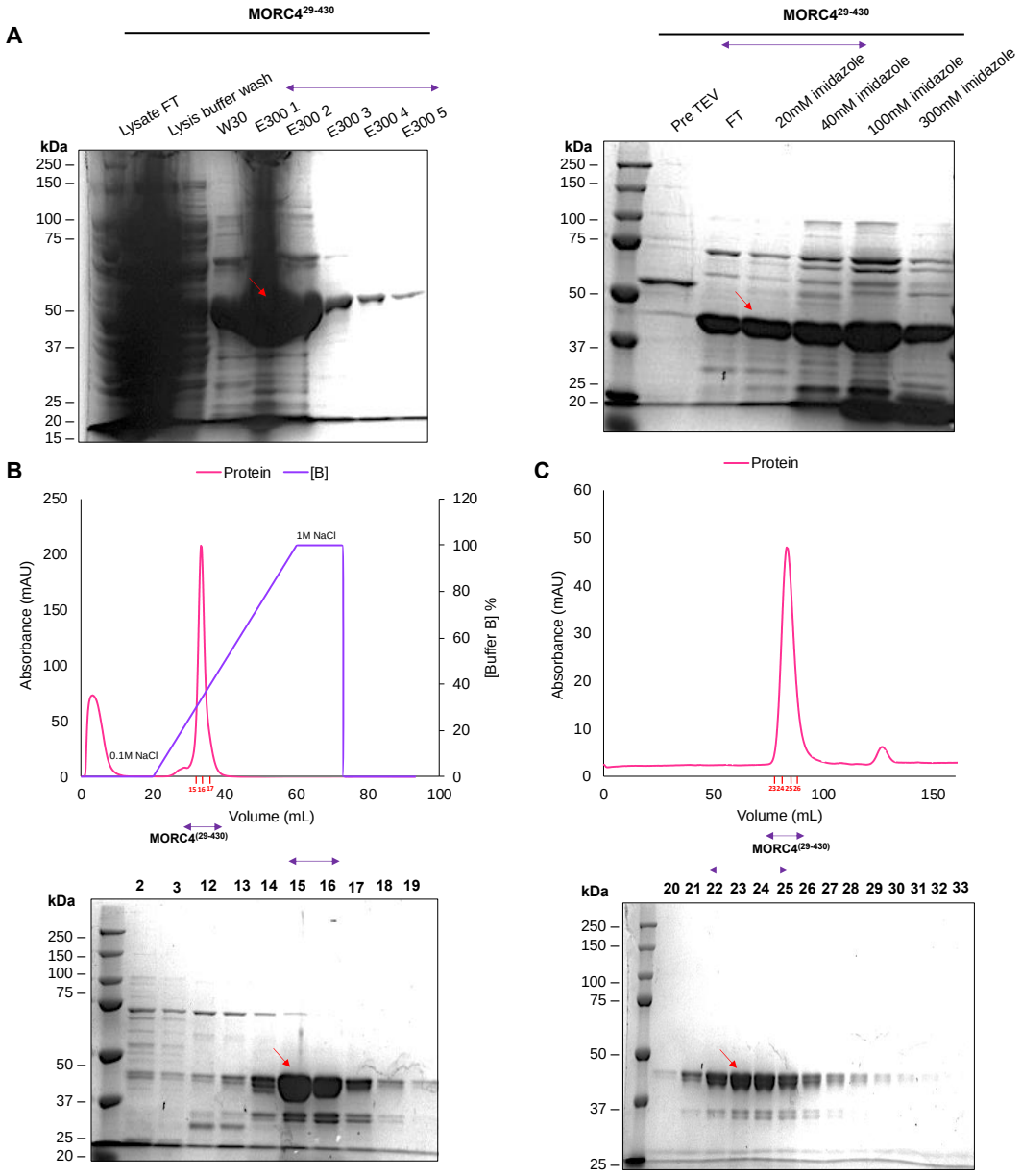


***Figure 3.27:* 4 L purification of MORC4$^{29\text{-}430}$ using Ni-IMAC, TEV rebind Ni-IMAC, SP-IEX and SEC.** MORC4$^{29\text{-}430}$ was expressed in 4 L TB, purified using a Ni-IMAC gravity flow column, cleaved using TEV protease and the TEV affinity tag was removed with an additional Ni-IMAC prior to SEC. Samples were resolved on a 10% tris-glycine SDS(TGS)-PAGE gel. **(A)** First Ni-IMAC purification gel showing W30 (wash 30 mM imidazole) and 5 E300 (elution 300 mM imidazole) washes. Second Ni-IMAC gel image following addition of TEV protease at 4°C overnight. Gel image shows a pre TEV sample and several imidazole washes. **(B)** Chromatogram of MORC4$^{29\text{-}430}$ and the corresponding gel image, following SP-IEX. SP-IEX was performed using a 1 ml HiTrap SP-IEX column and the ÄKTA pure. Samples 15 and 16 were pooled for SEC. (C) Chromatogram of MORC4$^{29\text{-}430}$ and the corresponding gel image, following SEC. SEC was performed using A HiLoad 16/600 Superdex 200 pg column and the ÄKTA pure. Samples 22-25 were pooled, concentrated, flash-frozen in LN_2 and stored at -80°. Purple arrows indicate which fractions were pooled and red arrows indicate target protein.

3.3.2.3.4.2.2 Size exclusion chromatography (SEC) of MORC4$^{29-430}$

Following SP-IEX, MORC4$^{29-430}$ was purified using SEC as the last purification stage. MORC4$^{29-430}$ was eluted from the SEC column at 83.3 mL and fractions 22 to 25 were pooled (Figure 3.27c). After expressing MORC4$^{29-430}$ in 4L of TB, there was a yield of 1.2 mL at 1mg/ml. However, MORC4$^{29-430}$ did contain some contaminants which were co-eluting with the target protein peak, suggesting that they could have been degradation products of MORC4$^{29-430}$. Following SEC, the protein was flash-frozen and stored at -80°C until required for ATPase assays.

3.3.2.4 Determination of MORC4 protein purity by mass spectrometry (MS)

3.3.2.4.1 ATPase-Zf-CW fragment of MORC4

To determine the size and purity of the ATPase-CW fragment, MORC4$^{29-480}$, a sample of purified protein was analysed by intact mass spectrometry (MS) (Structural Genomics Consortium (SGC), University of Oxford). Intact mass analysis confirmed that MORC4$^{29-480}$ was 51.6 kDa, which was within 1 Da of the expected size of 51.6 kDa (Figure 3.28a and b). This suggests that the protein was not degraded, oxidised or modified during the purification process. It also suggests that the protein sequence is correct (also confirmed by DNA sequencing analysis). However, the sample did show some additional peaks other than the target protein (Figure 3.28b), suggesting that they could be sodium adducts or oxidation products caused by the sample being left at room temperature throughout the transportation process.

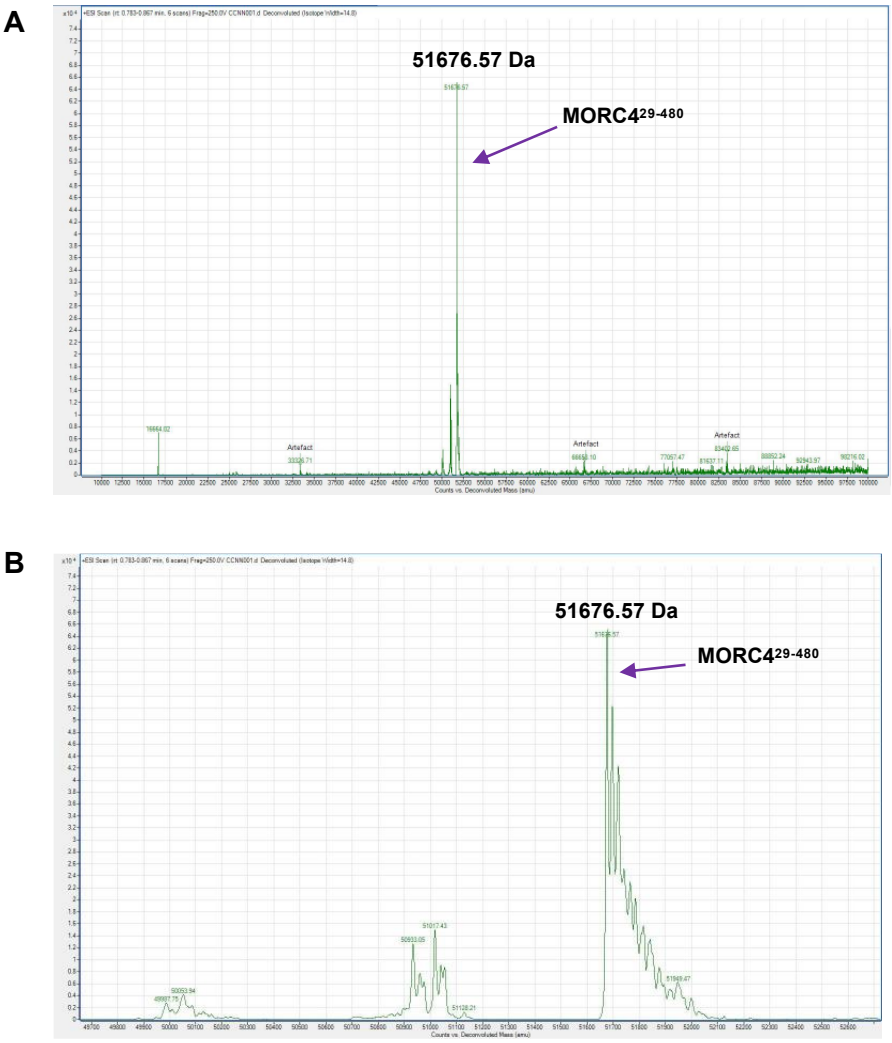
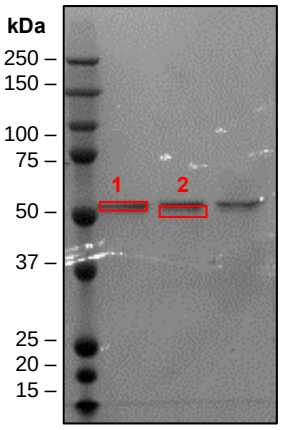


Figure 3.28: **Intact mass spectrometry (MS) determined the size of ATPase-Zf-CW MORC4$^{29-480}$.** Purified MORC4$^{29-480}$ was sent to the Structural Genomics Consortium (SGC), University of Oxford for intact MS analysis. **(A)** Chromatogram showing MORC4$^{29-480}$ peak at 51.67657 kDa. **(B)** Zoomed in chromatogram showing MORC4$^{29-480}$ peak at 51.67657 kDa.

An SDS-PAGE gel (Figure 3.29a) from SEC of MORC4$^{29-480}$ was also analysed by tryptic MSMS analysis to confirm the identity of the purified MORC4. Throughout several purifications of MORC4$^{29-480}$, there was often a doublet of bands present following SEC. Both bands were extracted from the SEC SDS-PAGE and sent for analysis to determine if MORC4$^{29-480}$ was present, and to assess if the other fragment was a degradation product or a contaminant. MSMS confirmed that the smaller product (Figure 3.29a and b) on the SEC SDS-PAGE gel was MORC4, suggesting that there was a small amount of proteolytic degradation. However, the larger protein band gave an insufficient MASCOT score (Figure 3.29b) for a confident

identification, although, this protein band could have been MORC4. Overall, these data confirmed that MORC4$^{29-480}$ was the correct size and that it had been purified successfully.

A



B

SDS-PAGE band	Mascot score	Peptide	Protein?
1	32	APOBEC3AA-c009	No protein
	91	Trypsin	
	51	β-galactosidase	
	46	β-galactosidase	
	46	ATP synthase subunit delta	
	45	Protein transport protein	
2	30	TLK2A-c020	MORC4$^{29-480}$
	404	MORC family CW-type zinc finger protein	
	63	MORC family CW-type zinc finger protein	
	61	DNA-directed RNA polymerase subunit gamma	
	59	Serine/threonine-protein kinase	
	51	Serine/threonine-protein kinase	

Figure 3.29: **Tandem mass spectrometry (MSMS) determined peptides present following SEC of MORC4$^{29-480}$.** Bands from an SDS-PAGE gel from SEC purification of MORC4$^{29-480}$ were extracted and sent for tryptic MSMS analysis at the Structural Genomics Consortium (SGC), University of Oxford. **(A)** SEC SDS-PAGE gel showing bands 1 and 2. **(B)** Table showing the peptides found from MSMS analysis with the mascot score and detection of protein.

3.3.2.4.2 C-terminal fragments of MORC4

Two C-terminal fragments of MORC4 were also analysed by MS (Figure 3.30). Intact mass of both MORC4$^{420\text{-}757}$ and MORC4$^{420\text{-}677}$ confirmed that the proteins were 38.3 kDa and 29.8 Da, respectively (Figure 3.30a and b). Both proteins were within 1 kDa of the expected size (MORC4$^{420\text{-}757}$ expected size 38.1 kDa and MORC4$^{420\text{-}677}$ expected size was 29.7 kDa) suggesting that, like the ATPase-Zf-CW protein MORC4$^{29\text{-}480}$, there had been no modifications to the proteins during expression or purification and that the proteins had the correct protein sequence, which was also confirmed with DNA sequencing.

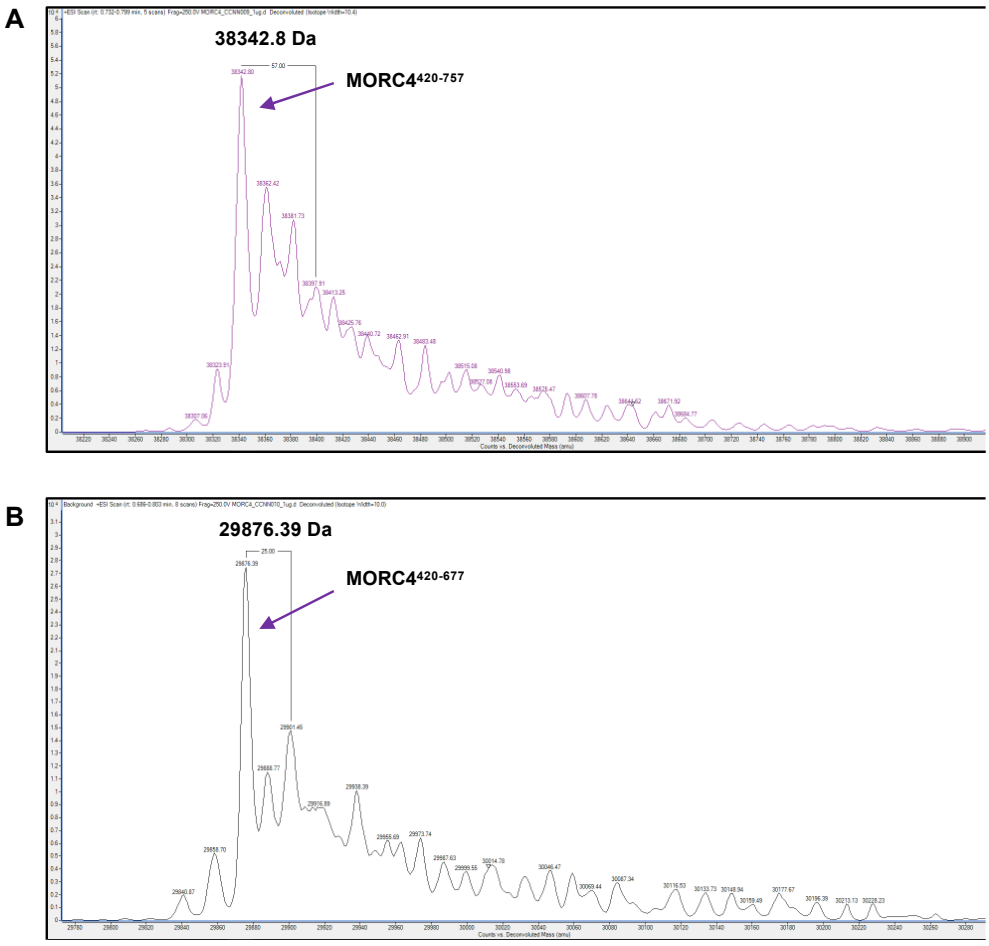


Figure 3.30: **Intact mass spectrometry (MS) determined the size of C-terminal fragments, MORC4$^{420\text{-}757}$ and MORC4$^{420\text{-}677}$.** Purified MORC4$^{420\text{-}757}$ and MORC4$^{420\text{-}677}$ were sent to the Structural Genomics Consortium (SGC), University of Oxford for intact MS analysis. **(A)** Chromatogram showing MORC4$^{420\text{-}757}$ peak at 38342.8 kDa. **(B)** Chromatogram showing MORC4$^{420\text{-}677}$ peak at 29876.39 kDa.

3.4 Discussion

To produce soluble MORC4 protein, high-throughput ligation independent cloning (LIC) was utilised to produce *E. coli* expression plasmids containing fragments of *MORC4*, which were used for subsequent protein expression analysis to determine soluble MORC4 proteins, for downstream biochemical and structural analysis. Prior to LIC, soluble domain boundaries of MORC4 were designed using protein sequence bioinformatics analysis. Sequence analysis revealed that MORC4 has an N-terminal ATPase domain, a central Zf-CW domain and a C-terminal coiled-coil domain. The analysis also suggested that MORC4 has a Mg^{2+} binding site, similar to MORC2 (Douse *et al.*, 2018) and MORC3 (Li *et al.*, 2016). Like MORC2 and MORC3, this suggests that the N-terminus of MORC4 could also be involved in the ATP hydrolysis mechanism (Frick *et al.*, 2007). In addition, both GLOBPLOT and PSIPRED analysis determined the disorder and secondary structure of MORC4, respectively. These bioinformatics analyses were used to identify suitable regions of the *MORC4* gene for the design of fragments for subsequent cloning.

Following the design of *MORC4* truncated fragments, LIC was used to create the fragments, which were cloned into several pET *E. coli* expression vectors such as pNH-TrxT, pNIC-Zb and pGTVL2. Each expression vector had an N-terminal protein tag, which in addition to the expression vector, increased the likelihood of creating a soluble protein fragment of MORC4. Each cloned fragment of MORC4 was initially used for a test expression in both LB and TB, which helped to identify high expressing fragments, which were used for multiple litre protein expression and purification following successful affinity tag cleavage.

Test expressions revealed that TB expressed a higher yield of protein in comparison to LB, suggesting that TB produces higher *E. coli* cell densities than LB, indicating that TB is a richer medium (Kram and Finkel, 2015). In addition, contaminants were found to be co-purifying in

several test expressions. An example of a contaminant which co-purified at 75 kDa in several test expressions, was likely to be the *E. coli* chaperone protein, DnaK (C. Cooper, *pers. comm.*). DnaK is an chaperone and ATPase (Barthel *et al.*, 2001) which plays a vital role in DNA replication (Jha *et al.*, 2017). Subsequent purification procedures were therefore important in the removal of DnaK, as it is an ATPase, suggesting that unless removed, it could interfere with subsequent MORC4 ATPase assays.

MORC4$^{29\text{-}480}$ comprises the ATPase-Zf-CW tandem module and produced a high yield of protein following test expressions and was therefore used for initial 1L protein expression and purifications to identify if the 6xHis-Trx-TEV affinity tag was cleavable with the addition of TEV protease. The TEV protease concentration was optimised, revealing that a concentration of 0.03mg/ml was required to remove the affinity tag. MORC4$^{29\text{-}900}$, MORC4$^{29\text{-}757}$ and MORC4$^{29\text{-}677}$ were MORC4 proteins which were cleavable, however, MORC4$^{29\text{-}480}$ had a higher yield of protein, suggesting it was more suitable for multiple litre protein expression and purification. Conversely, MORC4 proteins which had a different N-terminus and began at residue 51 (phenylalanine) did not cleave, however, it had the same C-terminus as the cleaved protein (MORC4$^{29\text{-}480}$). This suggests that it was the N-terminal starting position which influenced if the proteins were cleavable, presumably as a result of steric hindrance.

Failure to remove the 6xHis-Trx-TEV affinity tag could be due to a number of factors. There is some evidence to suggest that affinity protein tags can disrupt the structure and biochemical activity of some proteins such as in *Staphylococcus aureus* (Horchani *et al.*, 2009). Disruption of the structure of proteins following the addition of the affinity tag could therefore prevent the removal of the tag, if the structure no longer allows access to the TEV cleavage site (C. Cooper, *pers. comm.*). The inability of TEV protease to remove the affinity tag could also be caused by steric hindrance, suggesting that either the affinity tag or a region of the MORC4 structure

could be blocking access of TEV protease to the cleavage site (Waugh, 2011), consequently causing failure of affinity tag cleavage.

