Building the *Drosophila* Centriole:

A Structural Investigation of the Centriolar Proteins SAS-6, SAS-4 and Ana2

Matthew A. Cottee



Sir William Dunn School of Pathology

and

Brasenose College

A thesis submitted to the University of Oxford for the degree of Doctor of Philosophy

Trinity Term 2014

Abstract

Building the *Drosophila* Centriole: A Structural Investigation of the Centriolar Proteins SAS-6, SAS-4 and Ana2

Matthew A. Cottee

Sir William Dunn School of Pathology and Brasenose College

Submitted for the degree of Doctor of Philosophy

Trinity Term 2014

The centriole is a complex cylindrical assembly found in the cells of ciliated eukaryotes. It serves two important roles in the cell: templating the growth of cilia, and forming the basis of the centrosome, which is the major microtubule organising centre in the cell. Cilia and centrosomes are involved in many cellular processes, from signalling to cell division and differentiation. As such, defects in centriole assembly can have downstream consequences on these processes and are linked to a variety of human diseases including cancer and microcephaly.

The complex superstructure of the centriole has fascinated biologists for decades. It comprises a nine-fold, radially symmetric array of microtubule triplet blades attached to a central cartwheel structure. During the last two decades, proteomic analyses have identified many proteins that are associated with the centriole. However, genetic studies have shown that only a surprisingly small number of these proteins are essential for the biogenesis of the centriole.

In *Drosophila melanogaster*, three such essential proteins, SAS-6, Ana2 and SAS-4 are required in the early stages of centriole biogenesis. In this thesis I have investigated the assembly steps involving these key players from a structural perspective. I have identified and recombinantly expressed functional domains of these proteins in order to characterise them *in vitro*. Using X-ray crystallography and other biophysical techniques, I have been able to define mechanisms for several steps involved in the assembly of these proteins. In collaboration with colleagues in the laboratory I have been able to investigate the biological significance of these essential assembly steps *in vivo*. This information has provided novel insights into the molecular, and even atomic, detail of the initial steps of centriole assembly, including an explanation of a natural point mutation involved in human microcephaly.

Foreword and Acknowledgements

And so lies before you my thesis. Like everyone who writes one, for three months or so I gave everything to Madame la Thèse, and I saw her as a cruel mistress. I existed mainly in an unshaven and bedressing-gowned state in my room upstairs – like most weekends in fact – and I emerged (some say as before) unable to relate to the real world. Thank you Farv for insisting I get a desk-chair, monitor and keyboard to make the last month more comfortable! Although in retrospect the whole thing would have been much easier had I planned it better and started earlier, I now see this period of time as a very rewarding one.

I must declare my extreme gratitude to Jordan and Susan for co-supervising me on this project. When I asked Jordan if he was interested in hosting a structural biology student in his cell biology lab he was full of enthusiasm! I therefore had the unusual position of joining a cell biology group to conduct structural studies, and this allowed me to have general free-reign through both the Raff and Lea groups' vast resources – it has been a privilege that will be difficult to leave and I imagine difficult to find elsewhere! Jordan has been immensely positive and supportive, and always trusting that I was actually getting stuff done. Thank you for the intellectual freedom I've been allowed to enjoy, for never treating me as "just a student," and for the faith you've had in my general endeavours. Thank you for the opportunities to review and to supervise. Thank you Susan for expecting such high scientific standards, and providing the tools, guidance and know-how to achieve these things. I have found it incredibly valuable to learn about every step from construct design to structural solution – and be expected to be able to do it myself. Thank you both for the generous commitment of any resource required for any experiment. It felt like nothing was ever too expensive in the pursuit of results, and I must try not to take this for granted. Thank you to all the members of the Lea and Raff groups throughout my DPhil. It goes without saying you have all helped me in so many ways.

I must especially thank Steve, who (amongst teaching me basic lab tekkers before my DPhil) was always willing to advise and bend his ear to hash over the intricacies of crystallography or experimental design. Thank you for helping me turn my mad logic into (hopefully) scientific logic. Thanks also for helping carry out the vast amount of MALS experiments conducted over the last few years.

I want to convey especial thanks to my structural partner-in-crime in the Raff lab, (violent) Violet Feng. Working with you has brought me so much extra experience. When you turned up I suddenly had to ensure I was doing things as properly as possible, and there was a greater urgency to try and fill gaps in my knowledge. I daresay I'd have achieved much less without you, and I think working with you was one of the most valuable experiences in my DPhil. I am incredibly proud of you and you have my eternal gratitude. I must also apologise for my general madness, you have put up with me so well! Further thanks to Violet for allowing me to escape at the weekends by moving my plates into the fridge etc. etc.!

Thank you Saroj for your tireless dedication to everyone in the Raff lab and ensuring the lab was always running so smoothly. Thank you to Paul and Alan for always being obliging towards me and my philosophical chats. The importance of this goes without saying. Thank you to "aaaaaaaaah-Vicente!", for being a great bench partner, and by "donating" your benchspace to me after leaving. I apologise to both Vicente and Metta for being so messy and invading your

benches. On the topic of benches, thank you to whoever didn't tell me off for always making notes on the bench. Thank you to Nadine for our close collaborative work, which allowed us to get a fuller perspective on these scientific questions, and thank you to Jo Leveson for your dedication to your project and the valuable results it yielded. Thanks Chappell and Matthijs for the restorative breaks (we cordially invite you for a game of Wiff-Waff!) - every department should have a ping-pong table.

Thank you to Mark van Breugel, with whom I worked closely on the SAS-4/Ana2 project. We had an immensely satisfying and fruitful collaboration, and I strongly appreciated being part of such a great team. Sorry again for misspelling your name in the SAS-6 review…

Thanks to the BBSRC and the Dunn School EPA fund for financing my life for 4 years, and thanks to the Dunn School for taking a gamble on my being worth funding.

Thank you to Ashley and Emma for being great housemates for two years during this time, and for your solidarity in the face of house related issues. Thank you to all my family for your limitless support and for the holidays! Thank you Davey boy and G'ma for your persistent visits and associated provisions of supplies and meals out. Thank you Sophie, for loving such a nutcase, putting up with my boring self and late nights during the write up, and well, for everything really!

I acknowledge (although do not thank) Napoleon, the documentaries of whose life for some reason allowed me to unwind at the end of a long day of writing.

Lastly, to all others who have helped me but remain un-thanked, you have my gratitude.

Even if this thesis is only ever really read by Susan, Jordan and my examiners, I hope you at least enjoy the pictures.



Long live the venus fly trap (2011-): Although she didn't cope well with *Drosophila*, she's managed a few wasps in her time!

Publications resulting from this study

Cottee, M.A., Raff, J.W., Lea, S.M., and Roque, H. (2011). SAS-6 oligomerization: the key to the centriole? Nat Chem Biol 7, 650–653.

Cottee, M.A., Muschalik, N., Wong, Y.L., Johnson, C.M., Johnson, S., Andreeva, A., Oegema, K., Lea, S.M., Raff, J.W., and van Breugel, M. (2013). Crystal structures of the CPAP/STIL complex reveal its role in centriole assembly and human microcephaly. eLife 2, e01071.

(Post thesis submission)

Cottee, M.A., Muschalik, N., Johnson, S., Leveson, J., Raff, J.W., and Lea, S.M. (2015). The homo-oligomerisation of both Sas-6 and Ana2 is required for efficient centriole assembly in flies. eLife 4, e07236.

Abbreviations

Ana2	Anastral Spindle 2
Asl	Asterless
ASU	Asymmetric Unit
BLAST	Basic Local Alignment Search Tool
βME	β-mercaptoethanol
BSA	Bovine Serum Albumin
CC	Coiled-Coil
CD	Circular Dichroism
CeSAS-4	*Caenorhabditis elegans* SAS-4
CID	Cartwheel Inner Density
CrSAS-6	*Chlamydomonas reinhardtii* SAS-6
Cnn	Centrosomin
CoIP	Co-Immunoprecipitation
CPAP	Centrosomal P4.1-Associated Protein
CPB	Cryptic Polo Box
CR	Conserved Region
CSS	Complex Significance Score
CV	Column Volume
DDR	DNA Damage Response
DLC	Dynein Light Chain
DmSAS-4	*Drosophila melanogaster* SAS-4
DmSAS-6	*Drosophila melanogaster* SAS-6
DrCPAP	*Danio rerio* CPAP
DrSAS-6	*Danio rerio* SAS-6
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
EM	Electron Microscopy
ESI-MS	Electrospray Ionisation Mass Spectrometry
Fc	Flow channel
FPLC	Fast Performance Liquid Chromatography
FT	Flowthrough
GFP	Green Fluorescent Protein
H-bond	Hydrogen Bond
IEX	Ion-Exchange Chromatography
IMAC	Immobilised Metal Affinity Chromatography
IPTG	Isopropyl-β-D-1-thiogalactopyranoside
Kd	Dissociation constant

LB	Luria Bertani broth
MAD	Multi wavelength Anomalous Dispersion
MALS	Multi Angle Light Scattering
MBP	Maltose Binding Protein
MCPH	Microcephaly Primary Hereditary (Autosomal Recessive Primary Microcephaly)
MR	Molecular Replacement
MST	Microscale Thermophoresis
MT	Microtubule
MW	Molecular Weight
MWCO	Molecular Weight Cut Off
Ni-NTA	Nickel Nitrilotriacetic Acid
OD_{600}	Optical Density at 600 nm
PBS	Phosphate Buffered Saline
PCM	Pericentriolar Material
PDB	Protein Data Bank
PISA (motif)	Present In SAS-6 motif
PISA (server)	Proteins, Interfaces, Structures and Assemblies server
PIT	Protease Inhibitor Cocktail Tablet
PRM	Proline-Rich Motif
RMSD	Root Mean Square Deviation
RNAi	RNA interference
RT	Room Temperature
RU	Response Units
SAD	Single wavelength Anomalous Diffraction
SAPs	SAS-6/Ana2 Particles
SAS-4	Spindle Assembly Abnormal 4
SAS-5	Spindle Assembly Abnormal 5
SAS-6	Spindle Assembly Abnormal 6
SDS-PAGE	Sodium Dodecyl Sulphate - Polyacrylamide Gel Electrophoresis
SEC	Size Exclusion Chromatography
SeMet	selenomethionine
SRF	Self Rotation Function
STAN	STIL/Ana2 domain
STIL	SCL TAL1 Interrupting Locus
TCP (domain)	TCP10 or T-complex 10 domain
Ubq	Ubiquitin
WB	[protein purification] Working Buffer
WT	Wild Type
Y2H	Yeast 2 Hybrid
ZYG-1	Zygotic Defective 1

Table of Contents

Abstract i

Foreword and Acknowledgements iii

Publications resulting from this study v

Abbreviations vi

Chapter 1: Introduction 1

1.0: Preface 1

1.1: Centriole superstructure 2

1.2: Functions of centrioles: centrosomes and cilia 7

1.2.1: The centrosome and the centriole duplication cycle 7

1.2.2: Centrioles in cilia 10

1.3: Centrosomes and cilia in disease 12

1.3.1: Autosomal Recessive Primary Microcephaly (MCPH) 13

1.4 Centriole biogenesis 15

1.4.1 – The centriolar kinase Sak/PLK4/ZYG-1 16

1.4.2 – SAS-6 18

1.4.3 – SAS-5/Ana2/STIL 20

1.4.4 – SAS-4/CPAP 23

1.4.5 – Other core proteins 24

1.4.6: A core model for centriole duplication? 26

1.5: Thesis aims 28

Main Objectives: 28

Chapter 2: SAS-6 and the centriole cartwheel in *Drosophila* 29

2.0 – Introduction 29

2.0.1: SAS-6 29

2.0.2 Aims 29

2.0.3: Key findings 30

2.1: Producing recombinant DmSAS-6 31

2.1.1 Bioinformatic analysis of SAS-6 31

2.1.2: Attempts to produce full-length recombinant SAS-6 *in vitro* 33

2.1.3: Screening for soluble SAS-6 fragments 36

2.1.4: Large-scale purification of SAS-6^{1-241} 38

2.1.5: EM analysis shows that wild-type SAS-6^{1-241} can form large assemblies 39

2.2: Investigating the SAS-6 N-N interaction. 41

2.2.1: Purification of SAS-6^{1-241}(F143D) 42

2.2.2: SEC-MALS suggests that SAS-6^{1-241} can adopt a wide range of stoichiometries, while the F143D mutant is dimeric. 44

2.2.3: Generation of further CC truncations to test the effect of the CC on dimerisation 46

2.3: Crystallisation and data processing of SAS-6 49

2.3.1: Crystallisation of SAS-6^{1-216}(F143D) 49

2.3.2 Data collection 50

2.3.3: Data processing 51

2.4: Structure solution of SAS-6^{1-216}(F143D) 54

2.4.1 A molecular replacement solution for DmSAS-6 54

2.4.2: Refinement of the SAS-6^{1-216}(F143D) structure 55

2.5: Analysis of the DmSAS-6 structure 59

2.5.1: DmSAS-6 has a pseudo-orthorhombic crystal packing and forms a non-crystallographic spoke dimer 59
2.5.2: The DmSAS-6 N-CC orientation is consistent with orthologues thought to form a "canonical" cartwheel 61
2.6: *In vivo* analysis of the N-N interaction 66
2.6.1: Design of constructs and generation of transgenic lines 66
2.6.2: SAS-6(F143D) exerts a dominant negative sterility phenotype 68
2.6.3: Only GFP-SAS-6(WT) can localise to centrioles in the presence of endogenous SAS-6 in embryos. 70
2.6.4: GFP-SAS-6(F143D) appears to partially rescue the *Sas6*$^{-/-}$ centriole loss phenotype . 72
2.6.5: Perspectives on the SAS-6 *in vivo* experiments 74
2.7 An atomic model of the *Drosophila* SAS-6 cartwheel. 77
2.7.1: Oligomerisation of the central hub 77
2.7.2: Extension of the cartwheel spoke 80
2.7.3: Conservation analysis of SAS-6 in the context of the cartwheel model. 82
2.8: The conserved section of the SAS-6 coiled-coil 84
2.8.1: Design of SAS-6 CC fragments 84
2.8.2: Small-scale purification of CC fragments 85
2.8.3: Purification of SAS-6$^{313-432}$-GFP 87
2.8.4: Cleavage and purification of DmSAS-6 from GFP 89
2.8.5: SAS-6$^{313-432}$ is alpha helical but may be monomeric 90
2.9: Perspectives 92

Chapter 3: The interaction between Ana2 and SAS-4 97
3.0: Introduction 97
3.0.1: Evidence for an Ana2:SAS-4 interaction 97
3.0.2: Aims 98
3.0.3: Key findings 98
3.1: Bioinformatic analysis of SAS-4 and Ana2 99
3.1.1: Bioinformatic analysis of CPAP/SAS-4 99
3.1.2: Bioinformatic analysis of STIL/Ana2 101
3.2: Expression and analysis of soluble SAS-4 TCP domain constructs 103
3.2.1: Design of TCP domain constructs 103
3.2.2: Purification of SAS-4 TCP domain constructs 104
3.2.3: The SAS-4 TCP domain is monomeric at low concentration 108
3.2.4: The DmSAS-4 TCP domain has β-rich secondary structure 109
3.3: Detecting an interaction between the Ana2 CR2 and the SAS-4 TCP domain 110
3.3.1: Far western dot blot analysis 110
3.3.2: Initial ITC analysis of the Ana2:SAS-4 interaction. 112
3.3.3: Attempts to crystallise SAS-4$^{700-901}$ and Ana2$^{5-43}$ 115
3.4: Fusion of the Ana2 CR2 to the TCP domain of SAS-4 116
3.4.1: Design of Ana2 CR2 and SAS-4 TCP domain fusion proteins. 116
3.4.2: Purification of Ana2:SAS-4 fusion proteins. 117
3.4.3: Further purification of Ana2$^{1-47}$-SAS-4$^{700-901}$ using anion Exchange 119
3.5: Crystallisation and data collection of Ana2$^{1-47}$-SAS-4$^{700-901}$ 122
3.5.1: Crystallisation of Ana2$^{1-47}$-SAS-4$^{700-901}$ 122
3.5.2: Data collection and processing of Ana2$^{1-47}$-SAS-4$^{700-901}$ 123
3.5.3: Attempts at molecular replacement 124
3.6: Production of selenomethionine derivative crystals, and data collection 127
3.6.1: Production of selenomethionine-derivative Ana2$^{1-47}$-SAS-4$^{700-901}$ 127
3.6.2: Crystallisation of SeMet-Ana2$^{1-47}$-SAS-4$^{700-901}$ 129

3.6.3: Collection and processing of anomalous data for SeMet-Ana2$^{1-47}$-SAS-4$^{700-901}$ crystals 131
3.7: Solving the structure of Ana2$^{1-47}$-SAS-4$^{700-901}$ using multi crystal anomalous diffraction 136
3.7.1: Preparation of datasets and determination of HA substructure 136
3.7.2: Generation of phases and interpretation of experimental maps 137
3.7.3: Refinement of Ana2$^{1-47}$-SAS-4$^{700-901}$ 140
3.8: Structural Analysis of Ana2$^{1-47}$-SAS-4$^{700-901}$ 142
3.8.1: The SAS-4 TCP domain adopts a unique, extended β-sheet fold 142
3.8.2: The SAS-4 TCP domain acts as a proline recognition domain, and binds a proline-rich region of the Ana2 CR2 143
3.8.3: The SAS-4 : Ana2 interaction is highly conserved 146
3.8.4: The crystal packing of the SAS-4 TCP domain causes fibrils to form 147
3.9: Design of mutations to dissect the Ana2:SAS-4 interaction 151
3.9.1:Initial design of point mutant clusters 151
3.9.2: Design of single point mutants 152
3.10: Attempts to characterise the Ana2:SAS-4 interaction *in vitro* 154
3.11: The SAS-4:Ana2 interaction is essential for centriole duplication *in vivo* 155
3.11.1: Design of transgenic constructs 155
3.11.2: Ana2-GFP constructs can localise to centrioles 157
3.11.3: SAS-4-GFP can localise to centrioles, but localisation is much weaker when the interaction with Ana2 is compromised. 158
3.11.4: The *Sas4* or *Ana2* null phenotypes can only be rescued by constructs that maintain the SAS-4:Ana2 interaction 159
3.12: Perspectives 162

Chapter 4: Structure/function analysis of Ana2 *in vivo* 165
4.0: Introduction 165
4.0.1: Ana2 165
4.0.2: Aims 166
4.0.3: Key findings 166
4.1: Bioinformatic analysis of Ana2 167
4.1.1: Ana2 and STIL contain several conserved features 167
4.2: Designing a structure/function analysis 169
4.2.1: An mRNA injection assay to assess centriolar localisation 169
4.2.2: Construct design, cloning and RNA preparation 170
4.3: Localisation of GFP-Ana2 fusion proteins to the centriole 173
4.3.1 Quantification of construct localisation 173
4.3.2 Full-length GFP-Ana2 and Ana2-GFP localise strongly to centrioles 174
4.3.3 Truncation of Ana2 domains from the C-terminus reveals that the CC and STAN domains are involved in centriolar localisation 175
4.3.4 Truncation of Ana2 domains from the N-terminus confirms that the CC and STAN domains are involved in centriolar localisation 177
4.3.5 Deletion of the CC and STAN domains from Ana2 confirms that they are responsible for localisation of Ana2 to the centriole 179
4.3.6: Fusion proteins containing only the CC domain do not localise to centrioles 180
4.3.7: Point mutations of the Ana2 CC domain suggest that self-association of Ana2 is required for robust centriole localisation 182
4.4: Planning of experiments using Transgenic *Drosophila* 185
4.4.1: Construct design 185
4.4.2: Planned experiments 186
4.5: Perspectives 187

Chapter 5: *In vitro* characterisation of the Ana2 coiled-coil domain 189
5.0: Introduction 189
5.0.1: Insights from *in vivo* analyses 189
5.0.2: Aims 189
5.0.3: Key findings 189
5.1: Characterisation of the STAN domain 191
5.1.1: Purification and characterisation of Ana2$^{236-420}$, containing the STAN + CR4 domains 191
5.2: Analysis of the central coiled-coil domain of Ana2 192
5.2.1: Design of a recombinant Ana2 coiled-coil construct 192
5.2.2: Expression and Purification of diLipoyl-Ana2_CC 193
5.2.3: DiLipoyl-Ana2_CC is tetrameric in solution 194
5.2.4: Cleavage and Purification of Ana2_CC from the Lipoyl domains 195
5.2.5: Isolated Ana2_CC forms a tight tetramer in solution 197
5.2.6: Ana2_CC is alpha-helical 198
5.2.7: Mutations within the CC domain modulate stoichiometry 200
5.2.8: Investigation into homologous systems 201
5.3: Ana2_CC Crystallisation and Data Collection 204
5.3.1: Crystallisation of Ana2_CC 204
5.3.2: Data Collection and Processing of Ana2_CC 206
5.4: Solving The Crystal Structure of the Ana2 Central Coiled Coil Domain 208
5.4.1: A Molecular Replacement Solution for the Ana2_CC structure 208
5.4.2 Refinement of Ana2_CC 209
5.5: Analysis of the Ana2_CC structure 213
5.5.1: Ana2_CC is a parallel, symmetrical tetramer 213
5.5.2: Ana2_CC tetramerisation is driven by polar, charged, and hydrophobic interactions 215
5.5.3: An analysis of CC conservation 217
5.5.4: Rationalisation of Ana2_CC mutant constructs 219
5.5.5: Comparison of the Ana2_CC to other tetrameric coiled-coils 221
5.6 Perspectives 223

Chapter 6: Discussion 227
6.1: Summary 227
6.2: The temporal assembly pathway 227
6.3: Spatial organisation of centriole proteins 230
6.3.1: The Ana2 CC tetramer as a hub separator 232
6.3.2: The Ana2 CC tetramer as a radial spoke crosslinker 234
6.3.3: The Ana2 CC tetramer as a spoke-converger. 236
6.4: Conclusion 237

7.0: Appendix 239
7.1: Standard Laboratory Procedures 239
7.1.1: Primer design, PCR amplification and ligation cloning 239
7.1.2: Gateway (attBP/attLR) Cloning 240
7.1.3: Bacterial Transformation 241
7.1.4: Standard procedure for purifying soluble recombinant protein from the *E. coli* cytoplasm, and Ni-NTA purification details for all constructs. 241
7.1.5: High-Throughput Ni-NTA affinity purification 244
7.1.6: In vitro transcription, and purification of recombinant mRNA 245
7.1.7: Multi Angle Light Scattering (MALS) 246
7.2: X-ray crystallography practice 247

7.2.1: Data Collection 247
7.2.2: Data Processing 247
7.2.3: The phase problem 250
7.2.4: Phase Solution via Molecular Replacement (MR) 251
7.2.5: Single or Multi-wavelength Anomalous Diffraction (SAD/MAD) 255
7.2.6: Refinement 258
7.3: MSA of SAS-4 orthologue TCP domains 260
7.4: Purification of the Ana2 N-terminus + Coiled coil (Ana2$^{1-228}$) 262
7.5: Attempts to characterise the Ana2:SAS-4 interaction *in vitro* 264
7.5.1: Generation of MBP-fusion proteins 264
7.5.2: Production of non-tagged mutant TCP domains for circular dichroism analysis 267
7.5.3: Surface Plasmon Resonance (SPR) experiments. 270
7.6: Localisation of GFP-Ana2$^{1-229}$ and Ana2$^{230-420}$-GFP in RNA injected embryos 275
7.7: Characterisation of the Ana2 C-terminal domains 276
7.7.1: Purification of Ana2$^{236-420}$ containing the STAN + CR4 276
7.7.2: Circular Dichroism analysis of Ana2$^{236-420}$ 278
7.8: Trial MST experiment between the Ana2 coiled-coil and SAS-6$^{313-432}$ 280
7.9: Optimisation screen for Ana2_CC crystallisation 283
7.10: Processing and refinement of Ana2_CC anomalous signal 284

References 287

Chapter 1: Introduction

1.0: Preface

The centriole is a cylindrical supramolecular structure found in ciliated eukaryotes. Centrioles organise two important cellular organelles – the centrosome, and the cilium/flagellum. The centrosome is the main microtubule organising centre (MTOC) in the animal cell, and cilia, both sensory and motile, are absolutely essential for cellular signalling and motility. Unsurprisingly, aberrations in centriole assembly can adversely affect many cellular processes, such as mitotic spindle assembly. While centrosomes were first described more than a century ago, it was only with the advent of electron microscopy that their fascinating superstructure was highlighted – a nine-fold symmetric array of microtubule triplet blades, radiating out from the tips of a central cartwheel. It is only in the past decade or so that we have begun to understand how these beautifully complex structures are built. While various proteomic studies have shown that hundreds of proteins can be associated with centrioles, genetic studies have demonstrated that, surprisingly, only a handful of these proteins are essential for centriole biogenesis. These essential building blocks are now the targets of intensive research aimed at characterising the molecular, and even atomic, detail of centriole assembly.

1.1: Centriole superstructure

The centriole is renowned for its distinctive superstructure. At the simplest level the centriole can be described as a cylinder surrounded by blades of microtubules (MTs) (**Figure 1.1.i**). Dependent on organism and cell type, centriolar dimensions as observed in electron micrographs are in the region of 100-250 nm wide, and 150-500 nm long [Azimzadeh and Marshall, 2010; Gönczy, 2012; Winey and O'Toole, 2014]. Centrioles have a semi-conservative replication mechanism whereby a pre-existing mature "mother" centriole nucleates a new immature "daughter" orthogonally against its proximal wall. This immature centriole is commonly referred to as a "procentriole". Centrioles duplicate once, and only once per cell cycle, and, once formed, the mother centrioles persist throughout all subsequent cellular generations. The process of centriole duplication begins with the appearance of a central cartwheel, at the proximal end of the mother centriole [Allen, 1969; Anderson and Brenner, 1971; Cavalier-Smith, 1974; Gibbons and Grimstone, 1960; Guichard et al., 2010] (**Figure 1.1.i**). This cartwheel consists of a central hub with a fairly uniform diameter of ~22 nm, from which radiate nine spokes, to which the MT blades are attached (**Figure 1.1.i, + 1.1.ii**).

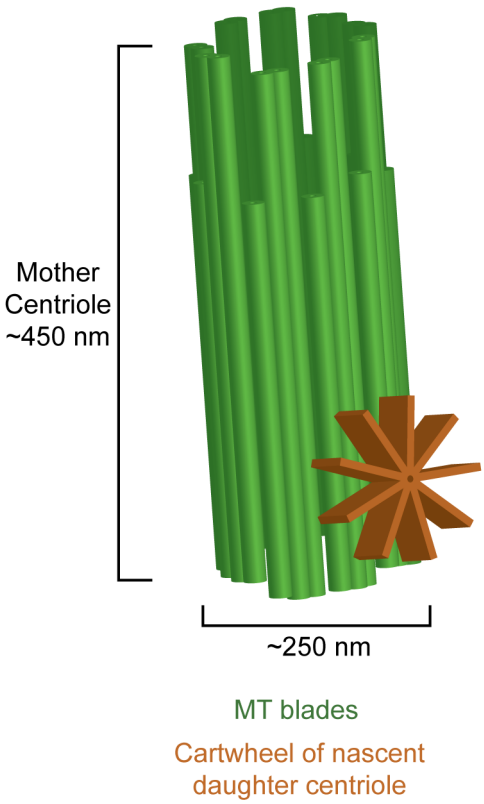


Figure 1.1.i: The centriole is a nine-fold symmetric barrel of microtubule blades. A mature "mother" centriole templates its own duplication, beginning with the nucleation of a cartwheel structure on its surface.

The cartwheel is very prominent in EM images of the well-defined basal bodies of various single celled organisms, such as *Chlamydomonas*, or *Tetrahymena* species, and so studies on these species dominated the characterisation of centriolar assembly early on. Centriolar study is now more widely influenced by the characterisation of centrioles from several model organisms, mainly; human, *Chlamydomonas reinhardtii*, *Drosophila melanogaster*, and *Caenorhabditis elegans*, which have subtly different superstructures. It is vital that these organisms' centrioles are well characterised so that insights from studies on a particular species can be evaluated in comparison to others. It is of particular interest to identify universal assembly mechanisms that define centriole biogenesis as a whole, as these conserved mechanisms are also likely to be the most functionally important. Human centrioles are of particular interest as they are most relevant for understanding the links between centriole dysfunction and human pathology; however, they are also more complicated than in some simpler model organisms. Superstructurally, human centrioles are similar to the well-defined centrioles of *C. reinhardtii*, and other such organisms characterised by EM study, and so I will begin by describing the general structure of these "canonical" centrioles, before explaining the differences in other model organisms. Many of these features have been reviewed in depth [Azimzadeh and Marshall, 2010; Bettencourt-Dias and Glover, 2007; Brito et al., 2012; Gönczy, 2012; Jana et al., 2014; Nigg and Stearns, 2011; Strnad and Gönczy, 2008].

Canonical centrioles contain MT triplet blades organised around a central cartwheel. The cartwheel itself comprises several separated stacks (when viewed in longitudinal section) (**Figure 1.1.ii, B**). Each central hub is separated by ~8 nm. An electron dense feature known as the "pinhead" then joins the cartwheel to the MT triplet blade. In vertebrate centrioles, the cartwheel is degraded shortly after the centriole has matured [Alvey, 1986]. The MT blade has recently been well characterised by structural cryo-electron tomography [Guichard et al., 2013; Li et al., 2012].

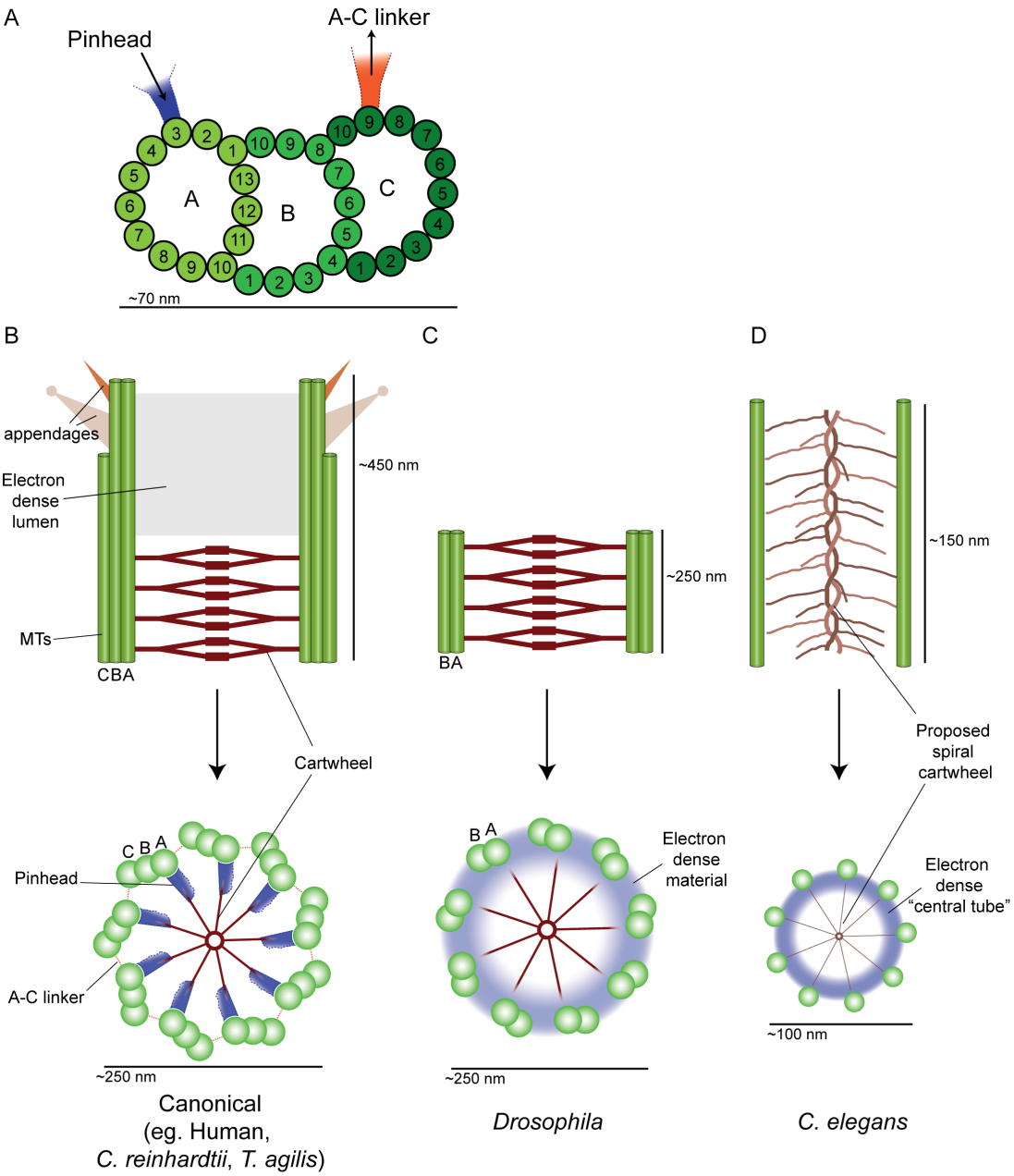


Figure 1.1.ii: Centriole superstructure. (**A**) A schematic of protofilament arrangement in the MT triplet blade according to EM analysis by Guichard et al., 2013, viewed along an MT blade, proximal to distal. **(B)** Schematics of the "canonical" centriole, as possessed by vertebrates and several unicellular ciliates. Above, viewed in longitudinal section, and below, proximal to distal. Note that the central cartwheel, viewed longitudinally, is displayed with paired spokes according to EM analysis of *T. agilis* (Guichard et al., 2013). Note also that the features in these schematics are not drawn to scale, although approximate overall length and width dimensions are indicated for each centriole. (**C**) Schematics of the *Drosophila* somatic centriole, which also contains a cartwheel, but has only MT doublet blades. The cartwheel is portrayed similarly to in **B**. (**D**) Schematics of the *C. elegans* centriole, which has microtubule singlets. Although a cartwheel is not visible by EM, a recent study has proposed a central nine-fold, double spiral, which is depicted in light and dark brown.

The MT triplet blade comprises a complete "A" MT, which is the most canonical MT with 13 protofilaments. The B and C MTs are incomplete, with only 10 protofilaments, and partially share protofilaments with the walls of the A and then B MTs respectively (**Figure 1.1.ii, A**). However, whether the B or C MTs have an 11^{th} protofilament is not obvious [Linck and Stephens, 2007]. The centriolar MTs are particularly stable, being resistant to various MT depolymerising agents, and containing post-translational modifications associated with stable MTs [Bobinnec et al., 1998]. Each protofilament consists primarily of αβ tubulin dimers. The wall of the centriole is completed by an "A-C" linker arm, which joins the C MT of one triplet, to the A MT of the adjacent triplet.

Several features break the symmetry of the centriolar structure. First, the MT blades slant in an anticlockwise fashion when viewed proximal to distal (as has become the conventional view) (**Fig 1.1.ii, B**). Second, the cartwheel is located at the proximal end, and only protrudes part way up the centriolar lumen. It is not known for certain how many cartwheel stacks exist, although the length of the cartwheel can vary dramatically in different organisms or tissue types [Hirono, 2014]. Third, the C MT terminates before the A and B MTs, leaving only an AB doublet at the distal end. Fourth, upon maturation, the centriole acquires distal and sub-distal appendages, that appear to be important for plasma membrane anchoring, required during ciliogenesis. Finally, recent cryoEM studies in the protist, *Trichonympha agilis* have shown that the cartwheel itself breaks symmetry via the pairing of its spokes in ordered doublet layers (**Figure 1.1.ii, B**) [Guichard et al., 2012, 2013]. This feature will be described in more detail later in this thesis.

Drosophila centrioles share many of the key features of canonical centrioles, containing MT blades nucleated by a central cartwheel. However, they differ in several ways (**Figure 1.1.ii, C**). In somatic cells, the centrioles are shorter, and the cartwheel extends the full length of the

centriole. The MT blade comprises an AB doublet, as opposed to the ABC triplet of canonical centrioles (although *Drosophila* spermatocyte centrioles have ABC triplet blades). Additionally, the cartwheel is not ultimately removed in mature mother centrioles as it is for human centrioles, but persists throughout the lifetime of the centriole; see [Sharp et al., 1999] for some especially illustrative electron micrographs. Although discrete pinheads are not obviously evident as they are for *C. reinhardtii* centrioles, the space around the MT doublets is filled with an electron dense material (personal observations).

C. elegans centrioles are widely studied but their structure appears to have significantly diverged from that of canonical centrioles. They contain only singlet MTs, and no cartwheel hub or spokes can be detected by EM; instead *C. elegans* centrioles have been described as containing a "central tube". This was the first discernable structure of the procentriole, evident as an electron-dense ring [Pelletier et al., 2006]. A recent study has provided evidence that a cartwheel of sorts might well exist, but rather than flat stacked rings, it would comprise a tall spiral with a 4.5-fold periodicity, or a pair of such spirals, intertwined [Hilbert et al., 2013] (**Fig 1.1.ii, D**).

1.2: Functions of centrioles: centrosomes and cilia

Analysis of the evolution of centrioles has suggested that the last eukaryotic common ancestor possessed a centriole [Carvalho-Santos et al., 2010; Hodges et al., 2010]. It has been suggested that the primary function of centrioles was to template cilia/flagella, as these were vital for the motility or feeding of unicellular organisms [Carvalho-Santos et al., 2011]. The centrosome may have evolved gradually from the association of centrioles and cell cycle regulated MTOC components as a means to reliably inherit cilia. However, centrosomes are now essential in several cell types for efficient mitotic spindle organisation. I will therefore briefly discuss how centrioles form these two important organelles.

1.2.1: The centrosome and the centriole duplication cycle

The centrosome comprises a pair of orthogonally arranged centrioles, which recruit a matrix of proteins known as the pericentriolar material (PCM) (**Figure 1.2.1.i**). Several PCM proteins are involved in nucleating and stabilising MTs, and the centrosome functions as the main MT organising centre (MTOC) in many animal cells. The most obvious function of the centrosome is in organising the mitotic spindle (**Figure 1.2.1.i**). In certain cell types, such as the early *Drosophila* embryo, centrosomes are essential for mitosis, and embryos arrest very early in development if centriole duplication and centrosome formation are blocked [Stevens et al., 2007; Varmark et al., 2007]. In other cells types, mitosis proceeds without centrosomes, as in mouse embryogenesis [Schuh and Ellenberg, 2007] or in female meiosis in *Drosophila* [Tavosanis et al., 1997]. Surprisingly, several somatic cells that normally have centrosomes appear to be able to undergo relatively normal cell division without centrosomes [Basto et al., 2006; Bazzi and Anderson, 2014]. Centrosomes appear to be particularly important in cells that undergo asymmetric cell divisions that are governed by the correct anchoring and orientation of the spindle relative to cortical cues [Basto et al., 2006; Siller and Doe, 2009]. Distinct from its role as

an MTOC, the centrosome is also thought to function as a coordination centre within the cell, where many proteins are concentrated, and cell processes are regulated. In particular, the centrosome is thought to act as a scaffold for mitotic kinases, and serves a role in recruiting or activating various cell cycle and checkpoint regulators [Doxsey et al., 2005a, 2005b; Ma and Poon, 2011].

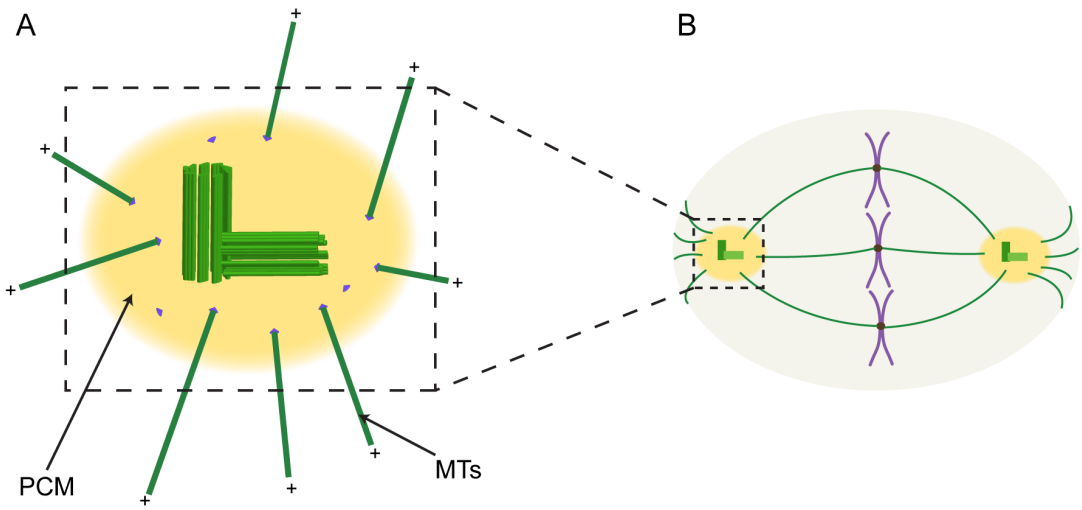


Figure 1.2.1.i: The centrosome. (**A**) A single centrosome, formed from a pair of orthogonally arranged centrioles surrounded by PCM (yellow). Microtubules (green) are nucleated in the PCM, anchored at their minus ends. (**B**) A pair of centrosomes, forming a mitotic spindle.

Centriole and centrosome duplication during the cell cycle has been thoroughly characterised [Bettencourt-Dias and Glover, 2007; Nigg and Stearns, 2011], (**Fig 1.2.1.ii**). After mitosis, each daughter cell inherits one centrosome, containing a mother and daughter centriole. The daughter centriole was synthesised during the previous cell cycle, and when it initially formed, it was orthogonally arranged, tightly against the surface of the mother centriole in an "engaged" configuration. Mother and daughter centrioles disengage at the end of mitosis, and this event is thought to be crucial in "licensing" the centrioles for the next round of duplication [Novak et al., 2014; Tsou et al., 2009]. During G1, the size of the PCM reduces, and the daughter centriole matures – although mothers and daughters remain attached by a fibrous linker connecting their proximal ends [Bornens et al., 1987]. The daughter centriole is now mature enough to act as a mother centriole, and during S-phase, each centriole templates the growth of an immature

procentriole, at their proximal end. This process is initiated by the centriolar kinase, (Sak/PLK4) discussed in **Section 1.4**. As the cell enters G2, the fibrous link between the centriole pairs is lost; the two centrosomes separate, and each starts to nucleate large amounts of PCM, thus forming two separate centrosomes. During mitosis, these centrosomes migrate to either side of the nucleus and ultimately form the poles of the mitotic spindle. Each centrosome is then segregated into a separate daughter cell following cytokinesis. Although I will not discuss this in detail, it is important to note that in certain cell types, (or under certain experimental conditions) centrioles can be assembled *de novo* in the absence of a pre-existing mother template [Loncarek and Khodjakov, 2009; Loncarek et al., 2008; Rodrigues-Martins et al., 2007a].

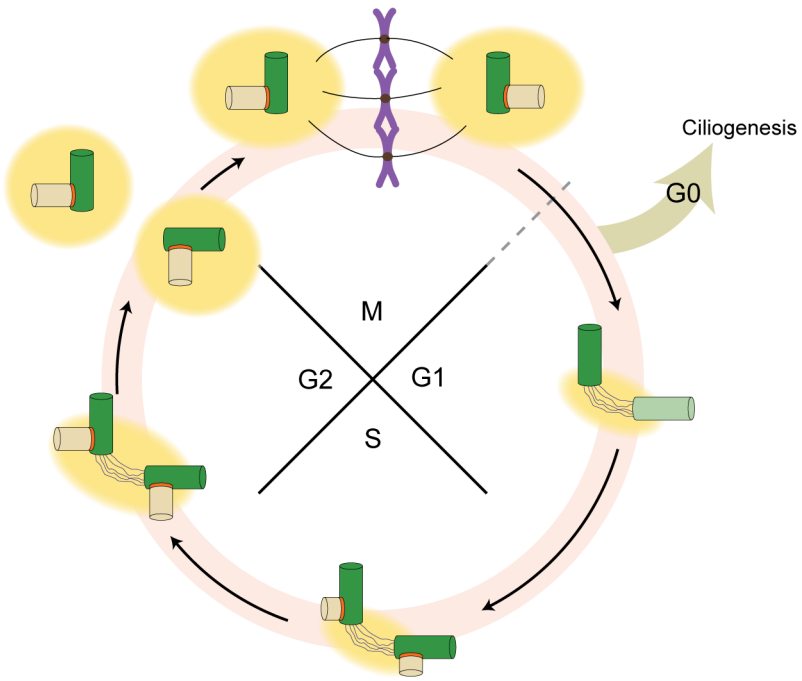


Figure 1.2.1.ii: A schematic explanation of the centriole duplication cycle. In G1, newly divided cells inherit one centrosome, containing a "disengaged" mother centriole (dark green) and daughter centriole (light green), connected by a fibrous linker. During S-phase both centrioles act as mothers, and template the formation of new daughter procentrioles (tan), which elongate during S and G2-phase. At the G2/M transition, the centriole pairs separate and each nucleates a large amount of PCM. There are now two centrosomes, which in mitosis organise the spindle.

1.2.2: Centrioles in cilia

After mitosis, cells can enter quiescence, or G0, and this is when ciliogenesis takes place. The linked mother and daughter centrioles migrate to the plasma membrane, where the mother can dock via its distal appendages. The A and B MTs of the mother centriole then extend to produce the wall of the cilium, contained within the ciliary membrane (**Figure 1.2.2.i**). There are two major classes of cilia; motile and primary. Motile cilia are used for beating/swimming. The axoneme of motile cilia contains two central singlet MTs in addition to the peripheral nine doublets – a so-called "9+2" configuration (**Figure 1.2.2.i, C**). Some cell types in multicellular eukaryotes possess many of these cilia, which beat in a coordinated fashion, often to move liquid. Such cell types line the respiratory epithelium, or the Fallopian tubes. Unicellular "ciliates" contain one or many motile cilia, which are often used for locomotion, and these are often referred to as flagella. Non-motile or primary/sensory cilia are used for sensing and cellular communication, and are defined by a "9+0" organisation of MTs (**Figure 1.2.2.i, B**). Most vertebrate cells contain a single primary cilium which is used for sensing of the extracellular environment, and intercellular signalling [Goetz and Anderson, 2010]. In particular, vertebrate primary cilia seem to be essential for hedgehog (Hh) signalling, one of the main signalling pathways required for development [Corbit et al., 2005; Huangfu et al., 2003]. Interestingly, Hh signalling is essential in Drosophila but this does not seem to require cilia [Roy, 2012]. Modified cilia can also perform specialised functions, such as the opsin-containing outer segment of eye photoreceptor cells [Ramamurthy and Cayouette, 2009].