Following 1 L protein expression and purification, several fragments of MORC4 proteins were used for 8 L protein expression and purification. MORC4$^{29\text{-}480}$ was expressed in 8 L of TB and cleaved using TEV protease. Following removal of the affinity tag, Ni-IMAC unsuccessfully removed all the protein form the column following imidazole washes, resulting in a substantial amount of protein remained bound to the column. This suggests that there was a strong interaction between the protein and the Ni^{2+} resin. MORC4$^{29\text{-}480}$ was then purified using cation IEX, which separated the protein and contaminants based on charge. Following IEX, MORC4$^{29\text{-}480}$ was purified using SEC, which separated the protein and contaminants based on size and is often the final 'polishing' step of protein purification.

Following the substantial amount of protein which bound to the Ni-IMAC column following the imidazole washes, optimisation of MORC4$^{29\text{-}480}$ included pooling 100-300 mM imidazole washes in an attempt to increase the overall yield of protein. The IEX SDS-PAGE gel suggested an increase in the yield of protein following the use of this pool. However, there was a small contaminant which co-purified with the protein, which could not be removed with SEC. This indicates that this protein could not be used for downstream biochemical assays or structural analysis. In addition, optimisation of MORC4$^{29\text{-}480}$ also included the use of 1 M NaCl and 500 mM KCl, in an attempt to reduce the amount of protein which binds to the Ni-IMAC column following imidazole washes. This optimisation was unsuccessful and neither 1 M NaCl or 500 mM KCl had any effect on protein yield.

The C-terminal fragments MORC4$^{420\text{-}757}$ and MORC4$^{420\text{-}677}$ were also successfully cleaved following the addition of TEV protease. They were also purified using anion IEX and final 'polishing step', SEC. Similar to ATPase-Zf-CW MORC4 proteins, C-terminal fragments also

appeared bound to the Ni-IMAC column following imidazole washes. As a similar result is seen for both the ATPase-Zf-CW and the C-terminal fragments, it could suggest that this strong binding could be due to the Zf-CW domain, more specifically, the Zn^{2+} finger motif. However, some of the ATPase only MORC4 fragments also remained bound to the Ni-IMAC column following imidazole washes, suggesting that it was perhaps due to patches of histidine residues (Bornhorst and Falke, 2000) in the MORC4 proteins, causing non-specific binding to the Ni^{2+} resin rather than the Zn^{2+} finger motif.

In an attempt to increase the protein yield and decrease the amount of protein which binds to the column following imidazole washes, both the ATPase-Zf-CW protein and C-terminal fragments were expressed in the presence and absence of exogenous Zn^{2+} (Langelier *et al.*, 2008). Exogenous Zn^{2+} was found to increase $MORC4^{29-480}$ protein expression, however, the quantity of protein which bound to the column following imidazole washes did not decrease. This suggests that Zn^{2+} is important for stabilising functional MORC4 protein, similar to other zinc finger proteins (MacPherson *et al.*, 2006). The addition of exogenous Zn^{2+} did not reduce the quantity of protein that binds to the column following imidazole washes. This indicates that the Zn^{2+} did not disrupt the interactions between the Ni-NTA resin and the protein. Conversely, exogenous Zn^{2+} had no effect on the protein expression of C-terminal fragment $MORC4^{420-480}$. This suggests that the MORC4 tandem domain (ATPase-Zf-CW) required an excess of Zn^{2+} in order for the protein to fold more efficiently, which consequently led to an increase in expression yield in *E. coli*.

To determine MORC4 protein purity, samples of MORC4 ATPase-Zf-CW fragments and C-terminal fragments were sent to The Structural Genomics Consortium, The University of Oxford, for MS analysis. Both the ATPase-Zf-CW and C-terminal MORC4 were both the correct size. This suggests that both were purified successfully and were not introduced to any

modifications during the purification process and they had not been degraded or oxidised. MS determined that the proteins could be used for subsequent structural and biochemical analysis.

Overall, the successful production of several MORC4 proteins covering both individual (ATPase or Zf-CW only) and tandem (ATPase-Zf-CW) domains produced proteins to both a high yield and purity. These truncated MORC4 proteins could be used for subsequent biochemical and structural analyses of MORC4.

CHAPTER 4

Results

Biochemical, structural and sequence analysis of MORC4

Chapter 4 Biochemical, structural and sequence analysis of MORC4

4.1 Introduction

Biochemical and structural characterisation of MORC4 is essential for determination of function. Chapter 4 utilised the MORC4 proteins purified in Chapter 3, towards biochemical and structural analysis using an *in vitro* protein approach. In Chapter 4, site-directed mutagenesis (SDM) of *MORC4* fragments were used to create mutants, which were then expressed, purified and used for biochemical experiments such as chemical crosslinking. Unfortunately, since this project was initiated, structures of both MORC2 (Douse *et al.*, 2018) and MORC3 (Li *et al.*, 2016) have been determined, with MORC3 the paradigm for the family as it is the most characterised family member. In addition, MORC2 and MORC3 have been observed to dimerise, function as active ATPases and chromatin remodellers (D Q Li *et al.*, 2012; Zhang *et al.*, 2019). Thus, some of these methods utilised for characterisation of MORC2 and MORC3 will be approached in an aid to characterise MORC4.

MORC2 (D Q Li *et al.*, 2012) and MORC3 (Zhang *et al.*, 2019) have previously been shown to contain an active ATPase domain, suggesting that MORC4 may also be an ATPase. In addition, sequence analysis during fragment design of MORC4 (Chapter 3, section 3.2.2) suggested that MORC4 had an ATPase domain. However, this chapter was required to identify and characterise MORC4 as an ATPase, therefore, ATPase assays and chemical crosslinking could be utilised for this characterisation. Chemical crosslinkers were used to investigate the oligomeric state of MORC4 in the presence and absence of an analogue of ATP (AMP-PNP), using both wild type and mutant MORC4 proteins. Biophysical analyses were also utilised to investigate the potential dimerisation of MORC4 on AMP-PNP binding, as both MORC2 and MORC3 have been previously found to form dimers in the presence of the ATP analogue AMP-PNP (Li *et al.*, 2016; Douse *et al.*, 2018). Size exclusion chromatography-multiangle light

scattering (SEC-MALS) and analytical size exclusion chromatography (SEC) was used to determine the oligomeric state of pure MORC4 in the presence and absence of ligand, AMP-PNP. These experiments could help to contribute to a greater understanding of MORC4 biochemical function and will help to assess if MORC4 acts similarly to MORC2 and MORC3.

Structural characterisation studies were also attempted using expressed and purified fragments of MORC4. Purified N-terminal (ATPase/CW domains) MORC4 protein was used for protein crystallisation trials and resulting crystals were screened by X-ray diffraction. In addition, bioinformatics analyses were used for homology modelling as suitably diffracting crystals were yet to be found, and to perform multiple sequencing alignments. Homology modelling was important to predict structures of proteins with challenges faced during protein crystallisation methods. In addition, multiple sequence alignments were required to investigate human protein sequences of proteins and compare them to other species and classes of vertebrates. This helped to investigate evolutionary patterns and sequence conservation of MORC4, which could contribute to the understanding of the roles of domains and residues in MORC4 function. Overall, all of these methods are key for investigating the structure and function of MORC4.

4.2 Bioinformatic analysis of MORC proteins

4.2.1 Multiple sequence alignments (MSAs)

Multiple sequence alignments (MSAs) of protein sequences were useful for investigating the conservation of each MORC in several species. Each MSA was completed using multiple sequence comparison by log-expectation (MUSCLE) and were visualised using Jalview (Waterhouse *et al.*, 2009).

4.2.1.1 Human MORC proteins

An MSA of full-length human MORC paralogues, MORC1, MORC2, MORC3 and MORC4 protein sequences revealed both similarities and differences between the protein sequences of each MORC. The MSA revealed that the ATPase domains of all MORCs were highly conserved (Figure 4.1), particularly from residues 84 to 252. Key catalytical residues (N60, D88 and W435) previously found to be involved in ATP binding (N60A and D88A), hydrolysis (Walker B residue D88A) and H3K4me3 peptide binding (W435) in MORC2 (D Q Li *et al.*, 2012; Douse *et al.*, 2018) and MORC3 (H Andrews *et al.*, 2016) were also highly conserved throughout all MORCs, including MORC4 (Figure 4.1). Also found conserved in all MORC proteins was the Walker A catalytical residue K126, which is involved in ATP binding (Figure 4.1). This suggests that all MORC proteins may play a similar role in terms of ATP binding and peptide binding, indicating that MORC4 may also bind and hydrolyse ATP in addition to peptide binding.

Furthermore, Zf-CW residues 400 to 547 were also relatively well conserved. However, the C-terminus of all MORC proteins were less conserved (Figure 4.1), suggesting an independent function for each MORC paralogue. It was also evident that MORC4 contained an N-terminal extension in comparison to MORC1, MORC2 and MORC3. MORC1 and MORC2 also contained the same insertion from residues 306-362 and 379-399 (Figure 4.1), indicating that MORC1 and MORC2 are similar in protein sequence and potentially function. In addition, MORC3 and MORC4 did not have this insertion (Figure 4.1), suggesting that MORC proteins perhaps group in 2 subsets, MORC1/MORC2 and MORC3/MORC4.

Figure 4.1: **Multiple sequence alignment of protein sequences of all human MORC proteins (MORC1-4).** Human MORC protein sequences were aligned using MUSCLE and visualised in Jalview. The conservation score across all MORCs was shown below the alignment (Yellow is the most conserved and brown or no colour is the least conserved). Catalytic residues are highlighted with a red *. Walker B and Walker A motif residues are highlighted in black and purple, respectively. The multiple sequence alignment residues were coloured using ClustalX colours.

4.2.1.2 MORC4 in vertebrates

Protein sequences of MORC4 in vertebrates (mammals, reptiles and birds) were also utilised for an MSA, in an attempt to identify similarities and differences between vertebrate classes, and the numbering of the mutations e.g. N60A is the mutation found in human MORC4, not necessarily all species (Figure 4.2). The MSA revealed that generally, the ATPase domain was highly conserved across all vertebrate classes (Figure 4.2) except for a number of substitution

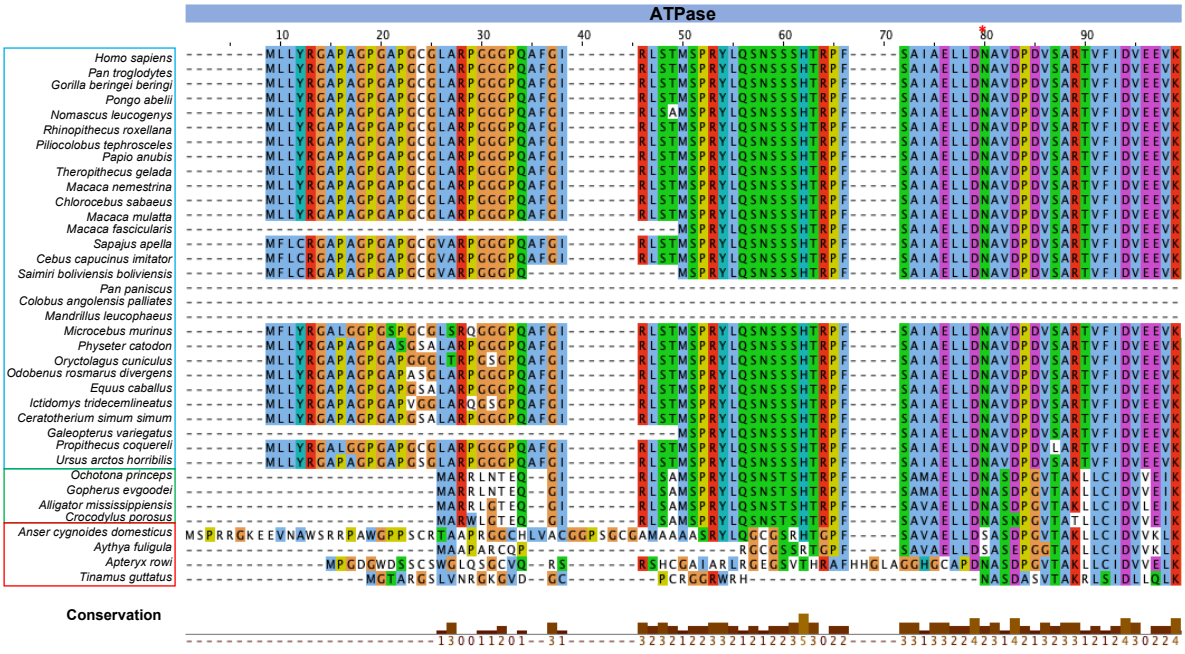
mutations. However, some insertions were found around residues 270-290 in reptiles. Key catalytical residues involved in ATP binding (N60A and D88A) and hydrolysis (D88A) in addition to a residue involved in H3K4me3 peptide binding (W435K), were also identified. The N60 residue involved in the binding of AMP-PNP and Mg^{2+} was highly conserved across all species, except for bird species, *A. domesticus* and *A. fuligula*, which was substituted to a serine (S) residue (Figure 4.2). However, both asparagine (N) and serine (S) are polar uncharged amino acids, suggesting that this substitution may not affect the overall function of MORC4 in these species.

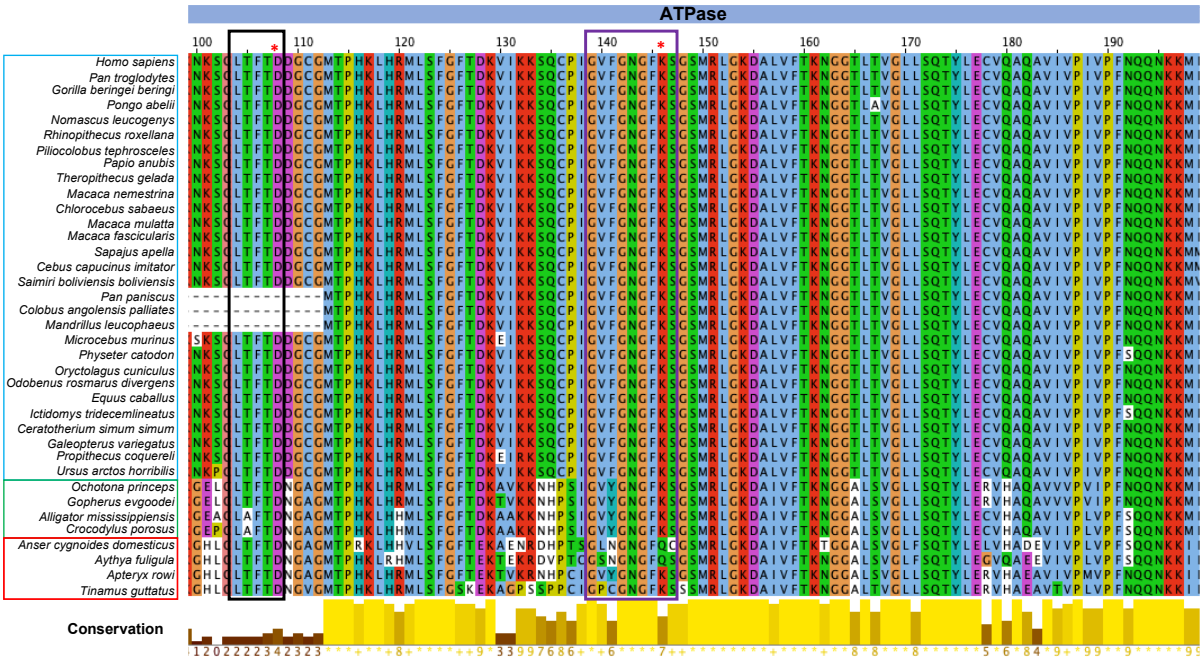
The Walker B motif ($L(H_y)_4D$) residue D88 involved in ATP hydrolysis (Rai *et al.*, 2006) was highly conserved across all species except for the primate species, *P. paniscus*, *C. a. palliates* and *M. leucophaeus* (Figure 4.2). These species also had an N-terminal deletion from 1-112 amino acids (Figure 4.2), suggesting a different function of MORC4 in these primate species or expression of a different splice variant of MORC4. However, the Walker A motif (GXXXXGKT/S) residue K126 involved in ATP binding (Ramakrishnan *et al.*, 2002) was highly conserved across all species including *P. paniscus*, *C. palliates* and *M. leucophaeus* (Figure 4.2), suggesting that these primate species are involved in ATP binding. Conversely, the bird species, *A. domesticus* and *A. fuligula* were found to have a substitution from a positively charged amino acid (K) to a polar amino acid (Q), suggesting that this substitution may also not have much effect on the function of MORC4 in these species.

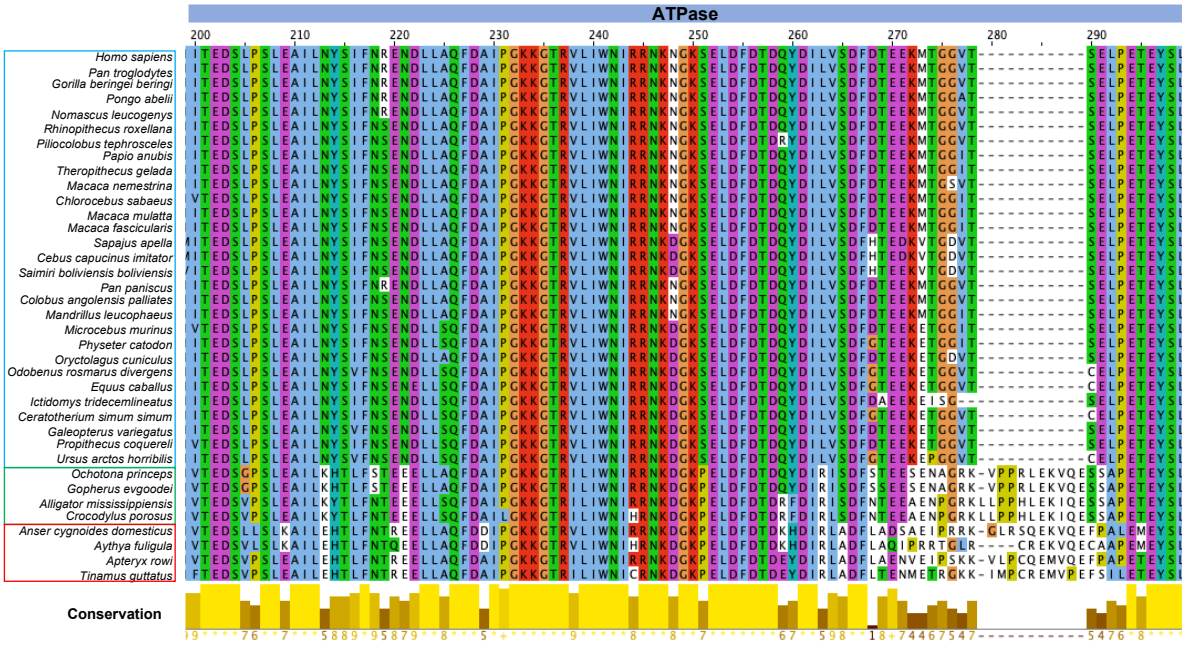
The W435 residue responsible for binding to histone peptides, specifically H3K4me3, was highly conserved across all species (Figure 4.2), suggesting that all species bind to the peptide and may be involved in chromatin remodelling. Analysis of the C-terminus of MORC4 showed that the C-terminus in mammals is highly conserved (Figure 4.2). However, in reptiles and birds the C-terminus was unconserved, suggesting that the C-terminus has a different function,

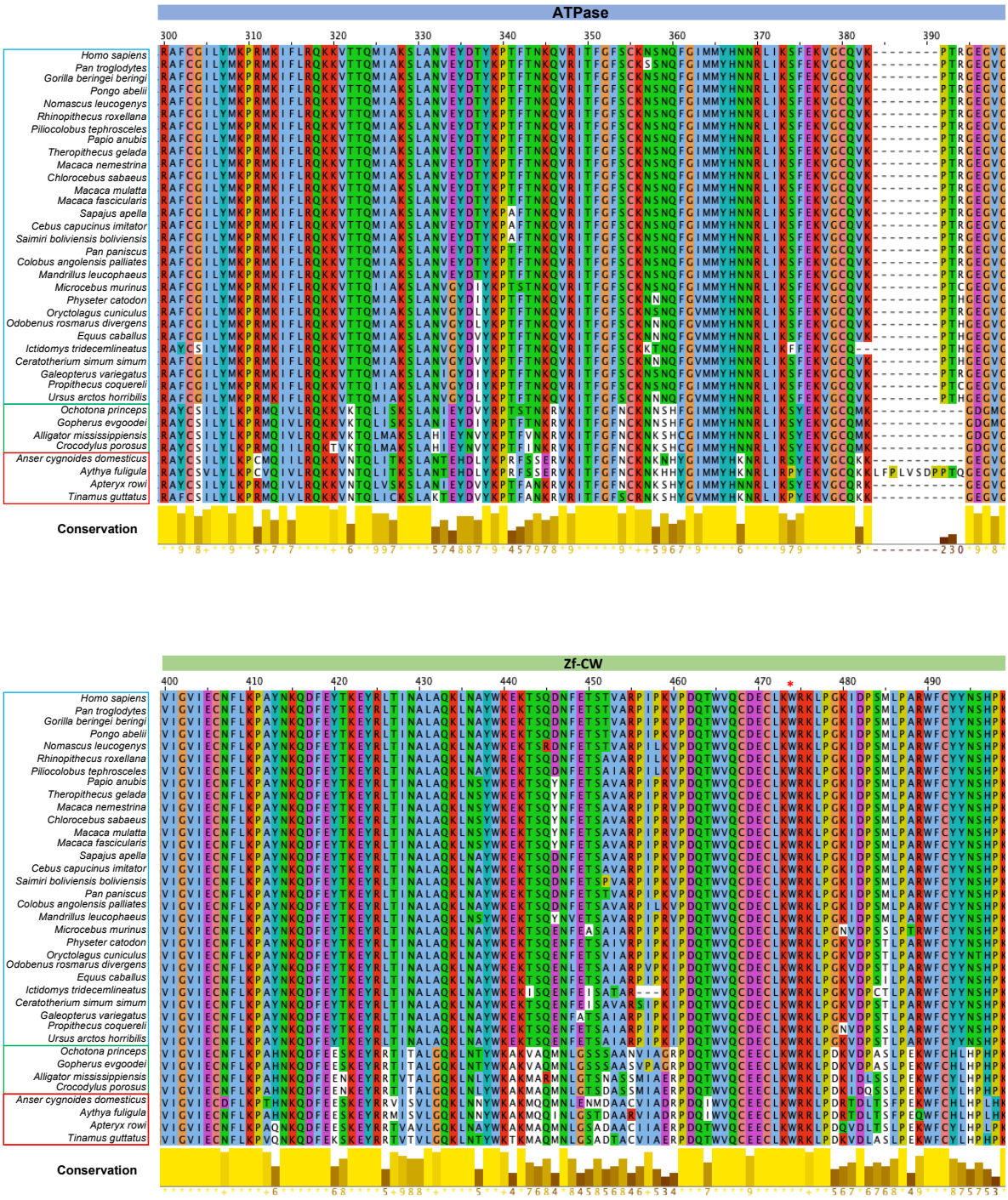
or expression of a different MORC4 splice variant in reptiles and birds but the function/splice variant in mammalian species is similar.

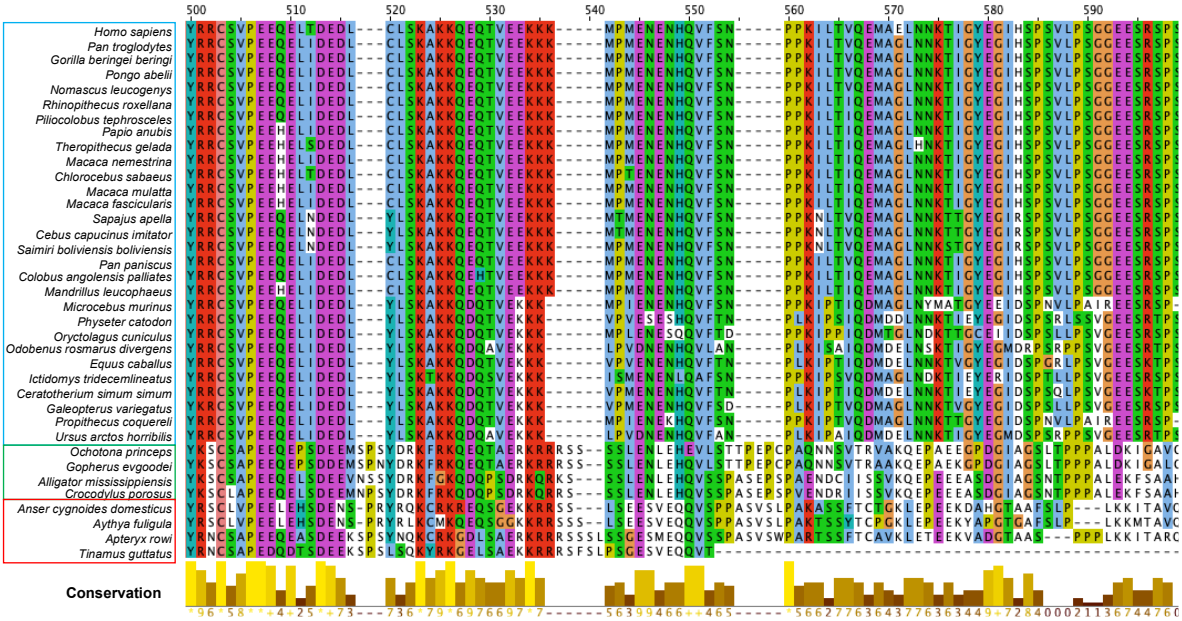
In addition to this MORC4 MSA, some phylogenetic tree analysis was performed to examine the MORC4 evolutionary relationship between species and classes (Appendix 8.5). This revealed that there was not much genetic change between the primate MORC4 protein sequences, as the branches were short in length (Appendix 8.5), suggesting that there was little variation in the MORC4 sequences in primates. This could also support the hypothesis that the function of MORC4 in mammals is similar. Phylogenetic analysis of MORC4 also supported known positions of the mammalian taxa (Nishihara *et al.*, 2009), including primates (Herlyn, 2016). However, the positioning of the birds and reptile classes was not as expected, indicating that perhaps not all MORCs are found in all species of vertebrates. Overall, this phylogenetic analysis concludes that the MORC4 protein sequence is highly conserved in mammalians.

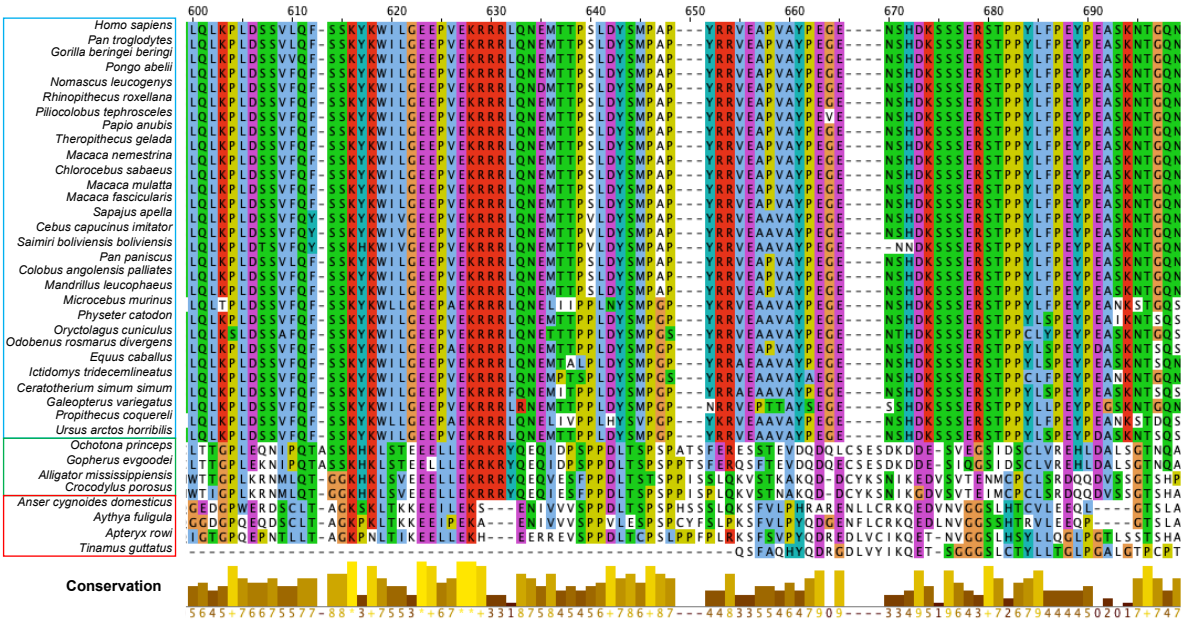


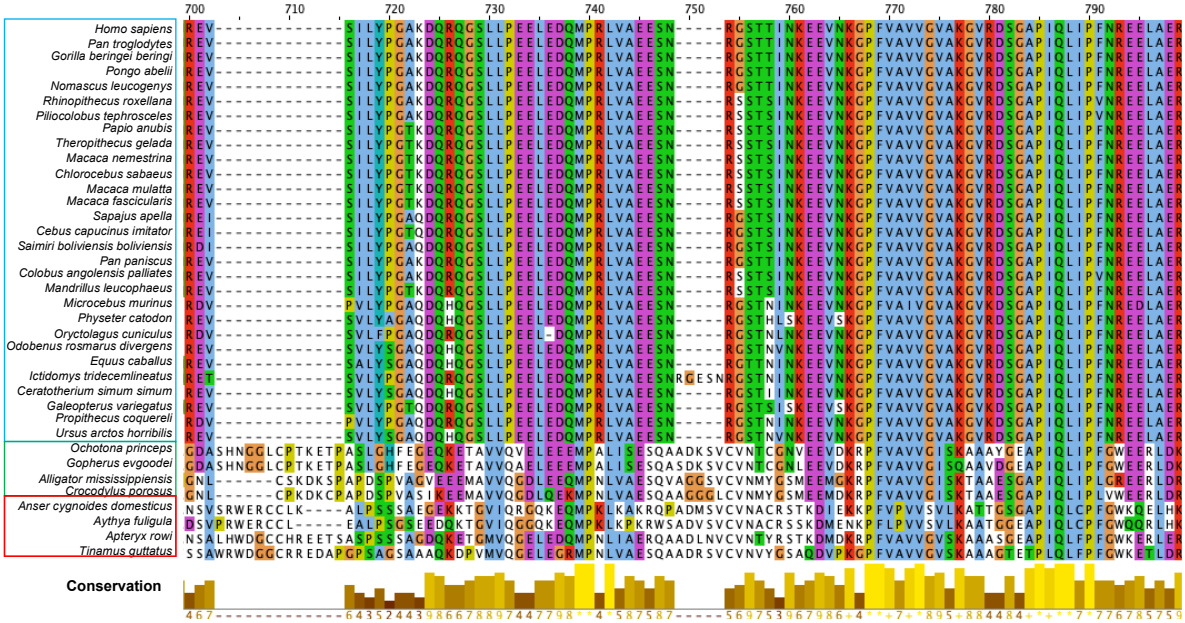

ATPase
100
110
120
130
140
150
160
170
180
190
Homo sapiens
Pan troglodytes
Gorilla beringei beringi
Pongo abelii
Nomascus leucogenys
Rhinopithecus roxellana
Piliocolobus tephrosceles
Papio anubis
Theropithecus gelada
Macaca nemestrina
Chlorocebus sabaeus
Macaca mulatta
Macaca fascicularis
Sapajus apella
Cebus capucinus imitator
Saimiri boliviensis boliviensis
Pan paniscus
Colobus angolensis palliates
Mandrillus leucophaeus
Microcebus murinus
Physeter catodon
Oryctolagus cuniculus
Odobenus rosmarus divergens
Equus caballus
Ictidomys tridecemlineatus
Ceratotherium simum simum
Galeopterus variegatus
Propithecus coquereli
Ursus arctos horribilis
Ochotona princeps
Gopherus evgoodei
Alligator mississippiensis
Crocodylus porosus
Anser cygnoides domesticus
Aythya fuligula
Apteryx rowi
Tinamus guttatus
Conservation

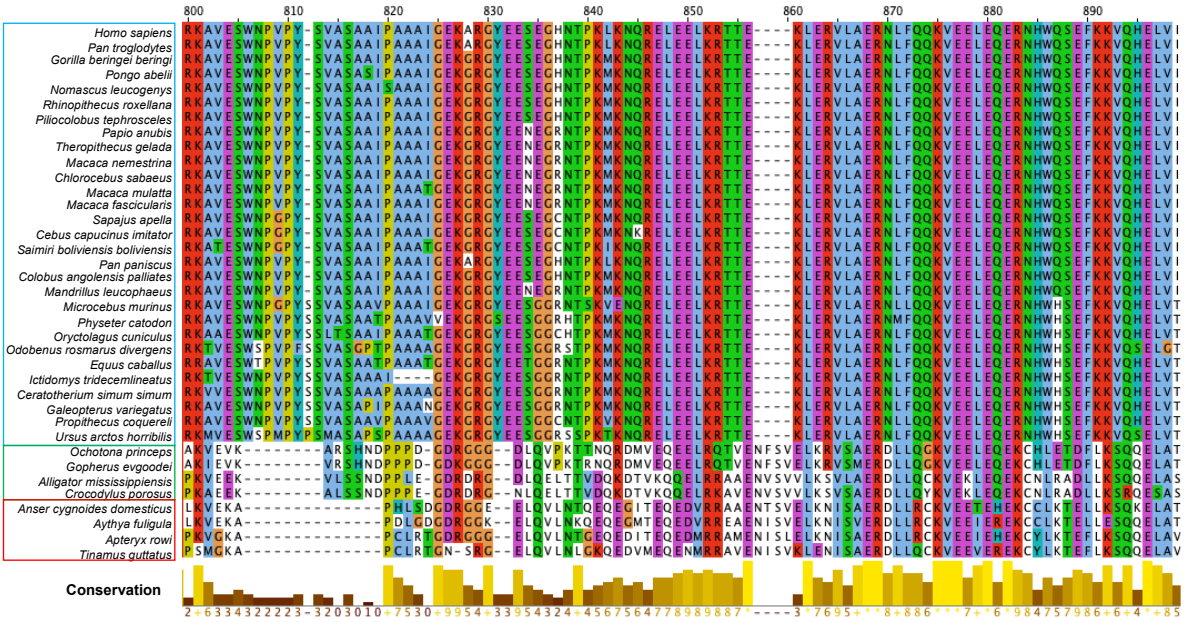

ATPase
200
210
220
230
240
250
260
270
280
290
Homo sapiens
Pan troglodytes
Gorilla beringei beringi
Pongo abelii
Nomascus leucogenys
Rhinopithecus roxellana
Piliocolobus tephrosceles
Papio anubis
Theropithecus gelada
Macaca nemestrina
Chlorocebus sabaeus
Macaca mulatta
Macaca fascicularis
Sapajus apella
Cebus capucinus imitator
Saimiri boliviensis boliviensis
Pan paniscus
Colobus angolensis palliates
Mandrillus leucophaeus
Microcebus murinus
Physeter catodon
Oryctolagus cuniculus
Odobenus rosmarus divergens
Equus caballus
Ictidomys tridecemlineatus
Ceratotherium simum simum
Galeopterus variegatus
Propithecus coquereli
Ursus arctos horribilis
Ochotona princeps
Gopherus evgoodei
Alligator mississippiensis
Crocodylus porosus
Anser cygnoides domesticus
Aythya fuligula
Apteryx rowi
Tinamus guttatus
Conservation


ATPase
Zf-CW
Homo sapiens
Pan troglodytes
Gorilla beringei beringi
Pongo abelii
Nomascus leucogenys
Rhinopithecus roxellana
Piliocolobus tephrosceles
Papio anubis
Theropithecus gelada
Macaca nemestrina
Chlorocebus sabaeus
Macaca mulatta
Macaca fascicularis
Sapajus apella
Cebus capucinus imitator
Saimiri boliviensis boliviensis
Pan paniscus
Colobus angolensis palliates
Mandrillus leucophaeus
Microcebus murinus
Physeter catodon
Oryctolagus cuniculus
Odobenus rosmarus divergens
Equus caballus
Ictidomys tridecemlineatus
Ceratotherium simum simum
Galeopterus variegatus
Propithecus coquereli
Ursus arctos horribilis
Ochotona princeps
Gopherus evgoodei
Alligator mississippiensis
Crocodylus porosus
Anser cygnoides domesticus
Aythya fuligula
Apteryx rowi
Tinamus guttatus
Conservation


Homo sapiens
Pan troglodytes
Gorilla beringei beringi
Pongo abelii
Nomascus leucogenys
Rhinopithecus roxellana
Piliocolobus tephrosceles
Papio anubis
Theropithecus gelada
Macaca nemestrina
Chlorocebus sabaeus
Macaca mulatta
Macaca fascicularis
Sapajus apella
Cebus capucinus imitator
Saimiri boliviensis boliviensis
Pan paniscus
Colobus angolensis palliates
Mandrillus leucophaeus
Microcebus murinus
Physeter catodon
Oryctolagus cuniculus
Odobenus rosmarus divergens
Equus caballus
Ictidomys tridecemlineatus
Ceratotherium simum simum
Galeopterus variegatus
Propithecus coquereli
Ursus arctos horribilis
Ochotona princeps
Gopherus evgoodei
Alligator mississippiensis
Crocodylus porosus
Anser cygnoides domesticus
Aythya fuligula
Apteryx rowi
Tinamus guttatus
Conservation


Homo sapiens
Pan troglodytes
Gorilla beringei beringi
Pongo abelii
Nomascus leucogenys
Rhinopithecus roxellana
Piliocolobus tephrosceles
Papio anubis
Theropithecus gelada
Macaca nemestrina
Chlorocebus sabaeus
Macaca mulatta
Macaca fascicularis
Sapajus apella
Cebus capucinus imitator
Saimiri boliviensis boliviensis
Pan paniscus
Colobus angolensis palliates
Mandrillus leucophaeus
Microcebus murinus
Physeter catodon
Oryctolagus cuniculus
Odobenus rosmarus divergens
Equus caballus
Ictidomys tridecemlineatus
Ceratotherium simum simum
Galeopterus variegatus
Propithecus coquereli
Ursus arctos horribilis
Ochotona princeps
Gopherus evgoodei
Alligator mississippiensis
Crocodylus porosus
Anser cygnoides domesticus
Aythya fuligula
Apteryx rowi
Tinamus guttatus
Conservation


700
710
720
730
740
750
760
770
780
790
Homo sapiens
Pan troglodytes
Gorilla beringei beringi
Pongo abelii
Nomascus leucogenys
Rhinopithecus roxellana
Piliocolobus tephrosceles
Papio anubis
Theropithecus gelada
Macaca nemestrina
Chlorocebus sabaeus
Macaca mulatta
Macaca fascicularis
Sapajus apella
Cebus capucinus imitator
Saimiri boliviensis boliviensis
Pan paniscus
Colobus angolensis palliates
Mandrillus leucophaeus
Microcebus murinus
Physeter catodon
Oryctolagus cuniculus
Odobenus rosmarus divergens
Equus caballus
Ictidomys tridecemlineatus
Ceratotherium simum simum
Galeopterus variegatus
Propithecus coquereli
Ursus arctos horribilis
Ochotona princeps
Gopherus evgoodei
Alligator mississippiensis
Crocodylus porosus
Anser cygnoides domesticus
Aythya fuligula
Apteryx rowi
Tinamus guttatus
Conservation


800
810
820
830
840
850
860
870
880
890
Homo sapiens
Pan troglodytes
Gorilla beringei beringi
Pongo abelii
Nomascus leucogenys
Rhinopithecus roxellana
Piliocolobus tephrosceles
Papio anubis
Theropithecus gelada
Macaca nemestrina
Chlorocebus sabaeus
Macaca mulatta
Macaca fascicularis
Sapajus apella
Cebus capucinus imitator
Saimiri boliviensis boliviensis
Pan paniscus
Colobus angolensis palliates
Mandrillus leucophaeus
Microcebus murinus
Physeter catodon
Oryctolagus cuniculus
Odobenus rosmarus divergens
Equus caballus
Ictidomys tridecemlineatus
Ceratotherium simum simum
Galeopterus variegatus
Propithecus coquereli
Ursus arctos horribilis
Ochotona princeps
Gopherus evgoodei
Alligator mississippiensis
Crocodylus porosus
Anser cygnoides domesticus
Aythya fuligula
Apteryx rowi
Tinamus guttatus
Conservation