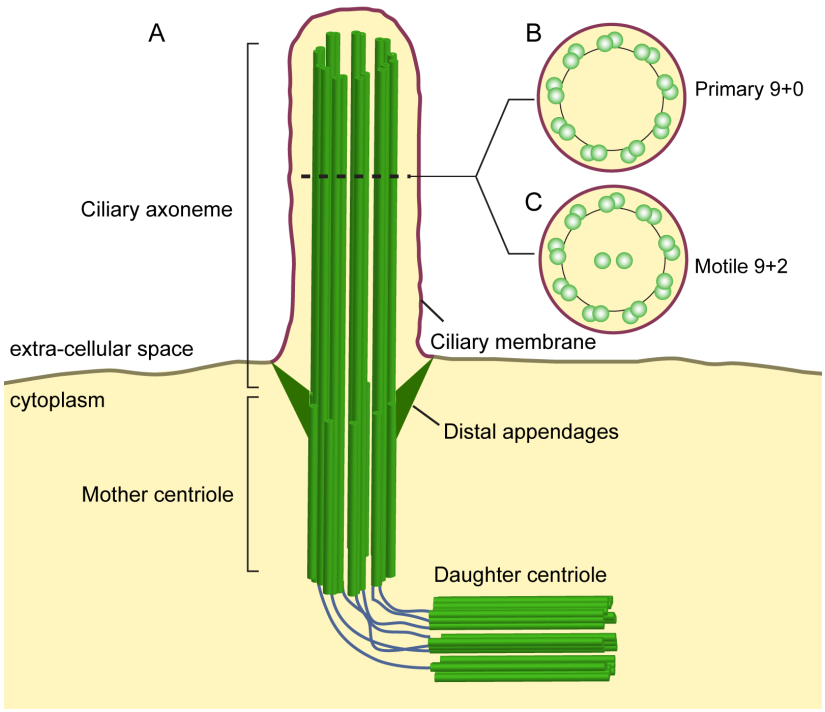


Figure 1.2.2.i: A schematic representation of a cilium. Mother centrioles can dock to the plasma membrane via their distal appendages to become ciliary basal bodies. The A and B microtubules extend from the basal body to produce the ciliary axoneme, which is contained within its own specialised membrane. (**B+C**) Top-down views of sections through different cilia. (**B**) Primary or sensory cilia contain 9 peripheral doublets (9+0 arrangement), whereas (**C**) motile cilia or flagella also contain a central pair of singlet microtubules (9+2 arrangement).

1.3: Centrosomes and cilia in disease

Improper formation of functioning centrosomes and/or cilia can have a multitude of disease consequences, which have been reviewed in depth [Afzelius, 2004; Bettencourt-Dias et al., 2011; Chavali et al., 2014; Goetz and Anderson, 2010; Ibañez-Tallon et al., 2003; Ishikawa and Marshall, 2011; Nigg and Raff, 2009; Nigg et al., 2014; Valente et al., 2014]. Historically, centrosome dysfunction has been linked to cancer. In 1914, Boveri first suggested that supernumerary centrosomes might account for multipolar spindle formation, and resultant chromosome mis-segregation in malignant tumours [Boveri, 2008]. Several studies have since shown that tumour cells often contain multiple centrosomes [D'Assoro et al., 2002; Pihan et al., 1998, 2003]; and some suggest that because centrosome amplification is apparent in early tumour tissue, it may be an underlying cause [Lingle et al., 2002]. In *Drosophila,* centrosome amplification appears to be sufficient to predispose cells to cancer, although the exact mechanisms remain unclear [Basto et al., 2008]. Perhaps the most common theory is that supernumerary centrosomes can cause multipolar spindles to form. This may increase the risk of merotelic attachment, where a single kinetochore becomes attached to multiple spindle poles (or opposite poles if the centrosomes cluster to form a bipolar spindle). Upon anaphase, this could lead to chromosome breakage, and so to aneuploidy, and genetic instability [Ganem et al., 2009]. Alternatively, it has been suggested that extra centrosomes can interfere with the asymmetric division and polarity of certain stem cells [Basto et al., 2008].

Another broad set of diseases now termed "ciliopathies" are linked to the malfunction of cilia. Such diseases include Bardet-Biedl, Meckel-Gruber and Joubert Syndromes, which are associated with a wide range of phenotypes including polydactyly, obesity and macular degeneration – almost certainly reflecting the large number of functions with which cilia have been associated [Bettencourt-Dias et al., 2011; Nigg and Raff, 2009].

1.3.1: Autosomal Recessive Primary Microcephaly (MCPH)

The developmental disorder MCPH has a particularly interesting link to centriole biogenesis. Patients with this hereditary disease have a structurally normal brain that is significantly smaller than average. As skull size is determined by the size of the brain, patients are usually diagnosed based on having a head circumference of more than 3 standard deviations below the mean for their age and sex [Woods et al., 2005]; however this symptom is usually accompanied by only mild to moderate mental retardation [Thornton and Woods, 2009]. Various genetic studies have been carried out on families of MCPH patients to identify the gene loci responsible [Awad et al., 2013; Bilgüvar et al., 2010; Bond et al., 2002, 2005; Genin et al., 2012; Guernsey et al., 2010; Gul et al., 2006; Hussain et al., 2012, 2013; Khan et al., 2014; Kumar et al., 2009; Nicholas et al., 2010; O'Driscoll et al., 2006; Sir et al., 2011; Yang et al., 2012; Yu et al., 2010]. To date, 14 loci have now been mapped to specific protein-encoding genes (**Table 1.3.1.i**). Strikingly, many of these proteins are essential or important centriole biogenesis components, suggesting that MCPH may arise from the malfunction of a centriole-dependent pathway. The essential centriole assembly components will be introduced in greater depth in the next section.

MCPH protein	Function
ZNF335	Neural differentiation regulator
Microcephalin/BRIT1	DNA damage repair
PHC1	Chromatin remodelling
CDK6	Cell cycle restriction point regulator
WDR62	Spindle pole associated
ASPM	Spindle pole associated/DNA damage repair
CASC5	Kinetochore associated
CDK5RAP2	PCM component
CEP63	Centrosomal/centriole protein
CPAP/CENPJ	Essential centriole protein
STIL	Essential centriole protein
CEP135	Essential centriole protein
CEP152	Essential centriole protein
SAS-6	Essential centriole protein

Table 1.3.1.i: The functions of proteins found to be mutated in Autosomal Recessive Primary Microcephaly

Some MCPH proteins are also associated with the PCM, spindle pole or kinetochore, leading to the suggestion that defects in centriole assembly or in mitosis may perturb the asymmetric divisions of neural progenitors that are thought to be essential for proper neurogenesis in vertebrates. [Megraw et al., 2011; Sun and Hevner, 2014; Wang et al., 2009]. However, some MCPH proteins are specifically involved in DNA damage response (DDR), or cell cycle check-points, suggesting that the DDR pathway may be defective in MCPH patients. Interestingly, defects in MCPH proteins have also been shown to cause phenotypes consistent with Seckel Syndrome – a pathology also associated with DDR defects [Al-Dosari et al., 2010; Kalay et al., 2011; McIntyre et al., 2012; Megraw et al., 2011]. Indeed the centrosome, spindle assembly, and the DNA damage response appear to be intricately linked [Barr et al., 2010; Shimada and Komatsu, 2009], and DDR and MCPH have also been linked to centrosome amplification [Bourke et al., 2007; Marthiens et al., 2013], as has a particular MCPH mutation of STIL [Arquint and Nigg, 2014]. Although the underlying causes of MCPH remain unclear, the striking genetic links between this condition and centriole biogenesis strongly suggest that a root cause is a failure to correctly assemble a functioning centriole.

1.4 Centriole biogenesis

Until approximately a decade ago, although the ultrastructure of centrioles was characterised in detail, only very few proteins that were associated with centrioles (such as centrins and some tubulin isoforms) had been identified [Beisson and Wright, 2003]. This list was significantly expanded via proteomic analyses. Mass Spectrometry analyses of the human centrosome, [Andersen et al., 2003] the basal bodies of *C. reinhardtii*, [Keller et al., 2005] and *Tetrahymena* [Kilburn et al., 2007], discovered hundreds of centriole-associated proteins. Comparative genomics studies also identified compatible sets of proteins [Li et al., 2004]. This "centrosomic" information is now well catalogued in an online database [Alves-Cruzeiro et al., 2013]. However, genome-wide, and targeted genetic analyses have refined this list down to only a small subset of proteins that are absolutely essential for centriole biogenesis.

Genetic manipulation of *C. elegans* was important in first defining this list. Genome-wide RNAi screens were used to deplete individual proteins, and assess whether they had a very characteristic phenotype that was associated with loss of centrioles – a mitotic arrest with monopolar spindles in the two cells produced after the first embryonic division [Dammermann et al., 2004; Delattre et al., 2004; Kemp et al., 2004; Kirkham et al., 2003; Leidel and Gönczy, 2003; O'Connell et al., 2001; Pelletier et al., 2004]. Despite the divergent morphology of *C. elegans* centrioles, this core of proteins appears to be conserved in other organisms, as exemplified by similar RNAi approaches in *Drosophila* [Dobbelaere et al., 2008], and human cells [Balestra et al., 2013] (although several of these *Drosophila* or human homologues were targeted individually before these screens). I will now introduce each of the major protein families that appear to be essential for centriole biogenesis.

1.4.1 – The centriolar kinase Sak/PLK4/ZYG-1

Studies have shown that centriole biogenesis appears to be initiated by the action of a centriolar serine/threonine protein kinase. A study in worms first identified ZYG-1 (zygotic defective: embryonic lethal 1) as being essential for centriole duplication [O'Connell et al., 2001]. The phosphorylation state of core centriole proteins has since been proposed as a mechanism for their recruitment to the centriole [Kitagawa et al., 2009, 2011a]. Although ZYG-1 is directly implicated in the recruitment of downstream centriole proteins, the issue of specific phosphorylations remains controversial [Kitagawa et al., 2009; Lettman et al., 2013]. A functionally homologous, although not strictly orthologous kinase, PLK4, (polo-like Kinase 4, or Sak in *Drosophila*) is also essential for centriole biogenesis in humans [Habedanck et al., 2005] and in *Drosophila* [Bettencourt-Dias et al., 2005].

PLK4 is defined by a conserved N-terminal Polo kinase domain, followed by 3 "polo box" domains. Two central tandem boxes are termed the "cryptic polo box" (CPB), and these are followed by a third, C-terminal polo box. Recently the structure of the Sak CPB was solved, and was shown to be a dimer, implicated in direct binding to another essential centriole protein Asl/CEP152 (described in **Section 1.4.5**) [Slevin et al., 2012]. The structure of the PLK4 CPB has since been re-determined to define a different interface [Park et al., 2014; Shimanovskaya et al., 2014], which, additionally, was also observed in the CPB structure of *C. elegans* ZYG-1 [Shimanovskaya et al., 2014]. This strongly suggests that while Sak and ZYG-1 may not be strict orthologues, they may share structurally similar mechanisms of action [Shimanovskaya et al., 2014].

PLK4 appears to localise to the wall of the mother centriole, where it is necessary for the nucleation of a procentriole [Kleylein-Sohn et al., 2007; Sonnen et al., 2012]. Whereas centrioles

usually duplicate once, and only once, per cell cycle, ectopic overexpression of Sak or PLK4 can drive centriole amplification – where a mother centriole nucleates multiple daughters [Kleylein-Sohn et al., 2007; Peel et al., 2007]. Sak-dependent centriole amplification can also occur *de novo*, without the need for a templating mother centriole [Rodrigues-Martins et al., 2007a]. These results suggest that the action or levels of Sak/PLK4 are usually limiting, but moreover Sak/PLK4 is one of the most upstream regulators of centriole biogenesis [Kleylein-Sohn et al., 2007]. Similarly, ZYG-1 has been shown to be the most upstream component that initiates centriole biogenesis in *C. elegans* [Delattre et al., 2006; Pelletier et al., 2006].

The function of the kinase activity of PLK4 in centriole biogenesis remains poorly understood. In *C. elegans* it is clear that ZYG-1 can recruit the downstream component SAS-6 (see **Section 1.4.2**) to centrioles independently of its kinase activity, but that kinase activity is essential for the proper incorporation of SAS-6 into centrioles and for centriole biogenesis [Lettman et al., 2013]. While an earlier report suggested that SAS-6 is an essential and direct target of ZYG-1 [Kitagawa et al., 2009] this has since been disproved [Lettman et al., 2013]. *In vitro*, PLK4 has been shown to phosphorylate CEP152 [Hatch et al., 2010] and CPAP [Chang et al., 2010]. Moreover, the kinase function of PLK4 was shown to be necessary for centriole amplification in human cells, as shown by a lack of amplification using “kinase-dead” constructs [Habedanck et al., 2005]. Nevertheless, the crucial downstream targets of Sak/PLK4/ZYG-1 kinase remain to be identified. In contrast, it is now well-established that PLK-4 regulates its own degradation by autophosphorylation, and that this mechanism is required to keep the levels of the protein low in order to suppress centriole overduplication [Cunha-Ferreira et al., 2013; Holland et al., 2010; Sillibourne et al., 2010].

1.4.2 – SAS-6

SAS-6 (Spindle Assembly Abnormal 6) was first discovered in *C. elegans* via RNAi-based screens, and was shown to be essential for centriole duplication [Dammermann et al., 2004; Leidel et al., 2005]. SAS-6 appeared to contain a conserved N-terminal region known as the "Present In SAS-6" (PISA) domain. Combined with a long, downstream section of coiled-coil, this allowed SAS-6 homologues to be identified in other species, where it was also found to be essential for centriole biogenesis [Kleylein-Sohn et al., 2007; Nakazawa et al., 2007; Peel et al., 2007; Rodrigues-Martins et al., 2007a; Yabe et al., 2007]. SAS-6 has also recently been identified as an MCPH protein [Khan et al., 2014]. Similarly to PLK4, ectopic overexpression of SAS-6 has been shown to drive the formation of supernumerary centrioles [Leidel et al., 2005; Strnad et al., 2007] or centriole-like structures that can recruit downstream centriole components [Peel et al., 2007; Rodrigues-Martins et al., 2007b]. However it has subsequently been shown that such centriole-like structures do not contain cartwheels without the overexpression of other factors [Stevens et al., 2010a]. This also supports evidence that SAS-6 is one of the first components recruited to the centriole [Delattre et al., 2006; Pelletier et al., 2006].

Pioneering studies in *C. reinhardtii* suggested that SAS-6 might be implicated specifically at the cartwheel, and is involved in templating the nine-fold symmetry of the centriole [Nakazawa et al., 2007]. A SAS-6 null mutant lacking flagella was studied using EM. The centrioles of these cells lacked a central cartwheel, and some had symmetries of 7-11 as opposed to 9. Immuno-gold EM showed that SAS-6 appeared to localise specifically to the central hub and spokes of the cartwheel. Similar localisation of SAS-6 in *Drosophila* [Gopalakrishnan et al., 2010], human [Kleylein-Sohn et al., 2007] and *Tetrahymena* [Culver et al., 2009] cartwheels has been observed.

During the course of my studies, the crystal structures of domains of several SAS-6 orthologues were solved, revealing that SAS-6 homo-dimerises via two interfaces. These interfaces enable SAS-6 to oligomerise into higher-order structures that likely form the cartwheel central hub and associated spokes [van Breugel et al., 2011, 2014; Kitagawa et al., 2011b]. However, purified SAS-6 can take on a variety of stoichiometries, and the majority of centrioles seen in SAS-6 null *C. reinhardtii* cells retain nine-fold symmetry [Nakazawa et al., 2007], suggesting that other symmetry-ensuring factors are likely to enforce nine-fold symmetry *in vivo* [Cottee et al., 2011]. Prior to these publications, I was working on solving the structure of *Drosophila* SAS-6, and so I will discuss the significance of these structural studies in greater detail, within the context of my own results.

In human cells, SAS-6 levels are regulated by protein degradation at the onset of anaphase [Puklowski et al., 2011], dependent on the presence of a KEN box degradation signal at the C-terminus of SAS-6 [Strnad et al., 2007]. This is consistent with the fact that the human cartwheel also disappears from centrioles at about this time. Interestingly, *Drosophila* SAS-6 contains no such degradation signal (personal observations) and the cartwheel does not disappear in flies [Loncarek and Khodjakov, 2009].

As discussed above, in *C. elegans*, a direct interaction between ZYG-1 and the SAS-6 coiled-coil appears to recruit SAS-6 to the centriole independently of the ZYG-1 kinase activity. The kinase activity is required for proper incorporation of SAS-6 into the centriole, but the phosphorylation target of ZYG-1 does not appear to be SAS-6 [Lettman et al., 2013]. One study has shown that the human SAS-6 CC domain was sufficient for recruitment to centrioles when ectopically expressed in U2OS cells. However, only the full-length protein permitted centriole duplication [Keller et al., 2014]. Addressing the question of SAS-6 localisation, many studies have focused on

SAS-6 interaction partners. It has been shown in *C. elegans* that SAS-6 directly binds to another essential centriole protein, SAS-5 [Hilbert et al., 2013; Leidel et al., 2005; Qiao et al., 2012]. *Drosophila* SAS-6 also interacts directly, but weakly, with the SAS-5 functional homologue, Ana2 [Stevens et al., 2010b; Tang et al., 2011]. However, an equivalent direct interaction has not been found for human SAS-6, despite attempts by several groups [Arquint et al., 2012; Tang et al., 2011; Vulprecht et al., 2012].

1.4.3 – SAS-5/Ana2/STIL

The SAS-5 (*C. elegans*)/Ana2 (*Drosophila*)/STIL (vertebrates) family of proteins are also core centriole components. These proteins have divergent sequence homology and differ significantly in length. However, all 3 proteins share a short, central CC domain, localise to the procentriole early in its biogenesis and are essential centriole assembly components.

SAS-5 (Spindle Assembly Abnormal 5) was originally identified in genetic screens where it was shown to be essential for centriole biogenesis [Dammermann et al., 2004; Delattre et al., 2004] and to be recruited downstream of ZYG-1 [Delattre et al., 2004, 2006; Pelletier et al., 2006]. SAS-5 and SAS-6 appear to be co-dependent for their localisation to the procentriole [Delattre et al., 2006; Pelletier et al., 2006]. SAS-5 was shown to robustly interact with SAS-6 in a Yeast 2 hybrid (Y2H) experiment [Leidel et al., 2005], possibly rationalising this co-dependence. It has been proposed that a phosphorylation mechanism may regulate the co-recruitment of SAS-6 and SAS-5 [Kitagawa et al., 2011a]. More recently, the SAS-5:SAS-6 interaction was dissected using mutagenesis, and mapped to a 15 residue, C-terminal fragment of SAS-5, and a central section of the SAS-6 coiled-coil [Hilbert et al., 2013; Qiao et al., 2012].

Homologues of SAS-5 remained unidentified for several years due to low sequence conservation. It was eventually shown that SAS-5 was the likely functional homologue of *Drosophila* Ana2 and human STIL, based on their functions, similar topology and shared limited sequence homology in a region known as the **ST**IL **An**a2 (STAN) domain [Stevens et al., 2010b]. Ana2 (Anastral Spindle 2) was first discovered in an RNAi-based screen to find proteins with roles in spindle formation [Goshima et al., 2007]. It was more specifically identified as an essential centriole component in a separate screen [Dobbelaere et al., 2008], before being shown to localise specifically to the centriole, and to weakly interact with SAS-6 in a Y2H assay [Stevens et al., 2010b]. In particular, Ana2 was incorporated into nascent daughter centrioles, and importantly, was also shown to drive the formation of centriole-like structures - this effect being intensified when co-overexpressed with SAS-6 [Stevens et al., 2010b]. Strikingly, when analysed by EM, particles induced by Ana2 and SAS-6 co-overexpression – known as **S**AS-6/**A**na2 **P**articles (SAPs) – contained structures reminiscent of the centriolar cartwheel [Stevens et al., 2010a]. Thus Ana2 may be important in the regulation of a SAS-6 based cartwheel. Ana2 null mutant *Drosophila* fail in centriole biogenesis, and it also appears that Ana2 may have a further role in spindle orientation that is dependent on an interaction with the Dynein Light Chain, LC8 or Ctp [Wang et al., 2011]. Recently this interaction has been confirmed by crystallographic analysis [Slevin et al., 2014], however it is not known whether this interaction contributes to centriole biogenesis.

STIL (SCL/TAL1 interrupting locus), was first identified as a gene associated with Leukaemia [Aplan et al., 1991], although it was later implicated in mitotic spindle formation [Pfaff et al., 2007]. After Ana2 was linked to centriole biogenesis in flies, further studies confirmed that STIL was essential for centriole biogenesis in human cells [Arquint et al., 2012; Tang et al., 2011; Vulprecht et al., 2012]. STIL specifically localised to the inner centriole lumen, although no interaction with SAS-6 was detected. [Arquint et al., 2012; Sonnen et al., 2012; Tang et al., 2011;

Vulprecht et al., 2012]. Similarly to Ana2, STIL overexpression can drive centriole amplification [Arquint and Nigg, 2014; Arquint et al., 2012; Tang et al., 2011; Vulprecht et al., 2012], and in humans the timely degradation of STIL is necessary to prevent this amplification [Arquint and Nigg, 2014]. Similar to human SAS-6, this is regulated by the presence of a KEN box degradation signal. STIL is an MCPH protein, and one particular MCPH mutation causes truncation of the KEN box, resulting in increased levels of STIL and subsequent centriole amplification [Arquint and Nigg, 2014]. This result is consistent with reports that link centriole amplification to MCPH [Marthiens et al., 2013].

Whether SAS-6 and STIL are interdependent for their localisation in human centrioles remains controversial. In one study, SAS-6 depletion had no effect on STIL localisation, suggesting that in humans, STIL localises to centrioles upstream of SAS-6 - although these data were not shown [Vulprecht et al., 2012]. However, in another study, RNAi-based SAS-6 depletion perturbed targeting of STIL to the centriole, as well as inhibiting STIL-induced centriole amplification [Arquint et al., 2012]. It is unclear why these results differ.

1.4.4 – SAS-4/CPAP

SAS-4 (Spindle Assembly Abnormal 4) was first discovered in a *C. elegans* RNAi screen where it was seen to be essential for centriole biogenesis [Kirkham et al., 2003; Leidel and Gönczy, 2003]. It appeared to be recruited downstream of SAS-5 and SAS-6, and was required for the addition of MTs to the procentriole [Dammermann et al., 2008; Delattre et al., 2006; Kohlmaier et al., 2009; Pelletier et al., 2006]. SAS-4 was found to have extensive C-terminal homology to a *Drosophila* protein, which was subsequently named SAS-4, and was shown to be essential for centriole duplication in flies [Basto et al., 2006]. Furthermore, this same region was identified in human CPAP (centrosomal P4.1-associated protein) [Hung et al., 2000]. It was also shown that depletion of CPAP did not prevent recruitment of SAS-6, although SAS-6 depletion did prevent CPAP depletion, suggesting that CPAP also acts downstream of SAS-6 in human cells, similar to SAS-4 in other systems [Kohlmaier et al., 2009]. SAS-4/CPAP appears to localise to the inner centriole wall [Kleylein-Sohn et al., 2007]. Recent super-resolution microscopy studies reveal that SAS-4 appears to form a discrete ring around the cartwheel structure in *Drosophila* cells [Mennella et al., 2012].

CPAP overexpression results in the production of overly long centrioles [Kitagawa et al., 2011c; Kohlmaier et al., 2009; Peel et al., 2007; Schmidt et al., 2009; Tang et al., 2009], and CPAP contains two N-terminal motifs that bind tubulin [Hsu et al., 2008; Hung et al., 2004]. One of these motifs stabilises MT polymerisation, while the other has a destabilising effect, so more study will be needed to rationalise these seemingly contradictory functions.

A recent study has shown that CPAP binds directly to STIL, and that this interaction is weakened by a particular CPAP MCPH disease mutation, E1235V [Tang et al., 2011]. Furthermore, CPAP was shown to lie downstream in the recruitment pathway to STIL, suggesting that STIL recruits

CPAP to centrioles. The same study also reported a weak interaction between *Drosophila* SAS-4 and Ana2 via CoIP, suggesting that this interaction may be conserved in the *Drosophila* system. Interestingly another study suggested that the same CPAP MCPH mutation disrupted wild-type spindle positioning, although the mechanisms underlying this remain poorly understood [Kitagawa et al., 2011c]. I examine the interaction between *Drosophila* SAS-4 and Ana2 in **Chapter 3.**

1.4.5 – Other core proteins

Several other proteins are implicated in the core centriole biogenesis machinery, but are perhaps not shared across all systems. I discuss these proteins collectively in this section.

The *Drosophila* protein Asl (Asterless) was originally identified in a genetic screen for male sterility mutants [Bonaccorsi et al., 1998] and was originally thought to be important in centrosome function [Bonaccorsi et al., 1998; Varmark et al., 2007]. A homologue, CEP152 (Centrosomal Protein 152 kDa) has since been identified in humans [Andersen et al., 2003; Blachon et al., 2008], and it is a known MCPH protein [Guernsey et al., 2010]. Asl and CEP152 are essential for centriole duplication in flies [Blachon et al., 2008; Varmark et al., 2007] and humans [Blachon et al., 2008] respectively, although there is no known homologue in *C. elegans*. Asl/CEP152 interacts with SAS-4/CPAP [Cizmecioglu et al., 2010; Dzhindzhev et al., 2010; Hatch et al., 2010], and SAS-4 recruits Asl to centrioles in flies [Dzhindzhev et al., 2010; Novak et al., 2014]. Asl/CEP152 localises to the outer wall of the centriole in a ring around SAS-4 [Mennella et al., 2012; Sir et al., 2011]. Another MCPH protein, CEP63 appears to cooperate with CEP152 to form this ring in humans [Brown et al., 2013; Sir et al., 2011]. Although Asl mediates the recruitment of PCM to the centriole [Bonaccorsi et al., 1998; Conduit et al., 2014; Varmark et al., 2007], it is also responsible for recruiting Sak/PLK4 to the mother centriole to initiate centriole

duplication [Cizmecioglu et al., 2010; Dzhindzhev et al., 2010; Hatch et al., 2010; Novak et al., 2014].

SPD-2 (Spindle Defective 2) was originally identified in *C. elegans* as a PCM protein, but was subsequently shown to recruit ZYG-1 to centrioles [Kemp et al., 2004; Pelletier et al., 2004]. *Drosophila* SPD-2 is important for PCM recruitment, but it is not required for centriole duplication [Conduit et al., 2014; Dix and Raff, 2007; Giansanti et al., 2008]. There is also a SPD-2 homologue in humans, CEP192 (Centrosomal Protein 192 kDa) [Andersen et al., 2003], but its role in centriole duplication is controversial [Gomez-Ferreria et al., 2007; Zhu et al., 2008]. Interestingly, CEP152 and CEP192 appear to cooperate to recruit PLK4 to centrioles in humans [Sonnen et al., 2013]. Thus, Sak/PLK4/ZYG-1 kinases appear to be recruited to mother centrioles by different methods in different organisms: in *C. elegans*, ZYG-1 is recruited exclusively by SPD-2; in *Drosophila*, Sak is recruited exclusively by Asl; in humans, PLK4 is recruited by both CEP152 and CEP192.

The centriolar component CEP135 (Centrosomal protein 135 kDa), was first characterised in CHO cells and shown to be a centrosomal protein [Ohta et al., 2002]. It has since been identified as an MCPH protein [Hussain et al., 2012]. Although absent in *C. elegans*, CEP135 is highly conserved, with orthologues present in many canonical-cartwheel bearing organisms [Carvalho-Santos et al., 2010], including *Drosophila* [Mottier-Pavie and Megraw, 2009]. CEP135 has been studied in detail in *C. reinhardtii*, where it is known as Bld10. Bld10 localises to the pinhead and distal tips of the cartwheel, and truncations or genetic knock-outs were shown to cause detachment of the cartwheel from the inner centriole wall, or indeed absence of the cartwheel [Hiraki et al., 2007; Matsuura et al., 2004]. Similar results in *Paramecium* suggested that CEP135 was essential for cartwheel stability [Jerka-Dziadosz et al., 2010]. Although CEP135 does not appear to be

essential for cartwheel formation in *Drosophila*, the cartwheel is similarly destabilised in its absence [Roque et al., 2012]. These authors also showed that *Drosophila* CEP135 localised similarly to Bld10 from *C. reinhardtii*. These results suggest that CEP135 may be especially important in joining the cartwheel to the MT blades. Consistent with this hypothesis, *Drosophila* CEP135 has been shown to bind MTs [Carvalho-Santos et al., 2012]. Human CEP135 has recently been shown to directly bind both SAS-6 and MTs and to be essential for centriole duplication [Lin et al., 2013]. Lin et al. also found that CEP135 depletion in human cells resulted in centrioles with MT triplet symmetry aberrations, suggesting that, similar to CPAP, CEP135 has downstream effects on MTs. However, SAS-6 localisation was not affected by CEP135 depletion, which suggests that CEP135 is recruited via SAS-6. This study highlights a potential problem in the field about the definition of what makes a centriole protein "essential". Both Lin et al., and Roque et al., observed morphologically perturbed centrioles upon loss of CEP135. One study concluded that CEP135 was "required for centriole assembly", while the other concluded that it was "not essential" for centriole duplication. Perhaps the knowledge that CEP135 is a known MCPH gene, and that its loss of function at the centriole may have severe consequences in human development, is what led to this conclusion in the human study. However, CEP135 is necessary for the formation of morphologically correct centrioles, and both studies concluded similarly: CEP135 joins the inner and outer centriole components.

1.4.6: A core model for centriole duplication?

Together, studies on core duplication proteins have enabled the broad definition of centriole biogenesis pathways in *C. elegans*, *Drosophila* and humans (**Figure 1.4.6.i**). In all cases a centriolar kinase is the most upstream element and is required for the recruitment of all other components. The kinase is initially followed by SAS-6, and SAS-5/Ana2/STIL. These two proteins appear to cooperate to some degree, probably to build the central cartwheel or spiral. This is

then followed by SAS-4/CPAP, possibly via an interaction with Ana2/STIL (maybe also SAS-5). SAS-6, SAS-5/Ana2/STIL and SAS-4 have roughly equivalent spatio-temporal roles in each system and these three proteins therefore appear to define a core structural module for centriole assembly, probably via the formation of a cartwheel structure. The actions of these proteins, or the interplay between them, might underpin universal assembly mechanisms for the centriole that are applicable across many systems.

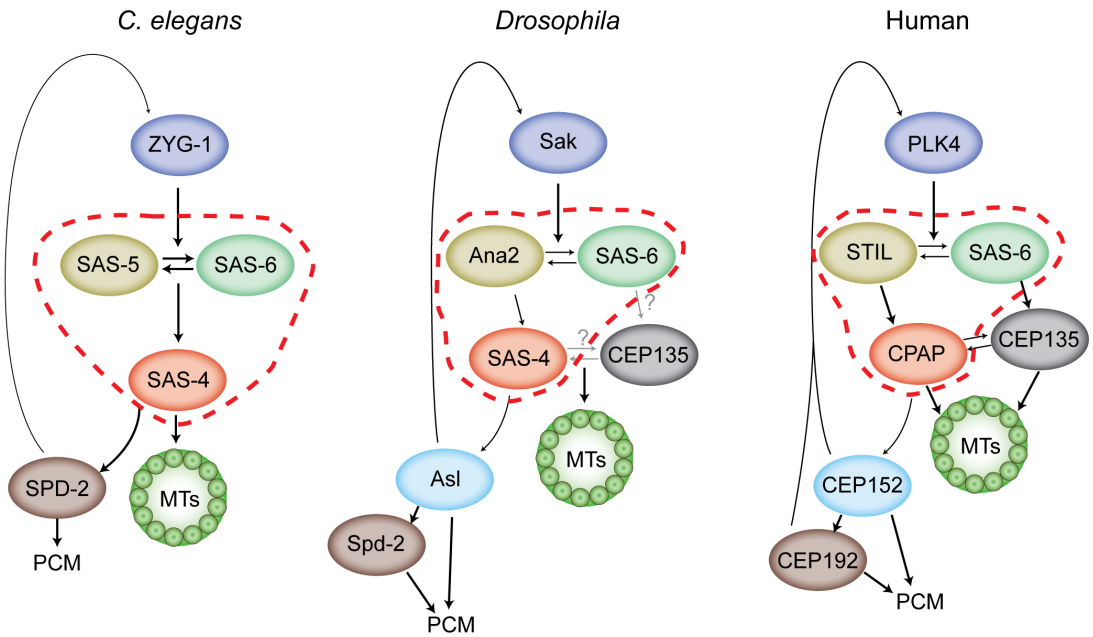


Figure 1.4.6.i: The core centriole assembly pathway in *C. elegans*, *Drosophila* and humans, summarised based on reports in the literature. The arrows do not necessarily represent physical interactions, but rather indicate sequential protein recruitment/dependence. In all cases, a centriolar kinase is the most upstream element. This is followed by a core module (dashed red circle) of SAS-6 and SAS-5/Ana2/STIL followed by SAS-4/CPAP. The mechanisms for PCM recruitment, and also re-recruitment of the centriolar kinase to begin another round of duplication, appear to differ.

1.5: Thesis aims

Prior to beginning this study, no high-resolution structure had been published for any of the essential centriole assembly proteins. I therefore wanted to target core centriole assembly proteins for structural characterisation in order to gain insight into, and possibly define, universal mechanisms for centriole assembly steps, at the molecular level. By targeting the *Drosophila* orthologues of SAS-6, SAS-4 and Ana2, I hoped to explore the functional relevance of my findings in an *in vivo* context.

Main Objectives:

- Identify functional domains within the target proteins, SAS-6, SAS-4 and Ana2
- Recombinantly express and purify fragments of these proteins
- Characterise these fragments and their potential interactions using biophysical techniques
- Obtain high-resolution structures to explain the functions of these proteins
- Explore the biological significance of these domains using *in vivo* techniques

Prefatory note:

The significance of the experimental results presented in this thesis are largely discussed throughout the Results section of each Chapter. This is in order to contextualise the subsequent experimental approaches. I therefore include a "Perspectives" summary, rather than a formal "Discussion" at the end of each chapter, and then a final Discussion Chapter. Supplementary methods and results are contained in the Appendix.

Chapter 2: SAS-6 and the centriole cartwheel in *Drosophila*

2.0 – Introduction

2.0.1: SAS-6

Spindle assembly abnormal 6 (SAS-6) is an essential centriolar protein, comprising a globular N-terminal domain followed by a long coiled-coil domain. Orthologues are present in all organisms possessing a centriole [Carvalho-Santos et al., 2010; Culver et al., 2009; Dammermann et al., 2004; Kleylein-Sohn et al., 2007; Leidel et al., 2005; Nakazawa et al., 2007; Peel et al., 2007].

Evidence suggested that SAS-6 was involved specifically in the building of the cartwheel. It localises to the cartwheel [Culver et al., 2009; Gopalakrishnan et al., 2010; Kleylein-Sohn et al., 2007; Nakazawa et al., 2007], and additionally, its removal, either genetically or via timed degradation, results in an absence of the cartwheel from centrioles [Nakazawa et al., 2007; Strnad et al., 2007]. Often, its genetic removal results in the complete absence of centrioles [Leidel et al., 2005; Peel et al., 2007; Strnad et al., 2007].

2.0.2 Aims

- Identify the functional domains of SAS-6.
- Biophysical and structural characterisation of *Drosophila* SAS-6, in order to provide insight into its role as a cartwheel protein.
- Test the significance of any structural features of SAS-6 *in vivo*.

When I started this project (in October 2010) no structural studies had been published on SAS-6. Subsequently, several structures have been published [van Breugel et al., 2011, 2014; Hilbert et al., 2013; Kitagawa et al., 2011b; Qiao et al., 2012] and, where appropriate, I discuss my work in the context of these structures throughout this Chapter.

2.0.3: Key findings

- *Drosophila* SAS-6 oligomerises via separate "N-N" and "CC" interfaces to form large oligomers.
- The structure of SAS-6 is compatible with it forming flat, nine-fold symmetric rings that are consistent with the central hub of the cartwheel.
- The N-N interface is necessary to ensure robust centriole assembly *in vivo*.

2.1: Producing recombinant DmSAS-6

In order to investigate the structural and biophysical properties of *Drosophila melanogaster* SAS-6 (DmSAS-6), large amounts of homogenous protein, or fragments thereof, would need to be obtained. I decided to use bacterial expression as this has the potential to produce large quantities of protein, and also allows for relatively fast, high-throughput screening of different constructs. In order to design fragments which were likely to be of structural importance, I used a bioinformatic approach to gain insight into the likely topology of SAS-6.

2.1.1 Bioinformatic analysis of SAS-6

In order to define conserved domains of SAS-6, a Multiple Sequence Alignment (MSA) of 44 species was carried out using ClustalX [Larkin et al., 2007], using sequences obtained by blast searches against known SAS-6 orthologues. The primary sequence conservation of SAS-6 proteins is not very high [Cottee et al., 2011], but it was possible to identify an enrichment of conservation in the previously identified, N-terminal PISA motif (**Figure 2.2.1.i, A+C**). Additionally, some enriched conservation was observed towards the C-terminus (discussed in **Section 2.7.3**). Due to the poor sequence conservation, it is difficult to judge how accurate such alignments are, especially for *C. elegans* SAS-6 (CeSAS-6), which has a particularly divergent sequence. However, further analyses yielded more obvious similarities between SAS-6 proteins. A PSIPRED [Jones, 1999] secondary structural prediction was carried out on several SAS-6 orthologues, which indicated that SAS-6 proteins are rich in secondary structure, and have a shared overall topology. The N-terminal ~150 residues comprise a mixture of predicted alpha and beta structure, followed by a long, relatively unbroken helix, which in DmSAS-6 extends almost to the end of the protein. The COILs server [Lupas et al., 1991] was used to analyse several SAS-6 orthologues (**Figure 2.1.1.i, B**), and the results suggested that the long predicted helix would form a continuous coiled-coil.

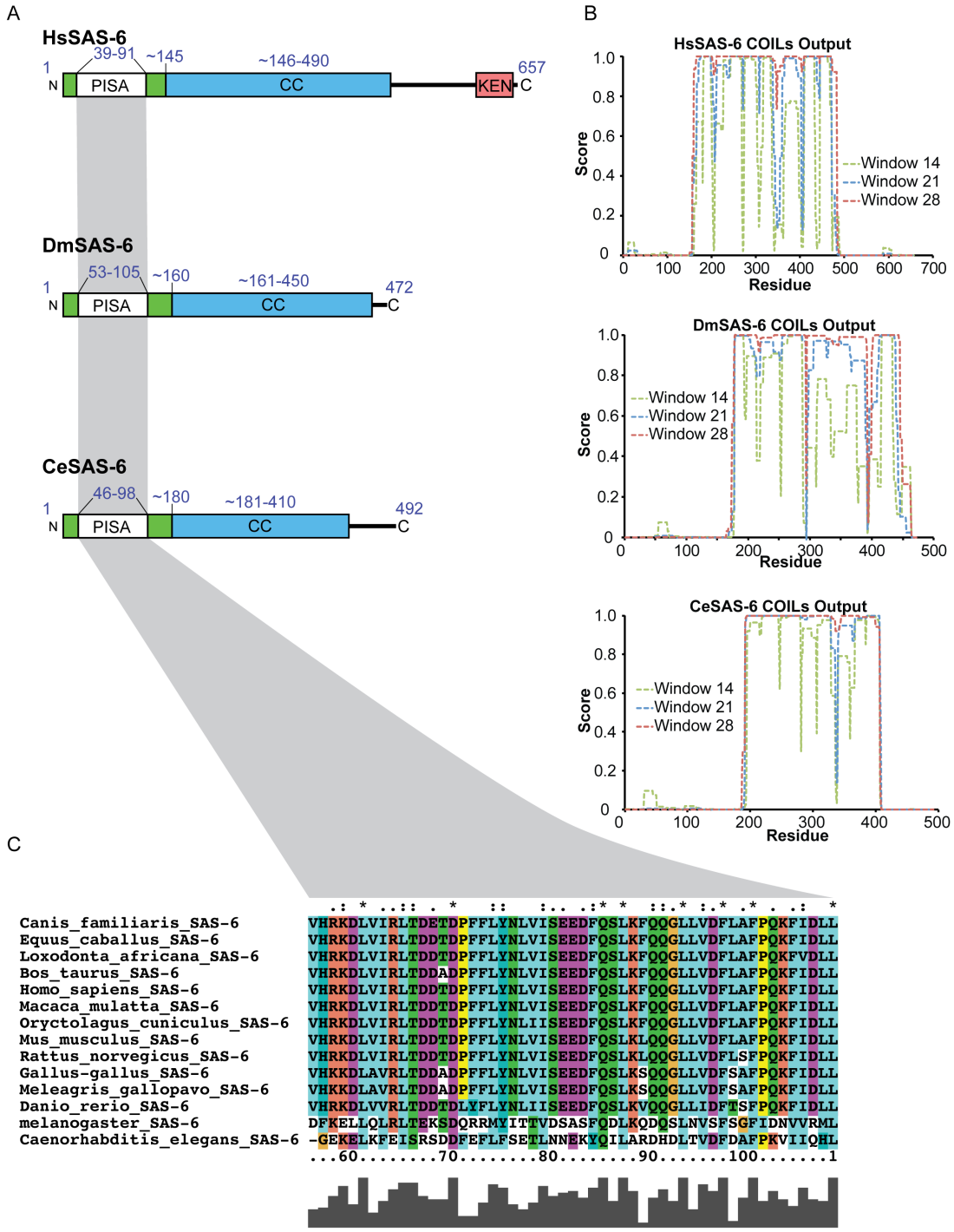


Figure 2.1.1.i: Bioinformatic analysis of SAS-6. (**A**) Schematics showing the overall topologies of SAS-6 orthologues, which share an N-terminal globular domain (green), followed by a long coiled-coil (blue). (**B**) Analysis of SAS-6 orthologues by the COILs server showing the predicted positions of coiled-coil domains. (**C**) MSA analysis showing the enriched conservation seen in the PISA motif. 44 sequences were used for this alignment, but only a selection of these is shown for clarity.

The length of this CC was shortest in CeSAS-6 (just over 200 residues), however in other orthologues including flies, unicellular organisms, and vertebrates, the length of the CC is ~290-350 residues – despite large variations in the overall length of the proteins. Despite slight differences in CC length, analysis of the MSA suggested that the beginning and end of the predicted vertebrate and DmSAS-6 coiled-coils roughly aligned, and that the vertebrate SAS-6 protein then extended beyond the CC. This extension contains what has been reported to be a KEN-box degradation signal (**Figure 2.1.1.i, A**) [Strnad et al., 2007]. PSIPRED analysis of vertebrate SAS-6 (not shown) indicated that this C-terminally extended region is largely unstructured. It is possible that the lack of a KEN box in DmSAS-6 explains why the *Drosophila* cartwheel persists, whereas in vertebrates, it is only visible in nascent/young centrioles.

In summary, despite the lack of strong sequence similarity, my bioinformatic analysis indicated that SAS-6 orthologues share an overall topology, consisting of an N-terminal globular head-group, followed immediately by a long coiled-coil.

2.1.2: Attempts to produce full-length recombinant SAS-6 *in vitro*

I first tested whether full-length SAS-6 fusion proteins could be expressed and purified from bacteria. Although SAS-6 could be expressed under certain conditions, almost no full length protein could be purified, as summarised in **Figure 2.1.2.i, A**. In brief, a cDNA encoding full length *D. melanogaster* SAS-6 (variant AAL68137) was PCR amplified, and cloned into various pET-based bacterial expression vectors, allowing IPTG-induced expression of recombinant protein under the control of the T7/lac system. The vectors, pETM-14, pET28b, pETM-44, and pETM-33 would each allow expression of His-SAS-6, SAS-6-His, His-MBP-SAS-6, and His-GST-SAS-6 respectively, in order to explore which tags would allow expression or purification of SAS-6. Small-scale expression trials using 5 ml cell cultures (not shown) suggested that SAS-6

could only be expressed when tagged at the N-terminus, and only when using certain *E. coli* strains, namely, B834 (DE3), BL21 (DE3), Rosetta2 (DE3) and Tuner (DE3). Small-scale Ni-NTA purification trials (data not shown, procedure described in **Appendix 7.1.5**) suggested that expressed protein could only be detected in the insoluble fraction of the cell lysate, and not in Ni-NTA elution fractions, suggesting that full-length SAS-6 was largely insoluble.

In order to see if a small amount of protein could be purified from a larger culture volume, His-SAS-6 was transformed into B834 (DE3) cells, which were grown in 1 L Terrific Broth. A detailed general procedure for soluble protein expression and purification can be found in **Appendix 7.1.4**, although it will be described once, here, briefly. Working Buffers (WB) for all constructs are listed in **Appendix 7.1.4**, **Table 7.1.4.i**. Inoculated cell culture was grown at 37 ˚C until an OD_{600} of ~0.6. Recombinant protein expression was induced by the addition of 1 mM IPTG, overnight at 21 ˚C. The cells were then harvested by centrifugation. His-SAS-6 was purified using a working buffer (WB) of 50mM Tris, pH 7.5, 150 mM NaCl, 1 mM DTT. The cell pellet was resuspended in WB, supplemented with lysozyme, DNaseI and a protease inhibitor cocktail (Roche). Cell resuspension was lysed via 5 passages through an Avestin C5 Emulsiflex homogeniser, at ~15,000 psi at 4 ˚C. The lysate was clarified by centrifugation at 26,600 x g at 4 ˚C for 30 mins. The lysate supernatant was supplemented to 25 mM imidazole, before being loaded onto a 1 ml Ni-NTA cartridge (Qiagen), pre-equilibrated in WB + 25 mM imidazole. The cartridge was washed for 20 CVs with WB + 50 mM imidazole before isocratic elution using WB plus 300 mM imidazole.

The elution fractions only contained ~1-2 mg total protein. SDS-PAGE analysis showed that although His-SAS-6 was expressed, most protein in the elution fractions was not full-length His-SAS-6. A Western blot was performed, using an α-His HRP conjugate antibody to specifically

detect the presence of His-tagged species in these fractions. Despite His-SAS-6 being barely detectable in the lysate supernatant, small amounts of intact His-SAS-6 were present in the Ni-NTA eluate, however full-length bands were very weak and the vast majority of His-SAS-6 was degraded (**Figure 2.1.2.i, B**). These results indicated that it would not be trivial to purify full-length SAS-6 under these conditions.

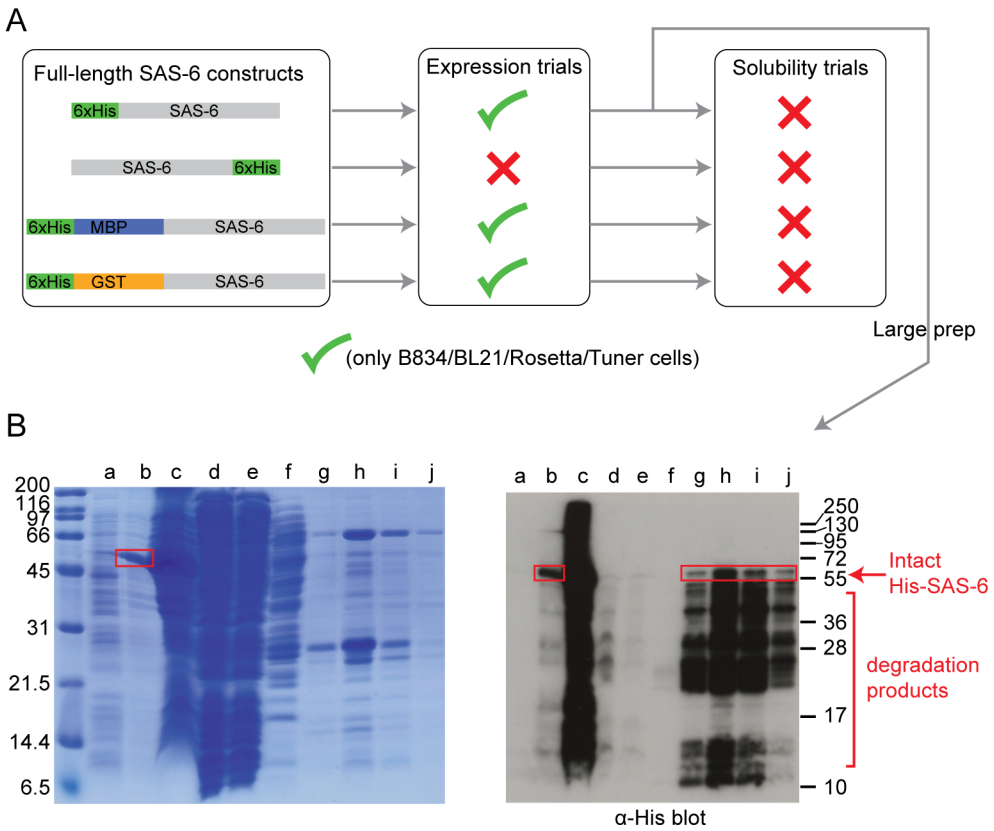


Figure 2.1.2.i: Attempts at producing full-length recombinant SAS-6. (**A**) Flow chart summarising small-scale expression and solubility trials of four different fusion proteins of full-length SAS-6. Different media and cell types were analysed for conditions that would enable expression of each construct, as judged by SDS-PAGE analysis of induced cell lysates. Small-scale Ni-NTA purification of such lysates did not yield any detectable soluble protein for any construct. (**B**) A larger scale purification of His-SAS-6. Left panel shows a coomassie stained SDS-PAGE gel of purification fractions. Although there was an induction band consistent with full-length His-SAS-6, this band was not highly enriched in the Ni-NTA elution fractions. Right panel shows a Western blot detecting His-tagged protein in these fractions, via the use of an α-His-HRP conjugate antibody (Qiagen). This confirmed the presence of His-SAS-6 in the induced cell sample, that most His-SAS-6 was present in the insoluble fraction of the lysate and that a very small amount of His-SAS-6 was detectable in Ni-NTA elution fractions. However the vast majority of this protein was severely degraded.