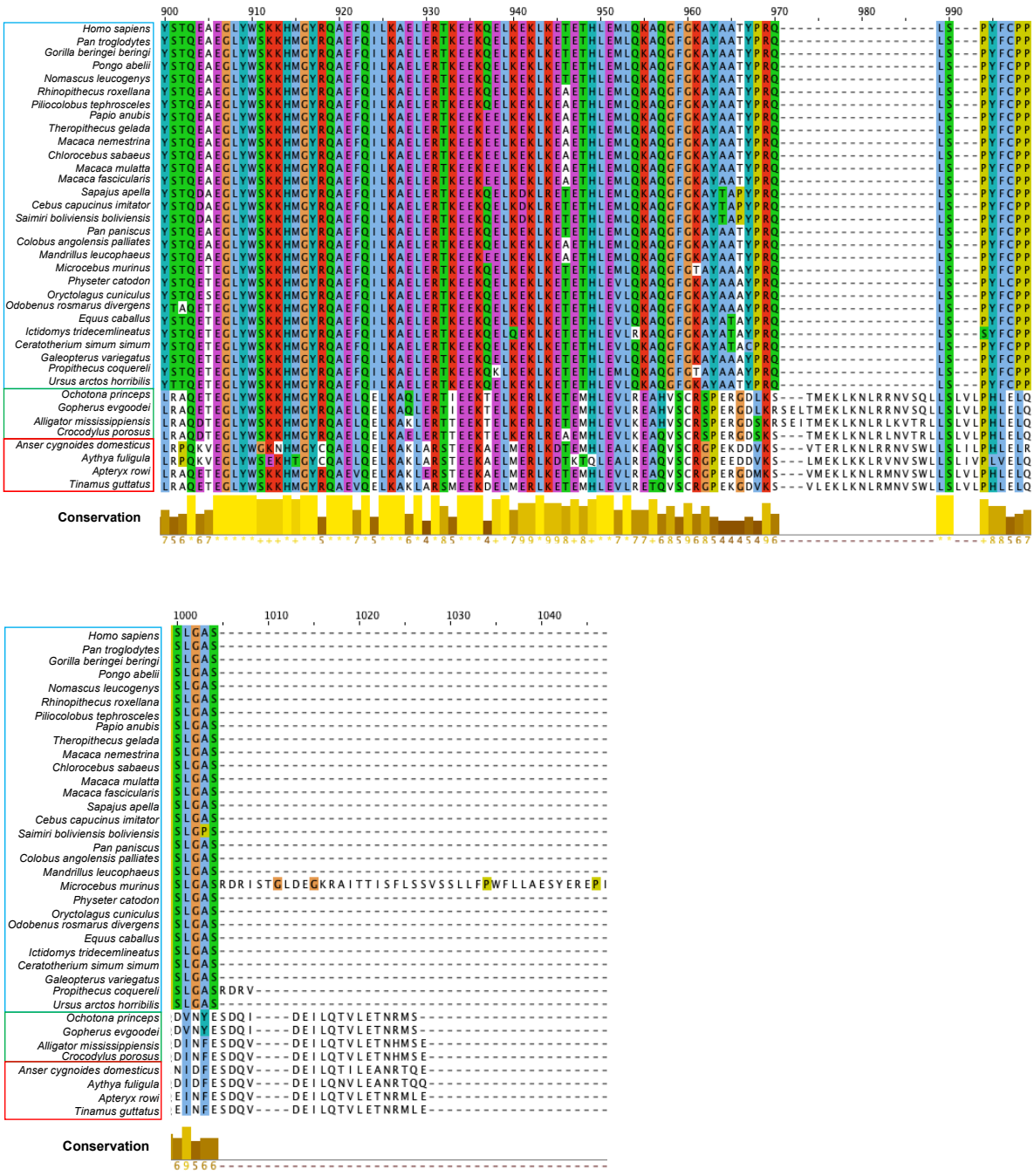


Figure 4.2: **Multiple sequence alignment of protein sequences of MORC4 from vertebrates.** MORC4 protein sequences from several vertebrates were aligned using MUSCLE and visualised in Jalview. The conservation score across all species was shown below the alignment (Yellow is the most conserved and brown or no colour is the least conserved). Mammals are highlighted in blue; reptiles are highlighted in green and birds are highlighted in red. Sequences highlighted in purple and black are the Walker A and Walker B motifs, respectively. Catalytical residues are highlighted with a red *. The multiple sequence alignment residues were coloured using ClustalX colours.

4.3 Structural characterisation of MORC4

Structural characterisation of proteins is important in helping to guide biochemical characterisation, towards gaining greater understanding of MORC4 function. In addition, structural studies will help to identify any differences or similarities of MORC4 to other MORCs. The structure of both MORC2 (Douse *et al.*, 2018) and MORC3 (Li *et al.*, 2016) were solved after this project was initiated, however, the MORC4 boundaries were independently designed, which were similar to the MORC3 crystallised fragment boundaries.

4.3.1 Protein crystallography trials

In an attempt to determine the structure of MORC4, purified protein (Chapter 3, section 3.3.2.3) was used to set up sitting drop vapour diffusion protein crystallography trials. Both ATPase-Zf-CW (MORC4$^{29\text{-}480}$) regions and Zf-CW (MORC4$^{420\text{-}757}$ and MORC4$^{420\text{-}677}$) pure proteins were used for protein crystallography trials. Using both recombinant proteins in crystal trials was expected to increase the likelihood of success in protein crystallisation. Furthermore, there were two potential outcomes of these crystallography trials including an investigation into a potential ATPase-Zf-CW relationship and the Zf-CW/C-terminus structure. The Joint Center for Structural Genomics (JCSG)-*plus*™ screen was initially used as it is a general sparse-matrix screen (Jancarik and Kim, 1991), indicating it statistically covers many conditions/crystallisation space that several proteins in the Protein Data Bank (PDB) have crystallised in. The JCSG-*plus*™ screen was designed based on several proteins from 480 conditions (Page *et al.*, 2003). The JCSG-*plus*™ screen was chosen as it covered all pH ranges and used conditions from both approved crystal screens and new screens containing less developed crystallisation chemicals (Newman *et al.*, 2005).

4.3.1.1 Crystallisation of MORC4 Zf-CW/C-terminal only protein

Both MORC4$^{420\text{-}757}$ and MORC4$^{420\text{-}677}$ Zf-CW/C-terminal domain only proteins were used for initial protein crystallography trials. Pure Zf-CW/C-terminal only proteins, MORC4$^{420\text{-}757}$ and

MORC4$^{420-677}$ were set up using the vapour diffusion sitting drop method with the JCSG-*plus*™ screen (Appendix 8.6). Following incubation of the crystal plate for 24 hours at 20°C, several conditions showed heavy precipitation, phase separation or no crystallisation of the proteins. Crystallisation was also not evident following extended incubation of crystal plates. This suggests that these Zf-CW/C-terminal proteins will not crystallise using this screen and that other methods of structural determination may be required to analyse the structure of C-terminal MORC4. No further crystal trials were attempted with Zf-CW/C-terminal proteins as there was insufficient protein and proteins were of lower purity than the ATPase-Zf-CW domain, suggesting that the tandem domain protein was more likely to crystallise.

4.3.1.2 Crystallisation of MORC4 ATPase-Zf-CW protein

4.3.1.2.1 MORC4$^{29-480}$ crystallisation screening: JCSG-plus™ screen

ATPase-Zf-CW MORC4 proteins were also used for crystallography trials. As with Zf-CW domain only proteins, an ATPase-Zf-CW MORC4 protein, MORC4$^{29-480}$, was first utilised with the JCSG-*plus*™ screen. Unlike the Zf-CW domain, the ATPase-Zf-CW MORC4$^{29-480}$ protein yield apparent rectangular crystals in the 0.2 M Li_2SO_4, 0.1 M Tris, 40% v/v PEG 400, pH 8.5 condition (Figure 4.3). It was also evident that MORC4$^{29-480}$ only crystallised in the presence of the ligand AMP-PNP/Mg^{2+} (Figure 4.3b), as no crystals were ever present in the wells containing protein only (Figure 4.4a). This suggests that, like MORC3 (Li *et al.*, 2016) and MORC2 (Douse *et al.*, 2018), MORC4 only dimerises in the presence of AMP-PNP/ Mg^{2+}.

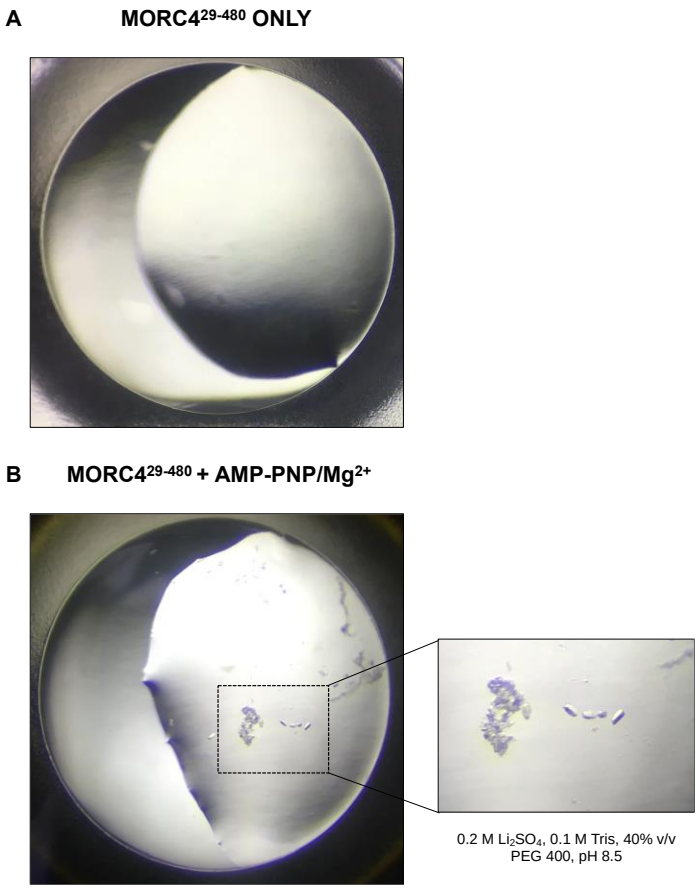


Figure 4.3: **MORC4$^{29-480}$ protein crystallography trial with the JCSG-*plus*™ screen.** Pure MORC4$^{29-480}$ protein was used for a sitting drop vapour diffuse crystal trial using the JCSG-*plus*™ screen. Pure protein was diluted into low salt (100 mM), incubated with 1.0 nM AMP-PNP/Mg^{2+} at 4°C for 1 hour and concentrated to 6.6 mg/mL. Protein which had been incubated with AMP-PNP/Mg^{2+} was mixed in the plate well in a ratio of 1:1 protein: mother liquor to give 2 µL drops. **(A)** MORC4$^{29-480}$ **(B)** MORC4$^{29-480}$ + AMP-PNP/ Mg^{2+}. Crystals were visible after 5 days. The conditions of each well were 0.2 M Li_2SO_4, 0.1 M Tris, 40% v/v PEG 400, pH 8.5. The dimension of each well was 2.76mm diameter.

4.3.1.2.1.1 JCSG-*plus*™ screen optimisation of MORC4$^{29-480}$ in the presence of AMP-PNP

In an attempt to improve the size and quality of the crystal formed from the initial JCSG screen, an optimisation screen was created using conditions around that of the initial the JCSG screen (Section 4.3.1.2.1). This included a range of 150-225 mM Li_2SO_4, PEG 400 was used in a range of 34-44% and pH range was from 7.5-9.0 (Appendix 8.7).

Following preparation of the crystal plate, the optimisation screen (JCSG_D07_FUP_1) was incubated at 20°C and after 24 hours crystals were visible in the well containing 150 mM

Li_2SO_4, 34% PEG 400 and 0.1 M Tris at pH 9.0. However, these crystals were smaller than the original crystals created by the JCSG screen. This follow up screen suggested that MORC4$^{29\text{-}480}$ protein was more likely to crystallise in lower concentrations of PEG 400, a low concentration of Li_2SO_4 and a higher pH i.e. pH 9.0.

Following the first optimisation screen (JCSG_D07_FUP_1), the conditions containing the crystals allowed an additional optimisation screen to be designed (JCSG_D07_FUP_2) (Appendix 8.8). This optimisation plate was designed so that the crystal conditions from the JCSG_D07_FUP_1 plate contained those conditions and varying conditions either side, in an attempt to increase the size and quality of the crystals. The JCSG_D07_FUP_2 optimisation screen appeared to produce a larger numbers of protein crystals that were slightly larger in size than the previous two screens (Figure 4.4), suggesting that optimisation was potentially successful. Crystals were formed in the conditions containing 100 mM and 125 mM Li_2SO_4, 34% and 36% PEG 400 with 0.1 M Tris at pH 9.5 (Figure 4.4a, b, c and d). This also confirms that MORC4$^{29\text{-}480}$ protein is more likely to crystallise in a lower concentration of Li_2SO_4 and PEG 400 with a high pH.

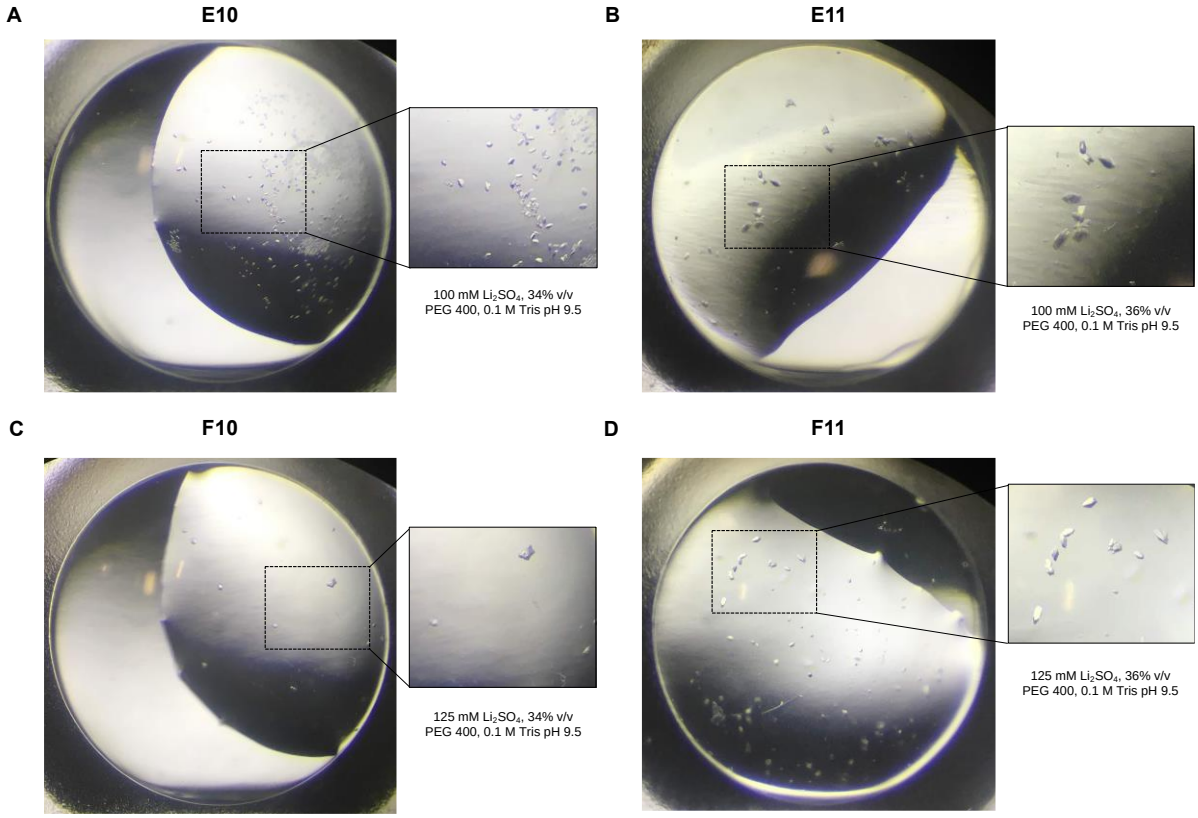


Figure 4.4: **MORC4$^{29-480}$ protein crystallography trial with the JCSG_D7_FUP_2 optimisation screen.** Pure MORC4$^{29-480}$ protein was used for a sitting drop vapour diffuse crystal trial using the JCSG_D7_FUP_2 optimisation screen. Pure protein was diluted to low salt (100 mM), incubated with 1.0 mM AMP-PNP/Mg^{2+} at 4°C for 1 hour and concentrated to 5.7 mg/mL. Protein which had been incubated with AMP-PNP/Mg^{2+} was mixed in the plate well in a ratio of 1:1 protein: mother liquor to give 2 µL drops. **(A)** E10 well containing 100 mM Li_2SO_4, 34% PEG 400, 0.1 M Tris, pH 9.5 **(B)** E11 well containing 100 mM Li_2SO_4, 36% PEG 400, 0.1 M Tris, pH 9.5 **(C)** F10 well containing 125 mM Li_2SO_4, 34% PEG 400, 0.1 M Tris, pH 9.5 **(D)** F11 well containing 125 mM Li_2SO_4, 36% PEG 400, 0.1 M Tris, pH 9.5 Crystals were visible after 5 days. The dimension of each well was 2.76mm diameter.

In addition to the design of optimisation plates, fresh (non-frozen) MORC4$^{29-480}$ protein was also used for the JCSG-*plus*™ screen as freezing the protein may affect crystallisation propensity due to protein quality (C. Cooper, *pers. comm.*). The procedure of AMP-PNP/Mg^{2+} incubation with protein was also optimised. Previously, AMP-PNP/Mg^{2+} was combined with the protein following dilution of the protein. However, this was found to affect the absorbance measurement when determining the protein concentration, thus AMP-PNP/Mg^{2+} was added at the final stage following dilution and then concentration of the protein. Both fresh and frozen

MORC4$^{29-480}$ protein was used for crystal trials for the JCSG-*plus*™ screen. Frozen MORC4$^{29-480}$ protein with the new procedure for adding the ligand, AMP-PNP/Mg^{2+}, created a crystal with 20% v/v Jeffamine® M-600, 0.1 M HEPES pH 7.5 conditions (Figure 4.5a). An investigation of the crystallisation of AMP-PNP/Mg^{2+} with the size exclusion chromatography (SEC) buffer was also carried out to eliminate the possibility of AMP-PNP/Mg^{2+} crystals. However, no crystals were identified under these conditions (Figure 4.5a). This suggests that the crystal observed in the sample with MORC4$^{29-480}$ and AMP-PNP/Mg^{2+}, was a protein crystal and not crystals formed from AMP-PNP/Mg^{2+} with the SEC buffer and no ligand. This being similar to previous crystal screens, supports the hypothesis that AMP-PNP causes dimerisation of MORC4$^{29-480}$, leading to protein crystals under optimal conditions, as with MORC2 (Douse *et al.*, 2018) and MORC3 (Li *et al.*, 2016). When the JCSG-*plus*™ screen was utilised with fresh MORC4$^{29-480}$ protein, the D7 well contained the best crystals, similar to when using frozen protein. However, this crystal was larger in size and would be easier to handle in terms of mounting and screening (Figure 4.5b). Well E8 contained large crystals under 1.0 M ammonia phosphate dibasic, 0.1 M sodium acetate, pH 4.5 conditions. However, crystals were also present in the sample containing AMP-PNP/Mg^{2+} and SEC buffer (Figure 4.5c), suggesting that the AMP-PNP ligand alone was able to crystallise under those conditions. Alternatively, it could also suggest that the salt, ammonium phosphate dibasic, formed crystals under these conditions as some evidence suggests that this is a frequent obstacle in protein crystallography trials (Raghunathan *et al.*, 2010).

In an attempt to optimise the F4 crystal created using the JCSG-*plus*™ screen (Figure 4.5a), a new optimisation screen was designed using varying concentrations of Jeffamine® M-600 precipitate and 0.1 M HEPES and Tris salt at pH 7.0-8.0 (Appendix 8.9). However, there were no crystal hits for this optimisation plate. Suggesting that the crystal found in the previous

JCSG-*plus*™ screen was unreproducible. In summary, several JCSG-plus™ screens and optimisation screens were used to screen MORC4$^{29\text{-}480}$ protein (Appendix 8.10).

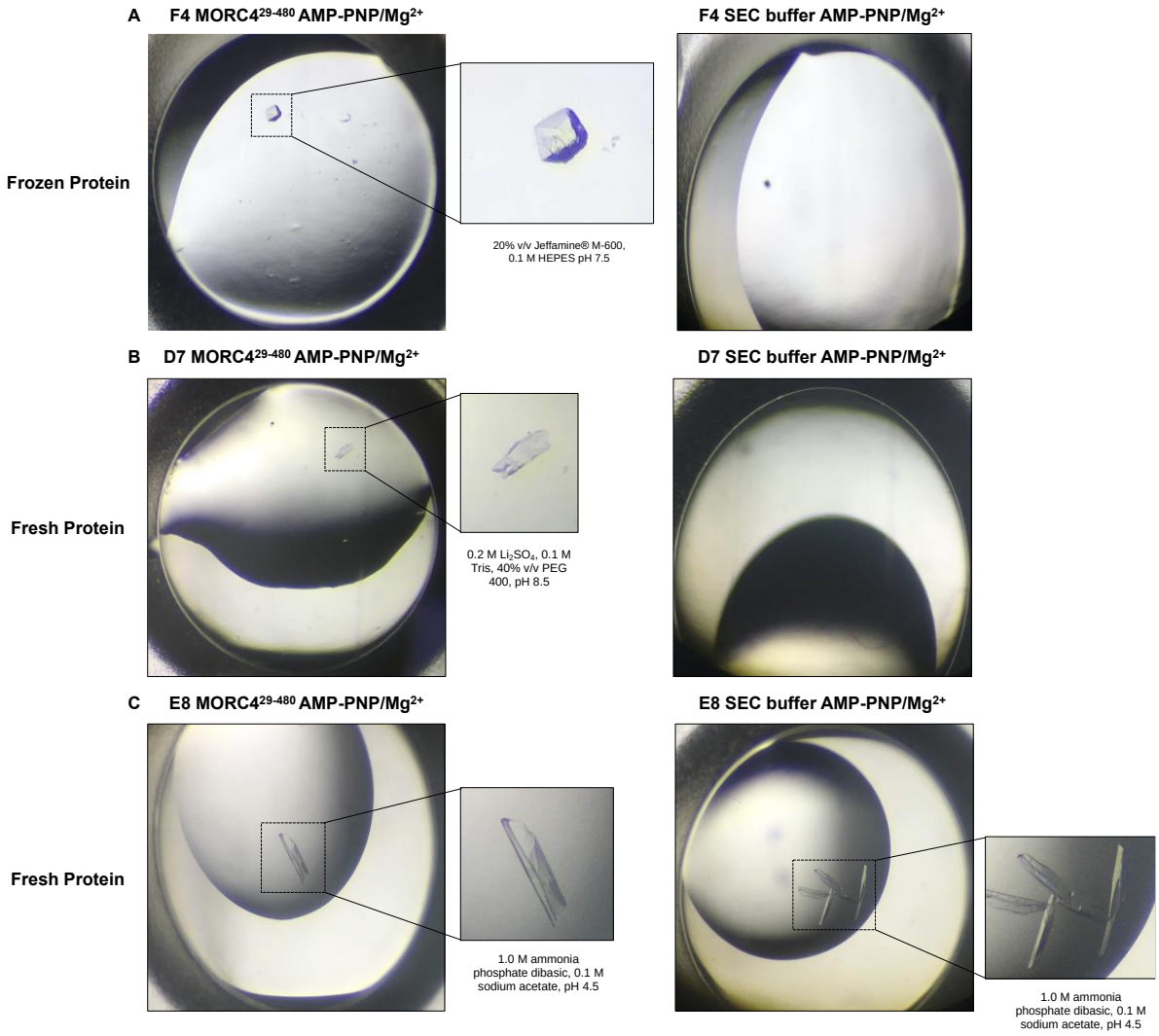


Figure 4.5: **MORC4$^{29\text{-}480}$ protein crystallisation trial with the JCSG-*Plus*™ screen.** Pure MORC4$^{29\text{-}480}$ protein was used for a sitting drop vapour diffuse crystal trial using the JCSG-*Plus*™ screen. Pure protein was diluted to low salt (100 mM), concentrated to 4.0 mg/ml and incubated with 1.0 mM AMP-PNP/Mg^{2+} at 4°C for 1 hour. Protein and SEC buffer which had been incubated with AMP-PNP/Mg^{2+} was mixed in the plate well in a ratio of 1:1 protein: mother liquor to give 2 µL drops. **(A)** Well F4 containing 20% v/v Jeffamine® M-600, 0.1 M HEPES pH 7.5. **(B)** Well D7 containing 0.2 M Li_2SO_4, 0.1 M Tris, pH 8.5. **(C)** Well E8 containing 1.0 M ammonia phosphate dibasic, 0.1 M sodium acetate, pH 4.5. Crystals were visible after 5 days. The dimension of each well was 2.76mm diameter.

4.3.1.2.2 MORC4$^{29\text{-}480}$ crystallisation screening: PACT premier™ screen

In addition to the JCSG-*plus*™ screen, the PACT *premier*™ screen was also utilised in the MORC4$^{29\text{-}480}$ crystallisation trials. The pH-, anion- and cation-testing (PACT) screen assessed the pH, anions and cations along with polyethene glycol (PEG) precipitant in an attempt to create protein crystals (Newman *et al.*, 2005). MORC4$^{29\text{-}480}$ protein crystals were observed in several PACT screen conditions including 0.1 M sodium malonate dibasic monohydrate, imidazole, boric acid (MIB)/DL-Malic acid, 2-(N-morpholino)ethanesulfonic acid monohydrate, tris (MMT) buffers with 25% v/v PEG 1500, pH 6.0-9.0. However, the crystals in these conditions were relatively small in comparison to the JCSG screen and optimisation. Despite the small crystal size, a PACT optimisation was designed in an attempt to increase crystal size and quality (Appendix 8.11). However, this PACT optimisation screen did not yield further MORC4$^{29\text{-}480}$ protein crystals.

4.3.1.2.3 MORC4$^{29\text{-}480}$ crystallisation screening: BCS screen

The basic chemical space (BSC) screen was designed to include several PEG smears (Newman *et al.*, 2005) (Appendix 8.12), allowing an extensive range of PEGs, compared to using single PEGs (Chaikuad *et al.*, 2015). PEG smears are organised by molecular weight, with low PEG smears including PEGs with a molecular weight of ≤ 1 kDa and high PEG smears consisting of PEGs with a molecular weight of ≥ 6 kDa (Chaikuad *et al.*, 2015). The BSC screen includes a range of PEG smears with buffers ranging from pH 4.5-9.5 and additives such as sodium bromide and glycerol, which have been found to increase crystallisation of proteins (Snell *et al.*, 2008).

Several MORC4$^{29\text{-}480}$ protein crystals were present in well F2 of the BCS screen, which contained 0.2 M NaCl, 0.1 M Tris, 28% v/v PEG smear low, pH 8.5 and 5% v/v glycerol (Figure 4.6a and b). This suggests that MORC4$^{29\text{-}480}$ protein is more likely to crystallise in a

low concentration of NaCl, low molecular weight PEG and high pH conditions. However, smaller crystals were present (Figure 4.6a) compared to the JCSG-*Plus*™ screen, which produced larger crystals. As with other initial screens, the BCS screen also resulted in no crystals present in the well containing no protein, size exclusion chromatography (SEC) buffer and ligand, AMP-PNP/Mg^{2+} (Figure 4.6b). This provides further evidence that the crystals present were protein crystals and not AMP-PNP/Mg^{2+} crystals.

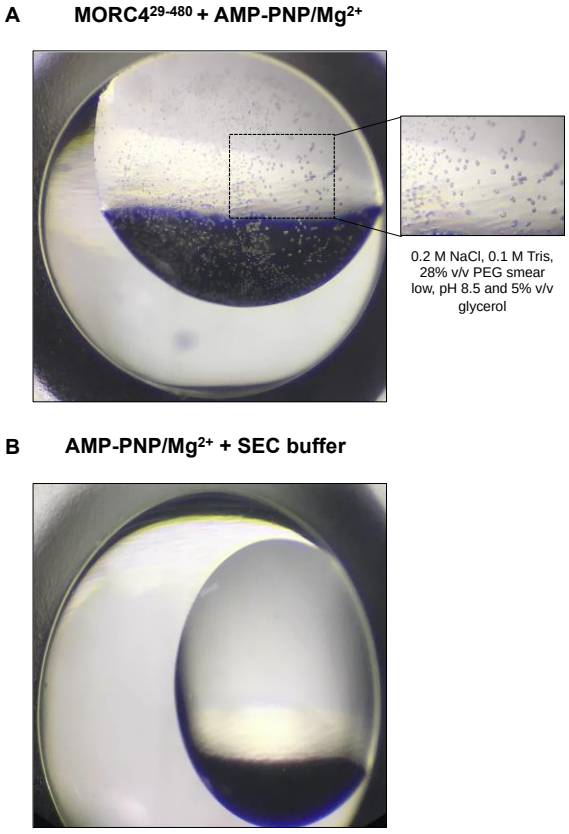


Figure 4.6: **MORC4$^{29-480}$ protein crystallisation trial with the BCS screen.** Pure MORC4$^{29-480}$ protein was used for a sitting drop vapour diffuse crystal trial using the BCS screen. Pure protein was diluted to low salt (100 mM), concentrated to 3.5 mg/ml and incubated with 1.0 mM AMP-PNP/Mg^{2+} at 4°C for 1 hour. Protein and SEC buffer which had been incubated with AMP-PNP/Mg^{2+} was mixed in the plate well in a ratio of 1:1 protein: mother liquor to give 2 µL drops. Each well shown contained 0.2 M NaCl, 0.1 M Tris, 28% v/v PEG smear low, pH 8.5 and 5% v/v glycerol. **(A)** MORC4$^{29-480}$ and AMP-PNP/Mg^{2+}. **(B)** SEC buffer and AMP-PNP/Mg^{2+}. Crystals were visible after 5 days. The dimension of each well was 2.76mm diameter.

4.3.1.2.3.1 BCS screen optimisation

In an attempt to increase the size and quality of crystals produced with the initial BCS screen (Figure 4.6a), a BCS optimisation was designed from the crystal conditions (Appendix 8.13). The BCS optimisation screen, BCS_F02_FUP_1, involved the separation of the condition which previously produced crystals (Section 4.3.1.2.3) in order to assess individual components for optimisation. This optimisation screen consisted of a range of single low PEGs and a low PEG smear, both in the range of 18-40%. This screen also included an additional additive screen. The additive screen is an optimisation screen consisting of 24 compounds and has the potential to increase the quality and/or size of a protein crystal (Gorrec and Löwe, 2018). Following the BCS_F02_FUP_1 optimisation screen and incubation of the screen for 5 days, wells D1, E1 and F1 conditions crystallised MORC4$^{29\text{-}480}$ protein (Figure 4.7a, b and c). All the conditions contained 0.2 M NaCl, 100 mM Tris pH 8.5 and 5% glycerol. However, each well contained a different type of 18% v/v PEG such as PEG 600 (Figure 4.7a), PEG 1000 (Figure 4.7b) and PEG smear low (Figure 4.7c). This suggests that MORC4$^{29\text{-}480}$ protein is more likely to crystallise in lower molecular weight PEGs.

In an attempt to optimise the MORC4$^{29\text{-}480}$ protein crystals further, a second BCS optimisation screen was designed. The BCS_F02_FUP_2 screen consisted of a base condition of 0.2 M NaCl, 100 mM Tris pH 8.5 and 5% glycerol. It also contained low molecular weight PEGs and one PEG smear low at concentrations between 10 and 20% v/v. Each PEG also had the addition of varying concentrations of glycerol including 5%, 10%, 15% and 25% (Appendix 8.14). Following this BCS_F02_FUP_2 optimisation screen, well A5 (containing 0.2 M NaCl, 100 mM Tris pH 8.5 and 5% glycerol base condition with the addition of 18% PEG 600) was the only well which contained MORC4$^{29\text{-}480}$ protein crystals (Figure 4.8). However, the conditions were the same as the previous optimisation plate, BCS_F02_FUP_1, suggesting that this optimisation plate was unsuccessful at producing new, larger protein crystals (Appendix 8.15).

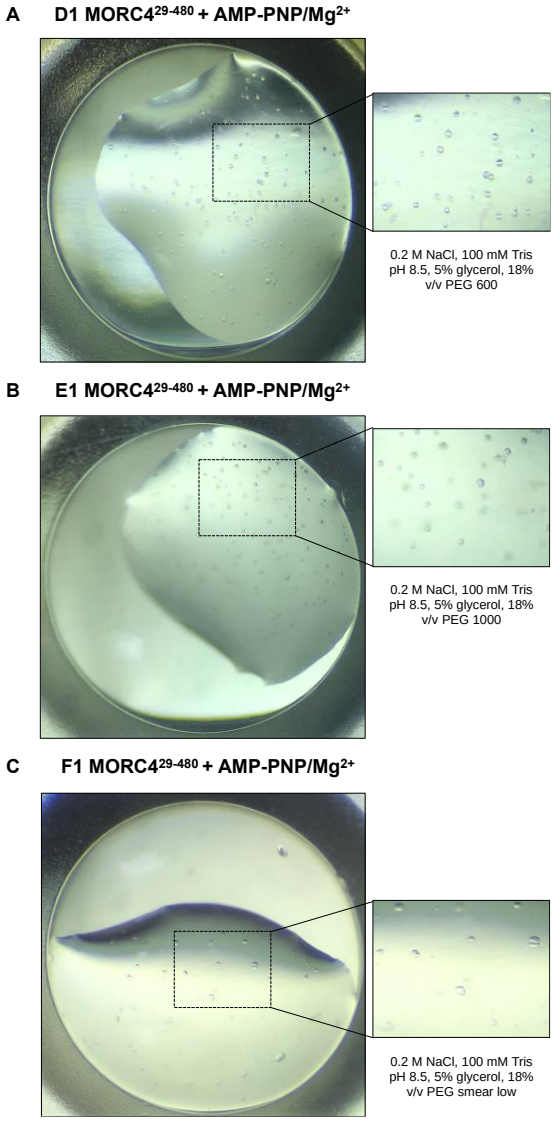


Figure 4.7: **MORC4$^{29-480}$ protein crystallisation trial with the BCS optimisation screen, BCS_F02_FUP_1.** Pure MORC4$^{29-480}$ protein was used for a sitting drop vapour diffuse crystal trial using the BCS_F02_FUP_1 optimisation screen. Pure protein was diluted to low salt (100 mM), concentrated to 4.0 mg/ml and incubated with 1.0 mM AMP-PNP/Mg^{2+} at 4°C for 1 hour. Protein which had been incubated with AMP-PNP/Mg^{2+} was mixed in the plate well in a ratio of 1:1 protein: mother liquor to give 2 µL drops. All wells contained 0.2 M NaCl, 100 mM Tris pH 8.5 and 5% glycerol. (**A)** Well D1 contained 18% v/v PEG 600. (**B)** Well E1 contained 18% v/v PEG 1000. (**C)** Well F1 contained 18% v/v PEG smear low. Crystals were visible after 5 days. Crystals were visible after 5 days. The dimension of each well was 2.76mm diameter.

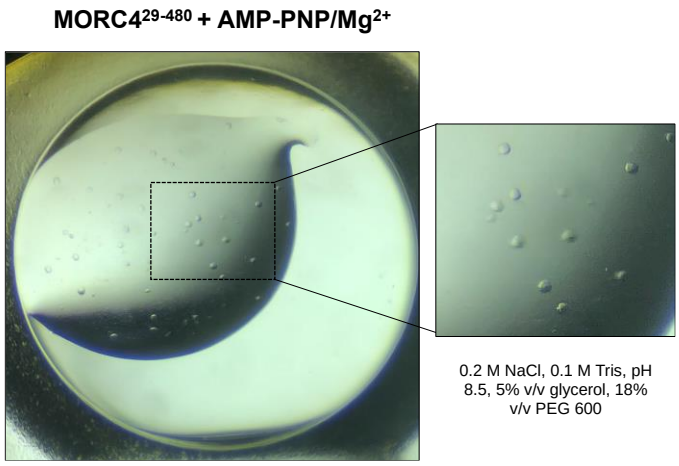


Figure 4.8: **MORC4$^{29-480}$ protein crystallisation trial with the BCS optimisation screen, BCS_F02_FUP_2.** Pure MORC4$^{29-480}$ protein was used for a sitting drop vapour diffuse crystal trial using the BCS_F02_FUP_2 optimisation screen. Pure protein was diluted to low salt (100 mM), concentrated to 4.0 mg/ml and incubated with 1.0 mM AMP-PNP/Mg^{2+} at 4°C for 1 hour. Protein which had been incubated with AMP-PNP/Mg^{2+} was mixed in the plate well in a ratio of 1:1 protein: mother liquor to give 2 µL drops. The well shown contained 0.2 M NaCl, 100 mM Tris pH 8.5 and 5% glycerol base condition with the addition of 18% PEG 600. Crystals were visible after 5 days. The dimension of each well was 2.76mm diameter.

4.3.1.2.4 MORC4$^{29-480}$ crystallisation: Morpheus® screen

The Morpheus® screen was created at the MRC Laboratory of Molecular Biology in Cambridge, England (MRC-LMB). The screen was designed with the use of 49 different low molecular weight ligands, which had been identified as abundant in several known crystal structures with the use of the Protein Data Bank (PDB) data mining (Gorrec, 2009). Each ligand used in this crystal screen was small in size, stable, economical and had to be present in at least five dissimilar known protein structures. This screen is important for examining original chemical space, which is not possible with more traditional types of crystal screens such as the JCSG-*plus*™ screen.

Following set up of the Morpheus® screen with MORC4$^{29-480}$ protein and incubation for 5 days at 20°C, several protein crystals were observed in multiple conditions (Figure 4.9). All crystals formed using this screen were present in alcohols (Figure 4.9a and b), ethylene glycols (Figure 4.9c) or monosaccharides (Figure 4.9d). Crystals were also formed in either Tris base buffer

(Figure 4.9b and c) or sodium HEPES buffer (Figure 4.9a and d). The precipitant mixes also contained either PEG 4000 (Figure 4.9a, c and d) or PEG 8000 (Figure 4.9b) at a pH of either 8.5 (Figure 4.9b and c) or 7.5 (Figure 4.9a and d). This suggests that for this screen, the MORC4$^{29\text{-}480}$ protein was more likely to crystallise in Tris or HEPES buffer, in a higher molecular weight PEG, at a higher pH.

All protein crystals identified in this screen were similar in appearance. They all appeared to be rod shaped crystals which were all joined at the centre (Figure 4.9a, b, c and d). However, this can be an issue when attempting to screen the protein crystals by X-ray diffraction as the crystals would need separating in order to examine them. Therefore, protein crystals created using this screen were not ideal for use in X-ray diffraction screening.

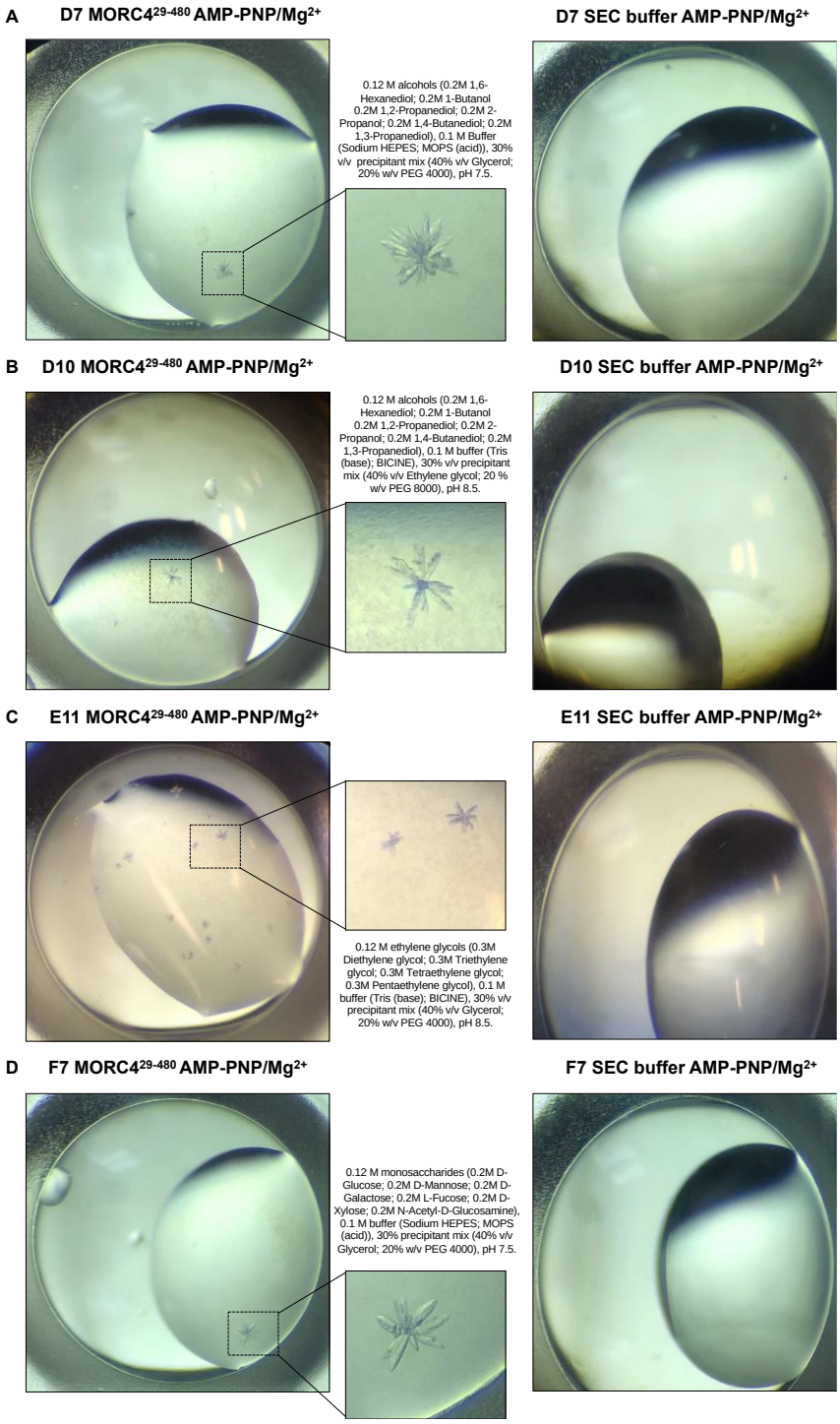


Figure 4.9: **MORC4$^{29-480}$ protein crystallisation trial with the Morpheus® screen.** Pure MORC4$^{29-480}$ protein was used for a sitting drop vapour diffuse crystal trial using the Morpheus® screen. Pure protein was diluted to low salt (100 mM), concentrated to 3.0 mg/ml and incubated with 1.0 mM AMP-PNP/Mg^{2+} at 4°C for 1 hour. Protein which had been incubated with AMP-PNP/Mg^{2+} was mixed in the plate well in a ratio of 1:1 protein: mother liquor to give 2 µL drops. All wells contained a ligand, buffer and a precipitant mix. **(A)** well D7. **(B)** Well D10. **(C)** Well E11. **(D)** Well F7. Crystals were visible after 5 days. The dimension of each well was 2.76 mm diameter.