2.1.3: Screening for soluble SAS-6 fragments

As it had not been possible to purify full-length constructs, I tried to identify soluble fragments of SAS-6. Because the N-terminus was predicted to form a conserved, N-terminal globular domain, all constructs contained the N-terminus. Truncated constructs were designed by successively trimming back from the C-terminus (**Figure 2.1.3.i, C**). As SAS-6^{1-472} had already been cloned into a pETM-14 vector with an N-terminal His-tag, cloning of these truncation constructs could be achieved in a high-throughput manner using site-directed mutagenesis. Primers (**Supplementary CD**) were designed for each construct which would insert "TAATAA" stop codons within the full-length ORF, thus producing a C-terminally truncated construct. (**Figure 2.1.3.i, B**). All constructs were generated using a QuikChange II XL kit (Agilent) according to the manufacturer's instructions.

Plasmids encoding the constructs were transformed into B834 (DE3) *E. coli* cells, which had already been shown to express the full-length construct. Protein was expressed using 5 ml cultures in TB medium. Gel analysis of induced cell samples (**Figure 2.1.3.i, D**) indicated that all constructs were expressed at their expected lengths. Protein was then purified from these cultures using high-throughput Ni-NTA affinity, (**Appendix 7.1.5**), eluting the Ni-NTA beads with 250 µl of WB + 500 mM imidazole. These elution fractions were then analysed by Western blot using an αHis-HRP antibody to detect the presence of soluble constructs (**Figure 2.1.3.i, E**). Soluble protein could be detected for constructs with C-termini ranging from 177-366, but, the bands were often weak and partially degraded. However, two constructs, SAS-6^{1-241} and SAS-6^{1-266}, had reasonable yields and appeared to be largely intact. These constructs were subjected to larger-scale purification.

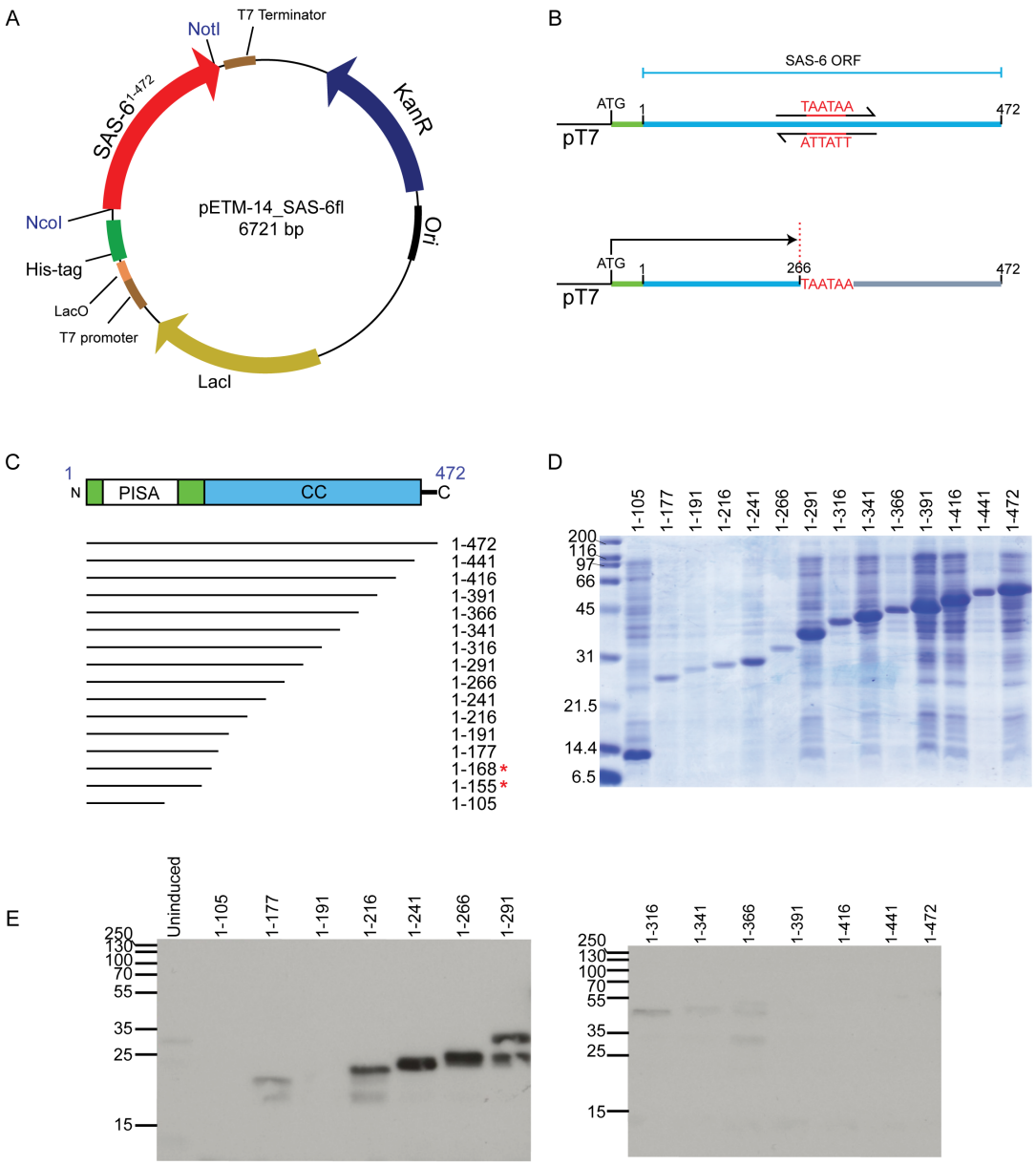


Figure 2.1.3.i: High-throughput trial to identify soluble SAS-6 constructs. (**A**) Schematic of the pETM-14 vector into which the full length ORF of *D. melanogaster* SAS-6 was cloned. (**B**) Schematic showing an example of the insertion of STOP codons via site-directed mutagenesis to generate a truncation construct (in this case SAS-6$^{1-266}$). (**C**) Schematic of the overall topology of SAS-6, and the different constructs that were designed. Constructs labelled with red asterisks were not trialled in the high-throughput test, but rather by large-scale purification, and no soluble protein could be obtained (not shown). (**D**) Gel analysis of induced cell samples, all of which expressed the indicated constructs at their expected lengths. (**E**) Western blot analysis of purified SAS-6 constructs. Following small-scale Ni-NTA affinity purification of SAS-6 constructs, eluates were probed with an α-His-HRP antibody to identify the presence of soluble SAS-6 constructs. It can be seen that SAS-6$^{1-241}$ and SAS-6$^{1-266}$ produced the highest yield, and the most intact products.

2.1.4: Large-scale purification of SAS-6$^{1-241}$.

Based on the results of the high-throughput solubility screen, SAS-6$^{1-241}$ and SAS-6$^{1-266}$ were chosen for large-scale production. Both proteins were expressed in 4-6 L of TB medium, and purified in the same way, using Ni-NTA affinity.

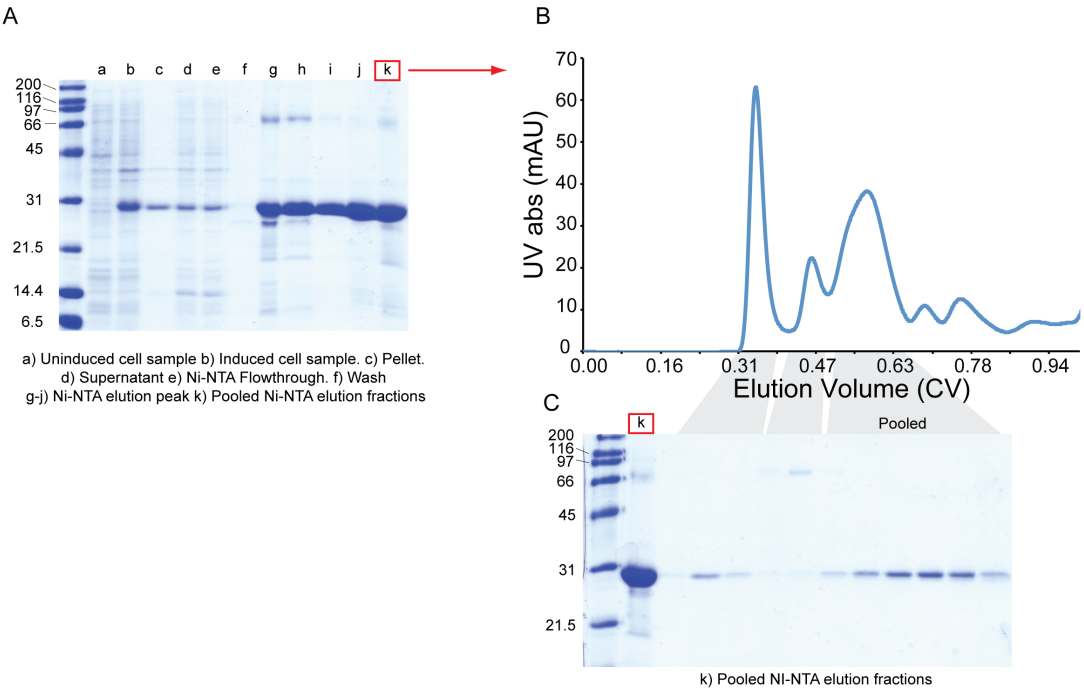


Figure 2.1.4.i: Purification of SAS-6$^{1-241}$. (**A**) Gel analysis of purification fractions. Pooled Ni-NTA elution fractions were highly enriched for full length SAS-6$^{1-241}$. (**B**) SEC trace of SAS-6$^{1-241}$ run on an S200 26/60 column (GE Healthcare). Blue = UV absorption. SAS-6$^{1-241}$ had a large-ranging elution volume, suggesting the sample is not monodisperse and contains large molecular weight species. (**C**) Gel analysis of SEC fractions confirms the presence of SAS-6$^{1-241}$ in these peaks.

It was noticed that for both constructs, material was lost over time due to precipitation, particularly when the samples were concentrated. This effect was greater for SAS-6$^{1-266}$, which also showed some minor degradation, so further work continued using SAS-6$^{1-241}$. Pooled elution fractions of SAS-6$^{1-241}$ were further purified by size-exclusion chromatography, using an S200 26/60 column (GE Healthcare), and using WB as a running buffer. Despite gel analysis showing that SAS-6$^{1-241}$ was relatively pure, it eluted in various peaks over much of the separating volume of the column (**Figure 2.1.4.i, B, C**). The void peak was fairly large, and only contained SAS-6$^{1-241}$. This result suggested that the sample was not monodisperse, and formed a range of species with

differing stoichiometries. Although it was expected that the sample might dimerise due to the predicted coiled-coil, this behaviour suggested that higher-order oligomerisation was occurring.

2.1.5: EM analysis shows that wild-type SAS-6$^{1-241}$ can form large assemblies

To investigate the appearance of these species, I analysed the sample using negative-stain electron microscopy (EM) with the kind help of Dr Errin Johnson. The sample was adjusted to 1 mg /ml in WB and then diluted 1:30 in water. 30 µl of sample was deposited for 2 minutes, onto a 200 mesh, glow-discharged carbon coated copper grid. To negatively stain the sample, a drop of 2% uranyl acetate was applied to the grid for 10 seconds before blotting, and air-drying. The sample was viewed using an FEI Tecnai 12 Transmission Electron Microscope, at 120 kV, at 43,000x magnification. Micrographs revealed that the protein formed large string-like assemblies (**Figure 2.1.5.i, A, B**).

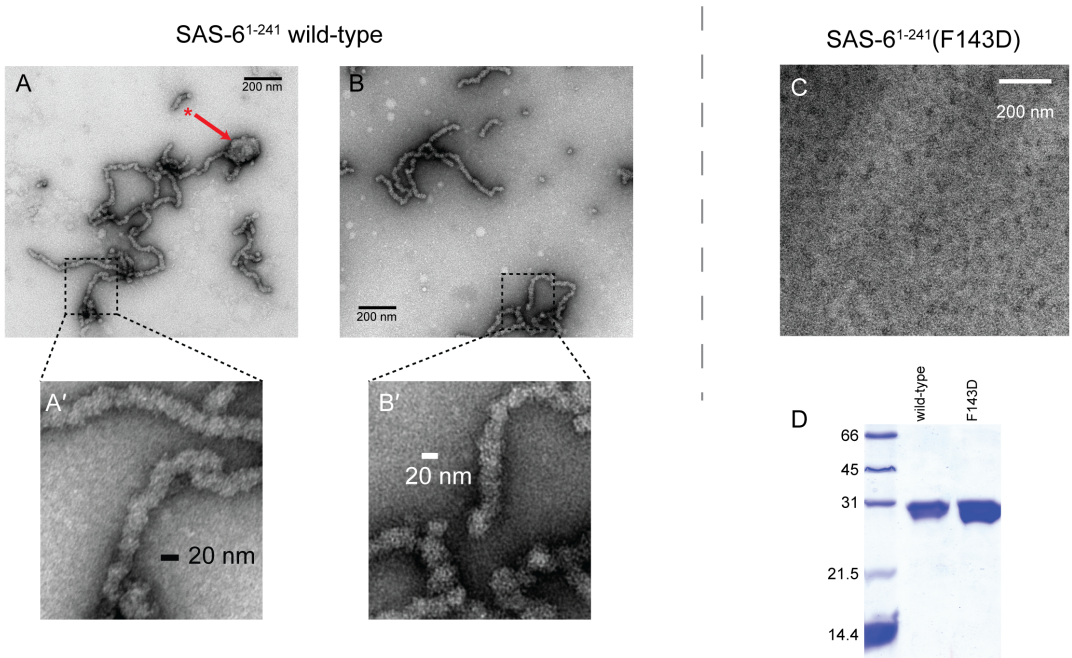


Figure 2.1.5.i: EM analysis of SAS-6$^{1-241}$(WT) and (F143D). (**A**+**B**) Representative micrographs showing string-like assemblies of SAS-6$^{1-241}$(WT). The majority of SAS-6$^{1-241}$(WT) assembled into such structures, which are unlike typical amorphous protein aggregate, (some of which can be seen in micrograph **A**, indicated by a red asterisk and arrow). The structures were variable in length, however they appeared to be fairly uniform in width. (**A'**+**B'**) Closeups of the indicated regions highlighting the approximate width of the string-like assemblies. (**C**) No string-like assemblies could be visualised for the F143D mutant (**section 2.2.1**), micrographs of which appeared to have a uniform granularity. (**D**) Gel analysis showing the purity of samples used for EM and subsequent MALS analysis (**Figures 2.2.3.i, ii**).

These were variable in length but typically hundreds of nm long, and although they crossed each other, they did not appear to be branching. Interestingly they had a fairly uniform width of ~20-30 nm.

2.2: Investigating the SAS-6 N-N interaction.

At this point in my studies, two papers were published reporting the structure of the N-terminal region of SAS-6 from *D. rerio, C. elegans* and *C. reinhardtii* [van Breugel et al., 2011; Kitagawa et al., 2011b]. (A review of these findings and their implications can be found in [Cottee et al., 2011]). These constructs were found to dimerise via an "N-N" interface. This is a symmetrical interaction, mediated by a hydrophobic finger residue from one monomer, inserting into a hydrophobic pocket on the other monomer (**Figure 2.2.i, A**). The pocket is formed largely of residues in the PISA motif, explaining why this region is particularly conserved, and the interaction could be strongly perturbed via mutation of the hydrophobic finger residue to a charged residue. Constructs harbouring this mutation, but containing several heptad repeats of the coiled-coil, could additionally form stable dimers via the coiled-coil (CC) interface (**Figure 2.2.i, C**). Structures of this mutant "N-CC" dimer were presented for the *C. reinhardtii* and *D. rerio* SAS-6 orthologues [van Breugel et al., 2011; Kitagawa et al., 2011b]. Together, these two interaction interfaces would allow SAS-6 to oligomerise into higher-order circular assemblies, consistent with the dimensions of the cartwheel central hub.

In light of these new findings, the higher order oligomerisation I observed for DmSAS-6 could be explained by the same CC, and N-N interactions. In my hands, head-group-only constructs could not be purified. Although SAS-6^{1-155} (with no predicted coiled-coil) and SAS-6^{1-167} (with 1-2 heptad repeats of coiled coil) could be expressed, they were not soluble (not shown). Nevertheless, an MSA analysis indicated that, similarly to vertebrates, DmSAS-6 had a phenylalanine hydrophobic finger residue: F143 (**Figure 2.2.i, B**). I reasoned that mutating this residue in my SAS-6^{1-241} construct might suppress the N-N interaction and so allow me to purify a stable dimeric form of the protein that would no longer form the higher-order structures I observed with the WT protein (**Figure 2.1.4.i, B** and **Figure 2.1.5.i, A, B**).

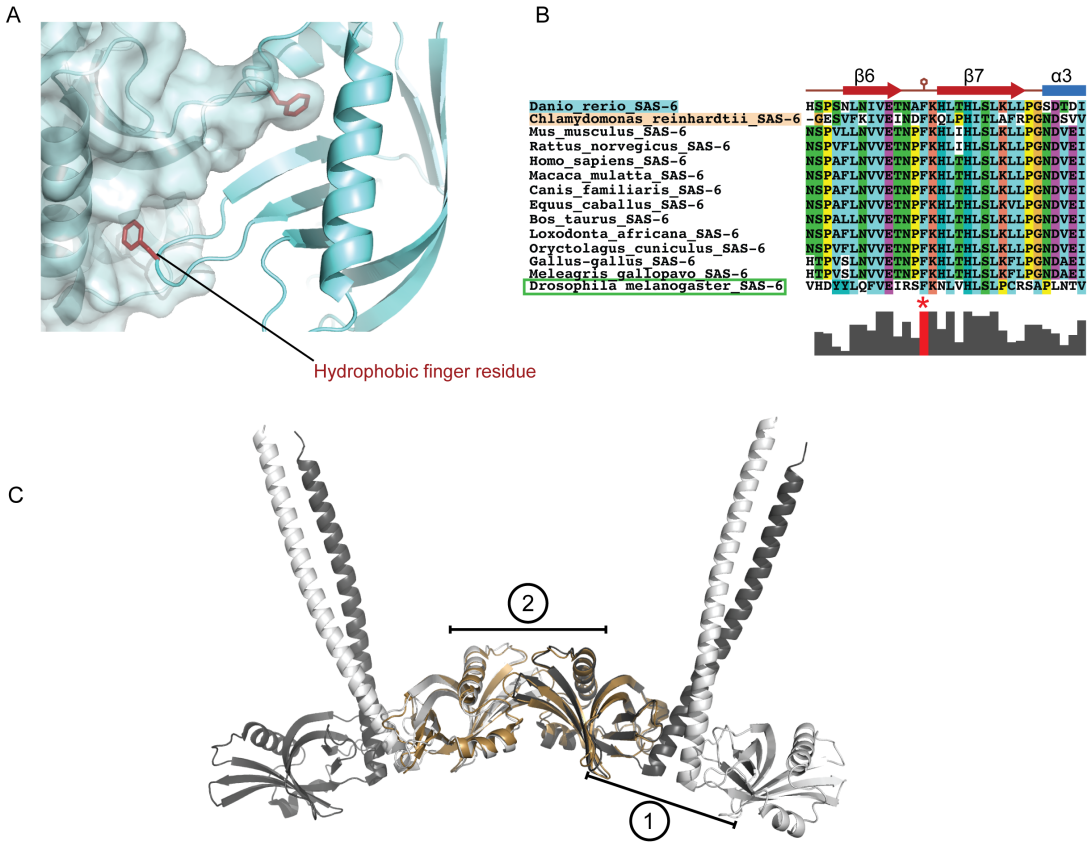


Figure 2.2.i: The N-N and C-C interactions of SAS-6 orthologues revealed by van Breugel et al., 2011, and Kitagawa et al., 2011b. (**A**) Closeup of the N-N interface of *D. rerio* SAS-6 (PDB entry 2Y3W). The left-hand monomer is shown in a space-filling representation. The phenylalanine finger residue is shown in red stick format. (**B**) MSA analysis around the β6-β7 loop, which harbours the hydrophobic finger residue. A schematic of the secondary structure is laid above the alignment, and the hydrophobic finger residue highlighted by a red asterisk and bar. (**C**) SAS-6 N-CC dimers (coloured light and dark grey for chains A and B respectively) are associated via the CC interaction (**1**). These dimers can then associate via the N-N interaction (**2**). Oligomerisation could continue in this fashion to form a 9-fold ring, consistent with the cartwheel central hub. In this panel, two *C. reinhardtii* N-CC dimers (PDB entry 3Q0X) have been superimposed together using the structure of the N-N dimer (underlaid as a gold cartoon, PDB entry 3Q0Y).

2.2.1: Purification of SAS-6$^{1-241}$(F143D)

The predicted DmSAS-6 N-N mediating hydrophobic finger residue was F143. I mutated this residue to Asp, similarly to the approach carried out in the *D. rerio* orthologue [van Breugel et al., 2011], using site directed mutagenesis, on the pETM-14_SAS-6$^{1-241}$ plasmid (QuikChange II XL, Agilent). The protein was expressed and purified in the same way as the wild-type construct (**Section 2.1.3**). The yield of protein was substantially greater than for the wild-type construct, and Ni-NTA eluate was much purer (**Figure 2.2.1.i**). Pooled fractions were purified via size

exclusion chromatography, as for the wild-type construct. Importantly, the protein eluted predominantly as one main peak, suggesting that the F143D mutation allows the protein to adopt one main conformation in solution. Interestingly, this main peak was preceded by a much smaller peak; gel analysis showed that both peaks consisted of pure SAS-6^{1-241}(F143D). Only fractions from the main peak were pooled for further study to ensure the sample was homogenous.

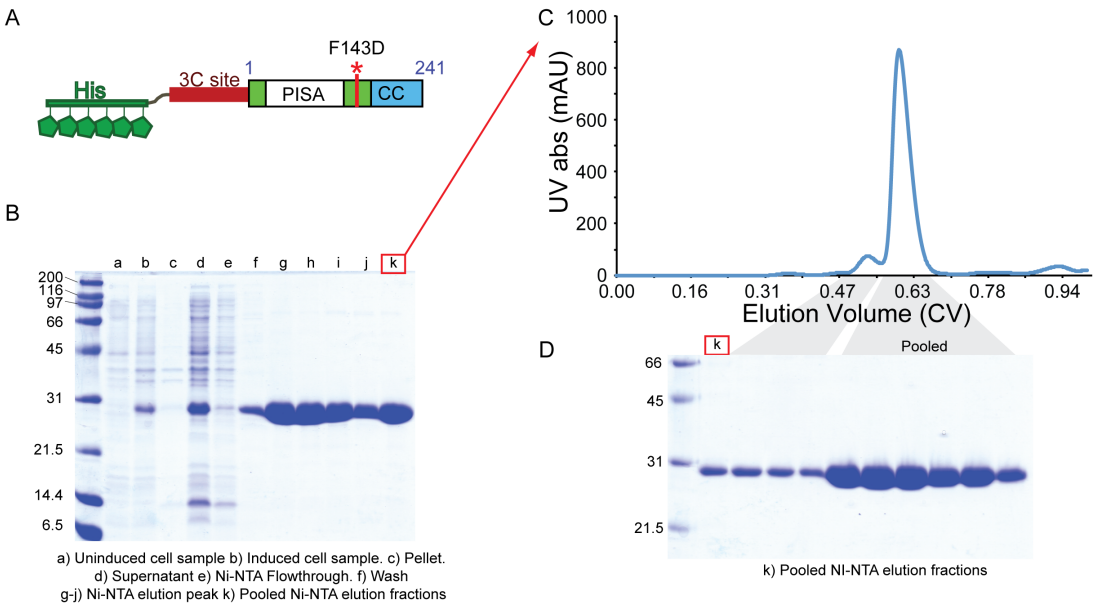


Figure 2.2.1.i: purification of SAS-6^{1-241}(F143D). (**A**) Schematic of the His-SAS-6^{1-241}(F143D) construct. (**B**) Gel analysis of purification fractions of SAS-6^{1-241}(F143D). (**C**), SEC trace from an S200 26/60 column (GE Healthcare). A main peak was evident, but was preceded by a small peak. (**D**) Gel analysis of SEC fractions showed that only SAS-6^{1-241}F143D was present in both the pre-peak and main peak fractions. Only the main peak was pooled for further study to ensure the homogeneity of the sample.

Similarly to the wild-type construct, this sample was analysed by EM. Conversely to the wild-type protein, micrographs of the F143D mutant were fairly featureless, exhibiting only a uniform granularity (**Figure 2.1.5.i, C**). These results suggest that only wild-type SAS-6^{1-241} can form large assemblies, and that these are specifically mediated by the N-N interface. The potential significance of these assemblies will be discussed in more detail in the Perspectives section.

2.2.2: SEC-MALS suggests that SAS-6$^{1-241}$ can adopt a wide range of stoichiometries, while the F143D mutant is dimeric.

My results at this stage were consistent with wild-type SAS-6 forming a range of high-order oligomeric species, while a mutation designed to knock out the N-N head-group interaction allowed SAS-6 to form a monodisperse species, presumably a dimer. In order to investigate the oligomeric nature of these constructs, a Multi Angle Light Scattering (MALS) (explained in **Appendix 7.1.7**) experiment was carried out. Purified His-SAS-6$^{1-241}$(WT) and His-SAS-6$^{1-241}$(F143D) were prepared at a range of concentrations: 33 µM, 66 µM, and 150 µM, using 10 kDa MWCO centrifugal concentrators (Amicon). Higher concentrations of 330 µM, and 1000 µM could be reached for the F143D mutant, but not for the WT protein, which invariably crashed out of solution at concentrations above ~150-200 µM. Samples were injected over an S200 10/30 column (GE Healthcare) using WB as the running buffer. The WT protein eluted in the void volume at all concentrations, suggesting that it was forming multimeric species too large to be separated by this column (**Figure 2.2.2.i, A**). (The effective separation range of this column is 10-600 kDa). If a central hub could be formed in solution it would theoretically contain 18 DmSAS-6 monomer copies and would therefore have a mass of ~540 kDa. However, the masses observed by MALS suggested that no single discrete species was forming. Rather, a range of multimers existed in this void peak, from 30-200 MDa, consistent with species containing thousands of monomer copies. Conversely, the F143D construct eluted much further into the separating volume of the column and, at all concentrations, a large predominant peak was observed with mass estimations consistent with a largely dimeric species. (**Figure 2.2.2.i, A**) At higher concentrations, it was possible to see the preceding peak that had previously been observed in SEC analysis (**Figure 2.2.2.i, A, peak at around 0.52 CV**). Fractions from the SEC pre-peak (**Section 2.2.1**) had not been pooled for MALS experimentation and so its reappearance suggests that it represents a dynamic interaction. Mass estimations across the pre-peak were not stable, however these estimations suggested the presence of a trimeric or tetrameric

species. It is possible that this dynamic interaction could represent weak residual N-N interface affinity, although the fact that a stable tetrameric mass is not reached at injection concentrations of 1000 µM suggests that if this were the case, it would be a weak interaction.

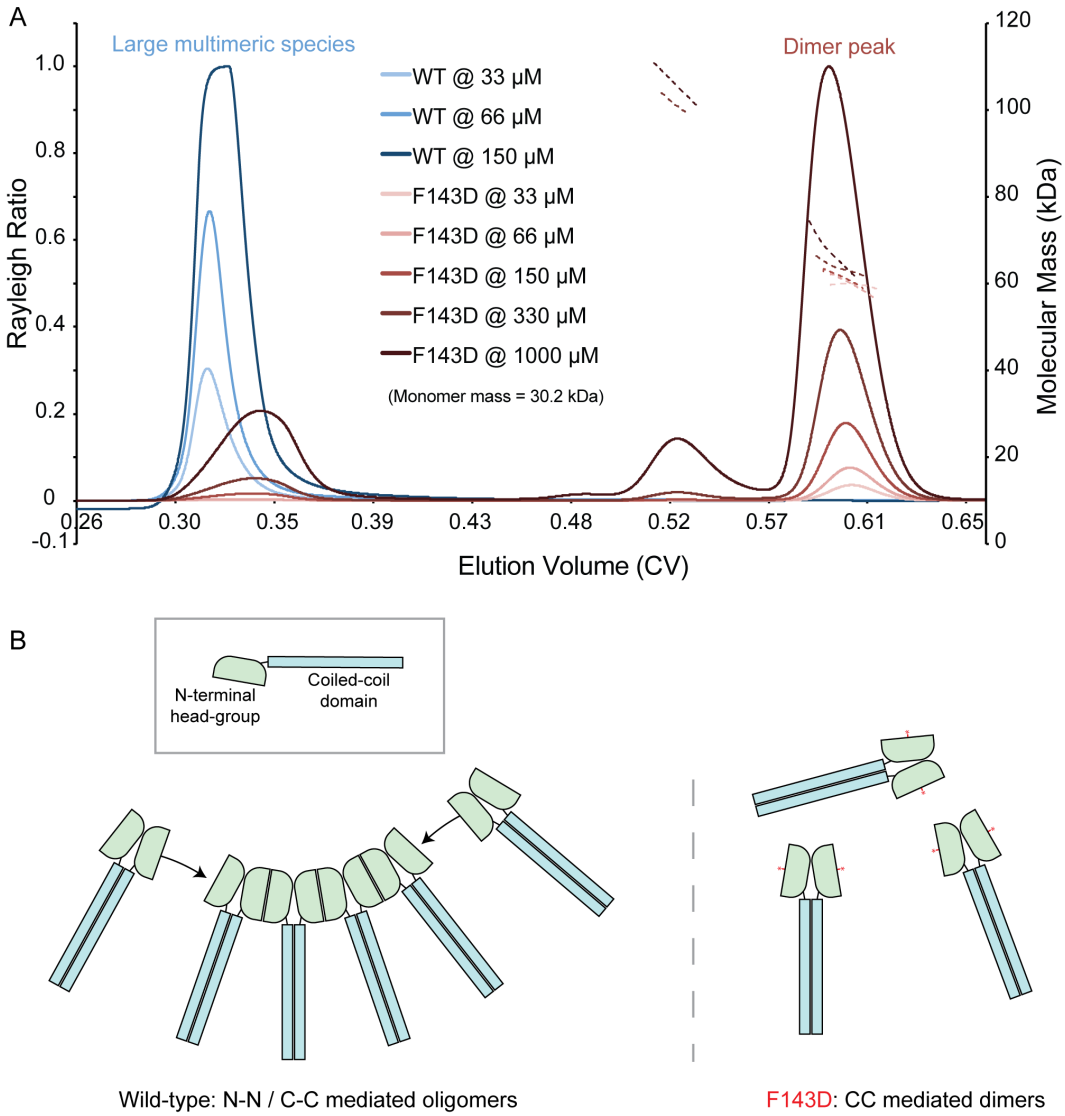


Figure 2.2.2.i: SEC-MALS analysis of SAS-6$^{1-241}$ WT and F143D. (**A**) SEC-MALS traces of both constructs injected at various concentrations. Solid lines show the relative Rayleigh ratio, dashed lines show the observed average mass, all coloured according to the legend. At all tested concentrations, the wild-type construct (blue) eluted in the void peak, whereas the F143D mutant (red) eluted predominantly in one peak, with an average mass consistent with a dimer. At higher concentrations the F143D mutant displayed some slightly higher order oligomerisation, characterised by a preceding peak, with masses ranging from trimeric to tetrameric. Some protein also eluted close to the void suggesting that the protein may be beginning to aggregate at these concentrations. (**B**) Schematics showing the formation of SAS-6 oligomers. Above, a schematic of a SAS-6 monomer. Left, high-order oligomerisation of CC dimers, oligomerising via the N-N interface. Right, N-N deficient (F143D) SAS-6 can only dimerise via the CC interaction.

2.2.3: Generation of further CC truncations to test the effect of the CC on dimerisation

Further MALS experiments were conducted on the F143D mutant, where the number of CC heptad repeats was reduced from ~11 (His-SAS-6^{1-241}), to ~8 (SAS-6^{1-216}) or only ~6 (His-SAS-6^{1-201}) to assess the role of the CC in dimerisation (**Figure 2.2.3.ii, A,** construct details in **Table 2.2.3.i**). The new constructs were created using site-directed mutagenesis to insert the F143D mutation, and also to insert stop codons to generate the new truncations.

Construct	Amino Acids	MW (Da)	pI	ε_{280}	Abs 0.1%
His-SAS-6^{1-241}(F143D)	261	30183.0	6.50	18910	0.627
SAS-6^{1-216}(F143D)	219	25137.4	6.25	17420	0.693
His-SAS-6^{1-201}(F143D)	221	25478.9	6.83	17420	0.684

Table 2.2.3.i: Details of constructs purified to test the effect of CC length on dimerisation, as calculated by Protparam.

Purification of SAS-6^{1-216}(F143D) and SAS-6^{1-201}(F143D) was carried out identically to SAS-6^{1-241}(F143D). Both purifications were highly similar, and so only the purification of SAS-6^{1-216}(F143D) is shown in **Figure 2.2.3.i**. The only difference between the two purifications was that the His-tag was removed from SAS-6^{1-216}(F143D) in order to aid crystallisation trials (**Section 2.3.1**). To remove the His-tag after Ni-NTA chromatography, 3C protease was added at a ratio of 1 mg 3C : 50 mg target. The reaction was dialysed against fresh WB overnight at 4 ˚C. The reaction mixture was run back over a 1 ml Ni-NTA cartridge in order to remove any un-cleaved SAS-6, and also over a 1 ml glutathione column (GSTrap, GE Healthcare) to deplete the GST-fused 3C protease. The flow-through was concentrated and run over an S200 26/60 column (GE Healthcare) using WB as the running buffer, and this resolved SAS-6 from residual protease. SAS-6^{1-216}(F143D) eluted as a small pre-peak followed by a main peak, as observed for SAS-6^{1-241}. Pure SAS-6 fractions were pooled from this main peak (**Figure 2.2.3.i**)

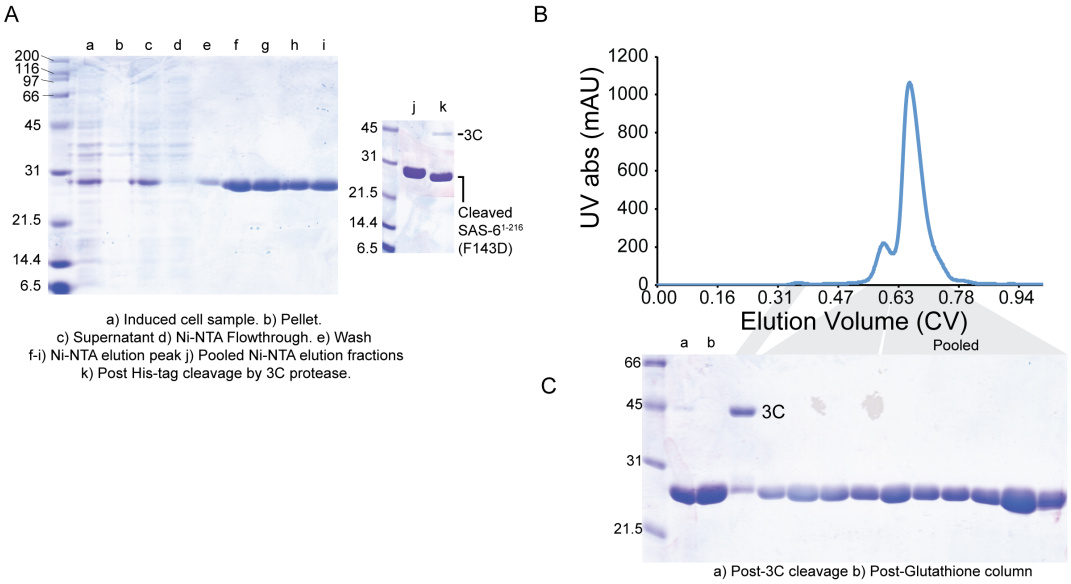


Figure 2.2.3.i: Purification of SAS-6$^{1-216}$(F143D) (**A**) Left, gel analysis of initial purification fractions. Right, gel analysis before and after His-tag cleavage by 3C protease. (**B**) SEC trace from an S200 column after His-tag cleavage. Despite depletion of 3C protease by reverse Glutathione and Ni-NTA chromatography, a small amount of 3C remained, but was separated by SEC. SAS-6$^{1-216}$ eluted similarly to previous observations (**Figure 2.2.1.i, C**), with a small preceding peak. (**C**) Gel analysis of peak fractions showed that they contained only pure SAS-6$^{1-216}$. Only the main peak was pooled for further work.

In order to assess the effect of truncating the coiled-coil on dimerisation of the constructs, MALS experiments were again carried out, using the same column and running buffer as before on His-SAS-6$^{1-241}$(F143D), SAS-6$^{1-216}$(F143D), and His-SAS-6$^{1-201}$(F143D). Under these conditions, reducing the length of coiled-coil appeared to reduce the strength of the dimeric interaction, as the observed mass of the "dimer" peak was lower – between that of a monomer and dimer. As shown in **Figure 2.2.3.ii**, even at higher molar concentrations, the shorter constructs could not dimerise as efficiently as the longer SAS-6$^{1-241}$(F143D) construct. This suggests that SAS-6(F143D) dimerisation is mediated by the CC interface. Furthermore, the fact that the wild-type protein can form high-order oligomers, while the F143D mutant forms predominantly dimers, strongly suggests that DmSAS-6 can also associate via an F143 mediated N-N interface.

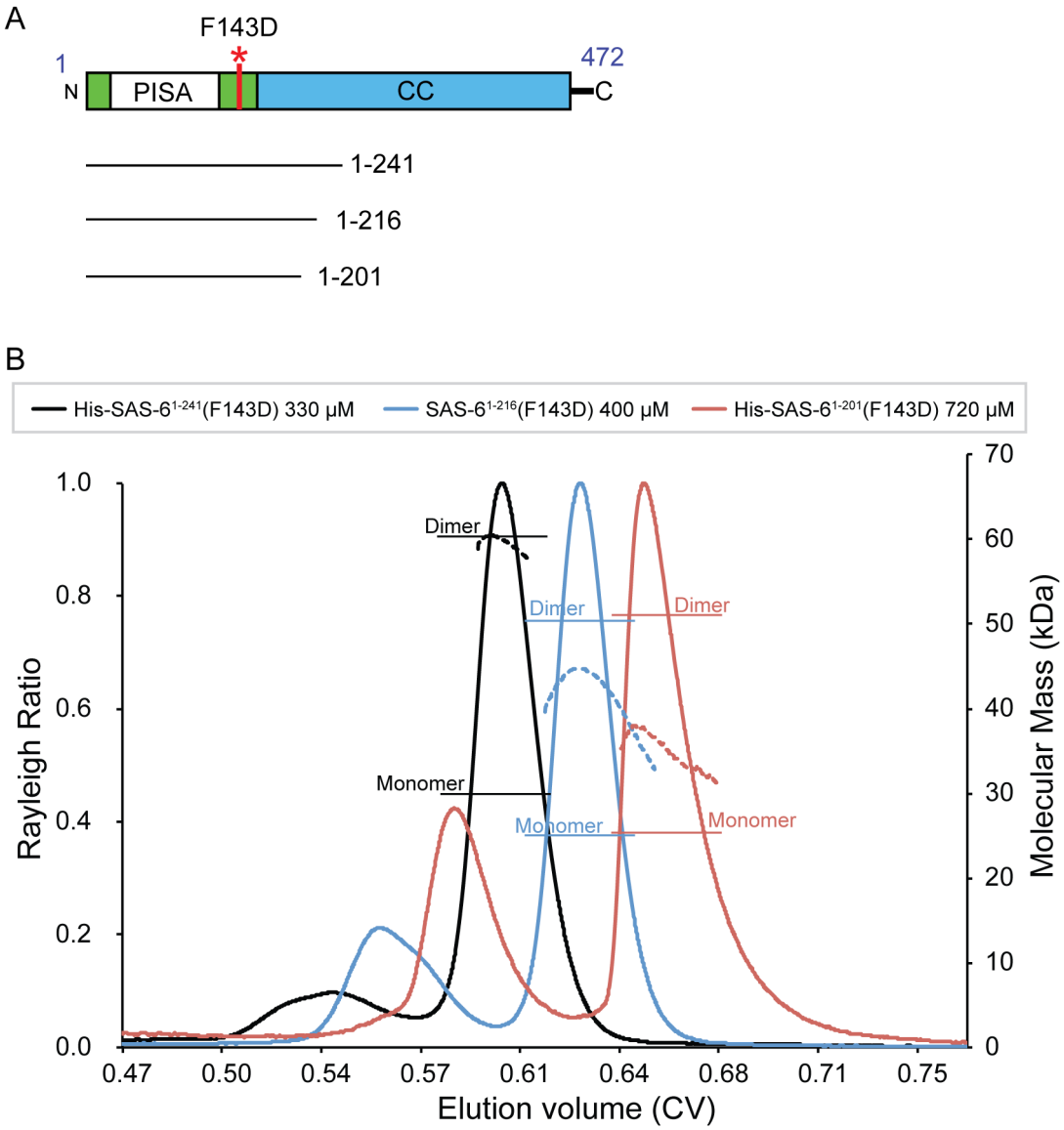


Figure 2.2.3.ii: SEC-MALS analysis to test the effect of the CC on dimerisation. (**A**) Schematics of the constructs used in this experiment – each one containing different amounts of the coiled-coil. (**B**) SEC-MALS traces coloured according to the legend above. Solid lines represent the Rayleigh ratio, while dashed lines show the observed average mass across the peak. The longest construct displays robust dimerisation when injected at 330 µM. However, even at higher concentrations, the constructs with less coiled-coil fail to dimerise as efficiently, with observed masses between dimer and monomer. Note that the SAS-6$^{1-216}$(F143D) construct had no His-tag, hence its similarity in MW to the shortest construct.

2.3: Crystallisation and data processing of SAS-6

My data so far suggested that *Drosophila* SAS-6 was likely to be similar in structure to SAS-6 in other species as it could dimerise by a CC mediated interface, which could further oligomerise into assemblies mediated by the N-N interface. Although the structure of the N-terminal region of SAS-6 had already been solved in several other species [van Breugel et al., 2011; Kitagawa et al., 2011b] – and more structures would follow in the years to come [van Breugel et al., 2014; Hilbert et al., 2013] – I decided at the time that it was worth attempting to solve the structure of the *Drosophila* protein, as this would allow us to use the powerful *Drosophila* model to further analyse SAS-6 function *in vivo*. I therefore attempted to crystallise the SAS-6(F143D) constructs.

2.3.1: Crystallisation of SAS-6^{1-216}(F143D)

Initial crystallisation trials were carried out on all three constructs described in **Section 2.2.3**, using the sitting drop approach, and trialling mother liquors from various high-throughput commercial screens (Molecular Dimensions). Trials were only successful for SAS-6^{1-216}(F143D). SAS-6^{1-216}(F143D) in purification working buffer, was concentrated to 41.6 mg/ml in a MWCO 10 kDa centrifugal concentrator (Amicon). Small rod-like crystals grew in 0.4 µl drops after ~3 weeks at 21 ˚C, in a mother liquor containing 0.1 M Bicine, 0.1 M Trizma Base pH 8.5, 20% w/v PEG550 MME, 10% w/v PEG20K, 30 mM $NaNO_3$, 30 mM Na_2HPO_4, 30 mM $(NH_4)_2SO_4$, at a protein:mother liquor ratio of 75:25 (**Figure 2.3.2.i, A, B**). The crystals appeared to stop growing after ~4 weeks, and were subsequently fished and flash frozen, with the mother liquor serving as cryoprotectant.

2.3.2 Data collection

Data were collected at the microfocus beamline, ID23-2, at the ESRF synchrotron. A microfocus beam was useful in this experiment due to the thin width of the SAS-6 crystals, as the brightness of the beam allowed maximum diffraction signal to be obtained from the relatively low number of unit cells that would be present in such a crystal. It also allowed easy targeting of the beam to certain parts of the crystals, which was instrumental in obtaining the highest resolution data. One crystal diffracted significantly better than others, although another, thinner crystal could not be separated from it, and remained in the loop (**Figure 2.3.2.i, B**).

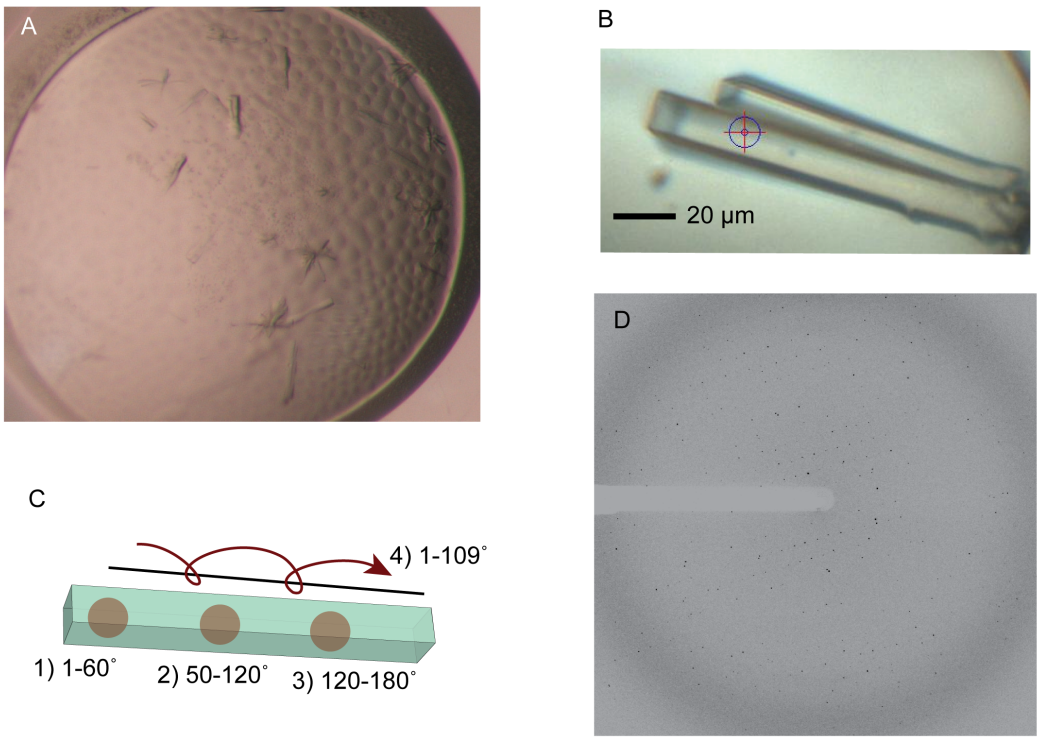


Figure 2.3.2.i: Crystallisation and data collection of SAS-6$^{1-216}$(F143D) (**A**) Rod-like crystals of SAS-6$^{1-216}$(F143D). (**B**) The crystal from which a complete dataset was collected. The crosshairs indicate the beam centering for position 1 detailed in **C**. (**C**) Schematic of the different sweeps used to collect data from the SAS-6$^{1-216}$(F143D) crystal. These consisted of three static sweeps (1-3) at different positions, covering 180˚, and a 109˚ helical sweep along the rod. (**D**) A representative diffraction image.