4.3.1.3 MORC4$^{29-480}$ crystals diffracted at low resolution

Following several crystal screens and optimisation screens, several crystals were chosen for X-ray screening with the assistance of Dr. Richard Bingham (University of Huddersfield). Several crystals were mounted, however, only two crystals were successful in diffracting. The JCSG screen F4 (Figure 4.5a) crystal and the D7 (Figure 4.5b) crystal, which diffracted to approximately 20 Å (Figure 4.10). However, the structure of MORC4$^{29-480}$ could not be solved using this diffraction data, as the resolution was too low. Some evidence suggests that structures of proteins are more typically solved between 2-2.7 Å (Wlodawer *et al.*, 2008), although some proteins have been solved outside these parameters. In addition, the diffraction data also showed that the diffraction spots in the central region were close together (Figure 4.10) indicating a large unit cell value (Smyth and Martin, 2000). Although it is not possible to determine the specific unit cell from this low resolution data, the dimensions could be approximately 100 Å (R. Bingham, University of Huddersfield, *pers. comm*.). This provides further evidence that the crystals were protein crystals, as a detergent or salt crystal is highly unlikely to have such a large unit cell value (Smyth and Martin, 2000). Futhermore, crystals of salts such as NaCl would have a small number of high-intensity diffraction spots seperated by large distances (Zhang *et al.*, 2006), providing further evidence that this diffraction pattern is not from a salt crystal.

Unfortunately, due to time constraints further crystal optimisation and crystal analysis could not be continued in this project. However, future work may include incubating crystal trials at 4°C and screening low resolution crystals with a stronger beam such as a synchrotron rather than a home X-ray source (Skarzynski, 2013). In addtion, further crystal trials were limited by the amount of protein available, following low expression yields of MORC4$^{29-480}$. Therefore, using robotic liquid handling and smaller volumes (150-200 nL) would enable trials with more screens and especially, multiple protein:mother liquor ratios such as 2:1 and 1:2 in addition to

the 1:1 ratio attempted in this project. Furthermore, more purified MORC4 protein fragments may have also been investigated in crystal trials in additon to further crystallisation methods such as surface entropy reduction (Cooper *et al.*, 2007) and chemical methylation (Kim *et al.*, 2008) in an attempt to achieve higher-quality protein crystals. Moreover, insect cells could be utilised for protein expression in an effort to obtain a higher yield of protein to allow a higher number of crystal trials.

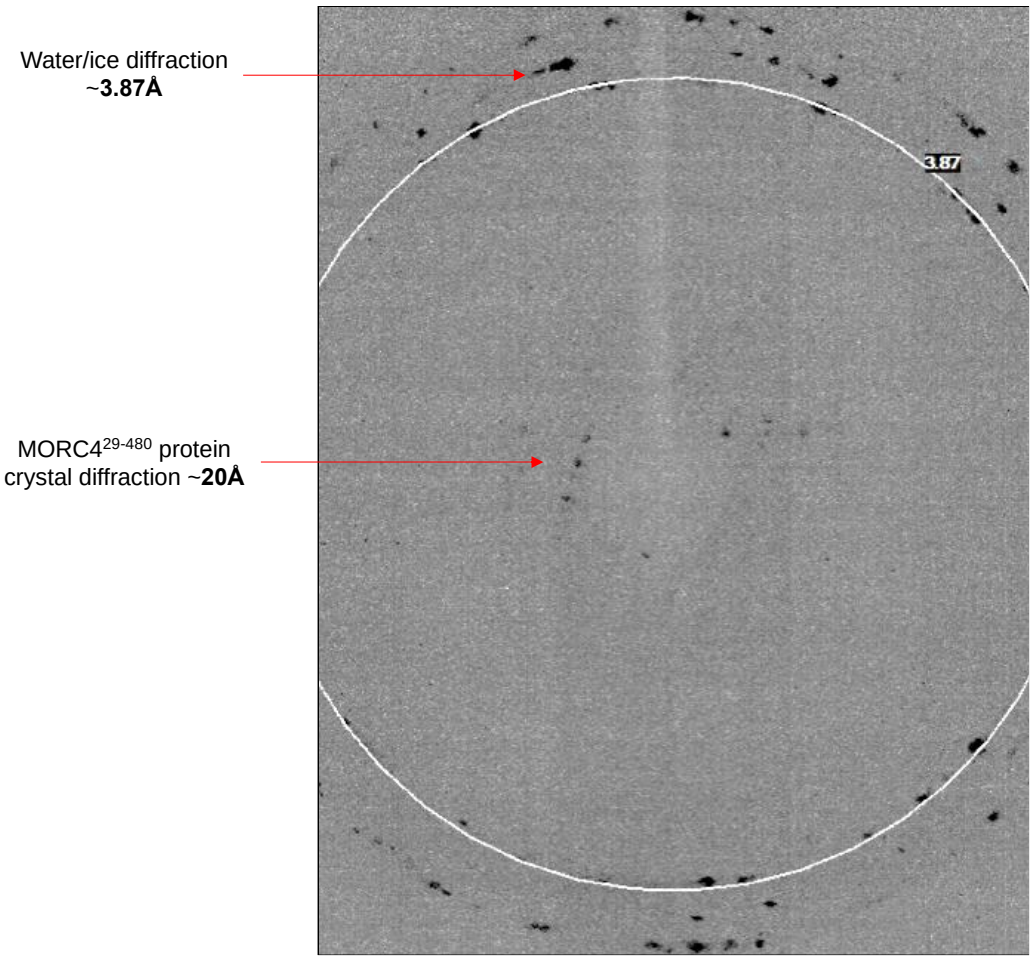


Figure 4.10: **Representative diffraction pattern of one of the low diffracting MORC4$^{29-480}$ protein crystals.** Data for this MORC4$^{29-480}$ crystal from well F4 (containing 20% v/v Jeffamine® M-600, 0.1 M HEPES pH 7.5) of the JCSG-*Plus*™ screen was collected using APEX3 software. Crystal diffraction data was taken in 10 second intervals at a distance of 140 mm, with a 2θ value of 0 mm with a 0.5° width.

4.3.2 Homology modelling

Homology modelling was used to predict the structure of MORC4 as it was not possible to determine a structure of MORC4 through protein crystallisation methods. Full-length protein sequence of MORC4 (1-900 residues) was input into an homology modelling server such as the protein homology/analogy recognition engine v2.0 (Phyre2) (Kelley *et al.*, 2015) and SWISS-MODEL (Waterhouse *et al.*, 2018; Bienert *et al.*, 2017; Guex *et al.*, 2009; Studer *et al.*, 2020; Bertoni *et al.*, 2017). Phyre2 is able to convert predictions of structures of the target sequence to a hidden Markov model (HMM), which is used to investigate against a database of known structures (Kelley *et al.*, 2015). MODELLER is then used to predict the most likely protein model (Lam *et al.*, 2017). SWISS-MODEL also uses HMM methods to predict a structure. However, this server assesses model quality by determination of a QMEAN value and also has the capability of modelling more homodimeric and heterodimeric proteins (Lam *et al.*, 2017). SWISS-MODEL was also utilised for Ramachandran analysis and quality assessments using the SWISS-MODEL structural assessment tool (Table 4.1).

Table 4.1: **Statistical analysis of all MORC4 homology models.**

Server	Model template	PDB code	QMEAN value	Overall model quality
Phyre2	MORC2	5OFB	-5.86	Poor
SWISS-MODEL	MORC2	5OF9	-4.51	Poor
Phyre2	MORC3	5IX1	-1.33	Good
SWISS-MODEL	MORC3	6O1E	-2.13	Moderate

4.3.2.1 Modelling of MORC4 using MORC2 as a structural template

4.3.2.1.1 Phyre2 analysis

Full length MORC4 was modelled using MORC2 as a template in Phyre2. This MORC4 model from MORC2 revealed 100% confidence in the target sequence and template sequence homology. However, this MORC2 template had a 37% identity to $MORC2^{1-603}$ at 2.02 Å (Figure 4.11a) (PDB: 5OFB) (Table 4.1). Unfortunately, the only MORC2 model available for Phyre2 engine analysis contained the spinal muscular atrophy mutation, S87I. This model of MORC4 was represented as a globular protein which included some of the N-terminus ATPase domain and some of the C-terminus Zf-CW domain. However, this model was also represented as a monomer. A sequence alignment of the MORC4 target sequence and MORC2 template sequence revealed that this model had a similar N-terminal sequence to the target (Figure 4.11b), suggesting some accuracy of this model. There was also a large insertion in the MORC2 template sequence from 283-360 residues, suggesting that this region was not present in the MORC4 target sequence. In addition, there were multiple insertions of residues present in the target sequence throughout (Figure 4.11b), which could affect the quality of the model.

Ramachandran analysis of this model also revealed several disallowed regions (Figure 4.12a and b), which were represented by a higher number of red residues (Figure 4.12b), suggesting that the MORC2 model was low quality. There were also more residues present in the light green and white space of the plot (Figure 4.12b), indicating that more residues had steric hindrance. Quality assessments using the SWISS-MODEL structural assessment tool also revealed that MORC2 was a poor-quality model (Figure 4.12c). The global quality estimate showed that QMEAN Z-score was -5.86, the Cβ Z-score was -2.37, the all atom Z-score -3.87 and the torsion Z-score was -5.31 (Figure 4.12c). All of these data suggest a poor-quality MORC4 model here.

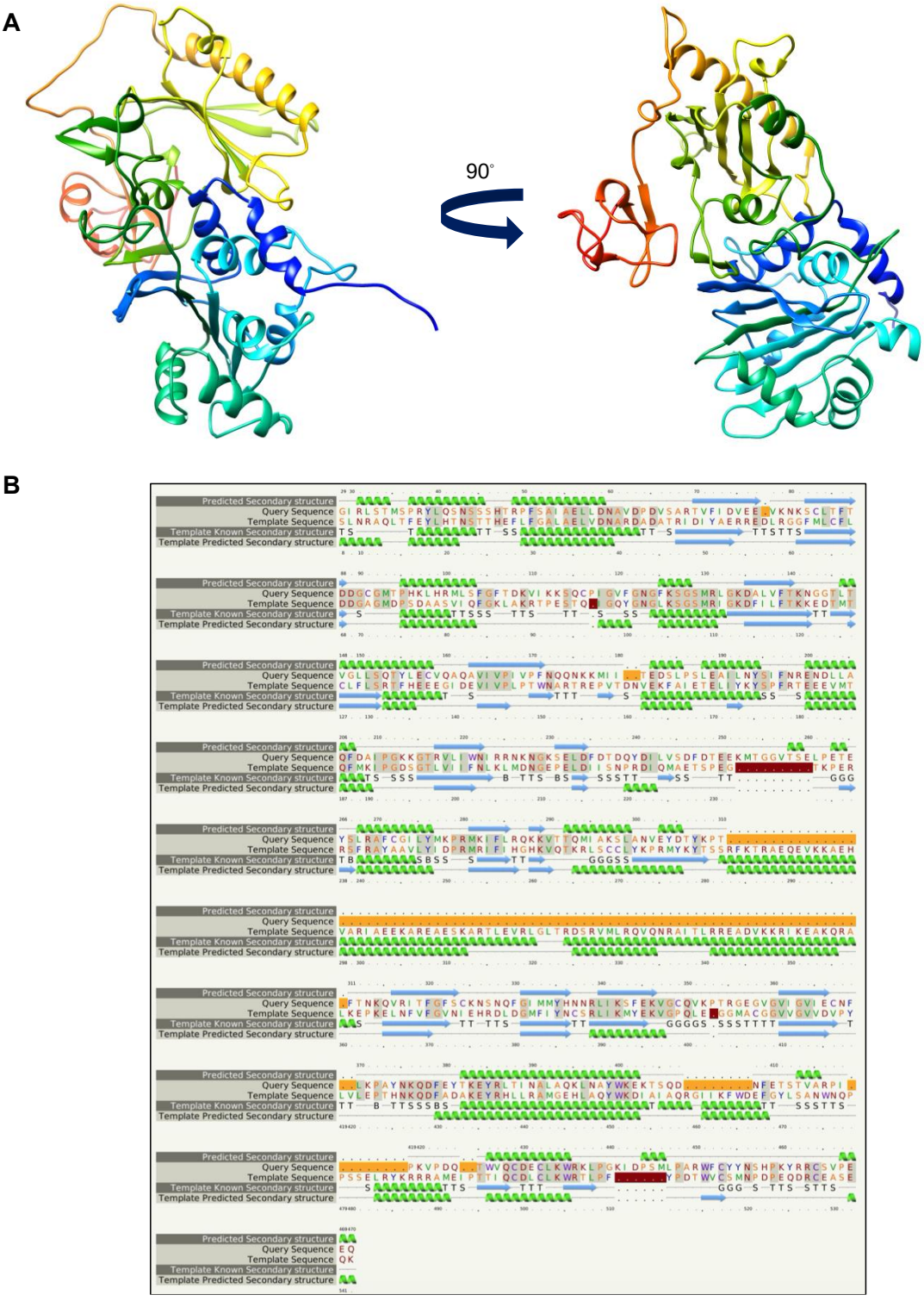


Figure 4.11: **Homology model of MORC4 using MORC2 as a template in Phyre2.** Phyre2 structural prediction software was used to create this prediction of MORC4 structure at a resolution of 2.02 Å. **(A)** The model of MORC4 is shown which was also rotated at 90°. PDB: 5OFB. **(B)** Sequence alignment of MORC4 sequence with MORC2 template sequence. Green coils and blue arrows represent alpha helices and beta sheets, respectively. G = 3-turn helix (3_{10} helix), I = 5-turn helix (π helix), T = hydrogen bonded turn, B = residue in isolated β-bridge and S = bend.

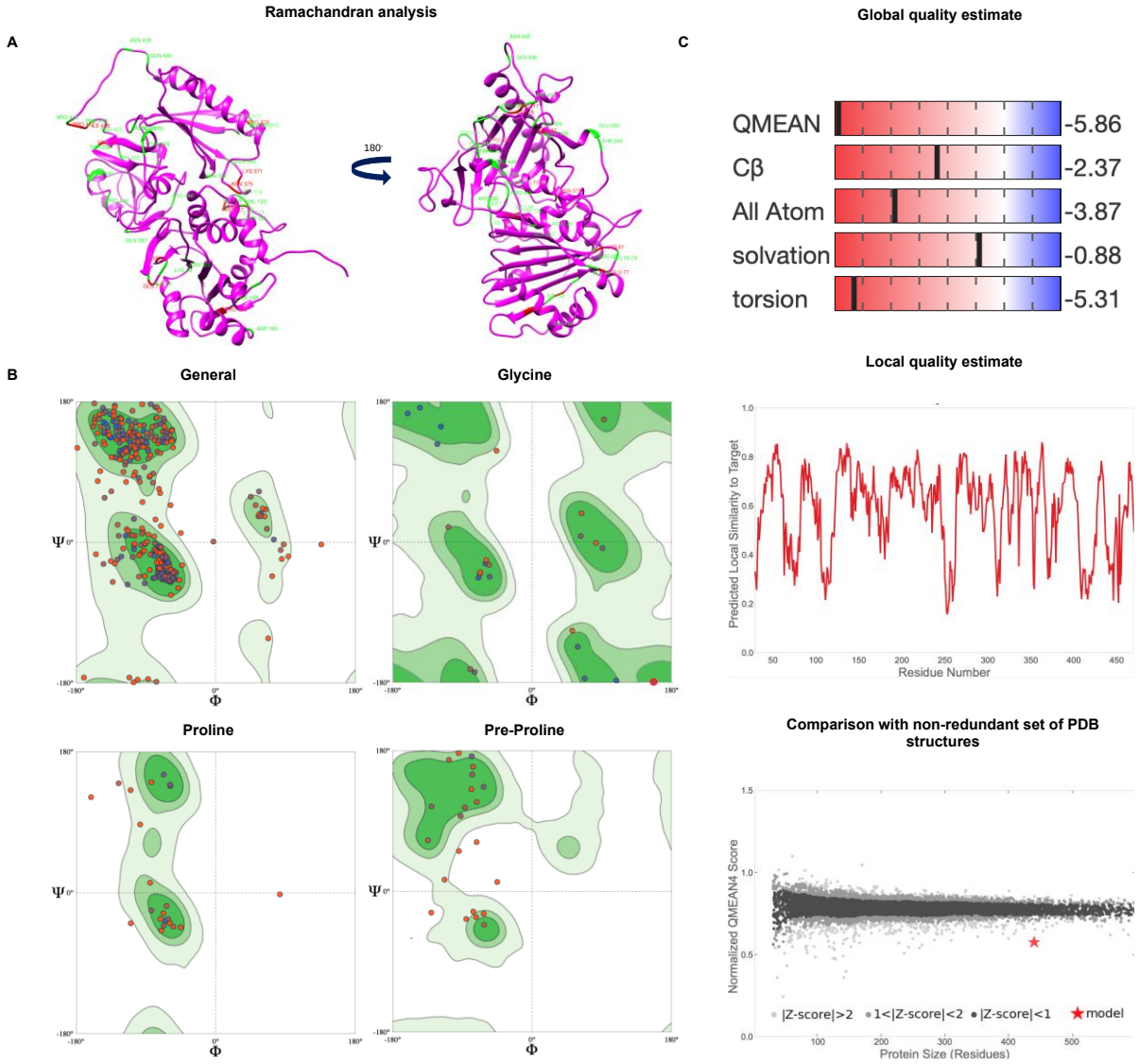


Figure 4.12: **Statistical analysis of the homology model of MORC4 using MORC2 as a template.** Phyre2 structural prediction software was used to create this prediction of the structure of MORC4 using MORC2 as a template. (**A)** The model of MORC4 is shown which was also rotated at 180° following Ramachandran analysis. Allowed residues are shown in green and disallowed residues are shown in red. (**B)** Ramachandran analysis of the MORC4 model using the SWISS-MODEL structural assessment tool. Modelled MORC4 residues are shown as red, blue or purple dots. Red and blue dots are residues which are predicted poor-quality and high-quality, respectively. The SWISS-MODEL structural assessment tool was used to perform a quality estimate. **(C)** Red represents poor-quality and blue represents high-quality. The red star represents this MORC4 model. PDB: 5OFB.

4.3.2.1.2 SWISS-MODEL analysis

SWISS-MODEL homology modelling of MORC4 used the MORC2 homodimer (39-309 residues) as a template (PDB:5OF9) (Figure 4.13a) (Table 4.1). This template structure was determined by X-ray crystallography and had a resolution of 1.81 Å. However, modelling of MORC4 to the MORC2 template resulted in a model with a 0.16 GMQE and a QMEAN value of -4.51. This model also had a sequence identity of 39.37% (Figure 4.13b), suggesting that less than half of the template is similar to the protein sequence of MORC4. As with Phyre2 homology modelling analysis, this MORC4 model was a globular protein (Figure 4.13a). However, unlike the Phyre2 analysis (Section 4.3.2.1.1), this model was represented as a homodimer and appeared to have a somewhat extended C-terminal tail (Figure 4.13a).

The poor-quality of this model was evident in the quality estimate plots (Figure 4.14b). The QMEAN and torsion Z-scores were -4.51 and -3.61, respectively, indicating a low-quality model (Figure 4.14b). Furthermore, the Ramachandran plot also indicates poor quality residues as there were several residues in the disallowed regions of the plot (Figure 4.14a), suggesting steric hindrance of some residues. All these data suggest that this model of MORC4 with MORC2 as a template is low-quality, suggesting that this model is not reliable for predicting an accurate structure of MORC4.

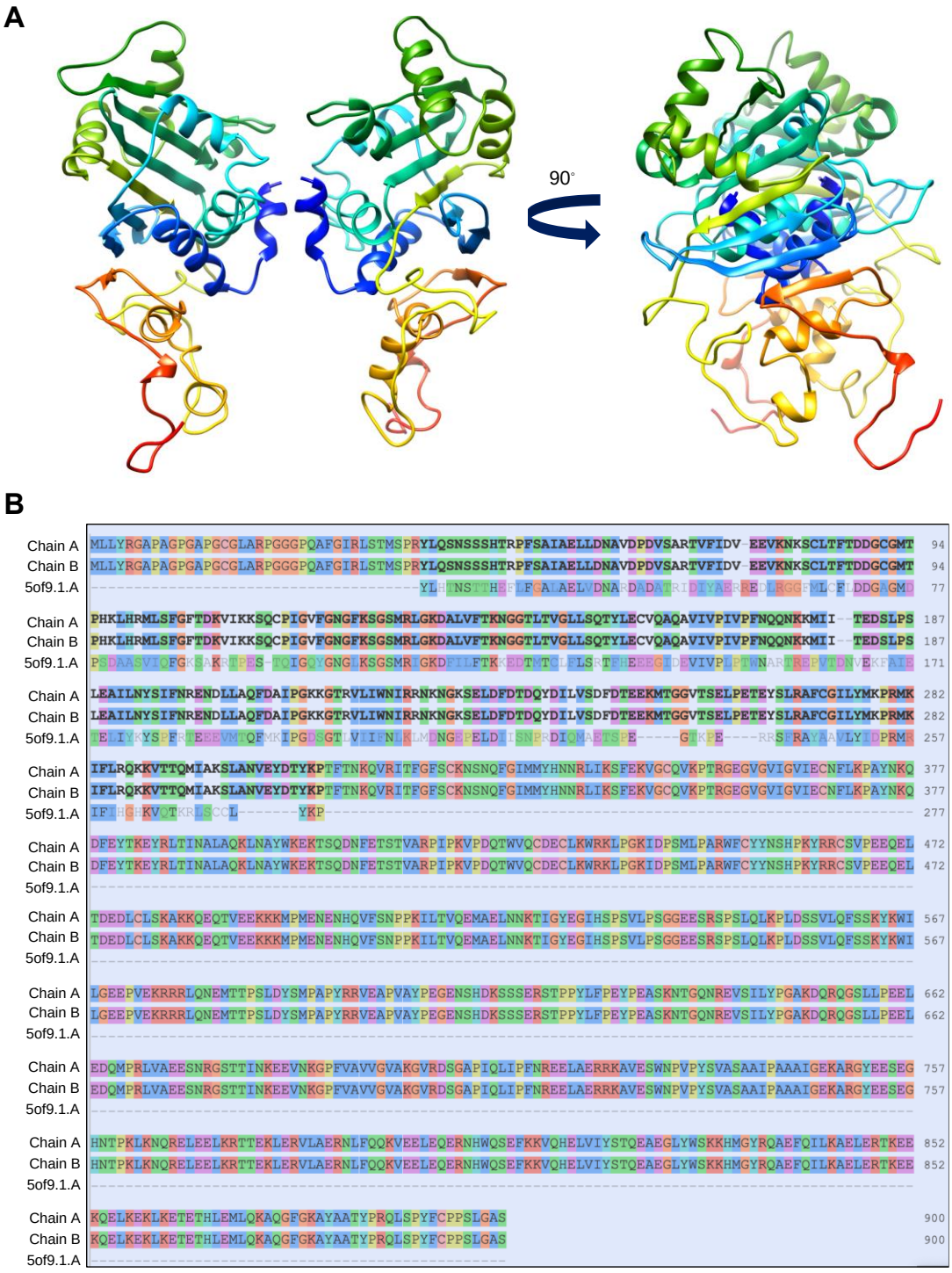


Figure 4.13: **Homology model of MORC4 using MORC2 as a template.** SWISS-MODEL structural prediction software was used to create this prediction of MORC4 structure at a resolution of 1.81 Å. **(A)** The model of MORC4 is shown which was also rotated at 90°. PDB: 5OF9. **(B)** Protein sequence alignment of MORC4 target sequence with MORC2 template sequence. The alignment was coloured using ClustalX colours.

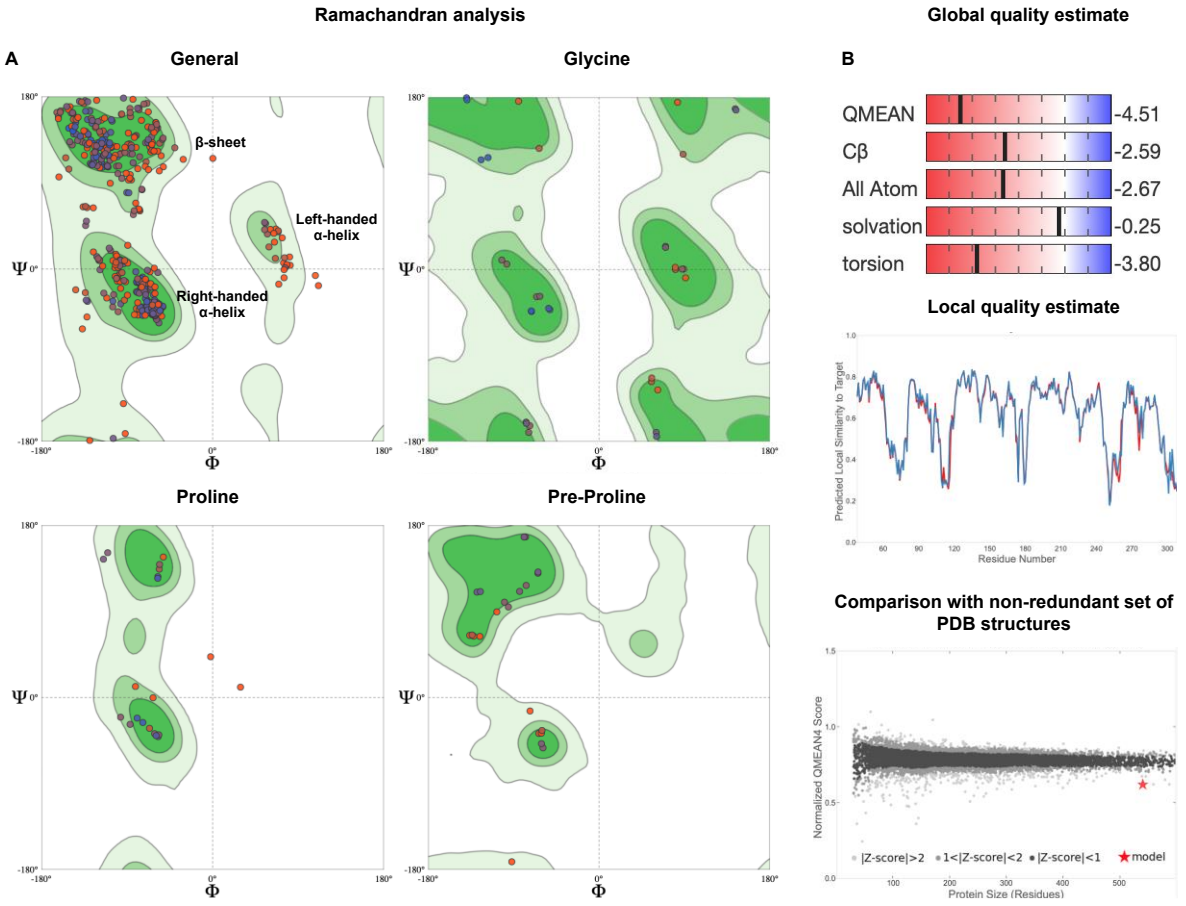


Figure 4.14: **Statistical analysis of the homology model of MORC4 using MORC2 as a template.** SWISS-MODEL structural prediction software was used to create this prediction of the structure of MORC4 using MORC2 as a template. **(A)** Ramachandran analysis of the MORC4 model using the SWISS-MODEL structural assessment tool. Modelled MORC4 residues are shown as red, blue or purple dots. Red and blue dots are residues which are predicted poor-quality and high-quality, respectively. **(B)** The SWISS-MODEL structural assessment tool was used to perform a quality estimate. Red represents poor-quality and blue represents high-quality. The red star represents this MORC4 model. PDB: 5OF9.

4.3.2.2 Modelling of MORC4 using MORC3 as a structural template

4.3.2.2.1 Phyre2 analysis

Following the input of MORC4$^{1\text{-}900}$ into the Phyre2 server, MORC4 modelled from MORC3 revealed 100% confidence that the target sequence and the template sequence were homologous. This model also had 60% identity to the structure of MORC3$^{29\text{-}473}$ with AMP-PNP and H3K4me3 peptide (PDB:5IX1) (Table 4.1), with a resolution of 2.60 Å (Figure 4.15a). A sequence alignment of this model with the MORC4 target sequence and MORC3

template sequence revealed that MORC4 was modelled was from 29-473 residues and the locations of multiple α-helices and β-sheets were predicted (Figure 4.15b). Unfortunately, a small quantity of gaps were present in the template sequence of MORC3 (presumably resulting from disordered regions), specifically between residues 405-419, leading to missing residues in the homology model (Figure 4.15a and b).

The overall fold of this model of MORC4 was globular (Figure 4.15a), similar to MORC3 (Li *et al.*, 2016). This model also included both the N-terminus ATPase domain and some of the C-terminus Zf-CW domain (residues 29-473) (Figure 4.15a). However, the coiled-coil region of MORC4 was not included in this model, suggesting that perhaps it is dissimilar from MORC3. In addition, this MORC4 model was also represented as a monomer when using the Phyre2 server, whereas, MORC3 is a homodimer (Li *et al.*, 2016). Furthermore, this model did not show any binding to Zn^{2+} at the C-terminal Zf-CW binding site or binding of the ligand, AMP-PNP (Figure 4.15a) as neither the zinc ion or AMP-PNP were identified in the structure, suggesting that the binding sites in this MORC4 model are not conserved in MORC3.

Ramachandran analysis revealed that there were several residues that were allowed (represented in green) and a small number of disallowed residues (represented in red, Figure 4.16a). The majority of glycine and proline residues were also allowed (Figure 4.16b), resulting from less steric hindrance of these residues. In addition, the quality assessment of this model using the SWISS-MODEL structural assessment tool revealed that the model was of moderate quality with a QMEAN Z-score of -0.31 (Figure 4.16c), indicating this model is higher quality than the MORC2 template model. Furthermore, a local quality assessment plot suggested a score above 0.6 for the majority of the model (Figure 4.16c), indicating a high-quality model.

Overall, these data all suggest that this Phyre2 MORC4 model utilising MORC3 as a template is a high-quality model.

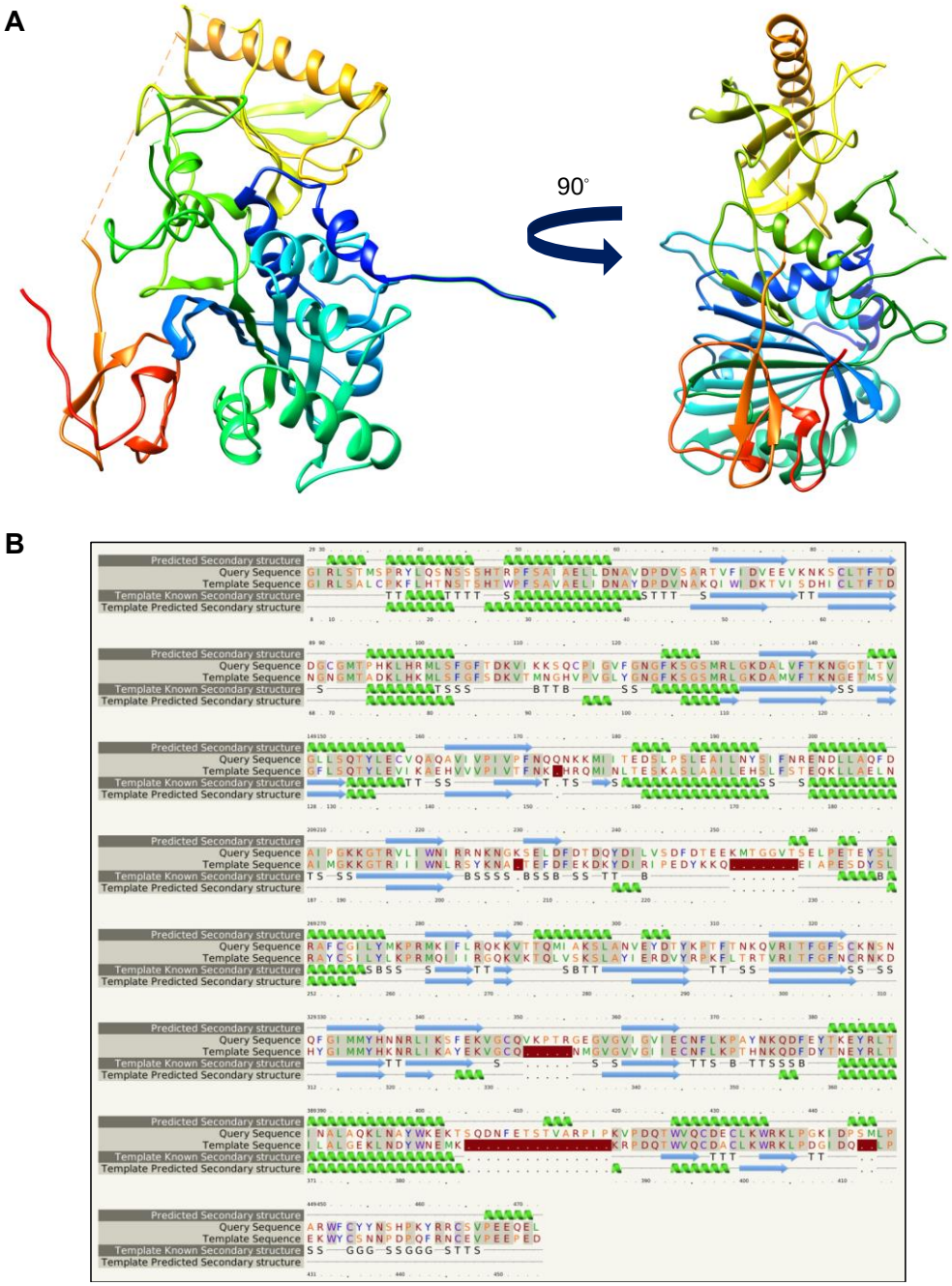


Figure 4.15: **Homology model of MORC4 using MORC3 as a template.** Phyre2 structural prediction software was used to create this prediction of MORC4 structure at a resolution of 2.60 Å. (**A)** The model of MORC4 is shown which was also rotated at 90°. PDB: 5IX1. **(B)** Sequence alignment of MORC4 sequence with MORC3 template sequence. Green coils and blue arrows represent alpha helices and beta sheets, respectively. G = 3-turn helix (3_{10} helix), I = 5-turn helix (π helix), T = hydrogen bonded turn, B = residue in isolated β-bridge and S = bend.

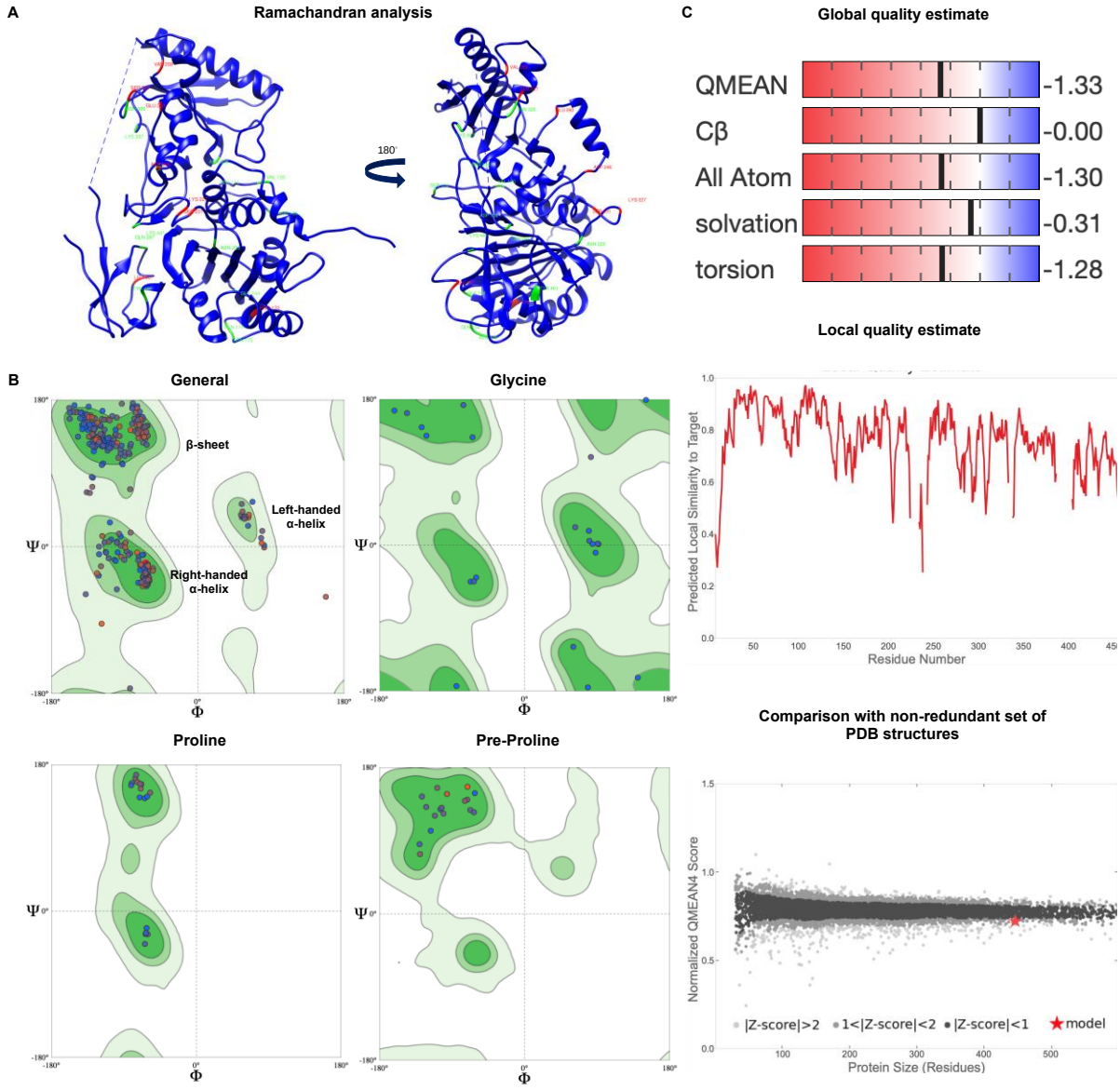


Figure 4.16: **Statistical analysis of the homology model of MORC4 using MORC3 as a template.** Phyre2 structural prediction software was used to create this prediction of the structure of MORC4 using MORC3 as a template. (**A)** The model of MORC4 is shown which was also rotated at 180° following Ramachandran analysis. Allowed residues are shown in green and disallowed residues are shown in red. (**B)** Ramachandran analysis of the MORC4 model using the SWISS-MODEL structural assessment tool. Modelled MORC4 residues are shown as red, blue or purple dots. Red and blue dots are residues which are predicted poor-quality and high-quality, respectively. **(C)** The SWISS-MODEL structural assessment tool was used to perform a quality estimate. Red represents poor-quality and blue represents high-quality. The red star represents this MORC4 model. PDB: 5IX1.

4.3.2.2.2 SWISS-MODEL analysis

The top hit for modelling of MORC4 was the MORC3 homodimer template (residues 30-471)

(PDB:6O1E) (Figure 4.17) (Table 4.1), which was determined by X-ray crystallography and

had a resolution of 2.41 Å. The global model quality estimation (GMQE) value was 0.39, with 1 representing the highest reliability model (Biasini *et al.*, 2014). This model also had a sequence identify of 59.13%, suggesting more than half of the target sequence was present in this model. The QMEAN score (Benkert *et al.*, 2008) was -2.05, a score of 0 would indicate good accuracy between the target and the model. A score of more than -4.0 would suggest a low-quality model (Studer *et al.*, 2020). This suggests that this MORC4 model was medium quality.