Initially this paired crystal was avoided by aiming the beam at the tip of the rod. Screening shots were taken and suggested a monoclinic unit cell, and 180˚ of data were collected. Autoindexing suggested that a complete data set had been collected to 3.5 Å, however, diffraction weakened after around 60 seconds of X-ray dose. In order to improve the resolution, a strategy was taken

to collect segments of data from along the length of the rod (**Figure 2.3.2.i, C**). Three overlapping sweeps of 60-70˚ were collected (Sweeps 1, 2 and 3), and an additional 180˚ helical collection (Sweep 4) was attempted along the length of the rod. Unfortunately, beamline functionality was lost during the helical collection and only 109˚ were collected.

2.3.3: Data processing

Data from sweeps 1, 2 and 3 were originally processed manually using iMosflm [Leslie and Powell, 2007] to index and integrate spots using the most likely lattice, which was monoclinic primitive. POINTLESS [Evans, 2005] was used to analyse the integrated data, and this suggested that the spacegroup was P2, due to the lack of systematic absences along the k axis. (**Figure 2.3.3.i, B**) Unavoidably, some images contained a second, offset lattice, due to reflections from the smaller paired crystal. However, this lattice was weak, and iMosfilm preferentially indexed the original, stronger lattice. The data were scaled and merged using SCALA [Evans, 1997]. A self-rotation function (SRF) was carried out using Molrep [Vagin and Teplyakov, 2010] to further analyse the pointgroup, and assess the presence of rotational non-crystallographic symmetry. No translational NCS was detected. Only two-fold rotational symmetry axes were detected, as shown by peaks in the Chi=180˚ section of the SRF (**Figure 2.3.3.i, A**). One crystallographic two-fold axis was evident along the *b* axis, as expected. Interestingly another strong peak (93% of the origin) lay orthogonally to the *b* axis (along the *a* axis), suggesting the presence of a non-crystallographic two-fold axis. By definition, an equivalent axis is generated orthogonally to the *ab* plane via operation of both the NCS two-fold, and the crystallographic two-fold. The combination of these orthogonal axes indicates that the molecules in the unit cell would exhibit a pseudo-orthorhombic arrangement.

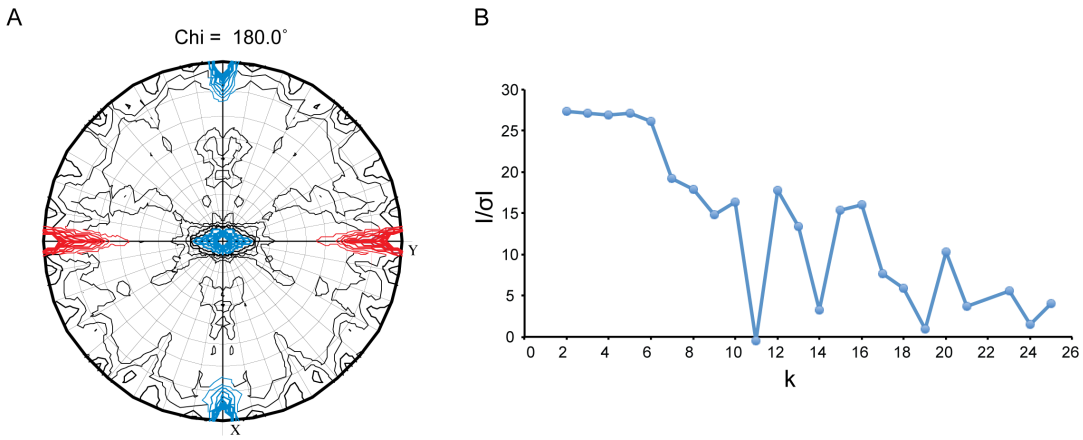


Figure 2.3.3.i: Data analysis of DmSAS-6$^{1-216}$(F143D) (**A**) Chi 180˚ section of the self-rotation analysis of the DmSAS-6$^{1-216}$ dataset using data to 4 Å, calculated in MOLREP. Peaks representing the crystallographic twofold symmetry axis lie along the *b* axis and are coloured red. The blue peaks are 93% of the origin and represent a symmetrically equivalent pair of NCS axes. These axes lie along the crystallographic *a* axis, and orthogonal to the ab plane. (**B**) A graph of I/σI measurements along the k axis. A twofold crystallographic twofold screw axis along *b* would result in absences for each odd index, however this is not the case.

The Matthews coefficient [Matthews, 1968] was used to estimate the number of molecules in the asymmetric unit (ASU). The cell volume was calculated to be 372979 Å^2, and with a monomer mass of 25137.4 Da it was estimated that the most likely number of chains in the ASU (P = 83 %) would be three (according to average crystal solvent content analysis of the PDB) (**Table 2.3.3.i**).

Chains/ASU	Matthews Coeff	% Solvent	P(2.92 Å structures)	P(total PDB)
2	3.71	66.86	0.09	0.07
3	2.47	50.29	0.83	0.80
4	1.85	33.72	0.06	0.10

Table **2.3.3.i:** Matthews Coefficient analysis calculated with the SAS-6$^{1-216}$(F143D) dataset, estimating a molecular weight of 25137.4 Da per chain.

However, there was only evidence for one unique two-fold NCS symmetry axis, no evidence for translational NCS, and SAS-6(F143D) constructs behaved as dimers in solution. It was therefore suspected that two molecules (P = 9 %) would be present in the ASU.

Sweeps 1, 2, 3 and the final 50˚ of sweep 4, were eventually re-processed using the XIA2 pipeline [Winter, 2010] under the 3daii mode, using XDS [Kabsch, 2010] to identify, index, and integrate spots, and AIMLESS [Evans and Murshudov, 2013] for scaling and merging. This produced a complete dataset to 2.9 Å, as detailed in **Table 2.3.3.ii**.

	DmSAS-6$^{1-216}$ (F143D)
Beamline	ESRF ID23-2
Spacegroup	P2
Wavelength	0.8726 Å
Unit cell dimensions (Å/˚)	a = 47.13 b = 64.74 c = 123.73 α = 90.00 β = 98.91 γ = 90.00
Resolution (Å)	41.43–2.92
Completeness (overall/inner/outer)	97.4/90.4/96.4
R_{merge} (overall/inner/outer)	0.128/0.035/0.668
R_{pim} (overall/inner/outer)	0.060/0.017/0.318
CC (½) (overall/inner/outer)	0.994/0.998/0.799
I/σI (overall/inner/outer)	12.1/48.4/2.6
Multiplicity (overall/inner/outer)	5.2/4.8/5.1
No. unique reflections	15,767

Table 2.3.3.ii: Processing statistics for the DmSAS-6$^{1-216}$(F143D) dataset.

2.4: Structure solution of SAS-6$^{1-216}$(F143D)

As it was likely that DmSAS-6 would adopt a similar fold to the published structures in the PDB, I used these to generate homology models for phasing via molecular replacement.

2.4.1 A molecular replacement solution for DmSAS-6

An early solution was found via molecular replacement using the CrSAS-6 N-CC dimer (PDB entry 3Q0X [Kitagawa et al., 2011b] ~21% sequence identity), with the coiled-coil trimmed to ~ten residues, using PHASER [McCoy et al., 2007] (not shown), and by enforcing the spacegroup P2. However, using a dimeric model inherently biases the solution by providing a presumed quaternary structure. Ideally, the same solution could be found using a monomeric model. An ensemble of monomeric search models was created, using the structures of SAS-6 from *D. rerio* (2Y3V, 2Y3W [van Breugel et al., 2011] ~26% sequence ID), and *C. reinhardtii* (3Q0X [Kitagawa et al., 2011b] ~21% sequence identity), each containing the head-group, and ~10 residues of α3 (which forms the CC) (**Figure 2.4.1.i, A**). The sequence of DmSAS-6 was mapped onto these structures, and sidechains were clipped back to the last common atom using CHAINSAW [Stein, 2008]. The models then contained only atoms expected to exist in the DmSAS-6 structure. Molecular replacement was carried out in PHASER [McCoy et al., 2007], and, as the Matthews coefficient suggested three monomer copies in the ASU, PHASER was tasked with finding two copies of the model, but expecting three. The best solution was in spacegroup P2, as expected from the analysis in POINTLESS, and placed two copies of the model in a two-fold NCS dimer. The symmetry axis of this dimer was orthogonal to the crystallographic two-fold axis along the *b* cell edge, thus explaining the strong NCS peak seen in the self-rotation function (**Figure 2.3.3.i, A**). PHASER indicated that this solution stood out above all others, and had a strong TFZ score of 10. Additionally, maps calculated from this solution displayed positive density suggesting that

the absent CC should be filled in. (**Figure 2.4.1.i, B**). No density for a further monomer was evident. Subsequent refinement of the structure confirmed that this solution was correct.

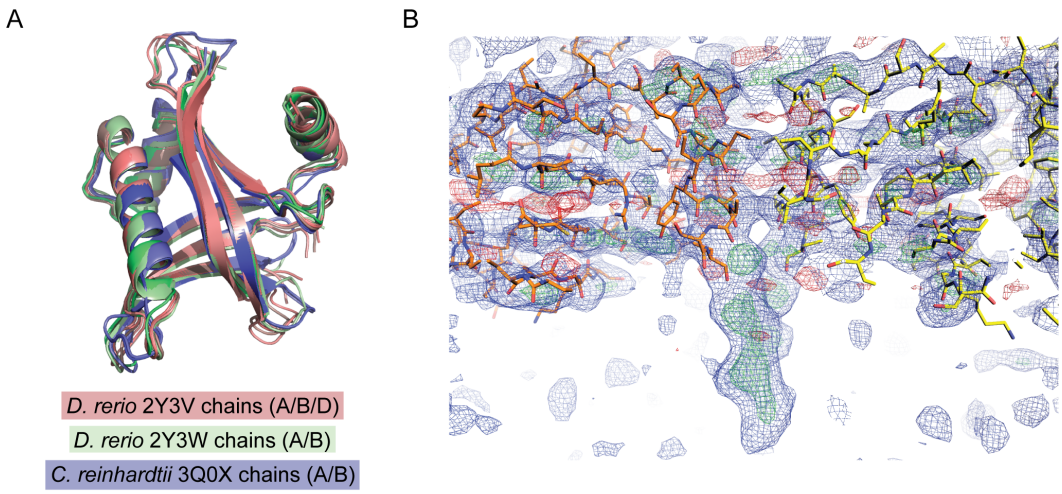


Figure 2.4.1.i: Molecular replacement solution of DmSAS-6$^{1\text{-}216}$(F143D) (**A**) The ensemble of search models, created from orthologous headgroup structures, with sidechains clipped back to the last common atom of DmSAS-6, according to MSA analysis. (**B**) Two copies were placed into the ASU, forming an NCS dimer. The chains are shown as yellow and orange sticks. The resultant density is shown. The 2Fo-Fc map is shown in blue, contoured at 1.0σ. The Fo-Fc difference map is contoured at 3.0σ, with positive density in green, and negative density in red. Positive density is evident for the coiled-coil, which was absent in the search models. Its position is consistent with the other structures of SAS-6 CC-dimers, 3Q0X and 2Y3W.

2.4.2: Refinement of the SAS-6$^{1\text{-}216}$(F143D) structure

Refinement of the SAS-6$^{1\text{-}216}$(F143D) structure was challenging due to the relative disorder of the CC domain. The program autoBUSTER appeared to deal with this issue the best in terms of map quality and was therefore used for refinement of the structure.

Initial steps focused on refining the head-groups. In order to reduce model bias, BUCCANEER [Cowtan, 2006] was used to autobuild the DmSAS-6 sequence into the maps. At this stage the maps were noisy due to the poor initial phases, and high solvent content of the crystal. Therefore solvent-flattening was carried out using CCP4-PARROT [Cowtan, 2010], to modify the density prior to autobuilding. These steps were iterated several times, after which the strongest

density was explained by the model, and further building was carried out manually using Coot [Emsley and Cowtan, 2004]. Refinement at this stage was aided using NCS restraints, (under the "-autoncs" argument), until the maps were of sufficient quality to distinguish differences between regions of each chain, mainly in the coiled-coil.

The CC was relatively disordered, which was probably exacerbated by the way the molecule packed into the lattice (shown later in **Figure 2.5.1.i, B, C**). The crystal consisted of rigid lattice layers in the *ab* plane, formed by the packing of the head-groups. These layers contained roughly 70% of the protein atoms, however they were quite separated. The only connections joining the layers were isolated CCs from each SAS-6 dimer, the tips of which contact the head-groups of the opposite lattice layer. This results in large solvent channels, which is consistent with the Matthews coefficient analysis (**Table 2.3.3.ii**), which predicted a relatively large solvent content of 66.86% for an ASU containing only 2 monomers.

Initially, the model had only ~10 residues of CC, however the space between lattice layers could be explained by a length of CC consistent with that which was missing from the model (~45 residues). Close to the head-groups, the CC was more ordered, and the density could initially be built into manually. This brought out additional density, however the density weakened further from the head-groups as the CC became more isolated and disordered. It was likely that some density was being lost due to the bulk solvent correction. The CC was gradually completed with the aid of a solvent mask, and by manually modelling in lengths of alpha-helix. During this process it became evident that the CC deviated slightly towards the B chain head-group. The CC helices of chains A and B are therefore not symmetrical, and so NCS restraints could no longer be used to aid refinement. In the final model, the CC density is only continuous at lower map contours (~≤0.6σ) (**Figure 2.4.2.i, A**). The decision was made to model the CC in its entirety in

order to join the lattice layers, and in accordance with *a priori* knowledge of its appearance and location. The relative disorder of the CC domain can be seen by the relatively high B-factors of this region with respect to the rest of the structure (**Figure 2.4.2.i, B**). The inclusion of the CC did not harm refinement statistics, and the R_{work}/R_{free} values eventually converged at 21.9/22.7 as detailed in **Table 2.4.2.i**.

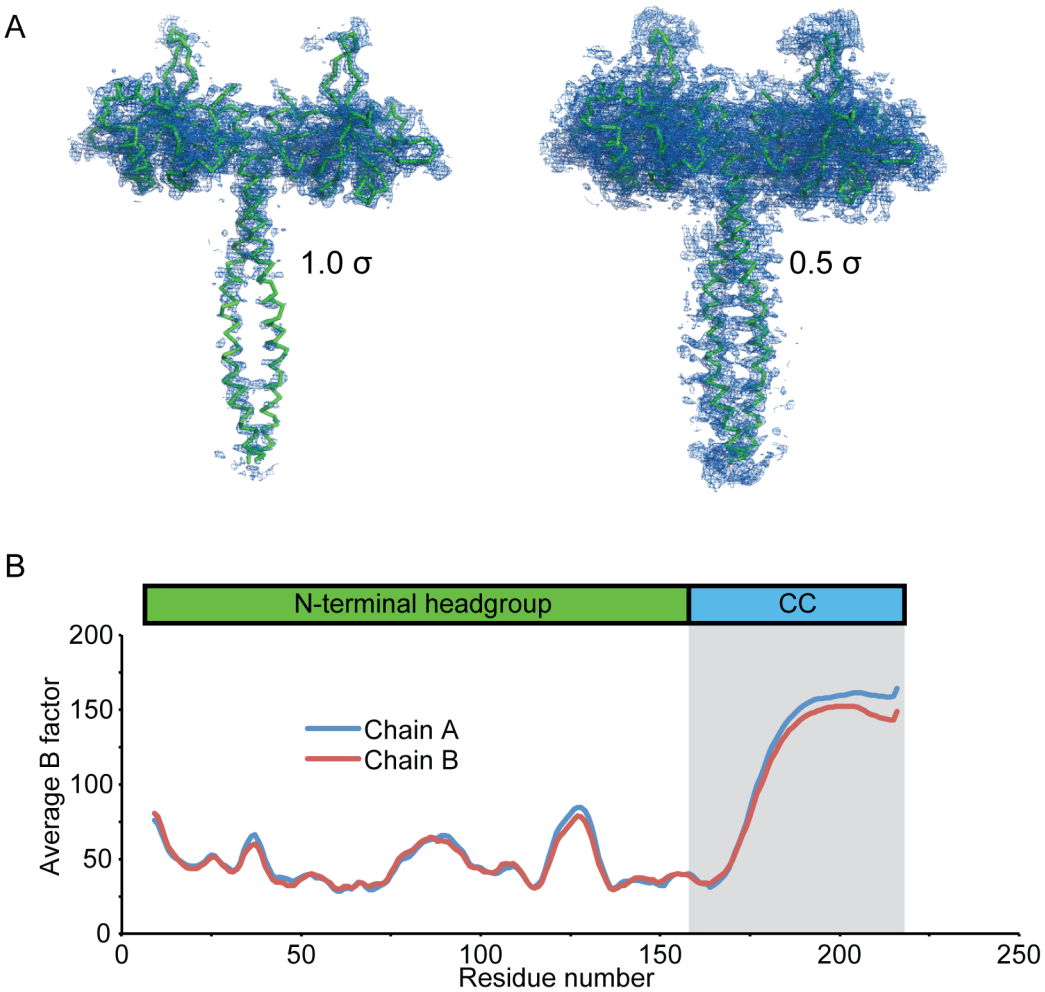


Figure 2.4.2.i: Refinement of SAS-$6^{1\text{-}216}$(F143D) (**A**) Two views of the final SAS-$6^{1\text{-}216}$(F143D) structure shown in green ribbon format. On the left, the 2Fo-Fc map is contoured at 1.0σ, whereas on the right it is contoured at 0.5σ where the CC density is continuous. (**B**) The disorder of the CC is illustrated in a B-factor plot over the length of the structure. The main chain average B-factors are shown for each residue, and are similar for chain A and B. CC residues have much higher B-factors than the head-group residues.

	DmSAS-6$^{1-216}$ (F143D)
Resolution Range (Å)	41.43-2.92 (3.12-2.92)
R_{work}/R_{free} /% size test set	21.9/22.7/5.00% (24.5/26.8/5.67%)
No. Reflections working set	14975 (2678)
No. Reflections test Set	788 (161)
Number of atoms (non-H)	3364
Waters	86
rmsd from ideal values: bond length (Å)/angles(˚)	0.007/0.95
Average B-factor	69.81
Correlation coefficient Fo-Fc/Fo-Fc free	0.853/0.854
Molprobity Score	2.03 - 99th percentile

Table 2.4.2.i: Refinement statistics for the DmSAS-6$^{1-216}$(F143D) structure. Values in parentheses are for the highest resolution shell.

2.5: Analysis of the DmSAS-6 structure

2.5.1: DmSAS-6 has a pseudo-orthorhombic crystal packing and forms a non-crystallographic spoke dimer

Two monomers of SAS-6 exist in the asymmetric unit, and as with other SAS-6 orthologues, can be seen to associate via the coiled-coil interface, generating an NCS dimer (**Figure 2.5.1.i, A**). Analysis by the Proteins, Interfaces, Structures and Assemblies (PISA) server scores this as the most biologically relevant interface, with a Complex Significance Score (CSS) of 0.996 and burying 1623.7 $Å^2$.

The coiled-coil deviates slightly from the NCS axis, introducing minor asymmetry between the chains. Apart from the coiled-coil, the two chains are remarkably similar. Indeed, omitting any coiled-coil beyond 2 heptad repeats, the A and B monomers superimpose with a very low RMSD of 0.014 Å across 988 atoms (**Figure 2.5.1.i, B**). The only other visible deviation is in the flexible β5- β6 loop, which is involved in crystal contacts.

The crystal has a pseudo-orthorhombic packing. The crystallographic two-fold symmetry operation along the *b* axis (**Figure 2.5.1.i, C**), results in 2 copies of the NCS dimer being present in the unit cell (**Figure 2.5.1.i, D**). As well as the CC NCS axis (which is orthogonal to the *ab* plane), chain A can also be mapped onto the symmetry copy of chain B via a two-fold screw operation along the *a* axis (**Figure 2.5.1.i, E, F**). The three orthogonal two-fold operations observed in the self-rotation function analysis (**Figure 2.3.3.i, A**), are therefore all explained by the observed crystal packing.

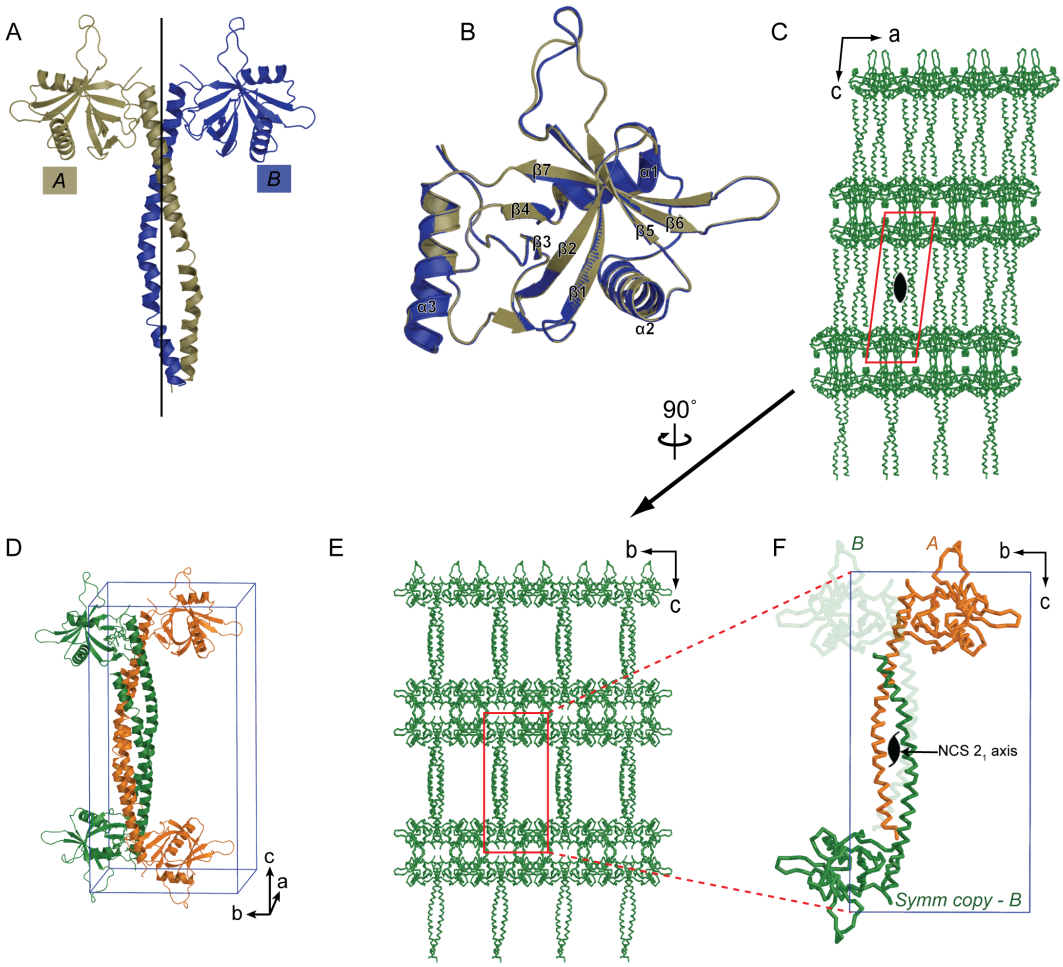


Figure 2.5.1.i: Crystallographic and non-crystallographic symmetry in the DmSAS-6$^{1-216}$(F143D) crystal. (**A**) The biologically significant NCS dimer, associated by the CC, with the two chains coloured gold and blue. The NCS axis is shown as a black line. (**B**) Superimposition of the head-groups of chains A and B, resulting in an RMSD of 0.014 Å over 988 atoms. (**C**) The crystal lattice viewed along the twofold crystallographic symmetry axis, *b*, with SAS-6 monomers shown in green ribbon format. (**D**) The unit cell contains two NCS dimers, related by the twofold crystallographic symmetry operation. The A and B chains are shown in orange and green respectively. (**E**) The NCS two-fold symmetry operation is also evident along the *a* axis. (**F**) Chain A is mapped onto chain B via a twofold screw axis along *a*.

2.5.2: The DmSAS-6 N-CC orientation is consistent with orthologues thought to form a "canonical" cartwheel

As described in **Section 2.2**, SAS-6 orthologues can oligomerise by two mutually exclusive, C-C and N-N interfaces [van Breugel et al., 2011; Kitagawa et al., 2011b]. Continuous oligomerisation via these interfaces is compatible with the assembly of a nine-fold symmetric ring, which would form the centriolar central hub. I refer to this arrangement as the "canonical cartwheel" arrangement. An additional study has recently shown that *L. major* SAS-6 (LmSAS-6) actually crystallised in this canonical arrangement [van Breugel et al., 2014]. Importantly, a study by [Hilbert et al., 2013] provided structures of the *C. elegans* SAS-6 (CeSAS-6) N-CC dimer and showed that a different assembly can be formed by the same two interactions. The architecture of the CeSAS-6 head-group remains the same as in other orthologues, however, the CC projects at a different angle with respect to the head-group – a feature common to two separate crystal forms. Remarkably, continuous oligomerisation of CeSAS-6 leads to a 4.5-fold symmetric spiral, rather than the closed circle compatible with the other published structures. Two turns of this spiral, or a double interlocked spiral, would generate a nine-fold symmetric structure. This alternative structure could be very significant, as EM studies appear to show that worms lack a canonical cartwheel structure (with nine spokes emanating from a central hub) and instead form a central, apparently hollow tube [Pelletier et al., 2006].

Interestingly, although *Drosophila* cartwheels clearly have a central hub structure in EM images, it has proven very difficult to image the nine spokes emanating from the central hub in any one plane – although a few spoke structures are usually visible, e.g. [Roque et al., 2012; Sharp et al., 1999]. I therefore wanted to analyse the geometry of the DmSAS-6 N-CC dimer in order to assess whether *Drosophila* centrioles are built in a canonical fashion, or whether they adopt an alternative arrangement.

Sequence conservation was mapped onto the DmSAS-6 structure, and this showed that conservation was mainly reserved to the PISA motif, which would lie at the N-N interaction interface (**Figure 2.5.2.i, A**). However, it could also been seen that conservation was enriched around strands β3 and β4, and at the top of a3, which together form the N-CC interface. To assess the structural similarity of the DmSAS-6 structure with other orthologues, the DmSAS-6 N-CC dimer was superimposed against the SAS-6 N-CC dimers from *D. rerio* (2Y3W), *C. reinhardtii* (3Q0X) *L. major* (4CKN, 4CKP) and *C. elegans* (4GFA, 4GFC), comparing the positions of mainchain atoms.

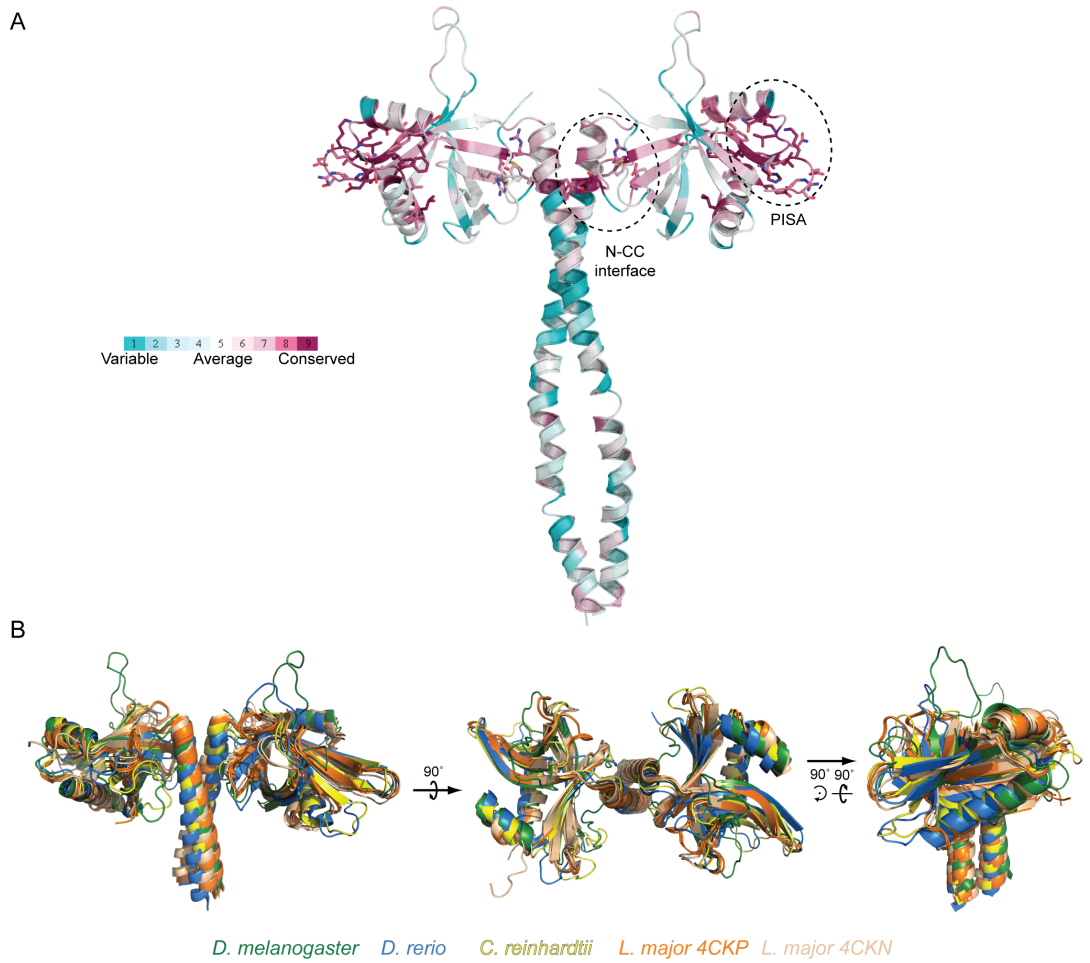


Figure 2.5.2.i: Comparison of DmSAS-6 to canonical SAS-6 orthologues. (**A**) The DmSAS-6 structure coloured according to the Consurf conservation score, generated from an alignment of 42 orthologues. Conservation localises mainly to the PISA motif and residues involved in the N-N interface. (**B**) Overlays of the DmSAS-6 CC dimer with solved SAS-6 orthologues from species with canonical cartwheels: *D. rerio*, *C. reinhardtii* and *L. major*.

DmSAS-6 only aligned well, across both monomers, with *D. rerio*, *C. reinhardtii* and *L. major* structures (**Figure 2.5.2.i, B**) (average pairwise RMSD = 1.85±0.34 Å over 601 ± 26 atoms). Conversely, although it was possible to align the headgroups in isolation, the DmSAS-6 N-CC dimer did not align with the CeSAS-6 N-CC dimer. The N-CC geometry of DmSAS-6 was therefore only consistent with "canonical" orthologues. All of these orthologues have the same N-CC orientation, despite having been crystallised in a multitude of different crystal forms.

To further investigate what defines this N-CC angle, the DmSAS-6 interface was examined (**Figure 2.5.2.ii, A**), along with the orthologues from *D. rerio*, *C. reinhardtii* and *L. major*, **(Fig 2.5.2.ii, B, C**). In agreement with the Consurf analysis (**Figure 2.5.2.i, B**), the N-CC orientation appears to be determined by a network of synergistic hydrophobic, polar, and cation-pi interactions, that tether β3 and β4 to the CC. The β3-β4 loop also appears to contact α3 of the opposing monomer. Analysis of all the "canonical" orthologues (**Figure 2.5.2.ii, B, C**) shows that the exact residues involved in these interactions are not always conserved, meaning an overarching mechanism is difficult to define, and therefore structural analysis appears to be the best way to confirm the N-CC orientation. Some features appear more conserved than others, such as the cation-pi interaction (R61 and Y164 for *Drosophila*), which is centrally located in the N-CC interface, linking the CC to β3 (**Figure 2.5.2.ii, B**).

Conversely to the canonical orthologues, *C. elegans* SAS-6 adopts a unique set of N-CC interactions that locks in its alternative conformation (**Figure 2.5.2.ii, D**), where the head-groups are twisted with respect to the CC. Additionally, CeSAS-6 has a beta-branched residue, valine, at the pivot between the head-group and the CC. It has been shown that the steric bulk of this residue would likely not allow a canonical N-CC orientation to be adopted by CeSAS-6 [Hilbert et al., 2013]. The consequence of the N-CC orientation differences is shown in **Figure 2.5.2.ii, E**. The

different N-CC interface in CeSAS-6 causes its CC to be displaced by ~30˚ with respect to the CCs of canonical orthologues.

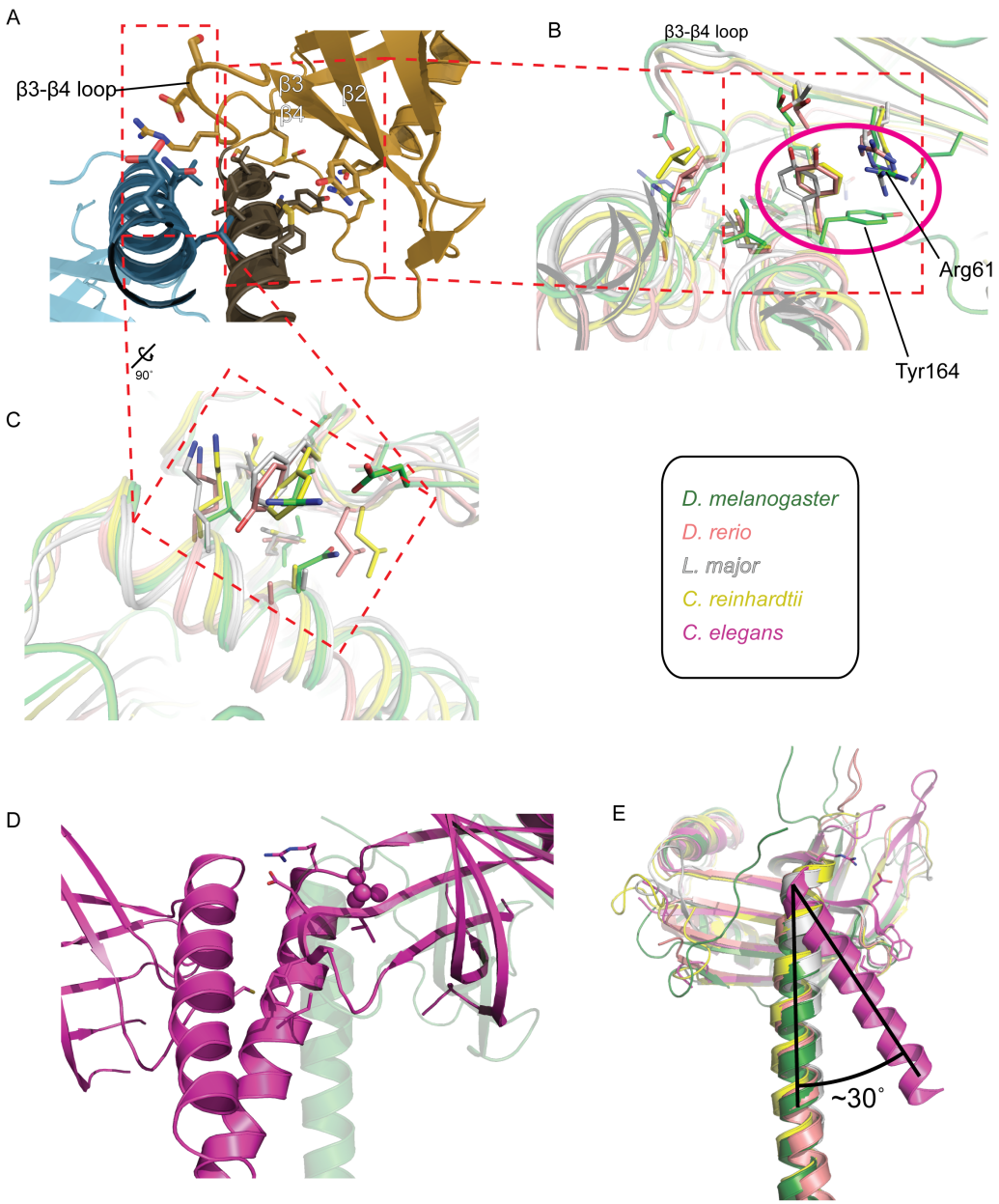


Figure 2.5.2.ii: Analysis of the N-CC interface (**A**) Closeup of the DmSAS-6 interface consisting of hydrophobic, polar and cation-pi interactions. (**B**+**C**) Overlays of different species at different regions of the interface, coloured as per the legend. The mainchain is shown as a ribbon and important residues are shown as sticks. (**B**) Closeup of β3 interacting with the CC, via a cation-pi interaction (circled). The Drosophila residues are labelled. (**C**) Closeup of the β3- β4 loop interacting with the opposing monomer's α3. (**D**) The N-CC interface of CeSAS-6 adopts a different orientation involving a different set of interactions. The existence of a valine (pink spheres) sterically hinders a canonical conformation. The DmSAS-6 structure is shown in translucent green for comparison (the N-terminal head-groups of both structures have been aligned). (**E**) Alignment via the head-groups of one monomer each, from all the orthologues labelled in the legend. While most of the CC helices align as well, the different N-CC interface in *C. elegans* causes a displacement of the helix by ~30˚ with respect to helices from canonical orthologues.

Taken together, my analysis suggests that the DmSAS-6 N-CC orientation is compatible with oligomerisation into a flat cartwheel hub, as is observed in the other “canonical” orthologues.

2.6: *In vivo* analysis of the N-N interaction

Results thus far suggested that DmSAS-6 (here simply SAS-6) associates via a canonical CC, and N-N interaction to form a cartwheel central hub. In order to fully assess the importance of SAS-6 self-association, experiments were designed to investigate the function of the N-N interaction in *Drosophila*. SAS-6 null flies display an uncoordinated, or "unc" phenotype; this is because SAS-6 is an essential centriole protein, so loss of SAS-6 leads to an absence of centrioles, and therefore primary cilia, which are required in type I sensory neurones to enable motor function [Basto et al., 2006; Kernan, 2007]. I wanted to test whether GFP-SAS-6(WT) and GFP-SAS-6(F143D) could rescue this centriole loss phenotype.

2.6.1: Design of constructs and generation of transgenic lines

Previous studies [Peel et al., 2007] showed that transgenic DmSAS-6 was functional when N-terminally fused to GFP. I therefore designed two constructs, GFP-SAS-6(WT) and GFP-SAS-6(F143D) to test the effect of knocking out the N-N interaction in *Drosophila* (**Figure 2.6.2.i, B**). cDNA encoding SAS-6 was PCR amplified so that the SAS-6 ORF was flanked by attB recombination sites. This allowed transfer of the insert via the Gateway system (Life Technologies) into a pDONR donor vector, using an attBP reaction (**Appendix 7.1.2**). The insert was then transferred into a pUbq N-terminal GFP DEST vector using an attLR reaction to form a pUbq_GFP-SAS-6 plasmid (**Figure 2.6.1.i, A**). In this plasmid the N-terminal GFP-SAS-6 fusion protein is placed under the control of the Ubq-promoter that drives medium levels of expression in all *Drosophila* tissues. This plasmid also has transposase sites, allowing the generation of stable transgenic *Drosophila* lines via P-element mediated transformation [Roberts, 1998].

The plasmids were injected into a w^{1118} (white-eyed) *Drosophila* strain, allowing for selection of transformants via the red eye colour conferred by the w^{+} gene in the plasmid (injection and

transformant screening was performed by BESTGENE inc, USA). Multiple transgenic lines were created for both constructs. The fly husbandry, specimen preparation, and microscopy analysis presented in **Sections 2.6.2-2.6.4** were carried out by my collaborator in the lab, Dr Nadine Muschalik.

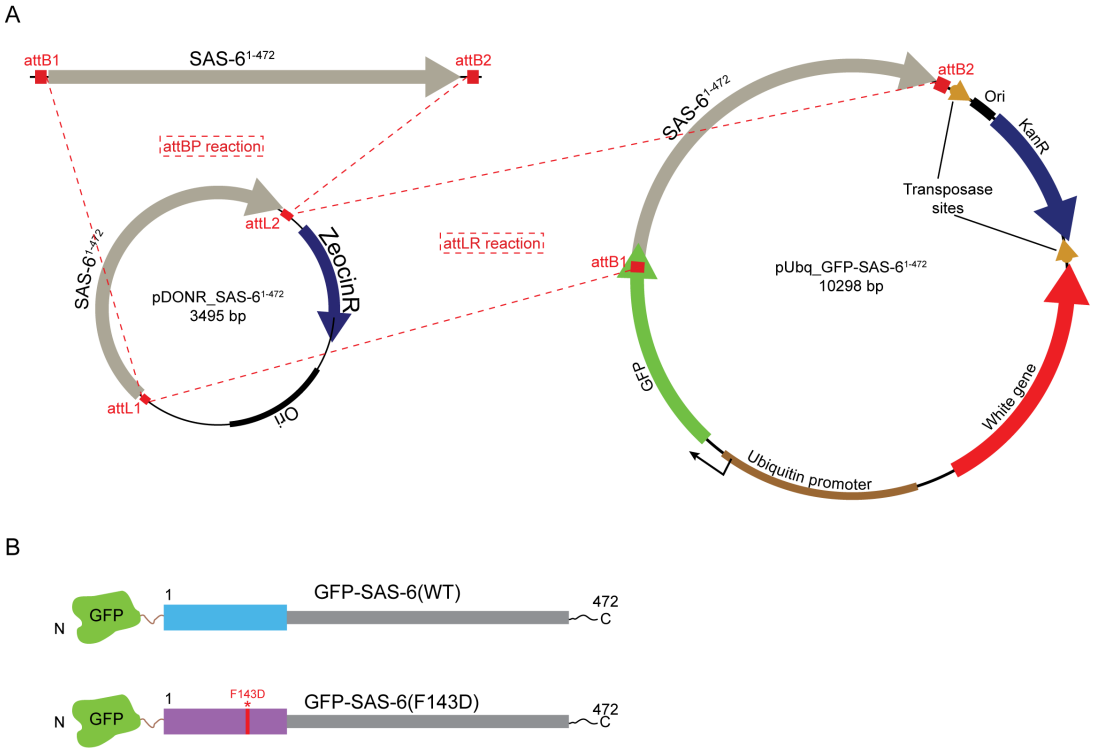


Figure 2.6.2.i: Design of SAS-6 transgenic constructs. (**A**) Generation of the pUbq_GFP-SAS-6 vector, via a pDONR_SAS-6 vector, using gateway cloning. The GFP-SAS-6 transgene is transferred into the *Drosophila* genome, along with the ubiquitin promoter, and white gene, via P-element mediated transformation. (**B**) Schematics of the GFP-SAS-6 fusion constructs used for *in vivo* experiments.

2.6.2: SAS-6(F143D) exerts a dominant negative sterility phenotype

When handling the transgenic lines expressing GFP-SAS-6(F143D), we immediately noticed that although these flies were not detectably uncoordinated (and so presumably had centrioles, which was later confirmed to be the case, see **Section 2.6.4**), the transgene caused dominant female sterility. All females carrying even a single copy of the transgene laid embryos that arrested very early in development, even though these females had two copies of the endogenous WT SAS-6 gene. This meant that these lines were difficult to work with: in each generation GFP-SAS-6(F143D)/+ males had to be selected and mated to WT females in order to maintain the lines, and it was impossible to generate homozygous stocks. We suspect that this dominant negative effect occurs because GFP-SAS-6(F143D), which is overexpressed by ~5-10-fold compared to the endogenous WT protein (confirmed by Western blotting – not shown), acts as an oligomerisation "terminator" (**Figure 2.6.2.i**). Presumably, somatic cells can tolerate the presence of SAS-6(F143D) because it only slows down the oligomerisation of WT SAS-6 into a complete cartwheel and therefore slows centriole biogenesis. This slowing of centriole duplication is tolerated in most somatic cells, but not in early syncytial embryos where centriole duplication has to occur every ~8-10 mins (see legend to **Figure 2.6.2.i**). It has been shown previously that these early embryos arrest very early in development if they cannot make centrioles and centrosomes [Stevens et al., 2007].



Figure 2.6.2.i: A model for the dominant negative effect of SAS-6(F143D). Wild-type SAS-6 is shown in green, SAS-6(F143D) in purple. Wild-type SAS-6 can oligomerise via the CC and N-N interaction to form the cartwheel central hub, while SAS-6(F143D) has a weakened N-N interface, so can oligomerise only via the CC. When both proteins are present, SAS-6(F143D) can form heterodimers with WT SAS-6 that can bind into growing hubs via the WT N-N interface, but the mutant protein cannot support further oligomerisation, so oligomerisation is effectively terminated. If the WT:F143D heterodimer could ultimately dissociate, this would allow a WT dimer to come in and promote further oligomerisation. In this way SAS-6(F143D) would slow hub assembly in the presence of the WT protein, but would not block it entirely. This may explain why the mutant protein only has a strong dominant negative effect in early embryos, where centrioles are required to assemble in only a few minutes. In somatic cells, centriole assembly can presumably occur over a longer time-frame, allowing sufficient time for WT homo-oligomers to ultimately assemble a cartwheel structure. This observation potentially has important implications for the molecular mechanism of cartwheel assembly.

2.6.3: Only GFP-SAS-6(WT) can localise to centrioles in the presence of endogenous SAS-6 in embryos.

Wild-type embryos carrying one copy of either transgene were fixed, and immunostained in order to assess the ability of these constructs to localise to centrioles. Embryos were fixed and stained as previously described [Stevens et al., 2009]. To detect the presence of the GFP-SAS-6 proteins, a GFP booster antibody was used (Chromotek). To identify if these SAS-6 foci were centriolar, the embryos were costained for the PCM protein Cnn, using sheep anti-Centrosomin, (Cnn, 1:1000), [Lucas and Raff, 2007]. Secondary antibodies were used at 1:1000, and were conjugated to either Alexa Fluor 488 or Alexa Fluor 568 (Life Technologies). DNA was visualised using Hoechst33258 (Life Technologies) at 1:5000. In embryos expressing GFP-SAS-6(WT), many well defined nuclei were observed (**Figure 2.6.3.i, C**), with different embryos at various cell cycle stages. Pairs of foci were associated with each nucleus and consistently co-stained for Cnn, Asl, and GFP-SAS-6(WT) suggesting that these were centrosomes. (**Figure 2.6.3.i, A, A', A''**). Conversely, in embryos expressing GFP-SAS-6(F143D) there were few, and often poorly-defined nuclei (**Figure 2.6.3.i, C'**), and across all observations, DNA appeared to be in an amorphous, condensed state – possibly all embryos had stalled at prophase. For most observations, although Cnn foci could occasionally be witnessed in association with these nuclei, GFP-SAS-6(F143D) did not obviously co-stain these foci, or only stained them very weakly. This suggests that this construct did not localise efficiently to centrosomes. Indeed, it is possible that these foci represent acentriolar MTOCS, and centriole biogenesis cannot occur due to the dominant negative effects of GFP-SAS-6(F143D) as described in **Section 2.6.2**.