This prediction of MORC4 structure was also modelled as a homodimer and showed binding of Zn^{2+} to the Zf-CW domain (Figure 4.17a), unlike the Phyre2 model but similar to the structure of MORC3 (Li *et al.*, 2016). This model also determined that the ligand, AMP-PNP and Mg^{2+}, which is bound to MORC3 to create homodimerisation (Zhang *et al.*, 2019), is not present in this MORC4 model. The binding site of AMP-PNP is not conserved in this model and Mg^{2+} is not in contact with this model (Figure 4.17a). The overall fold of this model was globular, similar to MORC3 (Li *et al.*, 2016) and only the N-terminus ATPase domain and C-terminus Zf-CW structure were predicted. This is similar to the structure of MORC3, where only the ATPase-Zf-CW domains have been crystallised, as evidence has suggested that the coiled-coil domain follows a unpredictable flexible linker region (Li *et al.*, 2016) resulting in an unknown structure of MORC3 for this C-terminal region. This similarity between the predicted structure of MORC4 and MORC3 could suggest that they have similar functions as proteins. In addition, the protein sequence alignment of the MORC4 target sequence and the MORC3 template showed this model begins from position 30 isoleucine and ends at position 471 glutamic acid. The residue deletion identified in the Phyre2 MORC3 homology model from residues 405-419 was not present in this SWISS-MODEL model (Figure 4.17b).

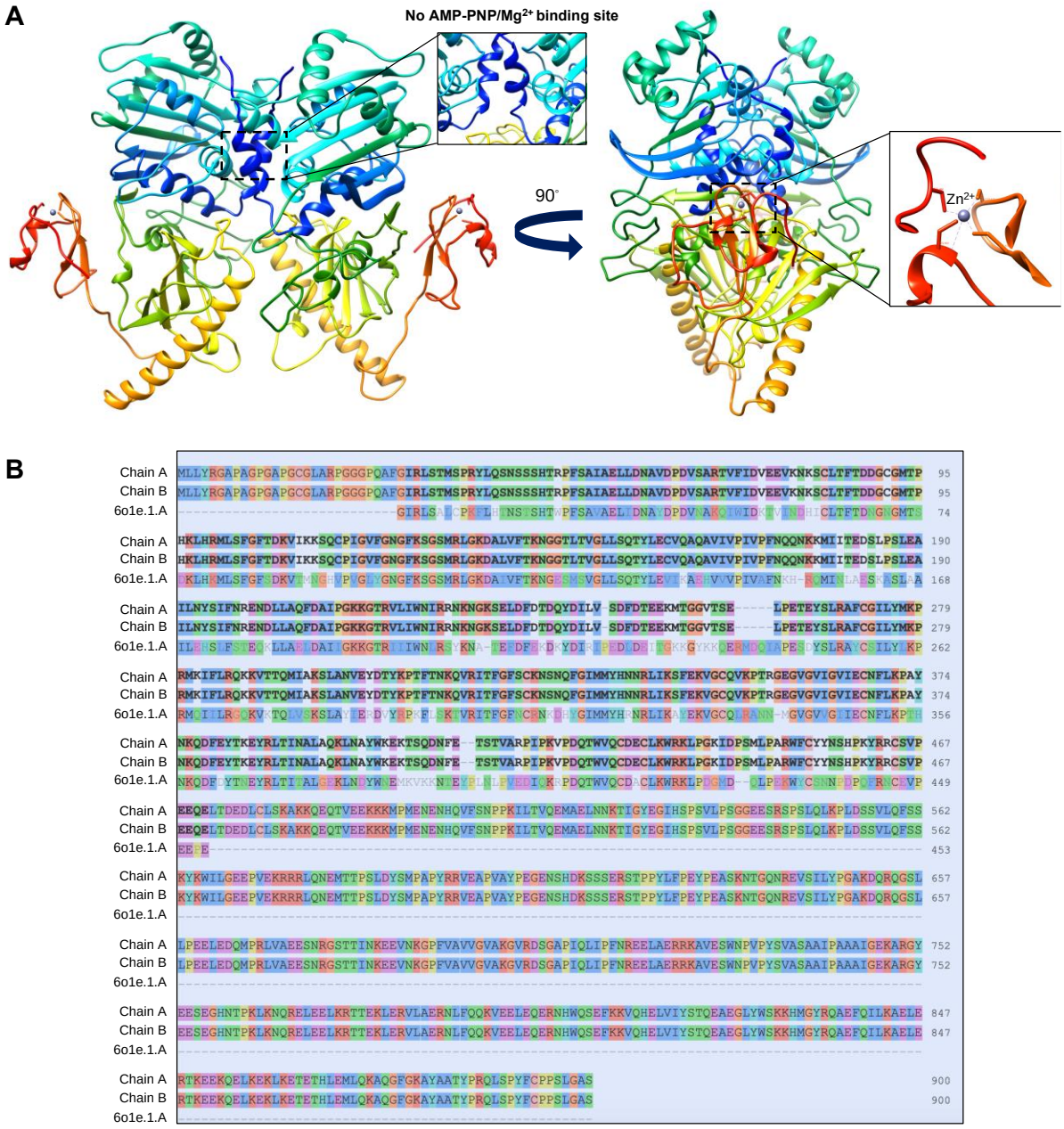


Figure 4.17: **Homology model of MORC4 using MORC3 as a template.** SWISS-MODEL structural prediction software was used to create this prediction of MORC4 structure at a resolution of 2.41 Å. (**A)** The model of MORC4 is shown which was also rotated at 90°. Zn^{2+} is shown as silver spheres. (**B)** Protein sequence alignment of MORC4 target sequence with MORC3 template sequence. The alignment was coloured using ClustalX colours. PDB: 6O1E.

Ramachandran analysis of this model also showed high-quality residues in general (Figure 4.18a). However, there were a higher number of low-quality residues in this model (Figure 4.18a) in comparison to the Phyre2 model, suggesting that the SWISS-MODEL model is of lower quality. In addition, quality output analysis revealed this model had a Z-score of between 1< and <2, suggesting a good quality model (Figure 4.18b). Although the Phyre2 model was higher quality (Section 4.3.2.21), this SWISS-MODEL homodimer model had an increased level of accuracy as it was not bound to the H3K4me3 peptide.

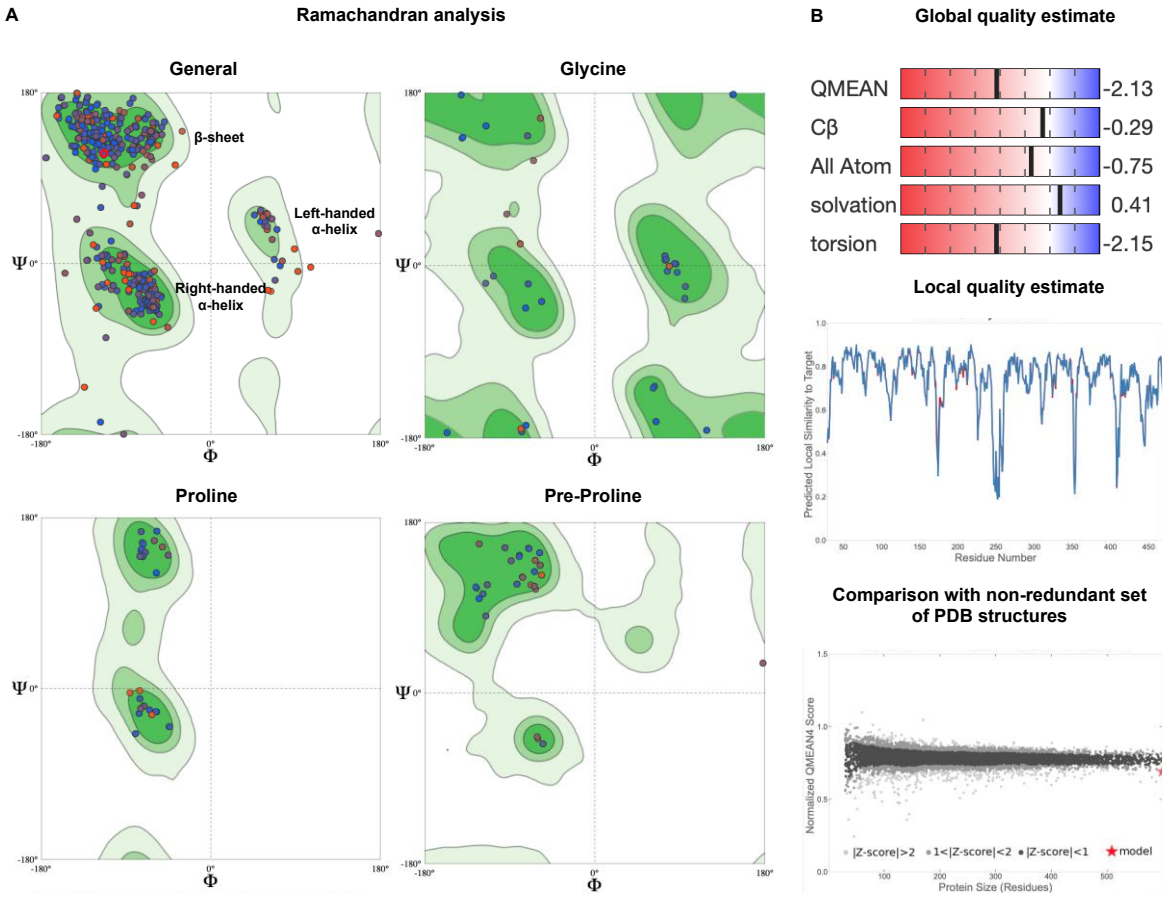


Figure 4.18: **Statistical analysis of the homology model of MORC4 using MORC3 as a template.** SWISS-MODEL structural prediction software was used to create this prediction of the structure of MORC4 using MORC3 as a template. **(A)** Ramachandran analysis of the MORC4 model using the SWISS-MODEL structural assessment tool. Modelled MORC4 residues are shown as red, blue or purple dots. Red and blue dots are residues which are predicted poor-quality and high-quality, respectively. The SWISS-MODEL structural assessment tool was used to perform a quality estimate. **(B)** Global quality estimate. Red represents poor-quality and blue represents high-quality. The red star represents this MORC4 model. PDB: 6O1E.

4.3.2.3 Superimposition of MORC4 structural models

4.3.2.3.1 Phyre2 analysis

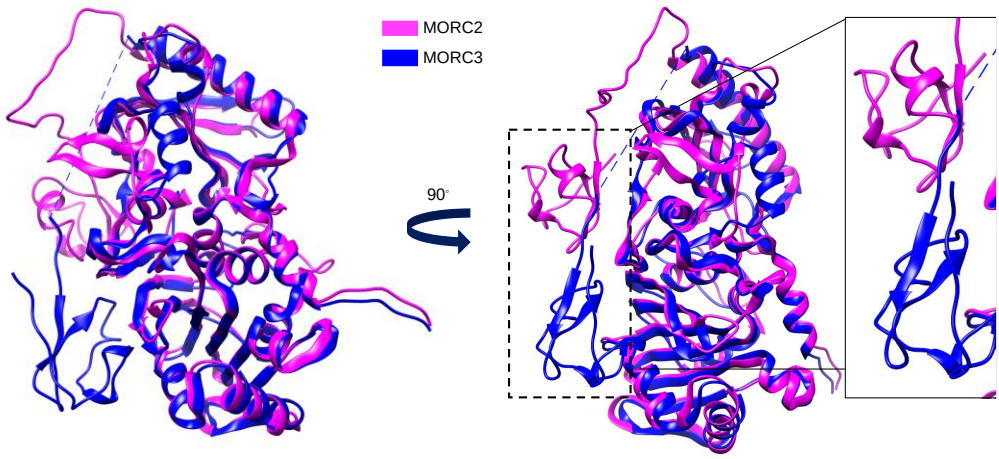
Superimposition of MORC2 and MORC3 models of MORC4 revealed that 351 pairs of residues were the same and a root mean square distance (RMSD) score of 2.1 Å (Figure 4.19a), suggesting that the average distance between atoms is 2.1 Å (Kufareva and Abagyan, 2012). In addition, the calculated template modelling (TM)-score was 0.77, suggesting that both models had a similar overall fold (Kelley *et al.*, 2015).

The superimposed models of MORC4 also revealed that both had dissimilar C-terminals structures, which was evident both in the model structures (Figure 4.19b) and protein sequences (Figure 4.19c). However, as determined by the model quality assessment and Ramachandran analysis, MORC4 models using MORC2 as a template was lower quality compared to the MORC3 template model, which could suggest that the MORC2 model of MORC4 is less accurate. Despite the MORC3 model being a higher quality model, the MORC3 template was bound to H3K4me3 peptide (Li *et al.*, 2016), thus, to create a more accurate representation the MORC2 model template would also need to be bound to this peptide. However, the Phyre2 server did not provide a MORC3 model without the H3K4me3 peptide. The ATPase domain of this superimposed model of MORC4 appears to be consistent with the MORC3 template (Figure 4.19b) except for the occasional point mutation and insertion/deletion (Figure 4.19b). In addition, the superimposed model of MORC4 produced using MORC3 as a template with the structure of MORC3 showed evidence of a change in structure of the C-terminal Zf-CW when bound to the H3K4me3 peptide (Figure 4.20a and b). This suggests that upon binding of the H3K4me3 peptide, MORC3 and perhaps MORC4 undergo a conformational change.

A

Rank	Model	Pairs	RMSD	TM-score
2	c5ix1A_ (Chain 'A')	351	2.186	0.77

B



C

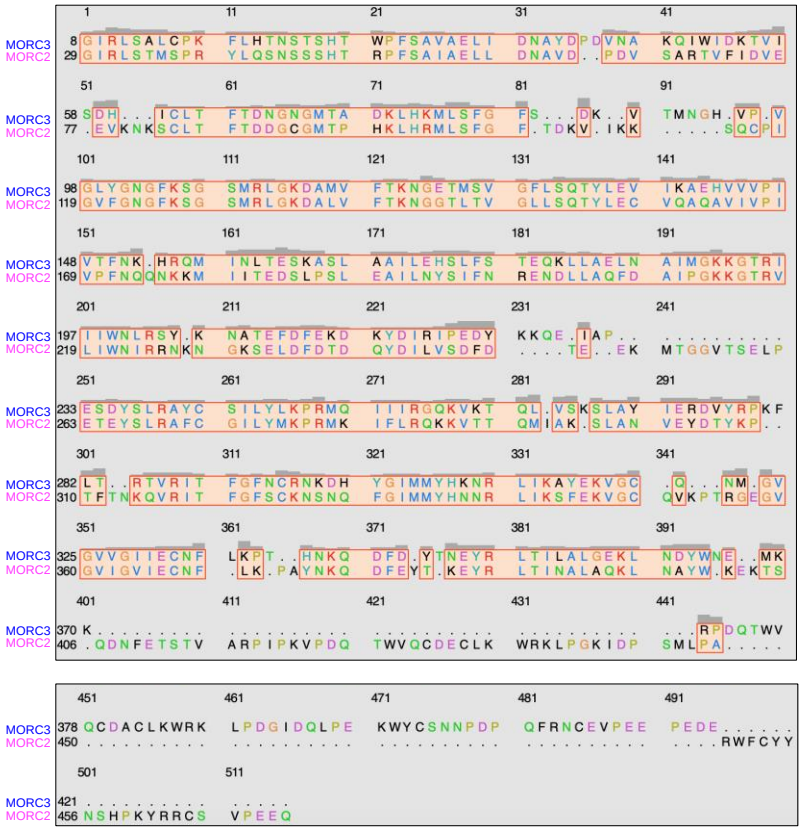


Figure 4.19: **Superimposed homology model of MORC4 using both MORC2 and MORC3 as templates.** Phyre2 structural prediction software was used to create this superimposition of the structure of MORC4 using MORC2 and MORC3 as templates. (**A)** Table showing the number of pairs found in the model, the root mean square distance (RMSD) between pairs of residues that were aligned and the template modelling (TM) score. (**B)** Superimposed models of MORC4 also rotated at 90° and created using UCSF Chimera software. (**C)** Aligned protein sequences of both MORC4 models. MORC2 PDB: 5OFB. MORC3 PDB: 5IX1.

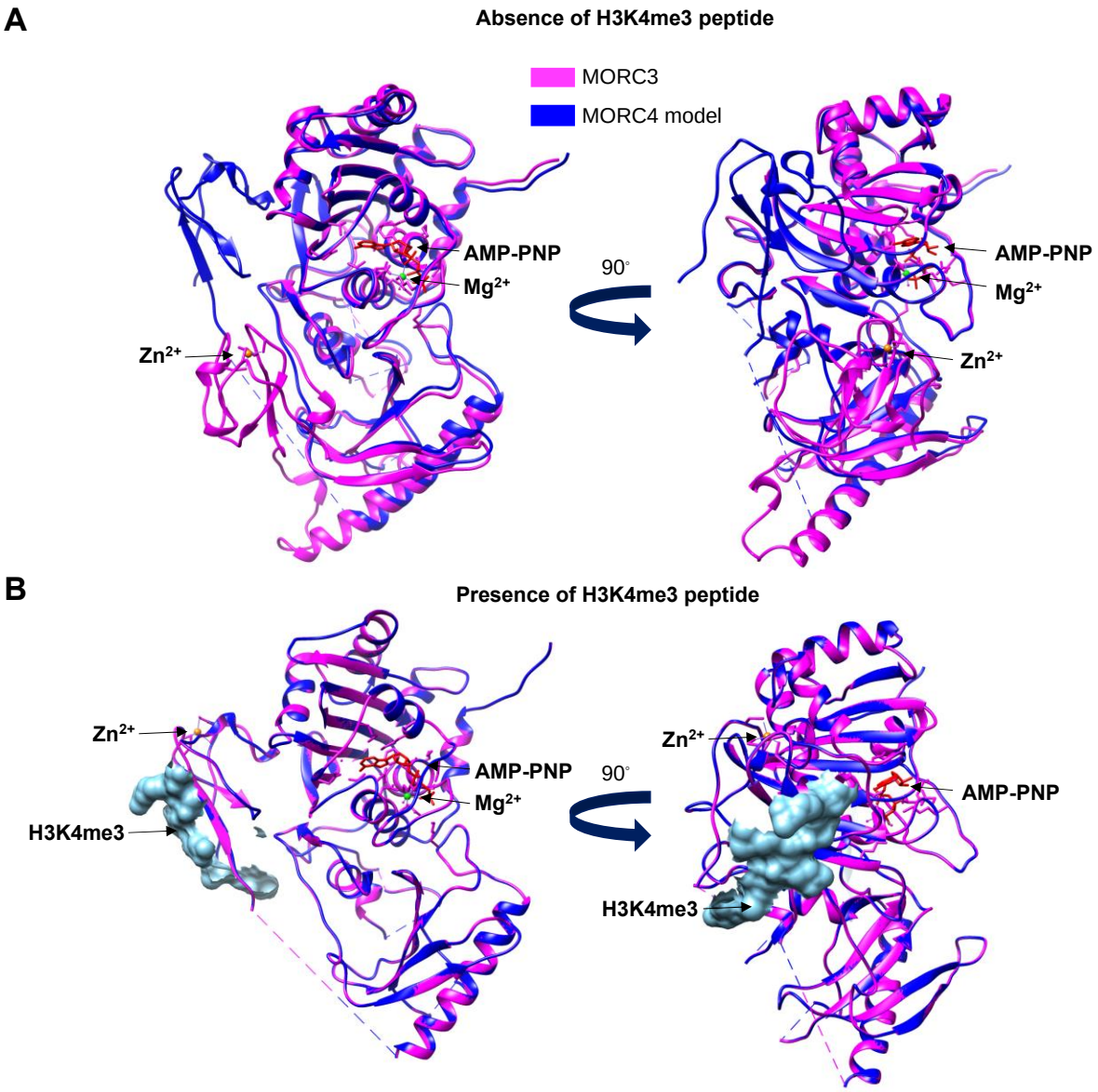


Figure 4.20: **Homology model of MORC4 using MORC3 as a template superimposed with the MORC3 structure.** UCSF Chimera software was used to create this superimposition of the homology model of MORC4 and the MORC3 structure. **(A)** Superimposed models of MORC4 also rotated at 90° in the absence of H3K4me3 peptide. MORC3 structure PBD: 6O1E. MORC4 model PBD: 5IX1. **(B)** Superimposed models of MORC4 also rotated at 90° in the presence of the H3K4me3 peptide. The MORC3 structure and MORC4 model is shown in pink and blue, respectively. MORC3 structure and MORC4 model PBD: 5IX1. AMP-PNP is represented in red; Mg^{2+} is represented in green; Zn^{2+} is represented in orange and the surface of the H3K4me3 peptide is represented in light blue.