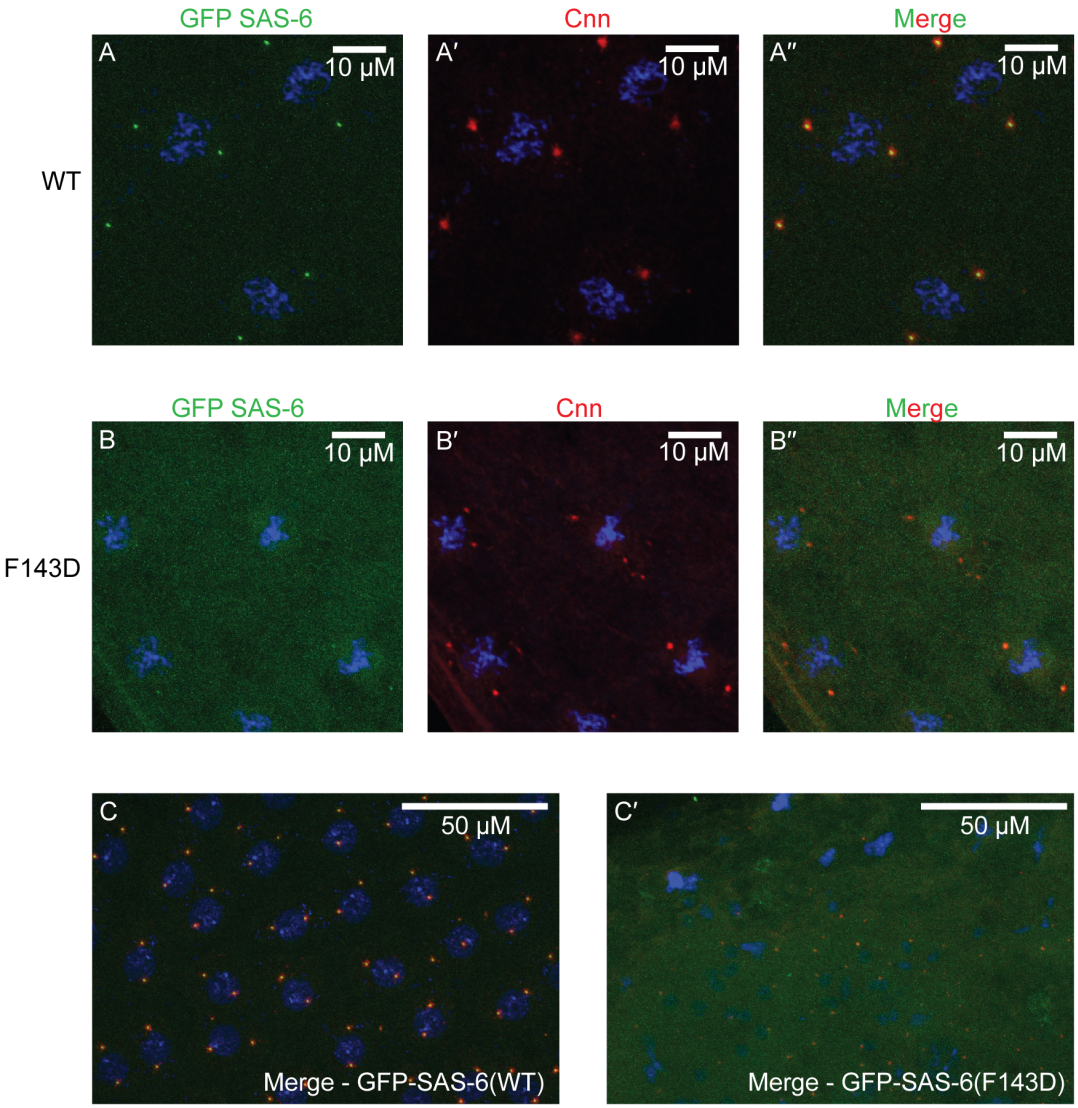


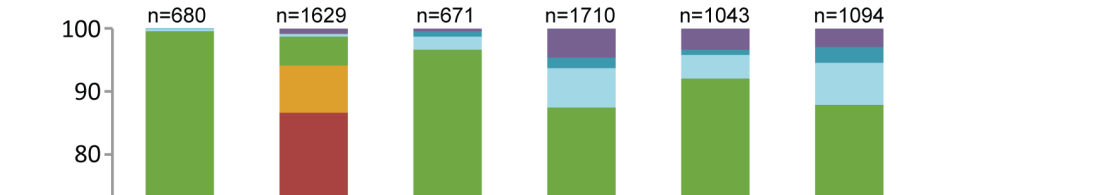
Figure 2.6.3.i: Localisation of GFP-SAS-6(WT) and GFP-SAS-6(F143D) in fixed wild-type embryos. In the micrographs shown here, GFP is costained with both the centrosomal protein Cnn (red) and DNA (blue) in ~prophase-prometaphase embryos. (**A**) GFP-SAS-6(WT) was seen to localise to pairs of foci, on either side of the nucleus. These foci can be seen to costain for the centrosomal protein Cnn. (**B**) GFP-SAS-6(F143D) did not obviously colocalise with Cnn, and instead a diffuse signal was observed. These images are ones in which the nuclei, and Cnn staining are particularly defined. (**C**) Typical merges for the WT (**C**) and F143D (**C′**) constructs. In the case of GFP-SAS-6(WT), there were numerous, well-defined nuclei. For the GFP-SAS-6(F143D), there were few, often poorly-defined nuclei.

2.6.4: GFP-SAS-6(F143D) appears to partially rescue the *Sas6*$^{-/-}$ centriole loss phenotype

The constructs were next assayed for their ability to rescue the centriole loss phenotype of *Sas6*$^{-/-}$ mutant flies. The GFP-SAS-6(WT) and GFP-SAS-6(F143D) were therefore crossed into a *Sas6*$^{-/-}$ background (*Sas-6*^{c02901}, [Peel et al., 2007]). In order to further assess the potential dominant negative effect of GFP-SAS-6(F143D), flies were also generated that contained one copy of GFP-SAS-6(F143D) in either a wild-type background (*Sas-6*$^{+/+}$), or a background with only one genetic copy of endogenous *Sas6*, (*Sas-6*$^{+/-}$).

In order to assay centriole loss, and assess whether it could be rescued with one genetic copy of either GFP transgene, a centrosome counting assay was performed. This assay involved staining brain cells for both the PCM marker, Cnn, and also the centriolar marker, Asl. In metaphase (where each cell should have two centrosomes), foci that stain for both proteins are counted as centrosomes. Brains were fixed and stained as previously described [Stevens et al., 2009], and results were acquired using 9-33 different brains.

In control flies (w^{67}) ~96% of mitotic brain cells had 2 centrosomes, while in *Sas6*$^{-/-}$ flies, ~85% of mitotic cells had no detectable centrosomes (**Figure and Table 2.6.4.i**). One copy of the GFP-SAS-6(WT) transgene robustly rescued this *Sas6*$^{-/-}$ phenotype (~88% of mitotic cells had two centrosomes), while the GFP-SAS-6(F143D) transgene partially rescued the phenotype (~24% of mitotic cells had two centrosomes). However, when allowed to eclose, these flies were uncoordinated, suggesting that the observed foci may not represent real centrioles (discussed in **Section 2.6.5**).



No. of Centrosomes	*Sas6* genotype	+/+	-/-	-/-	-/-	+/+	+/-
	transgene genotype			GFP-SAS-6(WT)	GFP-SAS-6 (F143D)	GFP-SAS-6 (F143D)	GFP-SAS-6 (F143D)
>4	% of cells	0.0	0.7	0.5	4.7	3.2	2.7
4	% of cells	0.2	0.1	0.8	1.8	1.3	2.9
3	% of cells	0.2	0.4	2.2	6.1	3.6	6.3
2	% of cells	95.9	4.7	88.1	24.1	78.7	65.3
1	% of cells	2.9	7.7	5.2	26.7	9.7	15.8
0	% of cells	0.9	86.4	3.3	36.6	3.6	7.0

Figure and Table 2.6.4.i: Quantification of centrosome number in third instar larval brain cells expressing GFP-SAS-6 transgenes. For six different genotypes, the chart displays the proportion of brain cells in which the indicated number of centrosomes were counted. Cells from WT animals (*sas6*$^{+/+}$) have predominantly two centrosomes per mitotic cell. Loss of *sas6* (*sas6*$^{-/-}$) results in most cells having no centrosomes. While the GFP-SAS-6(WT) transgene can return these animals to the wild-type phenotype, N-N deficient GFP-SAS-6(F143D) only provides a partial rescue. However, in the presence of WT endogenous *Sas6* (*+/+ or +/-*), the GFP-SAS-6(F143D) transgene provides only a minor dominant negative phenotype, with a slight increase in the number of cells with one or no centrosomes, as compared to WT or rescue with GFP-SAS-6(WT).

I conclude that the N-N interaction is essential for robust centriole biogenesis, but that some centriole duplication (or at least the formation of some centriole-like structures) can occur even when the N-N interaction is dramatically weakened.

As expected (see **Section 2.6.2**), the dominant negative effect of GFP-SAS-6(F143D) observed in embryos was not as severe in third instar larval somatic brain cells, which divide much more slowly, and therefore have a much longer time in which to build a centriole. In flies that had two endogenous copies of *Sas6* and one copy of GFP-SAS-6(F143D), ~80% of cells had 2 centrosomes (**Figure 2.6.4.i**). When the proportion of WT SAS-6 protein was lowered by having only one copy of the WT *Sas6* gene, then only ~65% of cells had two centrosomes (Figure 2.6.4.i), strongly supporting the idea that GFP-SAS-6(F143D) is acting in a dominant negative manner, as depicted in **Figure 2.6.2.i**.

2.6.5: Perspectives on the SAS-6 *in vivo* experiments

Taken together, these results indicate that the N-N interaction is important for centriole biogenesis in *Drosophila*, but that it may not be absolutely essential. In somatic brain cells lacking endogenous WT SAS-6, the expression of GFP-SAS-6(F143D) appeared to partially rescue the loss of centrioles.

There are at least two possible explanations for these observations. First, perhaps the F143D mutation does not completely prevent the N-N interaction *in vivo*, so centriole assembly is inefficient rather than abolished. Second, perhaps the structure of the central cartwheel is robust enough (due to a combination of other protein-protein interactions), that it can still partially assemble even if the N-N interaction interface no longer contributes to cartwheel assembly. These two possibilities are not mutually exclusive.

Observations by [Kitagawa et al., 2011b] in *C. elegans* embryos showed that similar to *Drosophila*, N-N deficient SAS-6 could not localise to centrioles. However, they also showed

absolutely no rescue of RNAi depletion of endogenous SAS-6. Partial rescue of a SAS-6 null phenotype was however shown in *C. reinhardtii* by [van Breugel et al., 2011]. They showed that, similarly to *Drosophila*, cells expressing only N-N deficient SAS-6 could form some centrioles. However, as this only occurred in the minority of cases, it was concluded that the N-N interaction was essential for centriole biogenesis.

Lastly, it is important to point out that in this experiment, structures containing both Asl and Cnn were scored as centrioles – however it is not clear whether these foci contain proper centrioles. Indeed, overexpression of DmSAS-6 alone in embryos has been shown to cause SAS-6 based aggregates known as centriole-like structures [Peel et al., 2007]. Although not centrioles, they are able to recruit other centriolar proteins (Raff lab unpublished observations).

It is highly likely that the rest of SAS-6 provides a scaffold for the association of downstream centriole factors. The partial rescue phenotype exhibited by N-N deficient SAS-6 may highlight the importance of these downstream functions of SAS-6, which may be sufficient to form centriole-like-structures that are capable of recruiting downstream centriole proteins. EM analysis of these structures would be necessary to conclude whether they are in fact centrioles.

During embryogenesis, the N-N interaction appears to be essential for nuclear divisions to continue, and the presence of N-N deficient SAS-6 in this context leads to developmental arrest. GFP-SAS-6(F143D) did not obviously localise to centrioles, presumably because it cannot associate via the N-N interface. Additionally, it interfered with the ability of endogenous SAS-6 to form centrioles, possibly described by the mechanism presented in **Figure 2.6.2.i**. Interestingly, in *Drosophila* larval brain cells, this dominant negative effect appeared less severe.

It is puzzling that the importance of the N-N interaction diminishes at this stage of development, but presumably this may be due to with the speed at which centriole biogenesis must occur.

2.7 An atomic model of the *Drosophila* SAS-6 cartwheel.

At this point in my studies, a paper was published reporting a 38 Å resolution EM map of the centriolar cartwheel from *Trichonympha agilis*, a multiciliated protist symbiont of the termite gut [Guichard et al., 2012, 2013]. The basal bodies of *T. agilis* cilia have particularly long cartwheels, which enabled Guichard et al. to carry out sub-tomographic averaging to reconstruct the map, from intact basal body centrioles, which had been purified from *T. agilis* cells. The maps confirmed the "canonical" cartwheel model, showing that the cartwheel is composed from discrete, stacked rings. An additional feature was that the stacked layers associated in pairs via the convergence of the spokes, thus providing the first evidence of symmetry breaking in the cartwheel architecture (**e.g. Figure 2.7.2.i, B**). I decided to try to model *Drosophila* SAS-6 into the *T. agilis* EM model, to gain further insight into the nature of the cartwheel. Although it is not known for certain which features from the *Trichonympha* map are conserved in other species, the authors suggest that the length of CC in canonical SAS-6 orthologues (including *Trichonympha* and *Drosophila*) is roughly consistent at ~300 amino acids. Furthermore, the spoke length from the central hub to the "**A**" MT in electron micrographs has been reviewed as being roughly consistent across evolution at ~70 nm, regardless of exact SAS-6 CC length [Jana et al., 2014]. I therefore deemed it appropriate to use this map for the modelling of full-length DmSAS-6.

2.7.1: Oligomerisation of the central hub

No N-N interface structure was available for DmSAS-6, however, due to the conservation of the PISA motif, it was likely that the N-N interaction would be very similar to that of other orthologues. DmSAS-6 monomers were therefore superimposed onto a crystallographically observed N-N dimer from *D. rerio* (PDB entry 2Y3V, [van Breugel et al., 2011]), to generate a model for DmSAS-6, N-N (herein termed DmSAS-$6^{2Y3V_N\text{-}N}$). This model was then analysed (**Figure**

2.7.1.i, A). It superimposed with the original 2Y3V dimer with an RMSD of 1.54 Å over 157 Cα atom pairs. Residues defining the DmSAS-6$^{2Y3V_N\text{-}N}$ interface overlaid and were highly conserved or similar to those in 2Y3V, suggesting that the DmSAS-6 N-N interface would indeed be similar to that observed in other orthologues. Initially, a circular oligomer of DmSAS-6 (not shown) was generated by association of DmSAS-6 N-CC dimers, via the DmSAS-6$^{2Y3V_N\text{-}N}$ interface. This yielded an almost flat spiral with a periodicity of 9.5 dimers and a pitch of ~one head-group per turn (not shown). This suggested that the structure would be compatible with a nine-fold symmetric, flat ring.

Next, I generated an idealised nine-fold symmetric ring by rotating the DmSAS-6 CC dimer structure around an origin in 40˚ steps. (**Figure 2.7.1.i, B, C**). This method generated a new N-N interface for DmSAS-6 (DmSAS-6$^{insilico_N\text{-}N}$). In order to validate DmSAS-6$^{insilico_N\text{-}N}$, it was compared to the DmSAS-6$^{2Y3V_N\text{-}N}$ model. The low RMSD of 1.4 Å over 299 head-group Cα atom pairs suggested that this was a realistic interface for the purpose of this model (**Figure 2.7.1.i, D**). The nine-fold DmSAS-6 ring was then manually docked into the *Trichonympha* cartwheel EM map (EMD-2329) produced by [Guichard et al., 2013] (**Figure 2.7.1.i, E**). It is worth noting that at the time of my modelling, the LmSAS-6 central hub structure (4CKP) [van Breugel et al., 2014] had not yet been published.

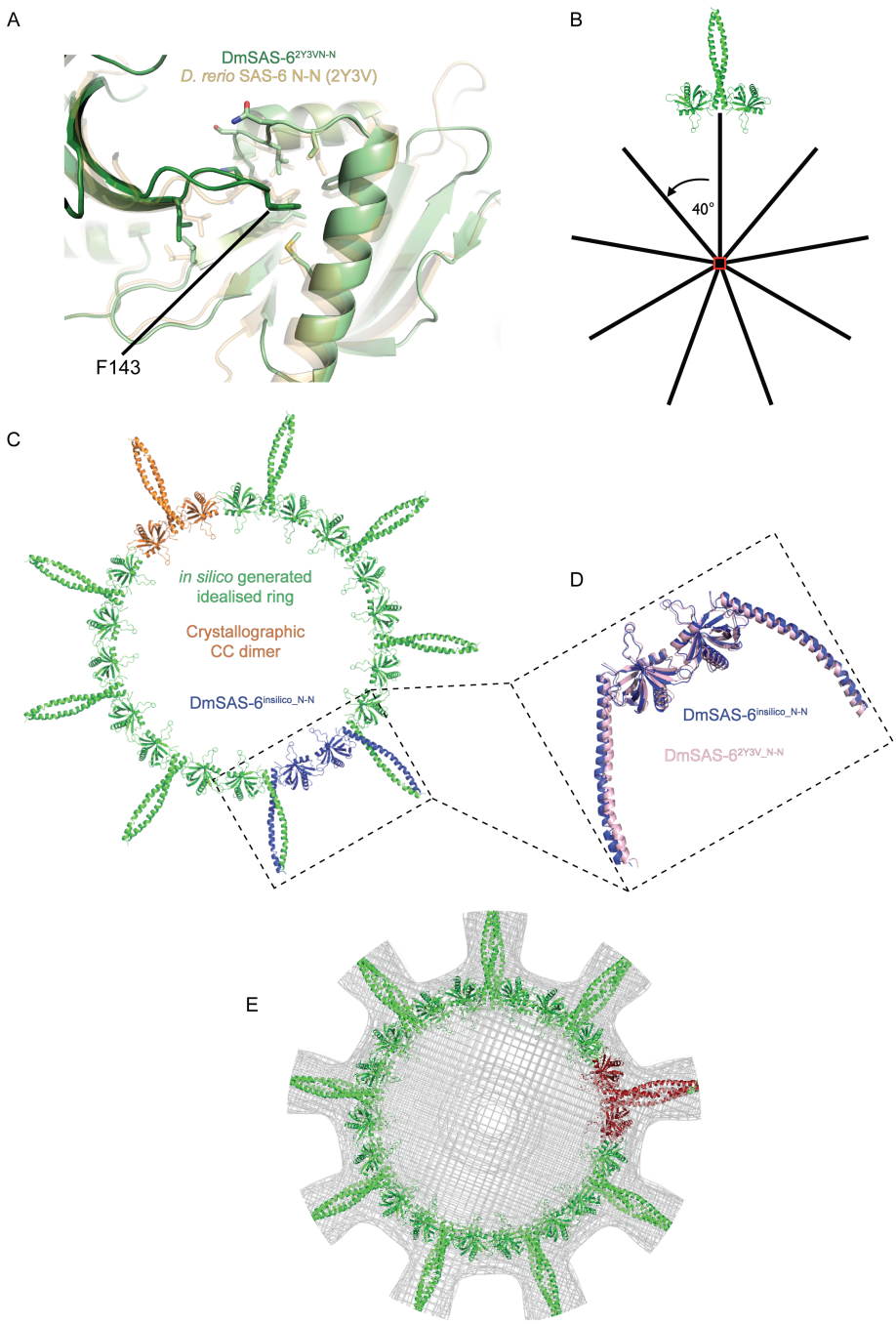


Figure 2.7.1.i: Building a DmSAS-6 cartwheel central hub. (**A**) Close-up of the DmSAS-6$^{2Y3VN-N}$ interface, generated by superimposition of DmSAS-6 monomers (dark and light green) onto the *D. rerio* crystallographic N-N interface observed in PDB entry 2Y3V (translucent beige). The DmSAS-6 F143D residue has been mutated back to the wild-type residue. Interface residues are shown as sticks, and are highly similar between the two models. (**B**) The DmSAS-6 central hub was generated by rotating the DmSAS-6 CC dimer in one plane, around an origin. (**C**) The idealised DmSAS-6 central hub. The N-N interface generated by this model (DmSAS-6$^{insilico_N-N}$) was compared in (**D**) to DmSAS-6$^{2Y3VN-N}$. The headgroups overlaid with an RMSD of 1.4 Å, suggesting only minor alterations were needed to generate a flat ring from the DmSAS-6$^{2Y3VN-N}$ interface. (**E**) The Idealised central hub fitted into the *Trichonympha* cartwheel EM density. One CC dimer is shown in red for reference.

2.7.2: Extension of the cartwheel spoke

I next wanted to extend my model by seeing whether I could fit the long coiled-coil spoke domain into the EM density. Due to the strong CC prediction given by the COILs server, the spoke was built as a canonical dimeric coiled-coil, extending one of the crystallographic DmSAS-6 N-CC dimers in the idealised central hub, docked into the EM density. In this way, one spoke dimer could be extended to full length using the EM density as a guide. In order to build the spoke, MSA and COILS analyses were used to assign a heptad repeat to the DmSAS-6 sequence (**Figure 2.7.2.i, A**). The heptad sequence was obvious in some separate segments, however these did not always run into each other in register, suggesting that stutters and stammers are present in the heptad repeat or that there could be localised unwinding of the CC in certain places. The sections for which the register was obvious were built by mutating the DmSAS-6 sequence onto models of canonical coiled-coil, such that the **a** and **d** position residues of the heptad faced inwards towards the CC axis. Where the register was more ambiguous I made a best estimate, and skips occurred in the heptad where these sections were included. The sections were generally built to overlap, such that they could be spliced together as fairly continuous CC, after manual positioning into the EM density. In the EM map, the spokes converge ~20 nm from the central hub. The density at this point becomes oval in shape (**Figure 2.7.2.i, B, right**), and so the CC was modelled such that it would only occupy half of this density, to leave room for the paired cartwheel spoke. However, it is very unclear what arrangement is taken at this point and whether other proteins may inhabit the density. Although COILs and PSIPRED predict that the CC would end at approximately residue 450-460 it is not certain where the CC would end, and so the entire protein, up to residue 472 was modelled as CC. Interestingly this length of CC reaches the junction in the EM map between the spoke and the “pinhead.” It is not known whether SAS-6 extends into the pinhead.

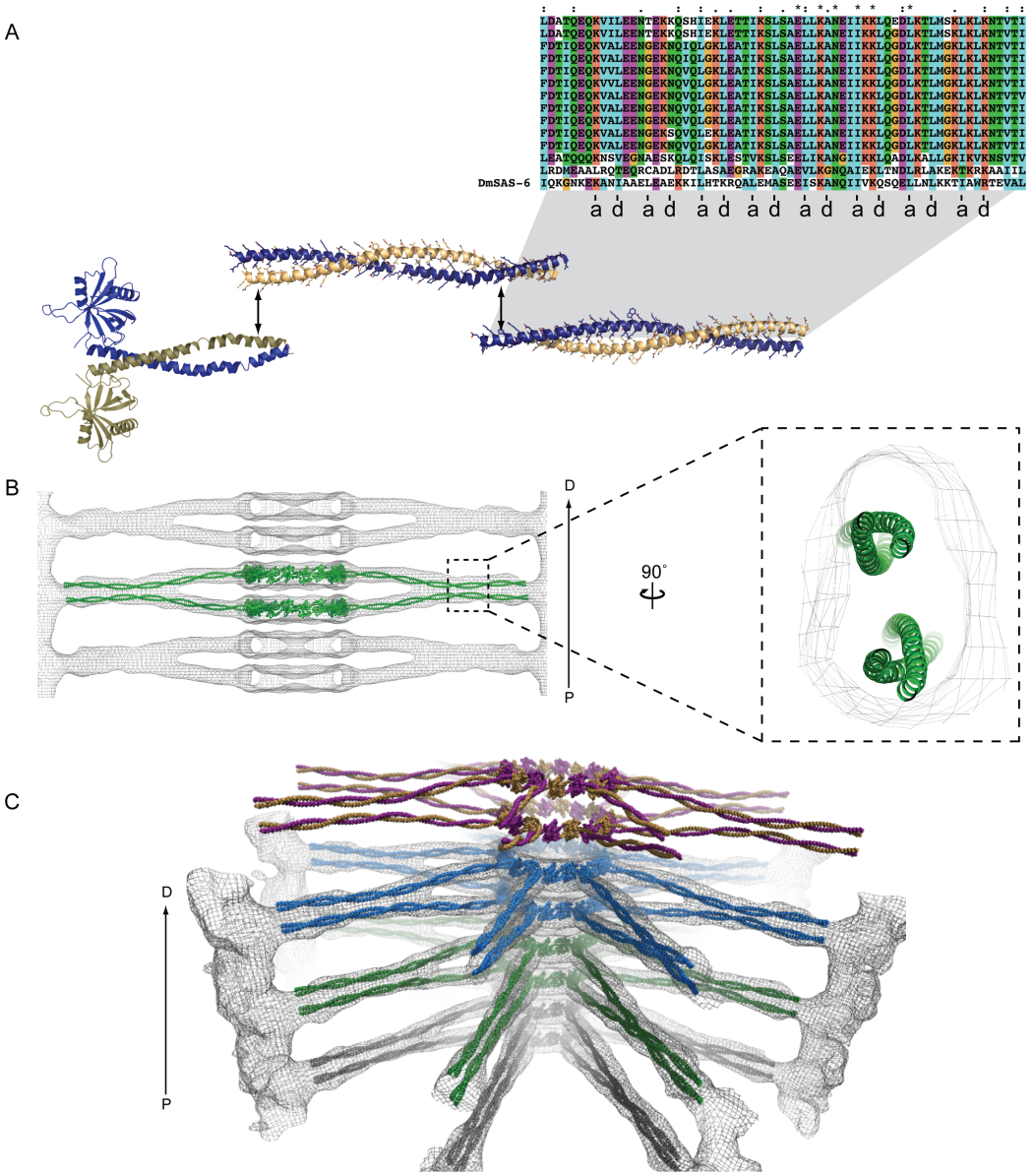


Figure 2.7.2.i: Building a full length DmSAS-6 cartwheel stack. (**A**) An example section of the MSA, used to assign the heptad repeat to the DmSAS-6 sequence. Segments of CC were modelled according to the heptad assignment, and spliced together to model full-length DmSAS-6. (**B**) Side-on view of a DmSAS-6 cartwheel pair docked into the *Trichonympha* cartwheel EM density (EMD-2329, Guichard et al., 2013). The proximal/distal axis of the cartwheel is indicated, as asymmetric features are evident in the map. The enlarged panel shows a cross section through the converged region of the spoke, exhibiting oval shaped density; note that the localisation of the CCs in this density is purely speculative, as the model was designed to accommodate the CCs within the density. (**C**) A stack of 4 cartwheel pairs in different colours. For reference, in the top pair, alternating DmSAS-6 monomers are coloured gold and magenta. Where shown, the EM density is contoured to enable visualisation of the hub, spokes and pinhead.

Once DmSAS-6 had been modelled to its full-length, it was superimposed onto each dimer in the idealised central hub model to make one complete nine-fold symmetric cartwheel. A second copy of this model could then be placed in the adjacent space in the EM map to make a model of a converged cartwheel pair. (**Figure 2.7.2.i, B, left**) Despite slight symmetry breaking features in the map along the proximal-distal axis, both cartwheels were well contained within the density, with paired spokes occupying the oval density (F**igure 2.7.2.i, B, inset**). The EM map contains three unique layers of cartwheel pairs. The cartwheel pair model was easily docked into these layers by translation of 1/3 along the map z axis (orthogonal to the central hub plane), (**Figure 2.7.2.i, C**).

2.7.3: Conservation analysis of SAS-6 in the context of the cartwheel model.

It is worth noting that each individual central hub is regularly separated from the hub either above or below it, and so neither individual hubs, nor hub pairs, appear to stack directly on top of each other. This makes sense, as the proximal/distal faces of SAS-6 (as docked into the model) are not particularly conserved (not shown), which would be expected if they formed an interface. Although the oligomerisation of the central hub is largely characterised, little is known about the spoke of SAS-6, and what its downstream functions might be. In order to gain more insight, MSA analysis was mapped onto the full-length DmSAS-6 structure to find conserved regions within the context of their location in the cartwheel model (**Figure 2.7.3.i**). This analysis highlighted three main regions along the spoke with enriched conservation. Both the “neck” region of the spoke, and an area near the very C-terminus were weakly conserved. Most notably, a central region in the converged region of the spoke was the most strongly conserved. It is possible that this region may be the binding site of a partner protein, or that its behaviour may be particularly important in determining the organisation of the cartwheel spoke at this point.

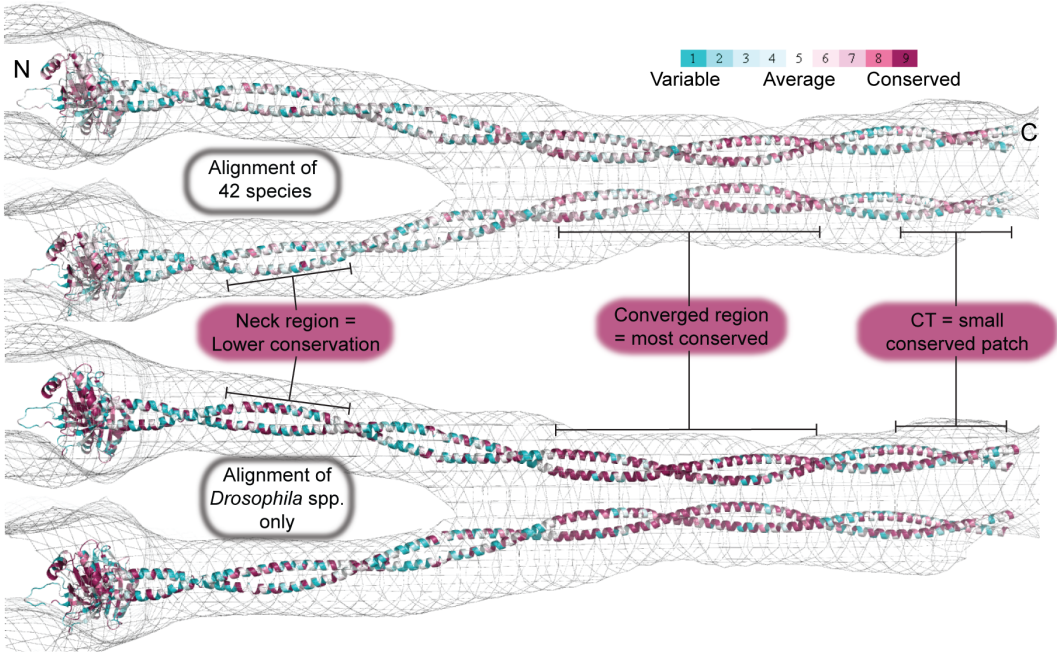


Figure 2.7.3.i: Conservation analysis across full-length DmSAS-6. Sequence conservation was mapped onto full length DmSAS-6 (shown docked into the EM map, and coloured according to the Consurf conservation score). The top structure illustrates two SAS-6 dimers, with mapped conservation from a large alignment of canonical SAS-6 orthologues, while the bottom structure uses an alignment of just *Drosophila* species sequences. Three regions of enriched conservation are highlighted, however the conserved area that stands out the most lies at the centre of the converged region of the spoke.

2.8: The conserved section of the SAS-6 coiled-coil

Due to the presence of higher sequence conservation in the converged part of the DmSAS-6 CC, I decided to recombinantly purify this region in order to characterise it *in vitro*.

2.8.1: Design of SAS-6 CC fragments

Constructs of the SAS-6 CC were designed, spanning the conserved region of the CC, and testing several N and C-terminal boundaries (**Figure 2.8.1.i, B**). I thought it possible that the conserved region of the SAS-6 CC might serve as a binding site for other centriole proteins and so I designed constructs that would allow me to eventually use these proteins in binding studies. Microscale thermophoresis (MST) is a relatively new technique for characterising protein-protein interactions that has several advantages mainly in saving time and material. This technique requires fluorescent labelling of one component, so these constructs incorporated a C-terminal GFP tag, via cloning into a pWALDO vector for bacterial expression (**Figure 2.8.1.i, A**). Wild-type GFP has a tendency to dimerise at high concentrations, with a Kd of ~110 µM, however this behaviour can be entirely abolished by the introduction of an A206K mutation [Zacharias et al., 2002]. In order to avoid any artefactual GFP-attributable dimerisation in MST experiments, I modified the pWALDO vector to introduce the GFP A206K mutation via site-directed mutagenesis (Quickchange, Agilent).

During the course of my studies I noticed that *Drosophila* codon optimised cDNAs often did not express well in bacteria when cloned into vectors without an N-terminal tag to initiate translation. Because these constructs had a C-terminal tag, primers were designed that codon-optimised the first few amino acids of the constructs for expression in bacteria. Plasmids were constructed using PCR and InFusion cloning methods (**Appendix 7.1.1**). The final fusion

protein contains a TEV cleavage site. This meant that the GFP could be removed if necessary in order to characterise the CC specifically.

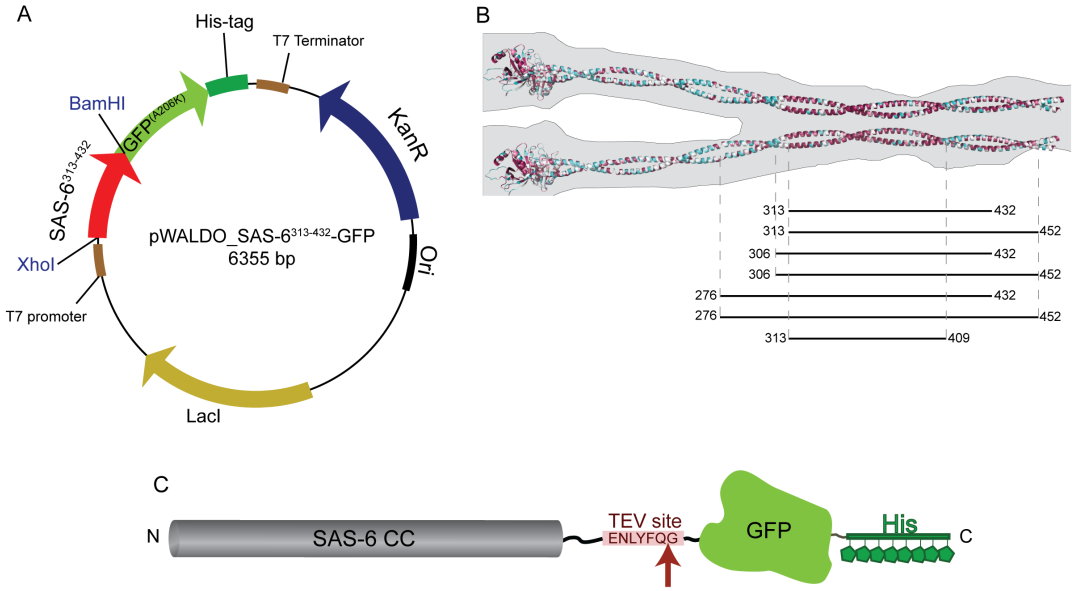


Figure 2.8.1.i: Design of SAS-6 CC fragments. (**A**) The pWALDO vector, encoding a fragment of SAS-6 CC fused to a His-tagged, modified GFP. (**B**) The different SAS-6 CC fragments, mapped onto the full-length DmSAS-6 docked cartwheel model (coloured as in **Figure 2.7.3.i, bottom**). The grey outline represents the extent of EM density from the map produced by Guichard et al., 2013. The fragments all span the core conserved region (~313-409). (**C**) Schematic of the designed constructs. The CC is fused to a TEV cleavable, C-terminal GFP-8xHis tag.

2.8.2: Small-scale purification of CC fragments

A small-scale purification trial was first conducted to assess which constructs would be suitable for large-scale purification. The SAS-6 CC constructs (**Figure 2.8.1.i, B**) were expressed using B834 (DE3) cells in 50 ml LB cultures. Usually, SDS-PAGE analysis would be used to gauge protein overexpression, however, due to the GFP fusion tag, it was possible to measure the fluorescence of the induced cell cultures. 1 ml of each culture was pelleted and resuspended in PBS before measurement using a fluorescence plate reader (SpectraMaxM5, Molecular devices) (**Figure 2.8.2.i, A**). Although the OD_{600} was not normalised, this approach took into account the overall yield for equally-sized cultures. This expression test revealed that 4 constructs (with N-termini of 276 or 313, and C-termini of 452 or 432) were expressed and were soluble. These four

constructs were taken forward for small-scale purification – a slightly modified protocol to that described in **Appendix 7.1.5**. Pelleted cells were resuspended in 7 ml lysis buffer, consisting of WB (**Appendix 7.1.4**) plus 2% tween-20, 1 PIT per 40 ml, 1 mg/ml lysozyme, and 0.05 mg/ml DNaseI. The cells were lysed by shaking for 30 mins at 4 ˚C, and the lysates were spun at 4000 x g to pellet insoluble material. Fusion protein was immobilised onto Ni-NTA agarose beads by incubating the supernatant with 0.5 ml Ni-NTA agarose bead suspension for 1 hour, whilst rotating at room temperature. The flowthrough was separated from the beads via gravity flow, after which the beads were washed twice with 3 ml wash buffer (WB with 20 mM imidazole, 0.05% tween-20). Elution was carried out via gravity flow using elution buffer (wash buffer with 250 mM imidazole.) All constructs eluted mostly within a ~1 ml peak and appeared to be relatively pure (**Figure 2.8.2.i, B**). At all stages of the purification the presence of fusion protein was clearly visible from its green colour.

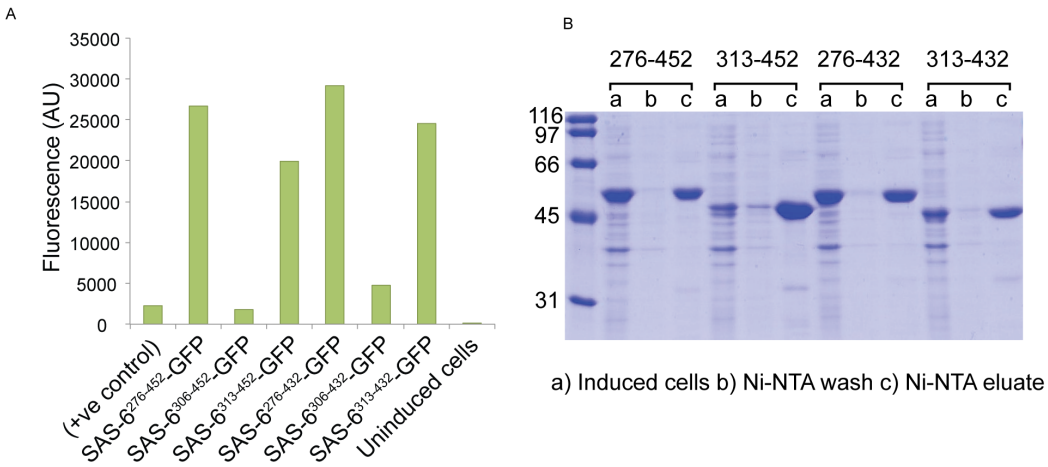


Figure 2.8.2.i: Small-scale purification of SAS-6 CC-GFP fragments. (**A**) Measurements of GFP fluorescence in induced bacterial cell cultures expressing DmSAS-6 CC-GFP fragments. The results reflect GFP-fusion protein yield between equivalently sized (50 ml) cultures. The positive control was an induced cell culture of TatC-GFP, kindly provided by Dr Sarah Rollauer. (**B**) SDS-PAGE analysis of the small-scale purification of the listed fragments.

Enough protein was obtained to run a trial cleavage of the GFP using TEV protease, which was added at a ratio of 1 mg TEV : 10 mg target. The reactions were dialysed overnight against WB

with 1 mM DTT. The results (not shown) indicated that in all cases, cleavage was successful, however in trying to remove the GFP by reverse Ni-NTA chromatography, the CCs also bound non-specifically to the matrix, and so could not be separated from the GFP in this manner. Some sample was left uncleaved, to analyse its stability over time. It was noticed that SAS-6$^{276-432}$-GFP and SAS-6$^{276-452}$-GFP precipitated out of solution over 24-48 hours, and so it was decided to only take SAS-6$^{313-432}$-GFP through to large-scale purification.

2.8.3: Purification of SAS-6$^{313-432}$-GFP

SAS-6$^{313-432}$-GFP was expressed as described above, using a 2 L culture of LB. Protein was purified via Ni-NTA affinity chromatography (**Figure 2.3.8.i, A**), according to the standard protocol (**Appendix 7.1.4**).

Amino Acids	Number Cys	MW	pI	ε_{280}	Abs 0.1%
382	1	43758.4	6.41	30370	0.694

Table 2.8.3.i: Details of the DmSAS-6$^{313-432}$-GFP construct calculated using Protparam.

Pooled Ni-NTA elution fractions of SAS-6$^{313-432}$-GFP were further purified by size-exclusion chromatography, using an S200 26/60 column (GE Healthcare), and using WB as a running buffer (**Figure 2.8.3.i, B**). The fusion protein was separated into two peaks.

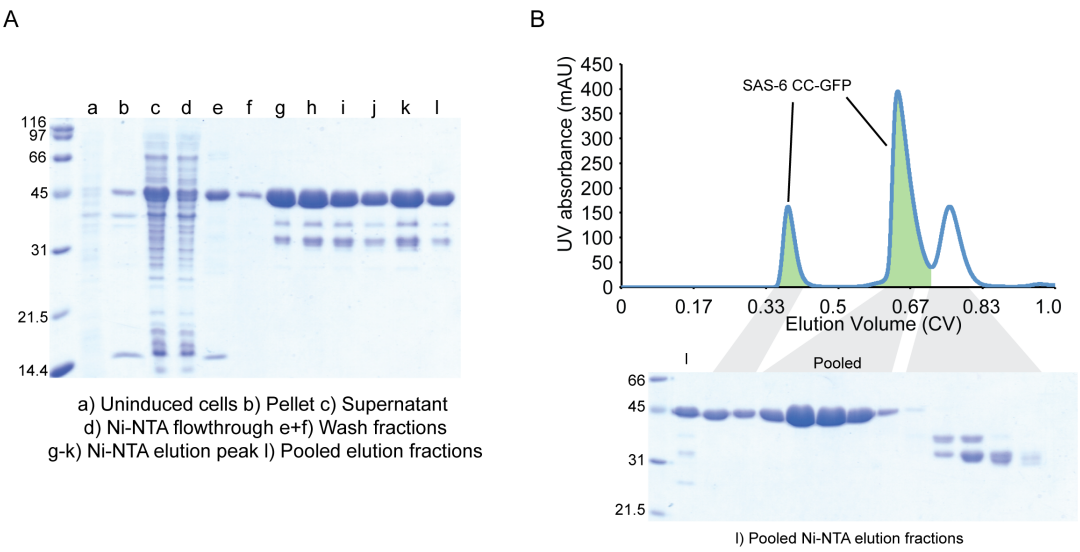


Figure 2.8.3.i: Large-scale purification of DmSAS-6$^{313-432}$-GF. (**A**) SDS-PAGE analysis of purification fractions. (**B**) Top panel, SEC trace from an S200 2660; bottom panel, SDS-PAGE analysis of SEC peak fractions. Two peaks were green and appeared to contain only the fusion protein, however the first peak was close to the void and so only the second peak was pooled for further experiments. The third peak contained contaminants.

Although these peaks could represent dimer and monomer, they were extremely well resolved, and the first peak was suspiciously close to the void volume of the column. Therefore only the second peak was pooled for further experimentation (including an MST experiment in **Appendix 7.8**). Protein was stored by snap freezing 100 µl aliquots in liquid nitrogen, before storage at -80 ˚C.

A trial MALS experiment was run on the SAS-6 CC-GFP fusion in order to assess whether it was dimeric. The experiment was carried out on a Superose 6 10/30 column, with the protein injected at ~60 µM (**Figure 2.8.3.ii**). The fusion protein was predicted to be a CC mediated dimer, but interestingly, it appeared to be largely monomeric in solution (although the observed mass was ~50% higher at the start of the peak, suggesting there was some tendency for self-interaction). However, even when considering dilution effects, this result suggests that the Kd of any dimerisation interaction is lower than expected; the Kds reported for other SAS-6 CC containing fragments are ~1 µM [van Breugel et al., 2011; Kitagawa et al., 2011b]. However, it is possible that the GFP fusion interfered with self-association, or that the CC section was not correctly folded.

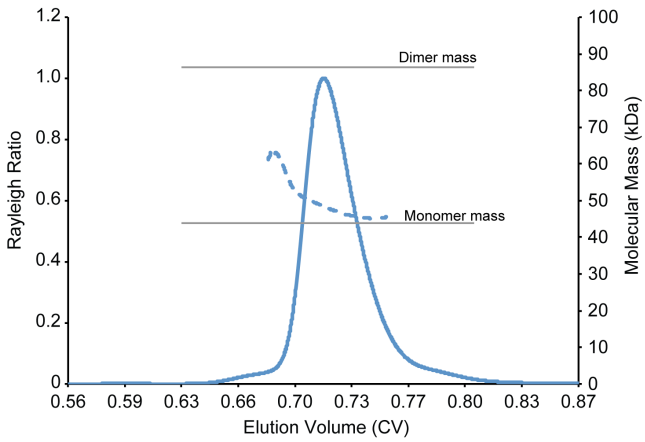


Figure 2.8.3.ii: MALS analysis of SAS-6 CC-GFP. SAS-6 CC fused to GFP was injected onto a superose 6 1030 column at ~60 µM. It eluted in a single peak that appeared to be predominantly monomeric, but showed some tendency towards self-association.

2.8.4: Cleavage and purification of DmSAS-6 from GFP

In order to investigate the SAS-6 CC in isolation, the GFP was cleaved, and the SAS-6 CC alone (**Table 2.8.4.i**) was purified.

Amino Acids	Number Cys	MW	pI	ε_{280}	Abs 0.1%
130	0	15261.4	6.35	8480	0.557

Table 2.8.4.i: Details of the cleaved DmSAS-$6^{313-432}$ construct calculated using Protparam.

TEV protease was added at a ratio of 1 mg : 10 mg target, and the reaction was dialysed into WB plus 2 mM DTT, overnight at 4 ˚C. The cleavage reaction was concentrated in a spin concentrator (MWCO 3 kDa) to a volume of <3 ml, before injection onto an S75 26/60 column (GE Healthcare), using WB as the running buffer (**Figure 2.8.4.i, A**). The SAS-6 CC was well resolved from the GFP peak, eluting ~0.1 CV earlier. This suggests that the CC has a larger hydrodynamic radius, which is in keeping with the prediction that it would form a long rod. (According to EMBL*, the SAS-6 CC (15.3 kDa) eluted at ~ the volume expected for ovalbumin, a 44 kDa globular protein). Residual TEV protease had to be removed via IEX chromatography. Pooled SEC fractions were dialysed into a low ionic strength buffer (25 mM Tris pH 9.0, 50 mM NaCl). Under these conditions the SAS-$6^{313-420}$ is negatively charged and will bind to an anion exchange resin. TEV protease was not expected to bind under these conditions. Anion exchange was carried out using two buffers, A (25 mM Tris pH 9.0) and B (buffer A plus 2 M NaCl). A MonoQ 5/50 GL column (GE Healthcare) was equilibrated to 50 mM NaCl for 20 CV before loading the protein sample (~40 ml). Some protein flowed through the column, which was then washed for a further 20 CV at 50 mM NaCl. A very shallow linear gradient of NaCl from 50-160 mM over 80 CV was then used to elute the column. DmSAS-$6^{313-432}$ eluted in one peak, fractions of which proved to be pure via SDS-PAGE analysis (**Figure 2.8.4.i, B**). Pooled fractions were analysed by Electrospray-Ionisation Mass Spectrometry (ESI-MS) (not shown, kindly performed by Dr David Staunton), which indicated that the sample consisted only of the expected fragment.

* http://www.embl.de/pepcore/pepcore_services/protein_purification/chromatography/hiload26-60_superdex75/

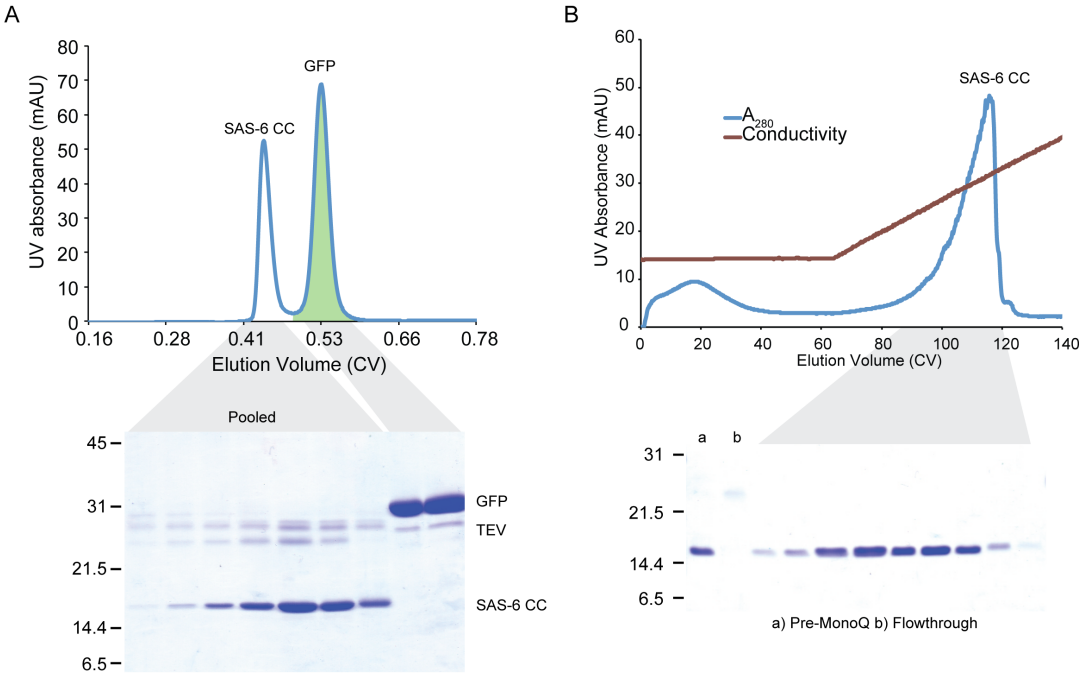


Figure 2.8.4.i: SEC and IEX purification of cleaved DmSAS-6$^{313-432}$. (**A**) Above, SEC trace from an S75 26/60 column. Below, SDS-PAGE analysis of SEC peak fractions. Cleaved DmSAS-6 CC had a larger hydrodynamic radius and eluted before GFP. Residual TEV was still evident in pooled fractions. (**B**) Anion exchange using a MonoQ column, and SDS-PAGE analysis of peak fractions. SAS-6 CC eluted in one peak over a very shallow NaCl gradient. Residual TEV protease appears to have flowed through the column without binding.

2.8.5: SAS-6$^{313-432}$ is alpha helical but may be monomeric

To assess whether SAS-6$^{313-432}$ was folded as expected, a Circular Dichroism (CD) experiment was carried out. Purified protein was diluted to ~200 µg/ml and dialysed at 4 ˚C using a Pierce Mini Dialysis Device (MWCO 10 kDa) against 10 mM $Na_xH_xPO_4$, pH 8.5. An average spectrum over 4 accumulations was measured at room temperature, between 190 and 260 nm, and normalised against a blank measurement taken with dialysis buffer (**Figure 2.8.5.i, A**). The shape of the spectrum strongly suggested that SAS-6$^{313-432}$ consists of helical secondary structure. To assess the oligomerisation state of the construct, purified protein was concentrated to 46 and 164 µM, and a MALS experiment was conducted on an S200 10/30 column (GE Healthcare), using WB as a running buffer (**Figure 2.8.5.i, B**). At both concentrations, SAS-6$^{313-432}$ eluted in a single peak, the average mass of which suggested it was monomeric in solution. Surprisingly, it therefore seems

that although the central portion of DmSAS-6 CC is helical, it is in fact not dimeric under these conditions, despite the strong prediction of CC formation for this region by the COILs server. Further work and the characterisation of other constructs will be needed to assess the significance of this result and to further characterise this region of SAS-6.

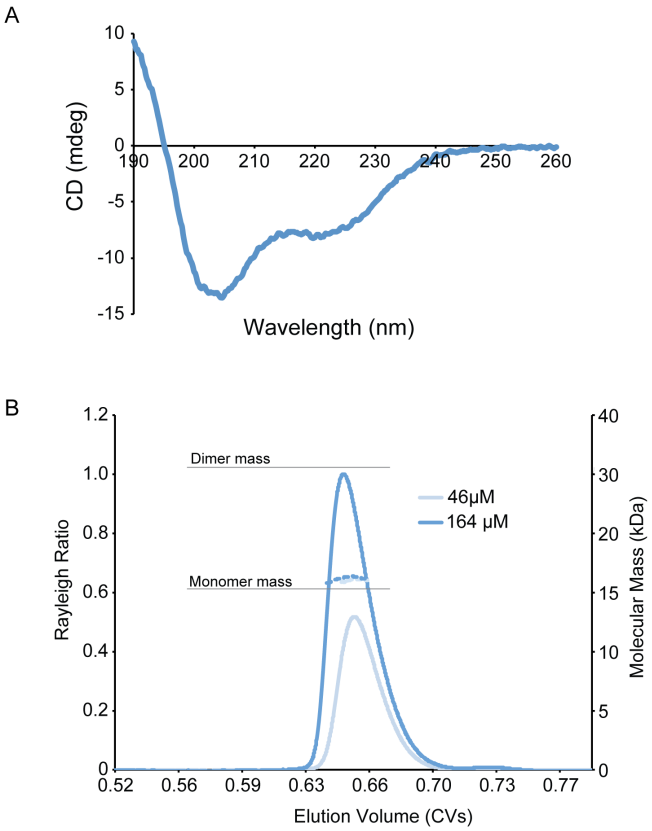


Figure 2.8.5.i: Analysis of isolated SAS-6$^{313-432}$. (A) CD trace of SAS-6$^{313-432}$. The double trough at ~205 and 222 nm are characteristics that suggest that the protein is helical. (**B**) MALS analysis of SAS-6$^{313-432}$ on an S200 10/30 column, injected at two different concentrations. The protein eluted in a single peak, with mass calculations showing it to be monomeric. (The mass of SAS-6$^{313-432}$ is 15.3 kDa).

2.9: Perspectives

At the beginning of my studies, relatively little was understood about the molecular nature of the centriole. The SAS-6 family of proteins had been discovered [Dammermann et al., 2004; Leidel et al., 2005] and it had been shown that these were likely to be structural proteins crucially involved in cartwheel assembly [Dammermann et al., 2004; Leidel et al., 2005; Nakazawa et al., 2007; Peel et al., 2007; Stevens et al., 2010a; Strnad et al., 2007].

In this Chapter I have characterised SAS-6 from *Drosophila melanogaster*, which like all SAS-6 orthologues comprises an N-terminal globular head-group, followed by a long section of coiled-coil. Using *in vitro* analyses of recombinant DmSAS-6 fragments, I demonstrated that SAS-6 self-oligomerises via two independent interaction interfaces, mediated by the CC, and an N-N interaction respectively. By designing constructs to dissect this oligomerisation *in vivo*, I showed that the N-N interaction is required both for efficient centriolar localisation of SAS-6 in embryos, and efficient centriole assembly. By determining the structure of the N-terminal region of DmSAS-6, I showed that this protein was highly similar to orthologous structures that were published during the course of my work, thus completing the “library” of SAS-6 structures available for the most popular model organisms used in centriole research. I generated an atomic model of the DmSAS-6 cartwheel that was highly compatible with an EM map of the cartwheel [Guichard et al., 2013]. Critically, in order to form a flat ring, a particular orientation of the SAS-6 N-terminal head-group must be adopted with respect to the CC. DmSAS-6 has the same domain orientation as SAS-6 from *C. reinhardtii* [Kitagawa et al., 2011b] *D. rerio* [van Breugel et al., 2011], and *L. major* [van Breugel et al., 2014]. It is likely that all of these orthologues can oligomerise to form a nine-fold symmetric, flat cartwheel.

Conversely, CeSAS-6 appears to behave differently, with results suggesting that it would form a spiral rather than a cartwheel [Hilbert et al., 2013]. This potentially rationalises the different appearance of *C. elegans* centrioles, where no spokes appear visible by EM [Pelletier et al., 2006]. The long-accepted model for *C. elegans* centrioles was that a "central tube" replaced the canonical "central cartwheel". Indeed, this new spiral model would result in a much lower density of spokes when observing transverse sections of centrioles by EM, perhaps explaining the lack of spokes observed by [Pelletier et al., 2006]. The implication of this *C. elegans* model is that the very first step of centriole assembly is different in this species. It would follow that downstream proteins and processes might also deviate from the majority of canonical organisms. Characterisation of the *C. elegans* system may, therefore, be less likely to reveal canonical processes for centriole assembly that are relevant in most other species. This highlights the importance of a thorough characterisation of centriole assembly within several different organisms.

Despite the likelihood of DmSAS-6 forming a canonical cartwheel *in vivo*, my EM analysis of recombinant wild-type DmSAS-6 revealed the formation of long string-like structures with a fairly consistent diameter of ~20-30 nm, compatible with the dimensions of the central hub. Intriguingly, these assemblies were N-N dependent. The string-like assemblies could be formed either by the stacking of cartwheel rings, or by the continued growth of a cartwheel ring into a spiral with a shallow pitch, although neither of these theories agree with the current model of canonical cartwheel arrangement – EM micrographs suggest that *in vivo* the central hubs are distinctly separated [Guichard et al., 2012, 2013] (**Figure 2.7.2.i, B, C**).

In the structure of the LmSAS-6 central hub (4CKP [van Breugel et al., 2014]), the rings are in fact stacked directly on top of one another, however the crystal contacts mediating this interaction

are weak and involve unconserved interfaces, and so this stacking interaction was judged to be non-specific. It is therefore possible that DmSAS-6 also uses a similar non-specific interaction to form these string-like structures *in vitro*, as observed in my EM analysis.

It is also possible that, *in vitro*, without the influence of regulatory factors, DmSAS-6 can oligomerise into a very shallow spiral. This would actually be thermodynamically favourable, as it would facilitate the continuous addition of SAS-6 into a large open-ended structure; it would require only very minor alterations to the N-N or CC interaction interfaces to form a spiral as opposed to a ring. Indeed, it would only take a few more than nine dimers to oligomerise a circular structure such that the start and end overlapped past each other. Such a structure would then easily be extended, without the possibility of ever resolving into a closed ring. This poses the question of how discrete rings are so rigorously enforced *in vivo*, something that is likely enforced by other proteins. Alternatively, although wild-type SAS-6 clearly oligomerises, the strings themselves could simply be an artefact of the EM preparation procedure.

Another important question regarding the cartwheel relates to its symmetry-breaking features. In the EM model, the cartwheel itself is an inherently symmetrical object, composed from CC mediated “spoke” dimers [Guichard et al., 2012, 2013]. The monomers in each dimer can be mapped onto one another via a two-fold rotation around the CC axis. This symmetry is broken at the point where the outward path of the spoke deviates so that it can pair with its partner spoke from the adjacent cartwheel. As soon as a spoke deviates from a straight CC axis, the monomers of this spoke dimer are asymmetrical, and can no longer be mapped onto one another by a two-fold symmetry operation along the spoke. Clearly, this feature of the cartwheel must involve factors other than SAS-6 – in isolation, an inherently symmetrical object has no stimulus to

deviate from its self-symmetry, especially in such a regular manner (although it is possible that inbuilt flexible regions could facilitate this).

Importantly, the pinhead region of the spoke has a periodicity of double that of the central hub, and so, by the nature of this feature, the cartwheel spokes must converge in order that they can all be docked at the cartwheel periphery. How is it that the spokes pair in such a regular fashion, exclusively with only one partner cartwheel? A mere tendency for spokes to self-associate would result in spokes indiscriminately pairing with the cartwheel above and below, rather than just one or the other. This begs the question as to whether the cartwheels assemble first, and then their symmetry is broken in a regulated manner, or whether cartwheel assembly could be influenced from the periphery. In the latter case, regularly spaced pinheads first bind the ends of two SAS-6 spokes, thus forcing the formation of paired central hubs – one pair per radial layer of pinheads – as is the case in the EM map.

This idea is intriguing considering a recent study, which found that SAS-6 could be transiently recruited to the proximal lumen of mother centrioles (from which the cartwheels are degraded). Furthermore, this recruitment was independent of the N-N interface of SAS-6, and instead was dependent on the luminal protein CPAP, and on the presence of the SAS-6 C-terminus [Fong et al., 2014]. Rather than entirely self-assembling, the cartwheel might therefore be templated from its periphery. However this mechanism is difficult to reconcile with the *Drosophila* system where the mother centrioles retain their cartwheels. Additionally this transient recruitment may not be mechanistic, but simply an artefact of having empty SAS-6 binding sites in the mother centriole lumen.

Assuming that the cartwheel is self-assembled, evidence suggests that SAS-6 is recruited upstream of more peripherally located cartwheel proteins such as CEP135 and SAS-4. Unless the cartwheel is built layer by layer, how could SAS-6 based cartwheel asymmetry be influenced by factors which it has yet to recruit? This question remains unanswered, but future studies will no doubt focus on how SAS-6 interacts with these partner proteins.

It will be important to characterise the SAS-6 spoke in order to answer many of these questions. Although I could purify such a fragment from DmSAS-6, the results were ambiguous, suggesting it was a monomeric helix. Further experiments will be required to test whether this unexpected behaviour is the result of poor construct design, unfavourable experimental conditions, or if it really is the native behaviour of this part of the protein.

Chapter 3: The interaction between Ana2 and SAS-4

3.0: Introduction

3.0.1: Evidence for an Ana2:SAS-4 interaction

Ana2 is thought to be the *Drosophila* orthologue of vertebrate STIL, and is a core centriole duplication protein [Stevens et al., 2010b; Wang et al., 2011]. Ana2/STIL is recruited to nascent centrioles early on in the assembly pathway, and it appears to localise to the centriolar lumen [Arquint et al., 2012; Stevens et al., 2010a]. SAS-4 is the *Drosophila* orthologue of vertebrate CPAP, and is also essential for centriole biogenesis [Basto et al., 2006]. Studies show that the SAS-4/CPAP family of proteins is likely to localise towards the periphery of centrioles, and is involved in microtubule binding [Hsu et al., 2008; Hung et al., 2004; Kirkham et al., 2003; Mennella et al., 2012]. Interestingly, the core duplication proteins STIL and CPAP have associated mutations in humans that are linked to autosomal recessive primary microcephaly (MCPH) [Bond et al., 2005; Darvish et al., 2010; Gul et al., 2006; Leal et al., 2003]. This suggests that centriole biogenesis is in some way linked to brain development. One such particular missense mutation causes a E1235V mutation in CPAP [Bond et al., 2005]. In a recent study, Co-IP and Yeast 2 hybrid experiments showed that a ~450 residue fragment of CPAP directly interacted with a ~400 residue fragment of STIL. Interestingly, this interaction appeared to be less efficient in the presence of the CPAP E1235V mutation. This indicates that a STIL:CPAP interaction might represent an essential step in centriole assembly.

3.0.2: Aims

- Investigate whether a STIL:CPAP interaction represents an important (and therefore presumably conserved) step in centriole assembly.
- Identify and characterise the responsible domains in the *Drosophila* orthologues, Ana2 and SAS-4.
- Obtain a crystal structure of the interaction in order to understand its molecular detail.
- Dissect the interaction using *in vivo* methods.

Over the course of this study, I collaborated with Dr. Mark van Breugel, who conducted parallel *in vitro* analysis using the homologous *D. rerio* proteins.

3.0.3: Key findings

- A crystal structure of the interaction interface revealed that the TCP domain has a novel fold, and binds to a small, conserved, proline-rich region in Ana2.
- The MCPH residue is involved at the interaction interface.
- The interaction is a conserved step in centriole biogenesis, and is essentially identical in vertebrates.
- Dissection of the interaction *in vivo* shows that the Ana2:SAS-4 interaction is essential, however, it is not completely abolished by the MCPH mutation.

3.1: Bioinformatic analysis of SAS-4 and Ana2

To identify potential CPAP/STIL binding regions, I initially undertook a bioinformatics analysis, to try to further delineate the potential binding regions within the rather large fragments already identified [Tang et al., 2011]. This approach might also reveal potential binding regions in the *Drosophila* homologues SAS-4 and Ana2.

3.1.1: Bioinformatic analysis of CPAP/SAS-4

A variety of bioinformatics tools were used, with the results summarised in **Figure 3.1.1.i**. A multiple sequence alignment (MSA) analysis was carried out using ClustalX [Larkin et al., 2007], on 36 orthologous sequences from insects, vertebrates, and nematodes, found via BLAST search against known SAS-4 orthologues. In particular, six regions were identified that had an enrichment of conservation, and these are termed conserved regions (CRs). CRs 1-6 appear to be shared between both *Drosophila* and vertebrate species. Although nematode SAS-4 sequences could be aligned, only regions corresponding to CR5 and CR6 could be confidently assigned. Previously described microtubule binding regions (Hung 2004, Hsu 2008) correspond to CRs 2-3, and appear to be shared between *Drosophila* and vertebrates. As CC regions have been identified previously in SAS-4 and its orthologues [Hung et al., 2000; Kirkham et al., 2003], the COILs server [Lupas et al., 1991] was used to define CC regions in *Drosophila* SAS-4, *C. elegans* SAS-4, *D. rerio* CPAP, and *H. sapiens* CPAP, with results mapped onto the schematics shown in **Figure 3.1.1.i**. It can be seen that a ~150-200 residue, possibly broken CC exists in all orthologues, and corresponds to CR5, suggesting it is an important feature of this protein family. Other CCs are evident in HsCPAP but are largely absent in non-vertebrate orthologues. As described by [Hung et al., 2000], the region with the most enriched conservation is the C-terminal ~200 residues, known as the TCP domain or G-box. In particular, the TCP domain of *D. melanogaster* SAS-4 (DmSAS-4) had ~42% sequence identity to the TCP domain of *D. rerio* CPAP

(DrCPAP). A secondary structural prediction using Psipred [Jones, 1999] suggested that for all orthologues this domain contains roughly 20 successive beta strands. The MCPH residue, (E1235V in HsCPAP, E792V in DmSAS-4) is predicted to lie within one of these strands, and is highly conserved in all orthologues. (An annotated region of the MSA covering the TCP domain is shown in **Appendix 7.3**). Because the MCPH residue is thought to affect STIL binding, and lies in the centre of the TCP domain, it was suspected that the TCP domain was the most likely domain to bind STIL, and so this domain was targeted for *in vitro* study.

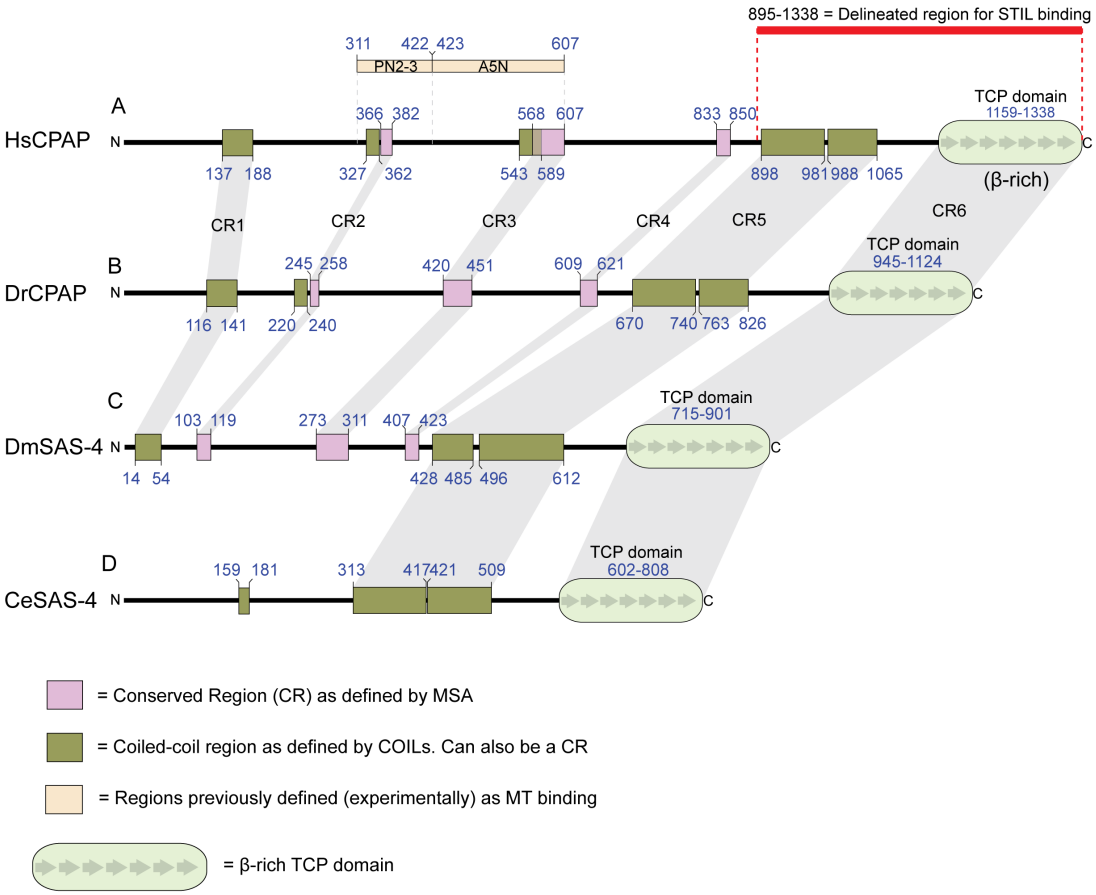


Figure 3.1.1.i: Summary of bioinformatic analysis on SAS-4. Shown are schematics of the topologies of (**A**) human CPAP, (**B**) *D. rerio* CPAP, (**C**) *Drosophila* SAS-4 and (**D**) *C. elegans* SAS-4. Various topological features are explained in the legend below. Using MSA analysis, it was possible to highlight and compare conserved regions (CRs, pink boxes). Some conserved regions mapped to, or contained predicted coiled-coils as defined by the COILs server (khaki boxes). Most obvious was the existence of a β-rich conserved region at the C-terminus of each orthologue, previously described as the TCP domain, or G-box. This region was included in fragments thought to bind to STIL in a study by [Tang et al., 2011]. MT binding regions were identified by [Hung et al., 2004; Hsu et al., 2008].

3.1.2: Bioinformatic analysis of STIL/Ana2

The region of human STIL thought to bind to CPAP was delineated to residues 236-619 by Tang et al 2011. If the binding mechanism were conserved through to *Drosophila* species, then Ana2 would be expected to share a conserved domain with this region of STIL. Both STIL and Ana2 protein families contain a C-terminal "STAN" domain [Stevens et al., 2010b], and a short predicted coiled-coil, located towards the centre of each protein (**Figure 3.1.2.i, A, B, C**). However the STAN and CC features were not present in the CPAP-interacting region of STIL, delineated by Tang et al 2011. By process of elimination, a homologous SAS-4-binding region in Ana2 would have to be located in the first ~195 residues, before the CC. STIL proteins are long (>1000 residues) compared to Ana2 proteins (~420 residues) and it is difficult to align the two full-length proteins. As such, initial MSA analysis was carried out on several vertebrate STIL proteins, and on several Ana2 proteins separately (not shown). MSA and Psipred analysis of STIL proteins suggested that there was a ~350 residue N-terminal conserved region, which was rich in secondary structure, denoted CR1 (**Figure 3.1.2.i, A**, **B**, Psipred analysis not shown). Its richness of predicted secondary structure would suggest that the STIL CR1 might fold into a globular domain, and so the sequence of the CR1 region was analysed using the Fold and Function Assignment System (FFAS) server, which uses a psi-BLAST approach to suggest structurally-homologous domains in the PDB. As the server failed to find any such matches in the PDB, the STIL CR1 may represent a novel domain. CR1 was followed by a short, conserved, proline-rich region with no predicted secondary structure, which was denoted CR2. Because the delineated region of 231-619 intersected CR1, and no topologically equivalent CR1 domain was evident in Ana2 proteins, CR1 was omitted from subsequent analysis under the assumption that a CPAP/SAS-4 binding region would be shared between the Ana2/STIL protein families. A large MSA was carried out using 7 full-length Ana2 sequences, and 30 STIL sequences from CR2 onwards. Deletion of CR1 now enabled the CR2 of STIL sequences to align convincingly with the extreme, proline-rich N-terminus of Ana2 sequences (**Figure 3.1.2.i**, **D**), which did not previously

stand out in MSA analysis of Ana2 proteins alone. It therefore appears that STIL and Ana2 proteins share homology from a proline-rich, "CR2" domain onwards, and that this region is the only shared domain that is contained within the CPAP-binding region of STIL. This region (consisting of the first ~44 residues of Ana2) was therefore a candidate for binding to SAS-4. (Note: a more comprehensive bioinformatic analysis of Ana2 is conducted in **Chapter 4, Section 4.1.1**). The CR1 domain may represent a unique feature that is only present in vertebrate STIL orthologues. Little is known about this region of the molecule, or whether it contributes in any way to centriole biogenesis.

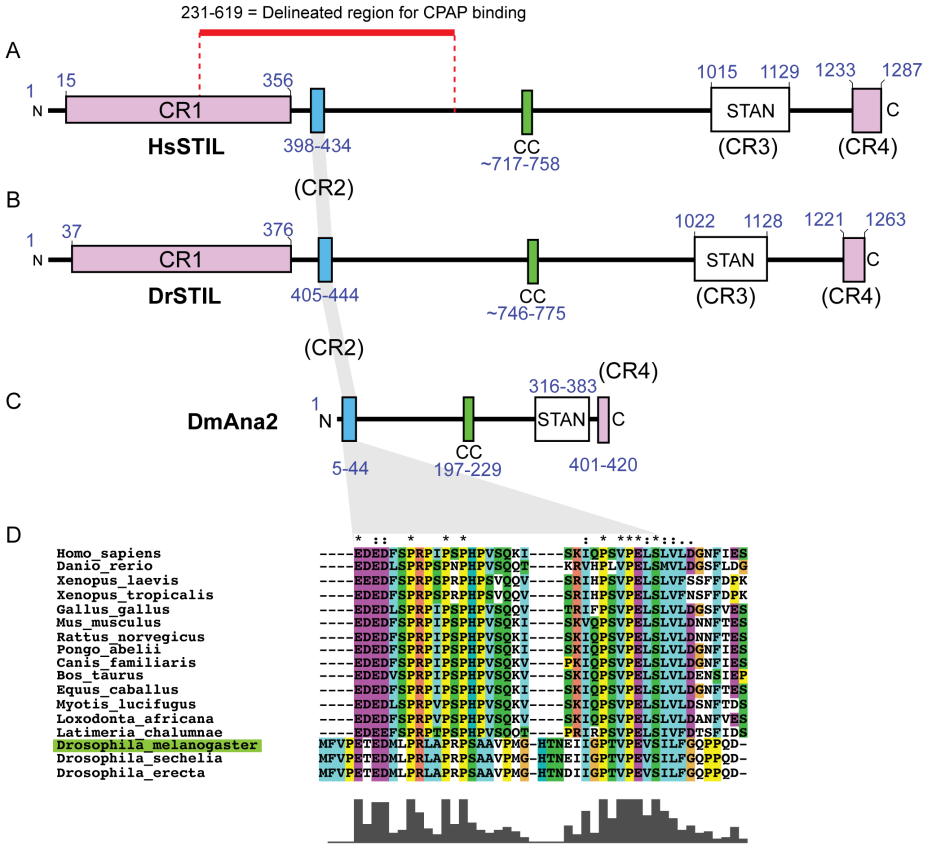


Figure 3.1.2.i: Delineation of the STIL/Ana2 Conserved Region 2 (CR2). (**A**) Schematic of the topology of human STIL, showing four conserved regions, and a central CC. The region delineated by Tang et al 2011 for binding to CPAP, is shown above by the red bar, and can be seen to intersect CR1. (**B**) Schematic of *D. rerio* STIL, showing similar features to HsSTIL. (**C**) Schematic of *D. melanogaster* Ana2, which aligns with STIL from the CR2 domain onwards. (**D**) Selected sequences from MSA analysis of STIL and Ana2, showing the proline-rich CR2.

3.2: Expression and analysis of soluble SAS-4 TCP domain constructs

3.2.1: Design of TCP domain constructs

To carry out *in vitro* characterisation of the Ana2:SAS-4 interaction it was necessary to recombinantly produce the SAS-4 TCP domain. Bioinformatic analysis (**Section 3.1.1**) indicated that the DmSAS-4 TCP domain shared a high level of homology with the DrCPAP TCP domain, with both domains predicted to contain 20 β-strands. My collaborator, Dr Mark van Breugel managed to solve the structure of the DrCPAP TCP domain. This revealed that the TCP domain folded into a continuous β-meander, with 19 of the predicted strands visible in the structure. (**Figure 3.2.1.i, B**). I therefore designed two constructs, one that included only the 19 strands visible in the DrCPAP crystal structure (SAS-4$^{725\text{-}901}$), and another that included the first predicted strand (SAS-4$^{707\text{-}901}$). A third construct (SAS-4$^{700\text{-}901}$) was later designed in an attempt to optimise the construct. According to sequence conservation and secondary structural prediction, the TCP domain was likely to continue until the extreme C-terminus of SAS-4, so all constructs terminated at the final SAS-4 residue, Y901. The constructs were cloned using conventional ligation methods into the bacterial expression vector, pETM-14 (EMBL) (**Figure 3.2.1.i, A**). The constructs would therefore contain a 3C protease-cleavable, N-terminal His-tag. DmSAS-4 cDNA was used as a template for SAS-4$^{725\text{-}901}$ and SAS-4$^{707\text{-}901}$, however the SAS-4$^{700\text{-}901}$ template was a synthetic *E. coli* codon-optimised sequence (GeneArt), which was used in an attempt to increase protein yield.

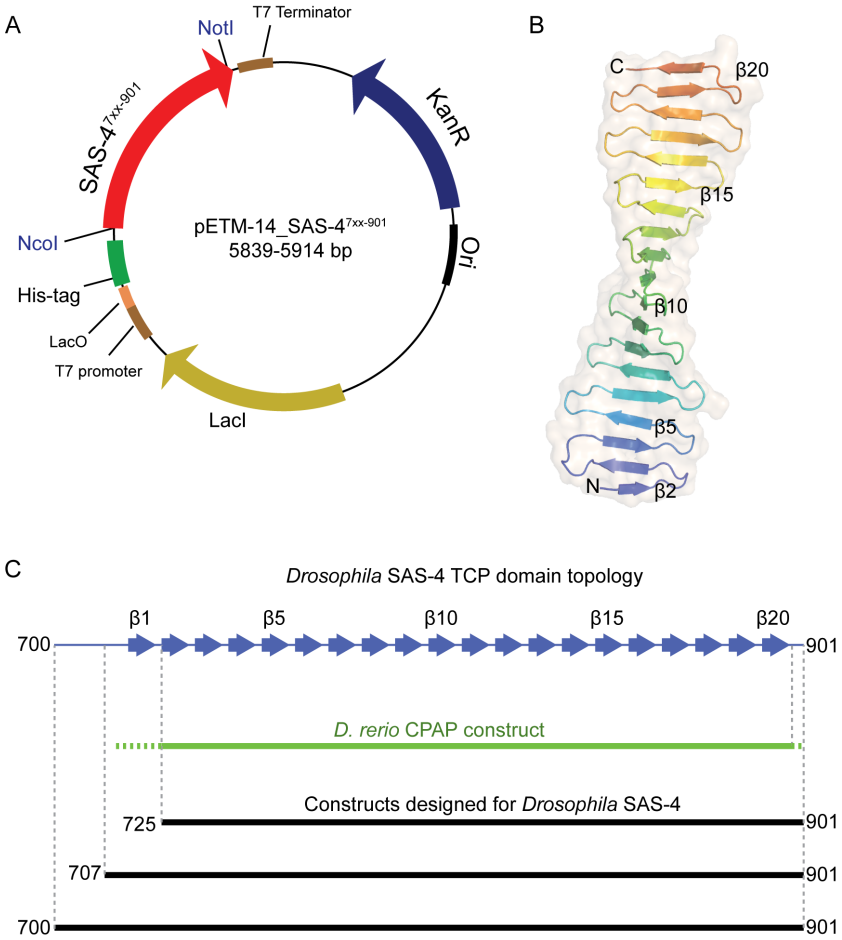


Figure 3.2.1.i: Design of DmSAS-4 TCP domain constructs. (**A**) The pETM-14 bacterial expression vector with the cDNA of SAS-4 TCP domain fragments cloned between the NcoI and NotI restriction sites. The product of this vector is a 3C cleavable, N-terminal 6 x His-tagged recombinant protein. (**B**) The structure of the *D. rerio* CPAP TCP domain, showing 19 of the 20 predicted strands, β2-β20. The structure is shown in cartoon format, coloured Blue-Red (N-C), inside a translucent beige surface representation. (**C**) Schematic of the 20 predicted β-strands of the DmSAS-4 TCP domain. The green line underneath represents the DrCPAP TCP domain construct (solid, visible in the structure, dashed, present in the construct). The black lines represent the extent of the DmSAS-4 TCP domain constructs designed for this study.

3.2.2: Purification of SAS-4 TCP domain constructs

Each construct was expressed in *E. coli* expression strain B834 (DE3) cells. Initially, 1 litre of TB medium was used for each construct, however later purifications used 4 L LB cultures in an attempt to achieve gentler expression. Successful expression was confirmed by SDS-PAGE analysis, which showed an obvious overexpression band for all constructs (**Figure 3.2.2.i**, **A, B,**

C'). All proteins were purified using the standard procedure (**Appendix 7.1.4**), with working buffer pH optimised according to the theoretical pI of each construct (**Table 3.2.2.i**).

Construct	Amino Acids	MW (Da)	pI	ε_{280}	Abs 0.1%	WB pH
His-SAS-4$^{725-901}$	198	23172.0	7.90	34380	1.485	7.5
His-SAS-4$^{707-901}$	216	25023.9	7.20	34380	1.375	8.0
His-SAS-4$^{700-901}$	223	25686.7	7.89	34380	1.340	8.5
SAS-4$^{707-901}$	199	22968.6	7.15	34380	1.497	8.0
SAS-4$^{700-901}$	206	23631.4	7.98	34380	1.455	8.5

Table 3.2.2.i Details of SAS-4 TCP domain constructs calculated using Protparam. No constructs contained cysteine.

Cell lysates were visibly white with precipitate, and it transpired that the constructs had a very limited solubility. In subsequent purification the volume of lysis buffer was significantly increased, to ~10-15 ml buffer per gram of cells, as opposed to the usual ~5 ml per gram. The SAS-4$^{725-901}$ construct was not pursued due to impurities and low yield (**Figure 3.2.2.i, A**); however, it was possible to purify SAS-4$^{707-901}$ and SAS-4$^{700-901}$, which each eluted from the Ni-NTA as fairly homogenous species. During Ni-NTA chromatography, an imidazole gradient was used to elute the protein, however using a steep gradient (15 CV) often caused a large amount of protein to precipitate towards the end of the elution peak. Therefore, in subsequent purifications, the length of the gradient was increased, eventually settling on 30 CV. Due to the limited solubility, there was a progressive loss of material during downstream purification steps. SAS-4$^{707-901}$ appeared to be less stable over time, with possible degradation products appearing in downstream purification steps (**Figure 3.2.2.i, B"**). The SAS-4$^{700-901}$ construct was designed in an attempt to increase yield, solubility and stability. Cleaner purification of SAS-4$^{700-901}$ suggested it was more stable, and so it eventually became the adopted construct. However, SAS-4$^{700-901}$ behaved similarly to SAS-4$^{707-901}$, suffering from poor yield and solubility, and as such, further discussion relates to both constructs. SDS-PAGE analysis (**Figure 3.2.2.i B, C'**) revealed that the vast majority of protein was in the insoluble fraction of the cell lysate, and this presumably explains the characteristically low yields from these purifications.

To remove the His-tag, Ni-NTA purified protein was digested with GST-3C protease, at an approximate ratio of 1 mg protease to 10 mg target. The reaction was then dialysed into fresh WB overnight at 4 ˚C. The SAS-4 TCP domain could be separated from the GST-3C and cleaved His-tag via size exclusion chromatography, using an S75 column (GE Healthcare), equilibrated in WB (**Figure 3.2.2.i, B', C'')**. Concentration of purified protein was carried out using 10 kDa MWCO spin concentrators (Pierce), and this allowed the protein to reach a maximal concentration of ~2-3 mg/ml, although it was not stable indefinitely at these concentrations. At ~ 1 mg/ml, 100 µl aliquots of the protein could be flash frozen in liquid nitrogen and stored at -80 ˚C. ESI-MS (kindly performed by Dr David Staunton) was used to confirm the identity and integrity of the constructs. The SEC profiles of thawed samples were unchanged (not shown), suggesting that freeze-thawed samples were still folded.

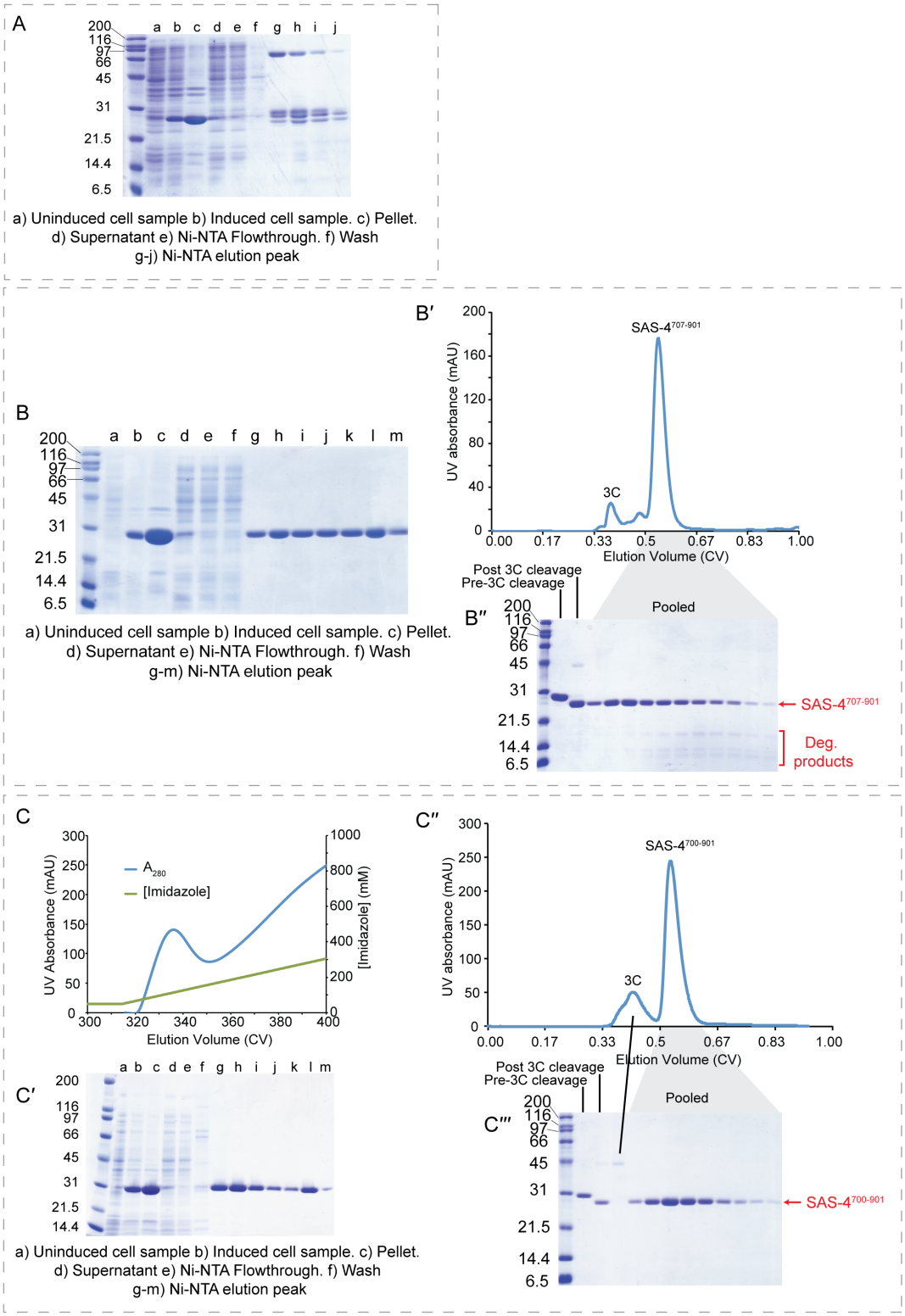


Figure 3.2.2.i: Purification of DmSAS-4 TCP domain constructs. (**A**) SDS-PAGE analysis of the unsuccessful purification of SAS-4$^{725-901}$. Despite being expressed, no soluble SAS-4$^{725-901}$ was obviously eluted from Ni-NTA resin. (**B**) SDS-PAGE analysis of the purification of SAS-4$^{707-901}$. Most protein was insoluble, as represented by the large band in the pellet fraction, however the soluble fraction was successfully purified using Ni-NTA chromatography. After His-tag cleavage, pure SAS-4$^{707-901}$ could be separated from 3C protease, by SEC (**B′, B″**), however SDS-PAGE analysis revealed the presence of contaminants. (**C**) Elution of SAS-4$^{700-901}$ from Ni-NTA, using a shallow imidazole gradient. (**C′**) SDS-PAGE analysis of purification fractions showing that, similarly to SAS-4$^{707-901}$, most of the protein was contained in the insoluble fraction. (**C″, C‴**) SEC trace, and subsequent SDS-PAGE analysis showing that SAS-4$^{700-901}$ could be purified to homogeneity.

3.2.3: The SAS-4 TCP domain is monomeric at low concentration

In order to characterise the oligomeric state of the SAS-4 TCP domain in solution, a SEC-MALS experiment was conducted on purified SAS-4$^{700-901}$. Experiments were carried out using an S200 10/30 column (GE Healthcare) equilibrated in WB. Unfortunately, due to the solubility limit of the domain, SAS-4 was only injected onto the column at 0.5 and 1.0 mg/ml (21 µM and 42 µM). At both concentrations, SAS-4$^{700-901}$ eluted as a discrete peak. The observed molecular mass was stable across the peak, and was consistent with the theoretical monomeric mass of 23.6 kDa (**Figure 3.2.3.i**). There were no indications of self-association at these concentrations, consistent with similar observations in the *D. rerio* system by Mark van Breugel [Cottee et al., 2013], where the DrCPAP TCP domain was monomeric even at 230 µM. These results are therefore consistent with the DmSAS-4 TCP domain being monomeric.

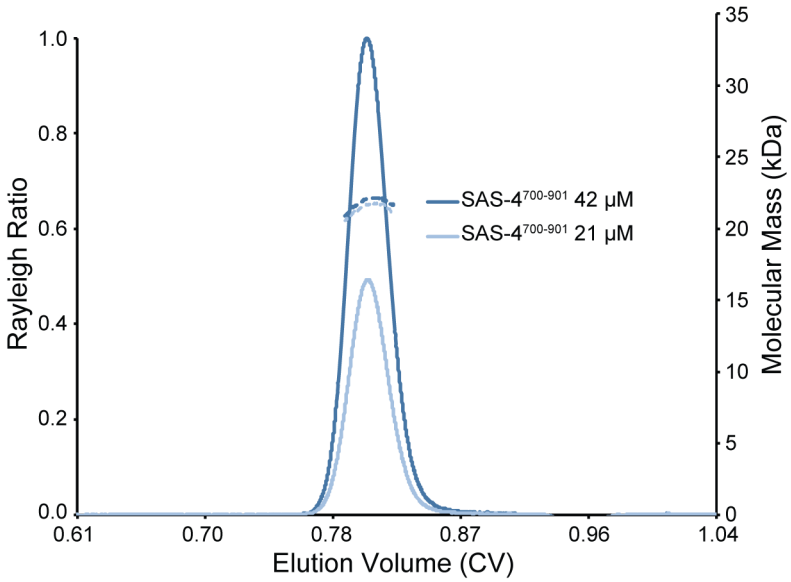


Figure 3.2.3.i: SEC-MALS analysis of the DmSAS-4 TCP domain. SAS-4$^{700-901}$ was injected over an S200 10/30 column (GE Healthcare). The solid lines show the elution profile, with the observed molecular mass shown as dashed lines. At either concentration SAS-4$^{700-901}$ eluted as a single species with a stable monomeric mass (~23.6 kDa).

3.2.4: The DmSAS-4 TCP domain has β-rich secondary structure

To assess the secondary structural content of the DmSAS-4 TCP domain, circular dichroism (CD) experiments were conducted on purified SAS-4$^{700\text{-}901}$. The protein was diluted to ~100 µg/ml, and dialysed overnight using a 10 kDa MWCO dialysis membrane (Pierce), into 10 mM $Na_xH_xPO_4$, pH 8.5. CD spectra were measured at room temperature on a j-720 machine (Jasco), measuring between 190 to 260 nm, at 0.5 nm intervals. Spectra were averaged over 4 accumulations, normalised against a blank measurement taken with dialysis buffer. Two essentially identical spectra, each from a different purification, are compared in **Figure 3.2.4.i**. The spectra show a trough at ~212 nm that is most characteristic of a β-type profile. This suggests that the SAS-4 TCP domain is rich in β-structure, consistent with the secondary structure prediction.

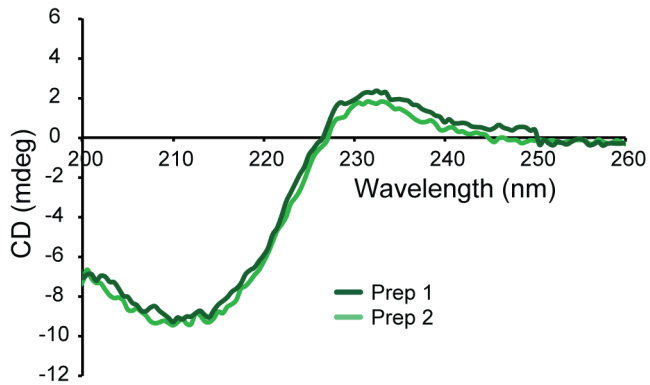


Figure 3.2.4.i: Background-subtracted circular dichroism spectra of SAS-4$^{700\text{-}901}$. Two spectra are shown, each from a separate purification of SAS-4$^{700\text{-}901}$.The single trough at ~212 nm is consistent with the SAS-4 TCP domain adopting a predominantly β-type fold.

3.3: Detecting an interaction between the Ana2 CR2 and the SAS-4 TCP domain

Following the successful purification of the SAS-4 TCP domain, it was necessary to confirm the interaction between the bioinformatically-delineated domains of SAS-4, and Ana2 using *in vitro* techniques.

3.3.1: Far western dot blot analysis

It was possible to purify a recombinant fragment of the N-terminal half of Ana2, (His-Ana2$^{1-228}$) including the CR2 domain, through to its predicted coiled-coil domain (**Figure 3.3.1.i, A**, purification discussed in **Appendix 7.4**). This fragment appeared to be pure, but was unstable in solution in the absence of imidazole. In order to assess whether this fragment could interact with the SAS-4 TCP domain, I carried out far-Western dot blot experiments. This involves the immobilisation of purified, native "bait" protein to a nitrocellulose membrane. The membrane is then exposed to "prey" protein. Bound prey protein can then be detected using an antibody. In this case, His-tagged Ana2$^{1-228}$ was used as prey, and was detected using an α-His HRP antibody. A dilution series of SAS-4$^{707-901}$ was prepared from 10 µM to 1 nM. A similar dilution series was prepared for recombinant SAS-6$^{1-216}$(F143D) (purification discussed in **Chapter 2, Section 2.2.3**). This reagent was readily available, and I also wanted to test whether Ana2$^{1-228}$ might interact with SAS-6. As a likely negative control, an MBP-fused fragment of the centrosomal protein Cnn, MBP-Cnn(C3) (courtesy of Eliana Lucas), was also included, at a concentration of 13 µM. Purified BSA (New England Biolabs) was included as a second negative control at 15 µM. Finally, His-tagged SAS-6$^{1-216}$(F143D) was included as a positive control for His-tag detection at 4 µM. 10 µl of each protein solution was pipetted slowly onto a nitrocellulose membrane (BioRad), allowing the membrane to fully absorb the solution throughout the application. The membrane was then allowed to dry for 10 minutes. The membrane was then washed in blocking buffer

(50 mM Tris, pH 8.0, 300 mM NaCl, 500 mM Imidazole, 1 mM βME, 0.1% Tween-20, 2% milk powder), overnight at 4 ˚C. A "prey solution" was then prepared, which consisted of blocking buffer, with the addition of His-Ana2$^{1-228}$ protein at a concentration of 12 µM. The blocked membrane was washed with prey solution for 1 hour at room temperature, before three washes in blocking buffer for 15 minutes at room temperature. The membrane was then exposed to an α-His antibody (Penta-His conjugate, Qiagen) solution, (1:5,000 dilution of antibody in blocking buffer). The membrane was then washed another three times in blocking buffer, before exposure to ECL substrate (Pierce) for 1 minute, and detection of fluorescence using X-ray film (Amersham ECL Hyperfilm, GE Healthcare). As shown in **Figure 3.3.1.i, B**, on a membrane not exposed to prey solution, only spots containing the positive control, His-SAS-6$^{1-216}$(F143D) gave any signal. However, on a membrane exposed to His-Ana2$^{1-228}$, signal was detected in spots containing higher concentrations of both SAS-4$^{707-901}$ and SAS-6$^{1-216}$(F143D). This result provides preliminary evidence that Ana2$^{1-228}$ might directly interact with both of these constructs.

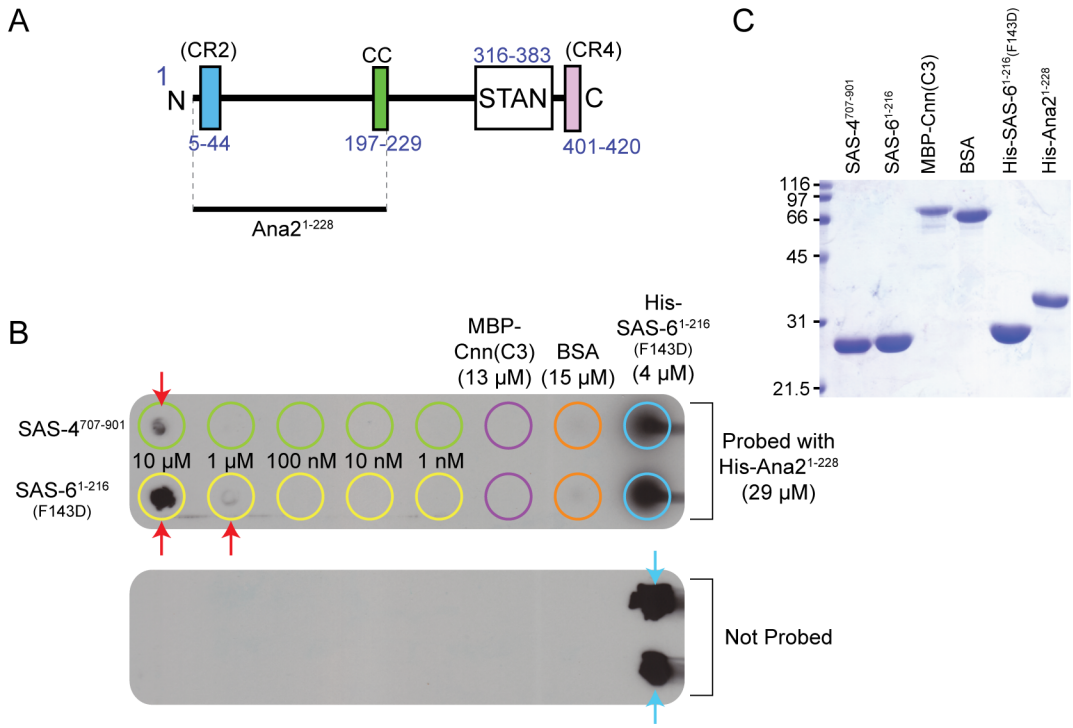


Figure 3.3.1.i: Far western blot analysis to search for interaction partners of His-Ana2$^{1-228}$. (**A**) Schematic showing the extent of the Ana2$^{1-228}$ construct, that includes the CR2 domain, up to and including the central coiled-coil. (**B**) Two identical nitrocellulose membranes onto which the indicated samples were immobilised. Note that SAS-4 and SAS-6 proteins immobilised in the green and yellow circles were untagged. The bottom membrane was not exposed to His-Ana2$^{1-228}$, whereas the top membrane was washed with His-Ana2$^{1-228}$, prior to His-tag detection using an αHis-HRP conjugate antibody. His-SAS-6$^{1-216}$(F143D) was included as a positive control for His-tag detection (lower blot, cyan arrows). There was no available positive control for Ana2 binding. His-Ana2$^{1-228}$ appears to have bound to higher concentrations of both the SAS-4 TCP domain, and SAS-6$^{1-216}$(F143D) (red arrows). Longer exposure times did not reveal binding to lower concentrations of these two constructs. (**C**) SDS-PAGE analysis of the proteins used in this experiment.

3.3.2: Initial ITC analysis of the Ana2:SAS-4 interaction.

In order to confirm that the CR2 domain of Ana2 directly interacts with the SAS-4 TCP domain, a peptide spanning only the CR2 domain (Ana2$^{5-43}$) was synthesised (Severn Biotech) (**Figure 3.3.2.i, A**). In order to assess whether the *Drosophila* MCPH mutation equivalent (E792V) affected binding, the mutant was cloned by conventional site directed mutagenesis (QuikChange, Agilent) of the wild-type plasmid. The recombinant mutated protein, SAS-4$^{700-901}$(E792V) was purified identically to the WT protein. Due to the low yield of these constructs, one purification of each variant generated only enough protein to carry out one

preliminary Isothermal Titration Calorimetry (ITC) experiment, in order to probe for a direct Ana2:SAS-4 interaction.

The SAS-4 proteins were dialysed into 50 mM Na_xH_xPO4, pH 7.3, 150 mM NaCl overnight, using 10 kDa MWCO dialysis membranes (Pierce). Dialysis buffer was used to dissolve the peptide to an approximate concentration of 1 mM, based on weighing out the lyophilised peptide, however this is unlikely to be accurate due to the small mass of peptide weighed out. Because the peptide does not absorb at 280 nm, its concentration (in the event of a successful experiment) would have to be experimentally assessed *a posteriori*, by amino acid content analysis.

ITC experiments were conducted using an iTC_{200} calorimeter (Microcal), which had an experimental cell size of 250 µl. Peptide was injected in 16 x 2 µl injections after an initial 0.5 µl pre-injection. A background experiment was first conducted, titrating $Ana2^{5-43}$ into buffer only. This was to measure the "heat of dilution" (HoD) that can result upon the dilution of concentrated ligands. This HoD is usually small in comparison to the experimental data, but is contained within it, and so must be subtracted. However, dilution of $Ana2^{5-43}$ was accompanied by a large endothermic response, which made interpretation of experimental data difficult (**Figure 3.3.2.i, B**). $Ana2^{5-43}$ was next titrated into SAS-$4^{700-901}$(WT) at a concentration of 80 µM in the experimental cell. Initial injections were accompanied by an exothermic response, but as the experiment progressed, these became endothermic (**Figure 3.3.2.i, C**). This result suggests that an exothermic interaction is taking place, and that binding is becoming saturated over the course of the experiment. However, the exothermic signal is masked by the endothermic effect of peptide dilution. Subtraction of the HoD from the experimental data resulted in very noisy data that could not be meaningfully processed in order to obtain a Kd (not shown). $Ana2^{5-43}$ was next titrated into SAS-$4^{700-901}$(E792V) at a concentration of 51 µM in the experimental cell. These

injections were all accompanied by an endothermic response, similar to the HoD, suggesting that no interaction was taking place (**Figure 3.3.2.i, D**). Although it is not possible to quantify or compare these preliminary results, they suggest that a direct interaction does indeed take place between the Ana2 CR2 domain, and the SAS-4 TCP domain, and that the E792V mutation weakens this interaction. Attempts made to optimise this experiment were unsuccessful.

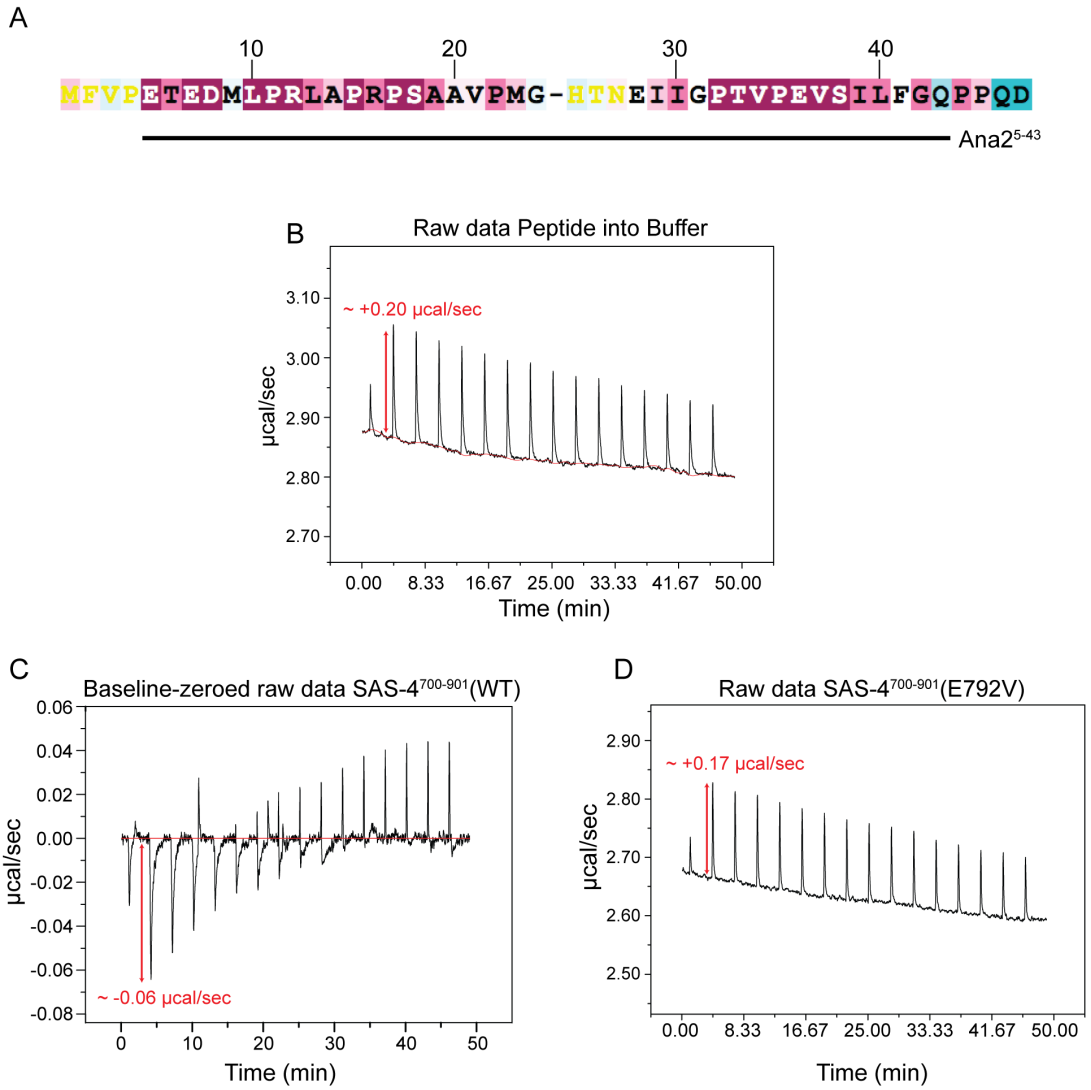


Figure 3.3.2.i: Preliminary ITC experiment to detect an interaction between the Ana2 CR2, and the SAS-4 TCP domain, (both the WT and MCPH (E792V) mutant forms). (**A**) The sequence of the Ana2 CR2 domain, coloured according to the Consurf conservation score. The black line represents the Ana2$^{5-43}$ peptide construct synthesised for this study. (**B, C, D**) Raw data resulting from the injection of 1 mM Ana2 peptide into **B**, buffer, **C**, 80 µM SAS-4$^{700-901}$(WT), and **D**, 51 µM SAS-4$^{700-901}$(E792V). The approximate magnitude of the first injection in each experiment is estimated and shown as a red bar, however to calculate this accurately, the area under the peak would have to be integrated. A large endothermic heat of dilution, evident in **B**, masks the experimental data, however the data suggest that an exothermic interaction is taking place when Ana2$^{5-43}$ was titrated into WT protein (**C**), which initially exhibited an exothermic response. (**D**) Titration into the E792V mutant was similar to titration into buffer.

3.3.3: Attempts to crystallise SAS-4$^{700-901}$ and Ana2$^{5-43}$.

In order to compare the SEC elution profile of the Ana2$^{5-43}$:SAS-4$^{700-901}$ complex, and apo-SAS-4$^{700-901}$, SEC was carried out on SAS-4$^{700-901}$ in isolation and SAS-4$^{700-901}$ to which an approximate 2.5-fold molar excess of peptide had been added (**Figure 3.3.3.i**). Experiments were carried out using an S75 10/30 column, using WB as the running buffer, and injecting ~200 µl of protein at ~2 mg/ml. This experiment might purify a homogenous complex of Ana2$^{5-43}$:SAS-4$^{700-901}$, and would also exchange the complex back into WB for crystallisation. Interestingly the SEC profiles were essentially identical. This result suggests either that the hydrodynamic radius of the complex is similar to the apo-SAS-4$^{700-901}$, or that the affinity of the complex was too weak to withstand the dilution experienced during SEC.

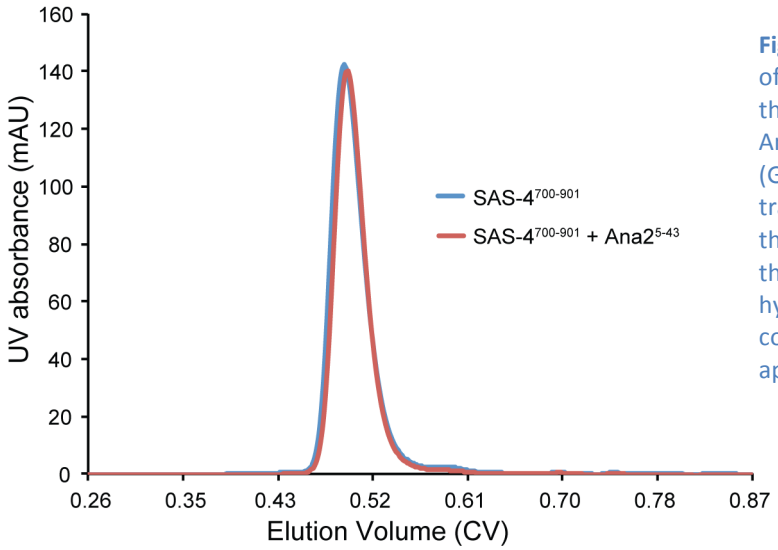


Figure 3.3.3.i: SEC chromatograms of SAS-4$^{700-901}$ in isolation, and in the presence of a molar excess of Ana2$^{5-43}$, using an S75 10/30 column (GE Healthcare). The fact that the traces are identical suggests that the Ana2$^{5-43}$ is not bound under these conditions, or that the hydrodynamic radius of the complex is very similar to apo-SAS-4$^{700-901}$.

It was not possible to detect the peptide in SDS-PAGE analysis of the SEC fractions (not shown). Crystallisation trials were conducted for SAS-4$^{700-901}$, at a maximal concentration of ~3 mg/ml. Additional peptide was also added to SAS-4$^{700-901}$, in an attempt to crystallise the complex, at a maximal concentration after peptide addition of 2.36 mg/ml. None of these trials produced crystals. Although some drops produced precipitate, most drops showed little activity, suggesting perhaps that the protein concentration was not high enough to promote nucleation.

3.4: Fusion of the Ana2 CR2 to the TCP domain of SAS-4

Due to the difficulty in detecting the peptide, and the ambiguity over the strength of the interaction, I decided to physically tether the Ana2 CR2 to the SAS-4 TCP domain. A direct fusion with the correct design might encourage complex formation with the sole objective of crystallographic characterisation of the interaction. This approach assumes, and would enforce, a 1:1 stoichiometry without the need to accurately titrate the two components or ensure that they remain associated during the purification procedure.

3.4.1: Design of Ana2 CR2 and SAS-4 TCP domain fusion proteins.

Several fusion constructs were designed to identify an arrangement that might allow the Ana2 CR2 to bind to the TCP domain. The Ana2 CR2 was to be fused at either the N-, or C-termini of SAS-4$^{700\text{-}901}$, which would provide both the N and C termini of the Ana2 CR2 to have the steric freedom to "search" for the correct binding site on the SAS-4 TCP domain. Two lengths of the Ana2 CR2 were trialled with the aim of optimising the amount of natural linker that might enable binding. In all cases, Ana2 and SAS-4 were directly fused, with no synthetic linker residues between them. The constructs were cloned by PCR amplifying Ana2 and SAS-4 inserts from codon-optimised cDNA, and performing a 3-way ligation into the pETM-14 vector, to generate constructs with a a 3C-cleavable, N-terminal His-tag. The internal restriction site between the two domains was then excised using site directed mutagenesis. The fusion proteins are detailed in **Figure 3.4.1.i, B**.

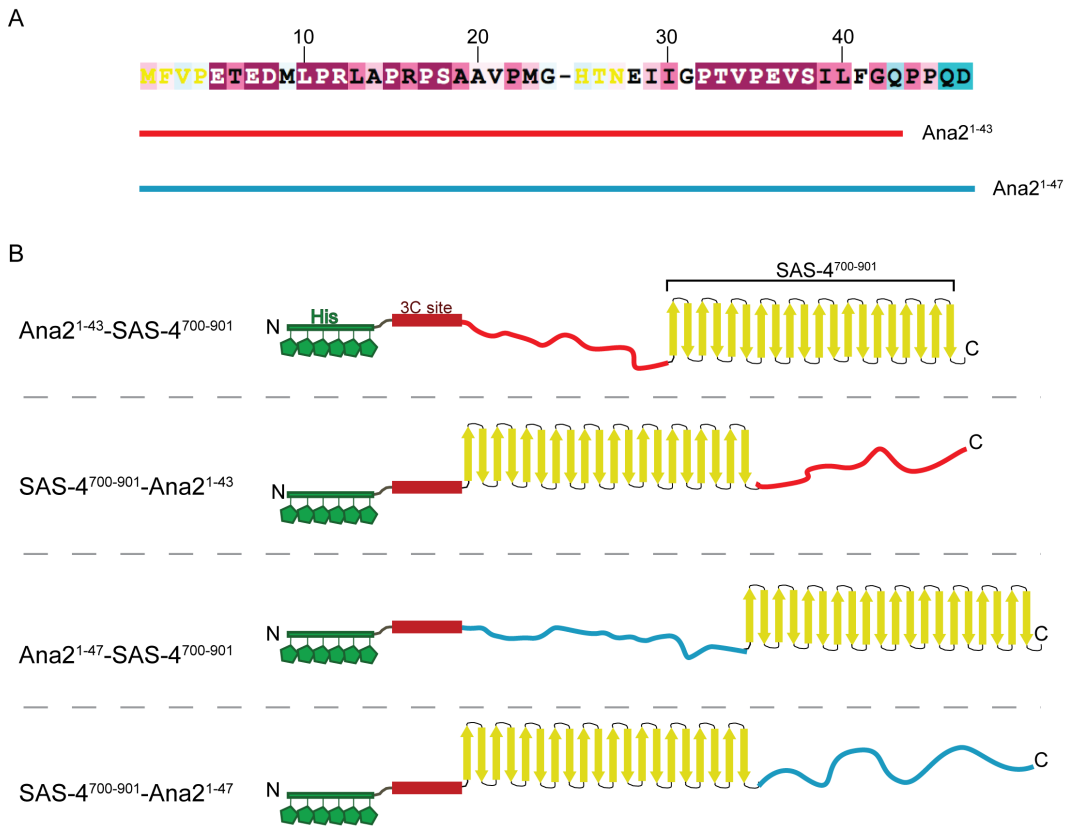


Figure 3.4.1.i: Design of the Ana2 CR2, SAS-4 TCP domain fusion constructs. (**A**) The sequence of the Ana2 CR2 domain, coloured according to the Consurf conservation score. The red line and blue lines represent the lengths of sequence that were included in the designed constructs. (**B**) Schematic representations of the four designed fusion constructs. SAS-$4^{700\text{-}901}$ is represented as a yellow β-sheet. The lengths of Ana2 CR2 that are fused are represented by the red and blue lines as described in **A**. All constructs contained a 3C-cleavable N-terminal 6 x His-tag.

3.4.2: Purification of Ana2:SAS-4 fusion proteins.

Ana$2^{1\text{-}43}$-SAS-$4^{700\text{-}901}$ could be purified, however this purification could not be easily replicated, and so will not be discussed. Efforts were focused on the fusion constructs containing Ana$2^{1\text{-}47}$. The details of these two constructs are given in **Table 3.4.2.i**. Both proteins were successfully purified using the same procedure, and so only the purification of Ana$2^{1\text{-}47}$-SAS-$4^{700\text{-}901}$ is shown in **Figure 3.4.2.i**. Plasmids were transformed into *E. coli* B834 (DE3) cells, which were cultured in

TB. Various culture volumes from 2-6 litres were used in different purifications. Later purifications used LB broth, which appeared to result in cleaner purification from Ni-NTA resin.

Construct	Amino Acids	MW (Da)	pI	ε_{280}	Abs 0.1%
His-Ana2$^{1-47}$-SAS-4$^{700-901}$	269	30621.3	6.46	34380	1.124
Ana2$^{1-47}$-SAS-4$^{700-901}$	252	28566.0	6.12	34380	1.204
His-SAS-4$^{700-901}$-Ana2$^{1-47}$	270	30752.5	6.46	34380	1.119
SAS-4$^{700-901}$-Ana2$^{1-47}$	253	28697.2	6.12	34380	1.198

Table 3.4.2.i Details of constructs in which Ana2$^{1-47}$ was fused at the N or C terminus of SAS-4$^{700-901}$, calculated using Protparam. No constructs contained cysteine. Both His-tagged, and His-cleaved versions of each construct are detailed.

Similarly to the SAS-4$^{700-901}$ construct, cell pellets were resuspended in a lysis buffer volume of ~10-15 ml per gram of cells and the lysate was rich in precipitate. SDS-PAGE analysis showed that the vast majority of recombinant protein was in the insoluble fraction (**Figure 3.4.2.i, B**). To cleave the His-tag, pooled elution fractions were subjected to digestion with GST-3C protease, which was added at a ratio of ~1 mg 3C per 10 mg target. This reaction was dialysed against fresh WB overnight at 4 ˚C. In initial purifications, the GST-3C was first depleted from the sample by flowing the reaction over glutathione resin (GSTrap, GE Healthcare). Although this worked to some extent, as evidenced by the ability to elute GST-3C from the glutathione resin, residual GST-3C could be observed in subsequent SEC traces, and so this step was abandoned in later purifications. Cleaved protein was then isolated using SEC, on either an S75 16/60, or 26/60 column (GE Healthcare). Interestingly, the two fusion proteins had different S75 SEC elution profiles. Where Ana2$^{1-47}$ was fused to the C-terminus of SAS-4$^{700-901}$ (SAS-4$^{700-901}$-Ana2$^{1-47}$), the elution peak appeared earlier, and was asymmetric, with a tail at the lagging edge of the peak. When Ana2 was fused to the N-terminus of SAS-4$^{700-901}$ (Ana2$^{1-47}$-SAS-4$^{700-901}$) the protein eluted slightly later, with a comparatively more symmetrical peak. The difference between these profiles might be explained by Ana2$^{1-47}$-SAS-4$^{700-901}$ being the most compact fusion protein. Possibly this is due to the binding of Ana2 against the SAS-4 TCP domain. I therefore focused further efforts on the Ana2$^{1-47}$-SAS-4$^{700-901}$ construct. CD analysis (not shown) confirmed that this construct folded similarly to the SAS-4$^{700-901}$ construct.

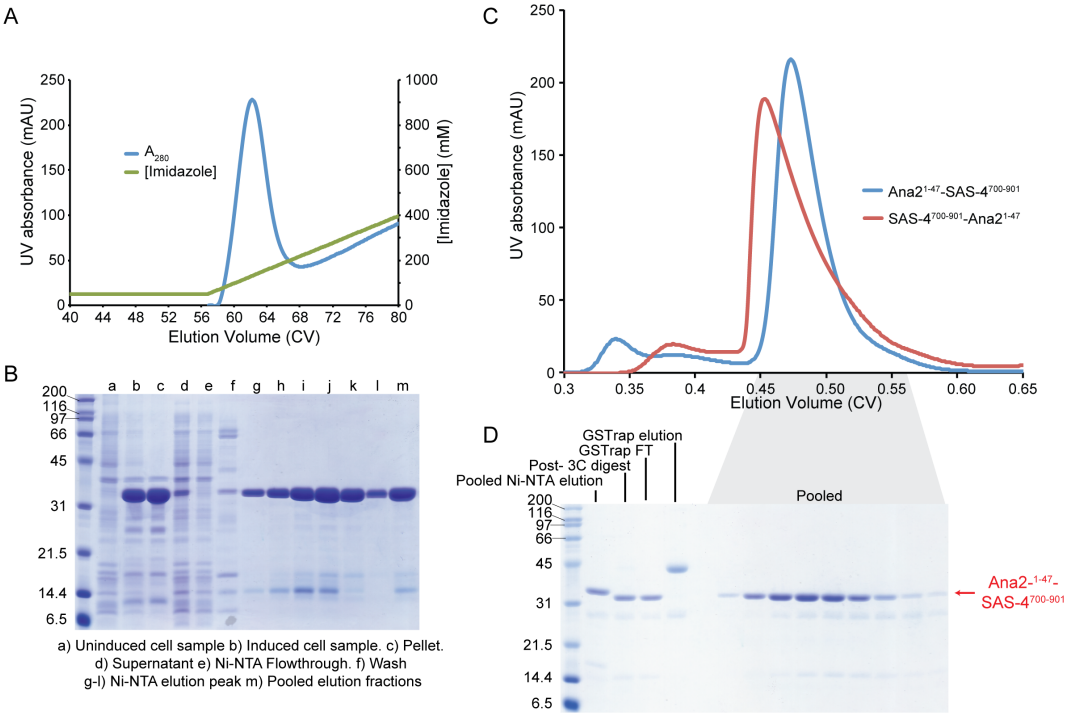


Figure 3.4.2.i: Purification of the Ana2$^{1-47}$-SAS-4$^{700-901}$ fusion protein, and comparison to the SAS-4$^{700-901}$-Ana2$^{1-47}$ fusion protein. (**A**) Elution of His-tagged Ana2$^{1-47}$-SAS-4$^{700-901}$ from Ni-NTA resin using a shallow imidazole gradient. (**B**) SDS-PAGE analysis of purification fractions of Ana2$^{1-47}$-SAS-4$^{700-901}$. (**C**) S75 SEC traces of two different fusion proteins, where Ana2$^{1-47}$ has been fused at the N-terminus (Ana2$^{1-47}$-SAS-4$^{700-901}$, blue trace) or the C-terminus (SAS-4$^{700-901}$-Ana2$^{1-47}$, red trace). The profiles are not identical. SAS-4$^{700-901}$-Ana2$^{1-47}$ elutes earlier, and the peak is asymmetric, showing an elongated tail. In comparison, Ana2$^{1-47}$-SAS-4$^{700-901}$, elutes slightly later, and has a more symmetrical profile, suggesting it is the more compact construct. (**D**) SDS-PAGE analysis of His-cleavage, GST depletion, and SEC elution fractions for Ana2$^{1-47}$-SAS-4$^{700-901}$. The contaminant bands present in some purifications prompted the introduction of an IEX chromatography step (**section 3.4.3**).

3.4.3: Further purification of Ana2$^{1-47}$-SAS-4$^{700-901}$ using anion Exchange

SDS-PAGE analysis of Ana2$^{1-47}$-SAS-4$^{700-901}$ suggested that the sample contained some contaminants or degradation products that might hinder crystallisation (**Figure 3.4.3.i, A**). In an attempt to remove these contaminants, Ana2$^{1-47}$-SAS-4$^{700-901}$ was purified and cleaved as in **Section 3.4.2**, and subsequently subjected to ion-exchange chromatography. According to the calculated pI of 6.12 for cleaved Ana2$^{1-47}$-SAS-4$^{700-901}$, it should be negatively charged when at a pH of 8.5. Purification via anion-exchange was therefore trialled, using a MonoQ 5/50 GL column (GE Healthcare), and using a series of different pH conditions into which the protein was

dialysed. Unexpectedly, attempts to bind the protein at pH 8.5 and 9.0 were unsuccessful or inefficient (**Table 3.4.3.i**, **Figure 3.4.3.i, B**), suggesting that the true pI of the construct may be higher than that calculated theoretically. The protein was therefore dialysed overnight into 25 mM CHES pH 10.0, 50 mM NaCl. This time, no protein was evident in the column flowthrough, and Ana2$^{1-47}$-SAS-4$^{700-901}$ eluted as a sharp double peak (**Figure 3.4.3.i, C**). Gel analysis showed that the fractions across this double peak appeared to be highly enriched with full-length protein – purer than could be achieved by SEC. This procedure therefore became the *de facto* method for future purifications. In all cases, the protein was loaded at 0.5 ml/min, onto a pre-equilibrated column, then washed for 40 CVs at 50 mM NaCl, before elution using a linear NaCl gradient.

IEX Buffer	pH	MonoQ binding?
25 mM Tris, 50 mM NaCl	8.5	No
25 mM Bicine, 50 mM NaCl	9.0	Partial
25 mM CHES, 50 mM NaCl	10.0	Yes

Table 3.4.3.i IEX buffer conditions trialled for Ana2$^{1-47}$-SAS-4$^{700-901}$.

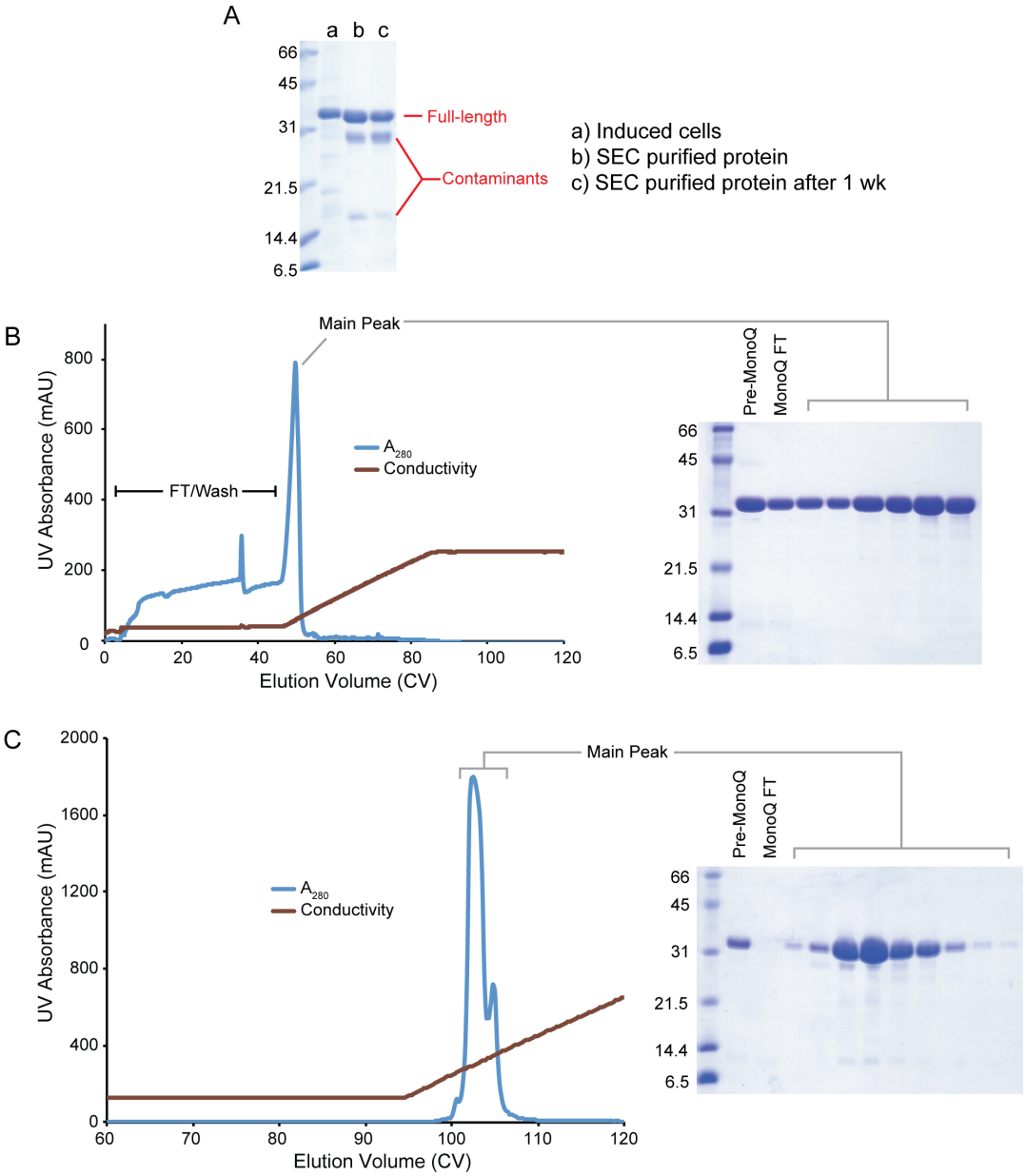


Figure 3.4.3.i: Ion exchange chromatography of Ana2$^{1-47}$-SAS-4$^{700-901}$. (**A**) SDS-PAGE analysis of SEC purified protein (**Section 3.4.2**). The sample contained contaminants, however it appeared to be stable, as no spontaneous degradation was evident after a week. (**B**) Non-optimal anion exchange, and subsequent gel analysis of Ana2$^{1-47}$-SAS-4$^{700-901}$ in Bicine buffer, pH 9.0. Much of the sample didn't bind the column and protein that did bind, eluted very easily. Note that protein used in this experiment had already been passed over anion exchange resin under different conditions, which might account for the apparent lack of contaminants observed in the "Pre-MonoQ" sample. (**C**) Successful anion exchange, and subsequent gel analysis of Ana2$^{1-47}$-SAS-4$^{700-901}$ using CHES buffer, pH 10.0. Very little protein flowed through the column under these conditions (not shown in the trace). Protein eluted as a double peak, but gel analysis suggested that both peaks contained the same species. Eluted fractions were highly concentrated, and although minor contaminants were still evident, the sample appeared much cleaner than the sample shown in **A.**

3.5: Crystallisation and data collection of Ana2$^{1-47}$-SAS-4$^{700-901}$.

Fusion of Ana2 to the N-terminus of the TCP domain appeared to produce a species that was more compact in size than a C-terminal fusion, and so I took this construct through to crystallisation trials.

3.5.1: Crystallisation of Ana2$^{1-47}$-SAS-4$^{700-901}$

I first tried to gauge the protein concentration that might produce favourable nucleation and crystallisation conditions. Purified protein was dialysed back into purification working buffer (50 mM Tris pH 8.5, 150 mM NaCl), and concentrated using 10 kDa MWCO spin concentrators (Amicon). Interestingly, the fusion protein had a higher solubility limit than the apo-SAS-4$^{700-901}$, and could be concentrated maximally to ~6 mg/ml, however fine precipitate could sometimes be observed at these concentrations which could be apparently resolubilised upon dilution. Trials were carried out with and without removal of this precipitate. Using the STURA/Macrosol screen (Molecular Dimensions), sitting drop experiments were set up at 2, 4 and 6 mg/ml. After several days, precipitative and phase-separating activity was only evident at the highest concentration of ~6 mg/ml, and so further trials were set up at 5.0–6.2 mg/ml, using various commercial screens. Although some trials were set up using SEC purified protein, only IEX purified protein eventually crystallised. After ~2-3 weeks, crystals grew in phase-separated conditions of the Morpheus screen (**Figure 3.5.1.i**) that contained a common precipitant mix of 20% Ethylene Glycol, 10%

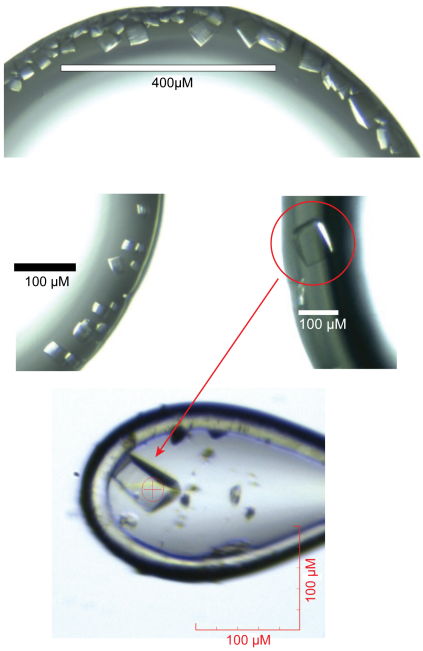


Figure 3.5.1.i: Selected crystals of Ana2$^{1-47}$-SAS-4$^{700-901}$. All crystals grew in Morpheus conditions containing 20% Ethylene Glycol, 10% PEG 8k. The circled crystal had the largest dimensions, and subsequently produced the strongest diffraction. It can be seen below mounted in a cryoloop at Diamond Beamline I04.

PEG 8k, but across a range of additives. Typically, crystals grew at pH 6.5, but were often overnucleated and small, apart from one, the conditions of which are described in **Table 3.5.1.i**. Crystals were harvested after ~4 weeks with the mother liquor serving as the cryoprotectant.

Crystal	Protein concentration (mg/ml)	Mother liquor	µl protein:µl Mother liquor	Temperature (˚C)
Morpheus_A2a	6.2	100mM MES/Imidazole mix pH 6.5, 30 mM $CaCl_2$, 30 mM $MgCl_2$, 20.0% Ethylene Glycol, 10.0% PEG 8000	0.15:0.05	21

Table **3.5.1.i:** Crystallisation conditions for native Ana2$^{1-47}$-SAS-4$^{700-901}$.

3.5.2: Data collection and processing of Ana2$^{1-47}$-SAS-4$^{700-901}$

Data were collected at Diamond Light Source, I04. After screening shots, no rotational crystallographic symmetry was obvious suggesting a spacegroup of P1. Most crystals diffracted to ~2.5-3.0 Å, and for the best crystal (**Table 3.5.1.i**), 288˚ of data were collected. The data were processed using the Xia2 pipeline [Winter, 2010] using the 3daii mode, which used XDS [Kabsch, 2010] to identify, index, and integrate spots, and AIMLESS [Evans and Murshudov, 2013] for scaling and merging. This produced a complete dataset with reliable statistics to 2.57 Å (**Table 3.5.2.i**).

	Ana2$^{1-47}$-SAS-4$^{700-901}$
Beamline	Diamond I04
Spacegroup	P1
Wavelength	0.9795 Å
Unit cell dimensions (Å/˚)	a = 58.64 b = 69.91 c = 69.98 α = 86.96 β = 88.64 γ = 67.69
Resolution (Å)	64.60–2.57
Completeness (overall/inner/outer)	97.6/93.6/97.2
R_{merge} (overall/inner/outer)	0.091/0.069/0.512
R_{pim} (overall/inner/outer)	0.061/0.035/0.449
CC (½) (overall/inner/outer)	0.993/0.993/0.710
I/σI (overall/inner/outer)	7.6/16.3/1.7
Multiplicity (overall/inner/outer)	3.1/2.9/3.1
No. unique reflections	31,911

Table 3.5.2.i: Processing statistics for the native Ana2$^{1-47}$-SAS-4$^{700-901}$ dataset.

Analysis of the unit cell was carried out using the Matthews coefficient [Matthews, 1968], and suggested that the ASU would most likely contain 4 copies of the construct (**Table 3.5.2.ii**).

Chains/ASU	Matthews Coeff	% Solvent	P(2.57 Å structures)	P(total PDB)
2	4.67	73.70	0.01	0.01
3	3.12	60.55	0.19	0.19
4	2.34	47.40	0.71	0.69
5	1.87	34.26	0.09	0.10

Table **3.5.2.ii:** Matthews Coefficient analysis calculated with the native dataset, estimating a total of 252 residues per chain.

A self-rotation function (SRF) was calculated in molrep in order to assess if there were any likely non-crystallographic rotational symmetry axes. The strongest peaks were both weak (~20-30% of the origin) and were evident in the chi = 180˚ section of the SRF, approximately along, and orthogonal to the z axis (**Figure 3.5.2.i**), suggesting a pseudo-orthorhombic arrangement. This was intriguing as the α and β angles of the unit cell were approximately 90˚. However there were many more weak peaks (~14-20% of the origin) around the xy plane, and so in the absence of stand out peaks, I concluded that there was no significant rotational NCS.

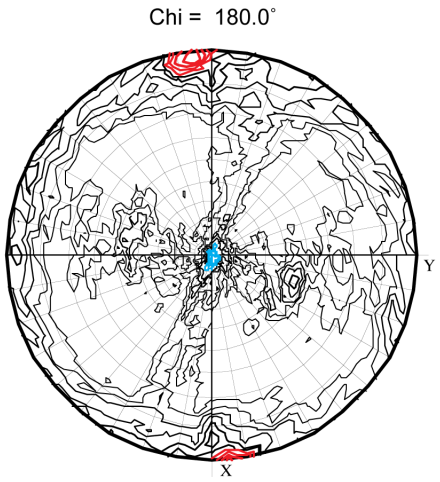


Figure 3.5.2.i: Chi=180˚ section of the self-rotation function of Ana2$^{1-47}$-SAS-4$^{700-901}$ plotted using Molrep (ref). Two weak peaks were identified ~orthogonal to (red) and ~along (blue) the z axis. However these did not stand out significantly above other peaks in the xy plane, suggesting that no obvious rotational NCS existed in the crystal.

3.5.3: Attempts at molecular replacement

I anticipated that a phase solution could be found relatively trivially using molecular replacement (MR), as I had access to the structure of the *D. rerio* CPAP TCP domain, provided by

my collaborator Dr Mark van Breugel (**Figure 3.2.1.i, B**). This domain has ~42% identity to the *Drosophila* SAS-4 TCP domain and was therefore expected to serve as an ideal search model. The domain was unusual, consisting of a single continuous beta sheet with a gradual twist, but the structure was consistent with the bioinformatic and *in vitro* characterisation of DmSAS-4 thus far, and so the *Drosophila* structure was predicted to be similar. I therefore generated a homology model by mapping the *Drosophila* SAS-4 sequence onto the DrCPAP structure, and trimming sidechains back to the last common atom using CHAINSAW [Stein, 2008] (**Figure 3.5.3.i, A**). As well as the full-length model, truncated models were made, that included roughly the first, middle and last 2/3 of the structure, in case any proteolysis or unfolding had occurred *in crystallo*. Additionally, ensembles of subtly perturbed models were generated by normal

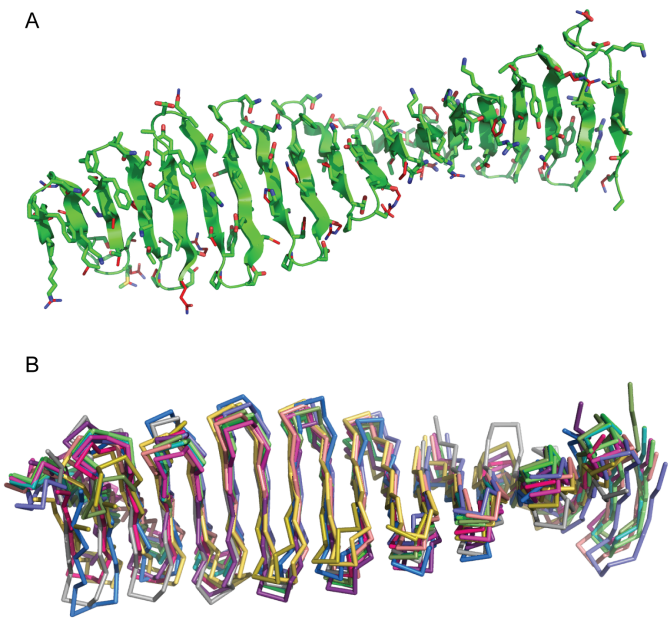


Figure 3.5.3.i: Molecular Replacement models for the SAS-4 TCP domain, based on the structure of *D. rerio* CPAP. (**A**) A model generated by CHAINSAW (Stein, 2008). The *Drosophila* SAS-4 sequence was mapped onto the DrCPAP structure according to sequence alignment analysis. Non-conserved sidechains were trimmed back to the last common atom. The *Drosophila* model is shown in green cartoon format, with sidechains as green sticks. The *D. rerio* CPAP structure is underlaid, with sidechains shown in red sticks to show which atoms have been trimmed. (**B**) A sample ensemble of full length models that have been perturbed according to normal modes. Note that many truncated models were also trialled.

mode analysis (NMA) using the ElNémo server [Suhre and Sanejouand, 2004a] (**Figure 3.5.3.i, B**). This is useful as many protein domain conformations and movements can be well modelled - simply by perturbing the structure according to low frequency normal modes [Suhre and Sanejouand, 2004b]. In this way, it can be possible to pre-empt minor structural flexing between the model, and the actual structure. However, despite these different approaches, MR results using PHASER [McCoy et al., 2007] were consistently ambiguous. Using various ensembles, many

solutions could be found that had a TFZ score of 7-10; scores that might normally be considered as statistically significant. These solutions were difficult to evaluate, as tens of similarly scoring solutions would usually be found in the same search. Rather than being obvious variations of the same solution, these joint solutions were often subtly different in terms of positioning and packing, and so it would not be possible for them to all be correct. The ambiguity was further amplified by the fact that several copies of the construct were likely present in the asymmetric unit. It was also not possible to convincingly improve any of these solutions by refinement – which would normally be the case for a correct solution. I suspected that, due to the repetitive nature of the structure, it was possible to place a model with the β-strands out of register in many different ways, but all these placements would similarly explain the data, and it would not be possible to discriminate a correct solution from others. As these molecular replacement results were highly ambiguous, it was decided that experimental phasing methods might offer an alternative way to obtain a phase solution.

3.6: Production of selenomethionine derivative crystals, and data collection

In the absence of a correctly placed molecular model from which to calculate phase information, phases would have to be experimentally determined. This requires the production of heavy atom derivative crystals in order to carry out isomorphous replacement or anomalous dispersion techniques. In either technique, the heavy atom scattering component in the dataset can be used to locate the heavy atoms, which in turn can be used to phase the contents of the entire unit cell (**Appendix 7.2.5**). Initially, heavy atom soaks were performed on native crystals in order to introduce heavy atoms, however the data collected from such crystals were very poor and difficult to process (not shown). It was decided to reproduce the native crystal form using selenomethionine (SeMet) derivatised protein, where the anomalous scatterer, selenium, replaces the sulphur atom of each methionine residue. There are six methionine residues in the Ana2$^{1-47}$-SAS-4$^{700-901}$ construct, and three of these are in the TCP domain, which is likely to be ordered in the crystal. SeMet derivation was therefore a realistic option to introduce ordered anomalous scatterers into the crystal.

3.6.1: Production of selenomethionine-derivative Ana2$^{1-47}$-SAS-4$^{700-901}$

Similarly to native Ana2$^{1-47}$-SAS-4$^{700-901}$, the SeMet derivative was expressed in *E. coli* strain B834 (DE3), which is a methionine auxotroph. Growing the culture in a medium lacking methionine, and containing only SeMet can therefore ensure 100% SeMet derivatisation. Transformed cells were washed in PBS to remove residual LB medium, before being cultured at 37 ˚C in 4 L of selenomethionine (+glucose) complete medium (Molecular Dimensions), supplemented with SeMet (Acros Organics, Belgium) at 40 mg per litre. Cultures were inoculated at a starting OD_{600} of ~0.05 such that the cultures would grow sufficiently using SeMet in place of endogenous methionine. After induction, cultures grew slightly slower compared to native conditions, and so

expression was induced at an OD_{600} of ~0.8 (rather than ~0.6), overnight at 21 ˚C, using 1mM IPTG. Purification was carried out using Ni-NTA chromatography, His-tag cleavage, and IEX methods (**Figure 3.6.1.i**), identically to the native protein, although typical yields were slightly lower for the SeMet derivative. ESI-MS analysis (kindly performed by Dr David Staunton) confirmed 100% selenomethionine incorporation into the protein.

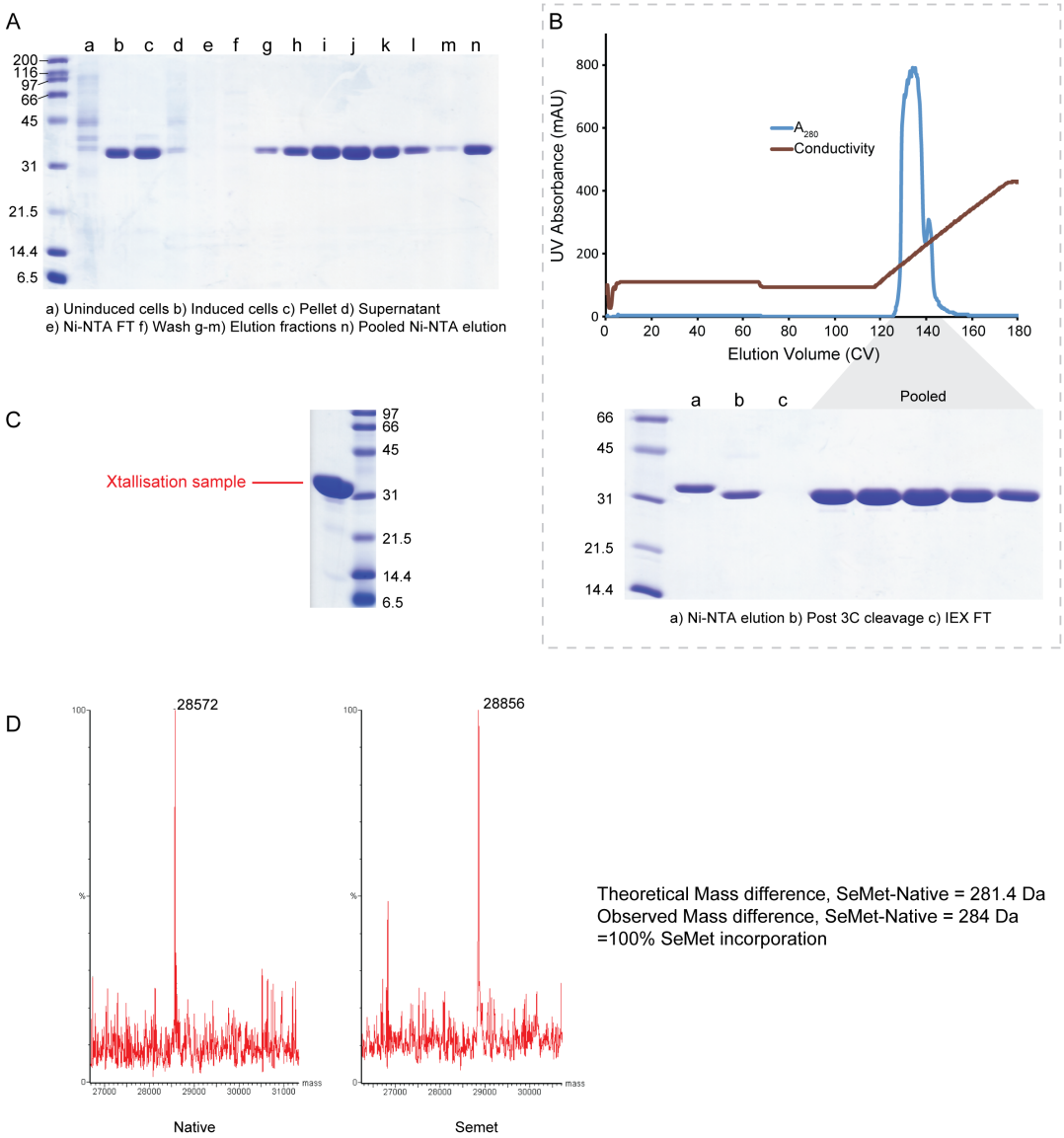


Figure 3.6.1.i: Purification of SeMet derivative Ana2$^{1-47}$-SAS-4$^{700-901}$. (**A**) SDS-PAGE analysis of fractions from NiNTA purification. (**B**) Anion Exchange in a CHES ph 10.0 buffer and subsequent gel analysis. (**C**) Representative SDS-PAGE analysis of a sample as used for crystallisation trials. (**D**) ESI-MS analysis showing the observed mass of purified native (left) and SeMet (right) Ana2$^{1-47}$-SAS-4$^{700-901}$. The data suggest 100% incorporation of selenomethionine.

3.6.2: Crystallisation of SeMet-Ana2$^{1-47}$-SAS-4$^{700-901}$

SeMet-Ana2$^{1-47}$-SAS-4$^{700-901}$ was concentrated similarly to native protein (in a 50 mM Tris pH 8.5, 150 mM NaCl buffer), although the maximal concentration that could be reached was now ~5 mg/ml, with visible precipitate again forming around this concentration. Trials were initially carried out identically to the native protein, but using only the Morpheus screen, or selected conditions thereof, in order to save material. Because the native crystals took several weeks to grow, repeat trials were set up where the drops were left open for increments of 5 minutes before sealing, in an attempt to accelerate evaporation, and hopefully crystallisation. Trials were also set up where the protein resided in the final buffer, as eluted from IEX (25 mM CHES ph 10.0, ~150 mM NaCl). At this pH, the protein was more soluble, and trials could be conducted at 8–10 mg/ml. Despite intense phase separation, none of the CHES trials produced crystals, however a few small crystals appeared in some of the Tris trials after ~3 weeks. Only one of these crystals was of diffraction quality. In an attempt to introduce crystalline nuclei to initiate crystal growth, streak seeding [Stura and Wilson, 1991] was carried out on some drops, using a black cat whisker, and some native crystals, although this did not produce any results. In order to produce more crystals, an optimisation screen was designed to fine-tune pH, and precipitant concentration (**Figure 3.6.2.i, A**). This trial also included the use of microcrystalline seeds, which were produced from small, non-diffraction quality SeMet crystals that had grown in the initial Tris trials. Dilutions of “seed stock” were produced using a seed bead kit (Hampton) (**Figure 3.6.2.i, B**) A small amount of this seed solution was then included in each drop of the screen to initiate crystal growth. In this manner, the optimisation trays produced large numbers of SeMet derivative crystals (**Figure 3.6.2.i, C, D**).

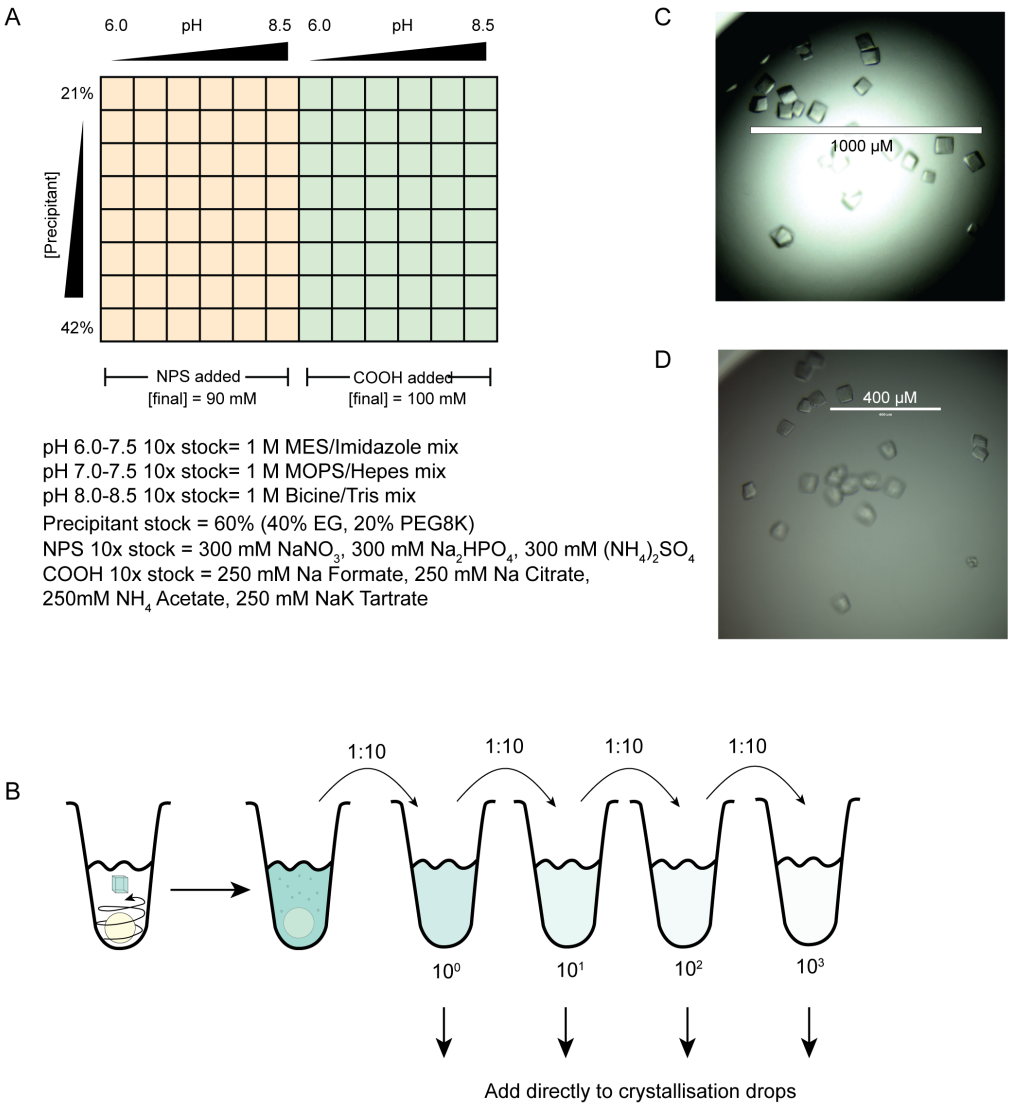


Figure 3.6.2.i: (**A**) Optimisation screen used to obtain SeMet crystals. The best crystals grew in the presence of the NPS additives. (**B**) Production of microcrystalline seeds using a seed bead kit. A drop containing crystals was resuspended in 50 µl mother liquor. Crystals were mechanically fragmented by vortexing in the presence of a PTFE bead. This stock is then serially diluted into fresh mother liquor to create varying concentrations of seed solution. Small amounts of these dilutions are added to crystallisation experiments. (**C**+**D**) SeMet derivative crystals grown using this approach.

Although many crystals diffracted poorly, together, these trials eventually provided four crystals that were subsequently used in structure solution. The crystallisation conditions in which they were grown are presented in **Table 3.6.2.i**.

Crystal	Protein concentration (mg/ml)	Mother liquor	µl protein:µl Mother liquor	µl Seed Stock
SeMet 1	5.00	100 mM MES/imidazole mix pH 6.5, 20% ethylene glycol, 10% PEG8000, 0.2 M racemic glutamic acid, 0.2 M glycine, 0.2 M racemic serine, 0.2 M racemic alanine, 0.2 M racemic lysine HCl	0.1:0.1	-
SeMet 2	5.29	100 mM MES/imidazole mix pH 6.5, 14% ethylene glycol, 7% PEG8000, 30 mM $NaNO_3$, 30 mM $NaPO_4$, 30 mM NH_4SO_4	0.3:0.1	0.05 (10^3)
SeMet 3	5.29	100 mM MES/imidazole mix pH 6.5, 14% ethylene glycol, 7% PEG8000, 30 mM $NaNO_3$, 30 mM $NaPO_4$, 30 mM NH_4SO_4	0.3:0.1	0.05 (10^3)
SeMet 4	5.29	100 mM MES/imidazole mix pH 6.5, 16% ethylene glycol, 8% PEG8000, 30 mM $NaNO_3$, 30 mM $NaPO_4$, 30 mM NH_4SO_4	0.3:0.1	0.05 (10^3)

Table **3.6.2.i:** Crystallisation conditions for SeMet Ana2$^{1-47}$-SAS-4$^{700-901}$ crystals.

3.6.3: Collection and processing of anomalous data for SeMet-Ana2$^{1-47}$-SAS-4$^{700-901}$ crystals

In order to use anomalous diffraction techniques for phase solution, it was necessary to measure data containing anomalous signal. Each reflection (h,k,l) in a dataset is related to another at (-h,-k,-l), resulting from reflection from either side of the lattice plane. The dataset therefore has an inherent centrosymmetry, which essentially halves the number of reflections one must observe in order to obtain a complete dataset. These reflections are known as “Friedel pairs” or “mates”, and normally, these reflections have equal amplitudes and inverse phases. Because only the intensity (I) is measured in a diffraction experiment, these reflections can be averaged in a dataset. However, in the presence of anomalous scattering components, the amplitudes differ, and the spots undergo a 90˚ phase shift, thus breaking the centrosymmetric relationship (**Appendix 7.2.5**). The reflections are now no longer equivalent, with one being defined “I+”, and its Friedel mate, “I-“, and it is this difference in amplitude that constitutes the anomalous content of a dataset. Due to the breakdown of centrosymmetry, both Friedel mates must be

measured, thereby doubling the amount of data necessary for a complete dataset (known as the anomalous completeness). Ideally, one would observe each Friedel pair multiple times (the anomalous multiplicity) to increase precision, and the differences observed should be seen to correlate between equivalent reflections. This is easier for crystals that exhibit a high symmetry, as symmetry axes multiply the occurrence of each reflection within the dataset. The differences between these symmetry related "Bijvoet" pairs can then be averaged to increase precision. However, in P1 there is no symmetry to increase this multiplicity.

Collection of anomalous data from SeMet-Ana2$^{1-47}$-SAS-4$^{700-901}$ crystals proved challenging. First, the spacegroup of these crystals was P1, which meant that in practise, even with a full 360˚ of data, the anomalous multiplicity was only ~1.6-1.8. Second, the large angular range of data required meant that crystals had to accumulate a relatively large dose of X-ray exposure, and some crystals noticeably deteriorated in diffraction quality or resolution with dose. This made it difficult to collect a high anomalous multiplicity, and for many crystals, a second 360˚ pass was not possible to process. Third, the SeMet crystals were comparatively smaller than the native crystal, and consistently diffracted more weakly. Data were therefore collected for a number of crystals at different beamlines in an attempt to obtain datasets with sufficient anomalous signal.

To collect maximum anomalous selenium signal it is necessary to collect at and around the wavelength of the X-ray absorption edge (where the f′ (dispersive/real) and f′′ (anomalous/imaginary) scattering components differ maximally). A "peak" wavelength is where f′′ is greatest, and an inflection wavelength is where f′ is greatest. Collecting at different wavelengths enables more accurate phase determination, however the optimal wavelength range is small, and it was therefore useful to conduct an X-ray fluorescence scan on the crystals to characterise the absorption edge (**Figure 3.6.3.i**). These scans guided the choice of

wavelength for data collection, although due to the poor quality of the crystals, wavelengths at or just past the peak wavelength were usually targeted to maximise anomalous signal, rather than maximising dispersive differences.

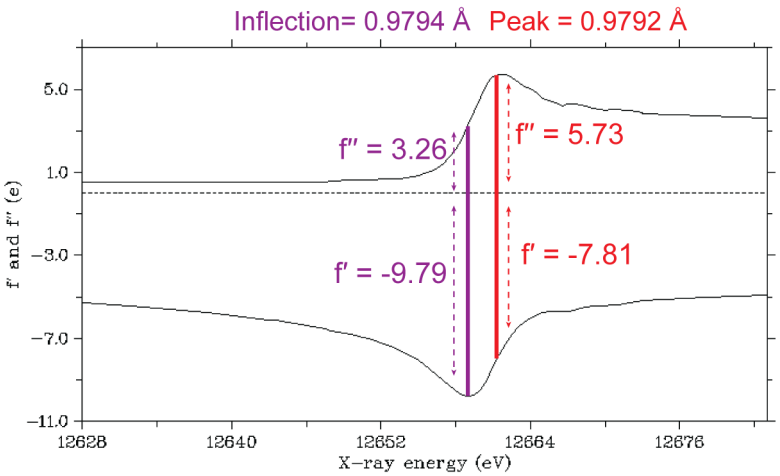


Figure 3.6.3.i: X-ray Fluorescence scan conducted on a SeMet Ana2$^{1-47}$-SAS-4$^{700-901}$ crystal at Diamond I03, at 0.8% beamline transmission. Using this profile, it was possible to determine "peak" wavelengths at which the f' and f'' components differ maximally.

Despite the difficulties in collecting anomalous data, it was eventually possible to generate four peak datasets (albeit at slightly different wavelengths from two separate beamlines), a low-resolution inflection and a low energy remote, which were used for phase solution (**Table 3.6.3.i**). These datasets were processed using data from four different crystals (although for several of these datasets, the data had to be cut at ~3.5 Å to give reliable outer shell statistics). The Xia2 pipeline [Winter, 2010] was used in 3daii mode as for the native dataset, but also searching for anomalous signal, so that Friedel mates were processed independently.

	SeMet 1-Peak	SeMet 1-Lrem	SeMet 2-Peak
Beamline	Diamond I04	Diamond I04	Diamond I03
Spacegroup	P1	P1	P1
Wavelength	0.9795 Å	0.9999 Å	0.9792 Å
Unit cell dimensions (a, b, c, α, β, γ)	59.31, 70.02, 70.01, 87.65, 89.24, 67.37	59.14, 70.24, 70.13, 87.62, 89.12, 67.35	58.47, 70.15, 69.99, 87.08, 88.41, 67.60
Resolution (Å)	54.74–3.50	64.77–3.50	64.80–3.44
Degrees viewed	360	360	720
Completeness (overall/inner/outer)	98.1/93.8/98.3	98.4/98.0/98.2	97.8/91.6/93.7
R_{merge} (overall/inner/outer)	0.093/0.055/0.188	0.086/0.041/0.238	0.170/0.076/0.518
R_{pim} (overall/inner/outer)	0.093/0.055/0.188	0.086/0.041/0.0.238	0.107/0.048/0.321
CC (½) (overall/inner/outer)	0.991/0.988 /0.972	0.995 /0.996/0.956	0.989 /0.993/0.919
I/σI (overall/inner/outer)	9.1/16.9/5.6	10.9/24.7/4.9	7.3/18.8/3.5
Multiplicity (overall/inner/outer)	3.9/3.8/3.9	3.9/3.8/3.8	7.0/7.0/7.1
No. unique reflections	12,832	12,884	13,315
Anom Completeness (overall/inner/outer)	94.7/85.9/94.7	94.4/93.5/93.8	94.7/91.6/90.9
Anom Multiplicity (overall/inner/outer)	1.9/1.8/1.9	1.9/1.9/1.9	3.4/3.5/3.5
Anomalous Correlation (overall/inner/outer)	0.149/0.259/-0.005	0.072/0.056/0.015	0.124/0.391/0.088
Overall Phasing Power	0.672	0.290	0.448
Resolution at which phasing power <1.0	6.12 Å	69.90 Å	6.70 Å
	SeMet 3-Peak	**SeMet 3-Infl**	**SeMet 4-Peak**
Beamline	Diamond I03	Diamond I03	Diamond I04
Spacegroup	P1	P1	P1
Wavelength	0.9791 Å	0.9794 Å	0.9791 Å
Unit cell dimensions (a, b, c, α, β, γ)	58.56, 70.03, 70.14, 86.93, 88.39, 68.09	58.72, 70.06, 70.28, 86.84, 88.47, 68.36	59.01, 70.17, 70.15, 87.16, 88.64, 67.58
Resolution (Å)	70.04–3.50	70.17–4.60	64.80–3.36
Degrees viewed	720	540	360
Completeness (overall/inner/outer)	97.6/93.6/97.2	97.8/79.1/89.9	97.6/91.0/97.4
R_{merge} (overall/inner/outer)	0.152/0.086/0.336	0.116/0.039/0.189	0.125/0.037/0.433
R_{pim} (overall/inner/outer)	0.100/0.058/0.224	0.116/0.039/0.189	0.125/0.037/0.433
CC (½) (overall/inner/outer)	0.992/0.996/0.954	0.992/0.989/0.974	0.988/0.994/0.889
I/σI (overall/inner/outer)	9.1/27.2/3.6	6.0/20.8/4.6	6.9/22.5/2.7
Multiplicity (overall/inner/outer)	6.0/6.7/5.3	3.5/3.6/3.5	3.5/3.4/3.6
No. unique reflections	12,797	5,641	14,461
Anom Completeness (overall/inner/outer)	93.8/93.9/89.4	90.4/77.3/81.7	92.2/86.7/93.2
Anom Multiplicity (overall/inner/outer)	2.8/3.3/2.4	1.7/1.8/1.7	1.7/1.7/1.7
Anomalous Correlation (overall/inner/outer)	0.081/0.280/0.085	-0.067/-0.059/-0.080	0.140/0.429/0.040
Overall Phasing Power	0.350	0.055	0.307
Resolution at which phasing power <1.0	6.70 Å	69.90	7.48 Å
OVERALL FOM for datasets		**0.270**	

Table 3.6.3.i: Processing statistics for the SeMet crystals. Note that because anomalous signal is searched for and Bijvoet pairs should differ, R_{merge} and R_{pim} are presented for reflections within Bijvoet pairs (within I+/I-). Note that in P1, R_{pim}=R_{merge} at low multiplicities. Semet 2-Peak was used as the reference in autoSHARP. Note also that the Infl and Lrem datasets appeared to offer no phasing power, however jobs in which these datasets were removed were unsuccessful, seemingly due to differences in the scaling together of the datasets.

The datasets contained weak anomalous signal, and had low anomalous correlations, suggesting that anomalous signal was not consistently measured. However, this statistic was difficult to interpret, as the anomalous multiplicity was often very low, meaning correlations must have been based on few observations. This can be seen in the variability of correlations as a function of resolution, rather than a smooth drop (**Figure 3.6.3.ii)**.

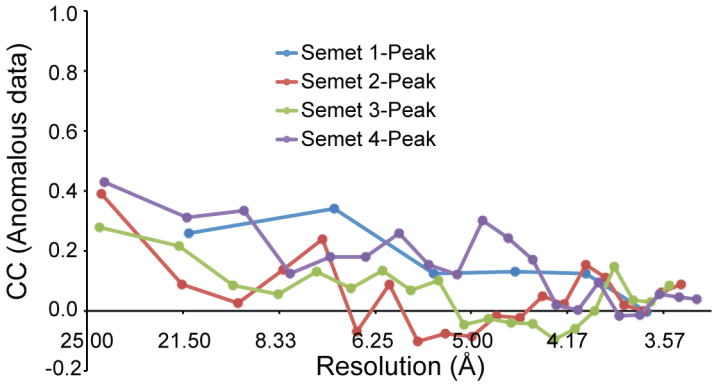


Figure 3.6.3.ii: The correlation of anomalous signal across different resolution shells for the four "peak" anomalous datasets, illustrating the difficulty in measuring strong and precise anomalous signal from these crystals.

3.7: Solving the structure of Ana2$^{1-47}$-SAS-4$^{700-901}$ using multi crystal anomalous diffraction

3.7.1: Preparation of datasets and determination of HA substructure

Phase information was derived from 6 datasets collected from four SeMet crystals, collected at two separate beamlines, and across different wavelengths, although four of these wavelengths were at or around the peak wavelength for anomalous selenium scattering (**Table 3.6.3.i**). The lack of symmetry in the P1 spacegroup means that there are several ways in which to define the unit cell axes, and so this required the datasets to be initially reindexed such that their reflections were equivalently defined. The datasets were then fed into the autoSHARP [Vonrhein et al., 2007] pipeline of programs for experimental phasing via Single or Multi-wavelength anomalous dispersion (SAD/MAD). First, the datasets were prepared for heavy atom detection, which involves scaling of the datasets relative to one another using SCALEIT [Howell and Smith, 1992]. The consistency between cell dimensions and reflection amplitudes was analysed, followed by rejection of outlying reflections. Reflections are also only included if they have been measured at least twice. SHELXC [Sheldrick, 2007] then defined the resolution cutoffs within which to find useful anomalous signal, based on the correlation of anomalous signal from combinations of datasets (MAD), or within single datasets (SAD). Heavy atom sites were then detected and determined using SHELXD [Sheldrick, 2007]. By analysing only the anomalous component of the amplitudes, the positions of the selenium atoms can be determined, via Patterson and direct methods, according to their interatomic vectors. The most correlative anomalous signal (CC(E) = 0.260) was provided by SAD analysis of dataset SeMet1-Peak, within which could be found a 5 site, heavy atom substructure. These sites were then refined using SHARP [Bricogne et al., 2003]. Phases are calculated from the heavy atom substructure, to generate density maps from the anomalous signal, using all datasets. Residual maps guide the

removal or addition of additional heavy atoms. Various parameters of each site are then refined, such as the positions, B-factors, occupancies, as well as the magnitude of the anomalous scattering components f' and f". This process is carried out in iterative cycles until the values converge. The progress of SHARP iterations is summarised in **Figure 3.7.1.i**, and the phasing powers of each dataset calculated during this process are summarised in **Table 3.6.3.i**.

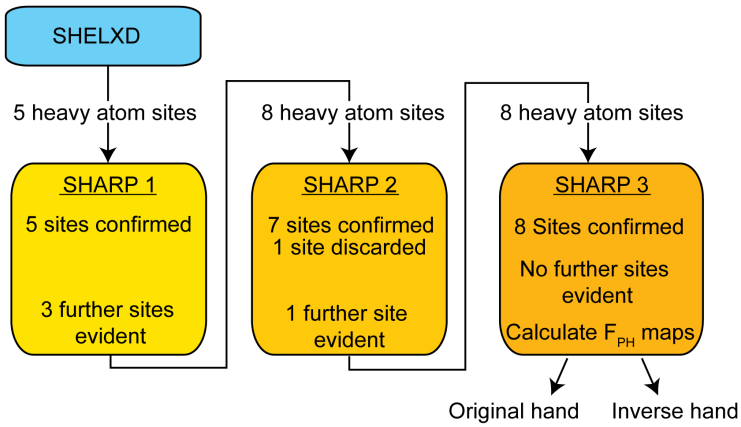


Figure 3.7.1.i: Progress of heavy atom site refinement by SHARP for the Ana2$^{1-47}$-SAS-4$^{700-901}$ phase solution.

3.7.2: Generation of phases and interpretation of experimental maps

Finally, SHARP used the finalised heavy atom sites to provide phase information and therefore electron density maps for the entire ASU, using a composite dataset. The heavy atom substructure is constructed from interatomic vectors. Each vector is centrosymmetric, and so the enantiomorph of the substructure will also present an equally satisfying solution. It is therefore necessary to calculate separate phase solutions and therefore maps, using both the original hand, and the inverted hand of the heavy atom substructure. Maps produced from these phase solutions are inherently noisy as the phases are estimates based on low and imprecise signal. The uncertainty of these phases is described by the overall figure of merit (FOM) which for these datasets was only 0.270. Additionally, in this case, according to the phasing power of each dataset (**Table 3.6.3.i**), only the peak wavelengths appear to have

contributed to the phase solution. As these peak datasets were collected at slightly different wavelengths, strictly speaking, this is a MAD solution, where additional wavelengths represent additional simultaneous equations with which to solve the phases. However, as peak or high remote wavelengths (and therefore similar f'' and f' values) were targeted, there may not have been significant dispersive differences to aid in MAD phase disambiguation. In order to interpret these maps, they can be modified according to *a priori* knowledge about electron density. Solvent regions can be identified and flattened, whereas protein regions can be sharpened, and also averaged (in the presence of NCS), thus allowing modified maps and correspondingly, more accurate phases. In this case, density modification was carried out using SOLOMON [Abrahams and Leslie, 1996], which iteratively flattened the solvent density around suspected protein regions. Density modification can be further guided by estimating the solvent content of the crystal, which can be estimated using the Matthews coefficient. Only a map produced from the correct enantiomorph of the heavy atom substructure should improve during density modification, and so this technique was used to decipher the correct hand of the heavy atom substructure, as well as flattening noise introduced from inaccurate phases.

The maps were of poor quality, however the inverse hand consistently gave marginally more favourable statistics than the original hand over a number of different solvent contents. Visual analysis of the maps also showed that the density was slightly more contiguous for the inverse hand. At this stage, "rungs" of density were evident, presumably representing the backbone chains of the beta-strands, and importantly the shape of the density was reminiscent of the structure of the SAS-4 homology model created in **Section 3.5.5**. To optimise this map, the density modification of the inverse hand solution was fine-tuned, with an optimal result produced by estimating a 56.0% solvent content. This provided a clearer map, exhibiting columns of semi-contiguous runglike density (**figure 3.7.2.i**), but it was not of sufficient quality to build a model of the protein *ab initio*.

However, a homology model of SAS-4$^{700\text{-}901}$ was created in **Section 3.5.3** using the structure of *D. rerio* CPAP**.** Using the heavy atom sites, which correspond to methionine sulphur atoms, a copy of this model could be manually placed into the map, providing a manual molecular replacement solution. Importantly, experimental density could be observed for the Ana2 peptide. Translation of the unit cell would cause the SAS-4 model to stack on itself end to end, and therefore it was clear that the model was too long at the C-terminus, and had to be trimmed (**Figure 3.7.2.i, A**).

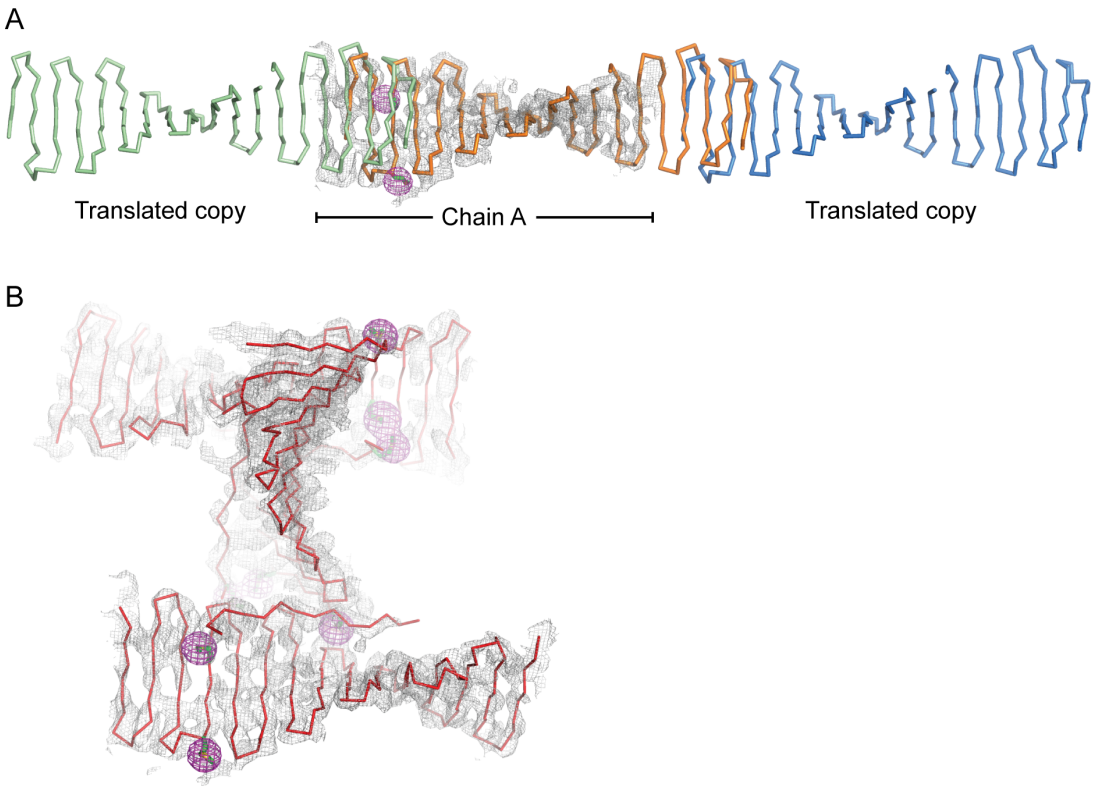


Figure 3.7.2.i: (**A**) The placing of a SAS-4 homology model (orange ribbon) into experimental density calculated from inverse hand of the heavy atom substructure (grey mesh, contoured at 1.2σ around the final refined chain A). Shown in purple mesh is the heavy atom map contoured at 3σ, showing the positions of selenium atoms, which were used to guide the placement of the SAS-4 homology model. Crystallographic translation of the chain along the C axis resulted in overlap with copies in adjacent cells (green and blue ribbons). Five strands therefore had to be trimmed from the C-terminus, and one strand (absent in the homology model) was built at the N-terminus. This was the case for all three chains that were eventually placed. (**B**) Experimental density for all three chains in the ASU, contoured at 1.2σ around protein atoms. The final refined structure is shown in red ribbon representation, with methionine residues shown as green sticks. Note that the methionine δ atoms are contained in the heavy atom density, and that experimental density can be observed along the Ana2 peptide.

Although four copies of the molecule were expected according to the Matthews coefficient analysis (**Section 3.5.2**), only 2 further columns of similar rung-like density could be observed in

the unit cell. A molecular replacement run using MOLREP [Vagin and Teplyakov, 2010] using the solvent flattened dataset and phase information was able to place a second copy of the full length model into one of these empty columns of density. Although the solution clashed, the methionine residues were in register with the heavy atom sites, and the beta strands fitted the density, and so this model had to be trimmed similarly to the first chain. Finally a third copy was manually placed, also requiring similar trimming (**Figure 3.7.2.i, B**).

3.7.3: Refinement of Ana2$^{1-47}$-SAS-4$^{700-901}$

Now that the models had been placed in the unit cell, the structure could be refined against the higher resolution native dataset, producing more detailed maps. These allowed completion of the structure by iterative rounds of manual model building in Coot [Emsley and Cowtan, 2004] and refinement using autoBUSTER (Bricogne et al., 2011). The three non-crystallographically related copies of the structure were very similar, and packed in the crystal in the same end-to-end manner. Therefore NCS restraints were used to aid refinement [Smart et al., 2012]. Apart from the invisible C-terminal strands, the SAS-4 TCP domain was well ordered, and could be built as a continuous β-sheet. Sidechains were well ordered and most had well defined pockets of density (**Figure 3.7.3.i, A**). Importantly, strong density was visible for a peptide section of the Ana2 CR2. Indeed, upon deletion of this peptide from the final model (along with the introduction of random noise to the remaining model (CCP4-PDBSET) to reduce model bias), a large positive residual is evident in the Fo-Fc omit map (**Figure 3.7.3.i, B**). Additionally, a heavy atom site was present in both chains B and C that corresponded to a methionine in the peptide, helping to confirm its register. Refinement was continued until the model adequately described the native dataset, according to the R_{work} and R_{free}, which finally reached 24.5% and 26.3% respectively. Molprobity was also used during refinement to monitor and validate the chemical feasibility of the model. Final refinement and validation statistics are presented in **Table 3.7.3.i**.

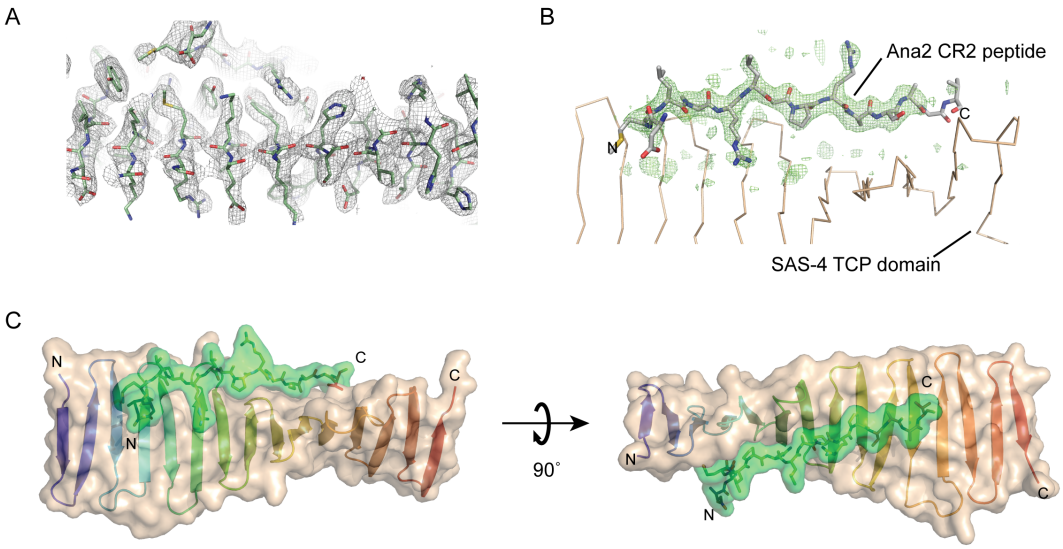


Figure 3.7.3.i: (**A**) Representative electron density (grey mesh, contoured at 1.2σ) of part of the refined model, shown as green sticks. (**B**) View of the positive residual (green mesh contoured at 3.0σ) in the Fo-Fc map (calculated using Phenix.maps), that results upon removal of the Ana2 peptide from the structure. Note that prior to map calculation, random noise was introduced to the rest of the model using PDBSET, in order to reduce bias in the model from the prior existence of the peptide. The complex structure is shown for reference, with the SAS-4 TCP domain as a beige ribbon, and the Ana2 peptide as white sticks, lying within, and clearly explaining the residual. (**C**) One chain of the final refined structure, showing the Ana2 peptide binding along one face of the SAS-4 TCP domain. The SAS-4 TCP domain is shown as a cartoon, coloured blue to red, (N to C), inside a semi-transparent beige surface representation. The Ana2 peptide is shown in stick format inside a green semi-transparent surface.

	$Ana2^{1-47}$-$SAS\text{-}4^{700-901}$
Resolution Range (Å)	64.60-2.57 (2.65-2.57)
R_{work}/R_{free} /% size test set	24.5/26.3/5.05% (32.1/36.0/5.07)
No. Reflections working set	30297 (2790)
No. Reflections Test Set	1611 (149)
Number of atoms (non-H)	3924
Waters	65
rmsd from ideal values: bond length (Å)/angles(˚)	0.007/0.900
Average B-factor	70.80
Correlation Coefficient Fo-Fc/Fo-Fc free	0.854/0.836
Molprobity score	1.40 (100^{th} percentile)

Table 3.7.3.i: Refinement and validation statistics for $Ana2^{1-47}$-$SAS\text{-}4^{700-901}$. Values in parentheses are for the highest resolution shell.

3.8: Structural Analysis of Ana2$^{1-47}$-SAS-4$^{700-901}$

3.8.1: The SAS-4 TCP domain adopts a unique, extended β-sheet fold

The SAS-4 TCP domain adopts an extended single β-sheet fold, exhibiting a slight twist (approximately 11˚ between consecutive strands). The sheet is formed from a continuous β-meander - a series of consecutive anti-parallel β-strands, connected by β-hairpins. The TCP domain was first identified by Hung et al., 2000 where it was observed that it comprised a series of ~nonameric repeats, with a regular occurrence of glycine residues and was therefore described as the "G-box". In the structure, these glycine residues can be seen to take part in a regular PDG motif, which forms the β–turns (**Figure 3.8.1.i**).

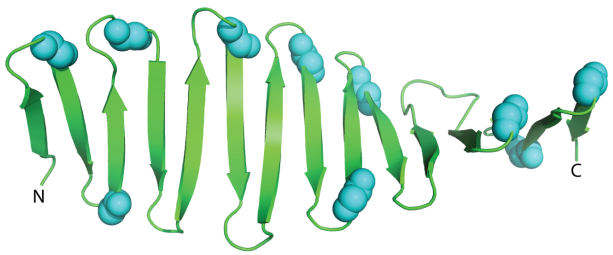


Figure 3.8.1.i: The Drosophila SAS-4 TCP domain shown in green cartoon format. Highlighted as cyan spheres are the regularly occurring glycine residues described by Hung et al., 2000. In the structure it can be seen that their role lies in the construction of β-turns.

β-sheets are highly common motifs found in many proteins, however this domain is particularly extended, and aside from the Ana2 peptide, is completely solvent exposed and lacks the hydrophobic core common in globular domains. (The apo-structure of the orthologous TCP domain from *D. rerio* was crystallised by my collaborator Dr Mark van Breugel, and this domain adopted the same conformation, and was monomeric in solution. This suggests that in solution, the domain is indeed entirely solvent exposed). A search of similar domains using the DALI server [Holm and Rosenstrom, 2010] shows that in other structures with extended β-meanders, they are usually capped with, or packed against other domains such that the sheet is not largely solvent exposed (**Figure 3.8.1.ii**).

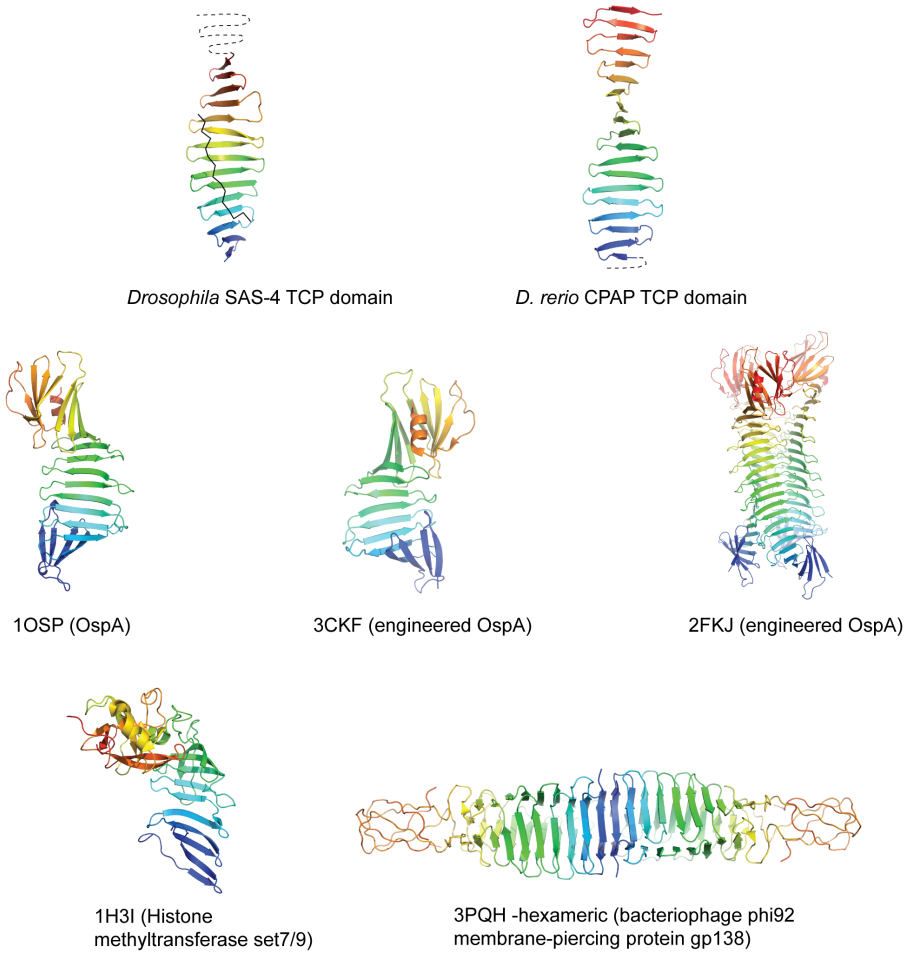


Figure 3.8.1.ii: Structures with extended beta-meanders found using the DALI server. All structures are shown in cartoon format, coloured N(blue)-C(red) in their predicted biological assemblies, as seen *in crystallo* and predicted by the PISA server. At the top is the *Drosophila* SAS-4 TCP domain with the Ana2 peptide depicted as a black ribbon. The predicted missing strands are denoted by a dashed black line at the C-terminus. To its right, is the orthologous TCP domain apo-structure, from *D. rerio* CPAP (PDB entry 4BXP, Cottee et al., 2013). Although several examples of the TCP domain are now available in the PDB, they all adopt this solvent-exposed conformation.

3.8.2: The SAS-4 TCP domain acts as a proline recognition domain, and binds a proline-rich region of the Ana2 CR2

The most important feature of this structure is that it reveals the molecular nature of the Ana2:SAS-4 interaction. The TCP domain acts as a proline recognition domain, and in all three copies of the complex, a proline-rich motif (PRM) of the Ana2 CR2 domain was visible, and bound in an identical manner, along one face of the TCP domain. This peptide region adopts a

left-handed “type II polyproline” helix with a lack of internal hydrogen bonding - the occurrence of prolines in this sequence reducing the available H-bond donors. Four main interaction mechanisms can be observed between the TCP domain, and the Ana2 CR2. First, the Ana2 prolines, P11, P15 and P17 pack against aromatic residues of the TCP domain, Y748, F749, Y767 and F786 (**Figure 3.8.2.i, A**). This packing mechanism is similar to that observed in other PRM binding domains (Kay et al 2000). It is also possible, due to the occurrence of downstream prolines in CR2, and the existence of further aromatic residues in the TCP domain, that this interaction continues along the length of the TCP domain (**Figure 3.8.3.i, A**). Second, TCP domain sidechains of Y765, Q790 and E792 make H-bond contacts with the Ana2 CR2 mainchain (**Figure 3.8.2.i, B, C**). This is important, as MSA analysis shows that E792 is the conserved MCPH residue that is mutated to valine in HsCPAP. A loss of this hydrogen bond could lead to a reduction in affinity between the two domains. Third, the Ana2 residue R12 forms a cation-pi interaction with the aromatic ring of Y765 of the TCP domain (**Figure 3.8.2.i, C**). Finally, R12 also takes part in a water mediated H-bond network with TCP domain residues H774, and S776 (**Figure 3.8.2.i, D**). Surface charge analysis was carried out on the SAS-4 TCP domain, and this revealed a positively charged patch towards the N-terminus, on the face that binds the Ana2 CR2. Interestingly, a conserved, acidic E(T/D)ED motif lies just N-terminal of the PRM, and would be predicted to lie over this positively charged patch (**Figure 3.8.2.i, E**). This suggests a possible mechanism for Ana2 CR2 recruitment, where a long-range charge-based attraction would recruit and orientate the PRM, before engagement of the close-range proline stacking interactions (**Figure 3.8.2.i, F**).

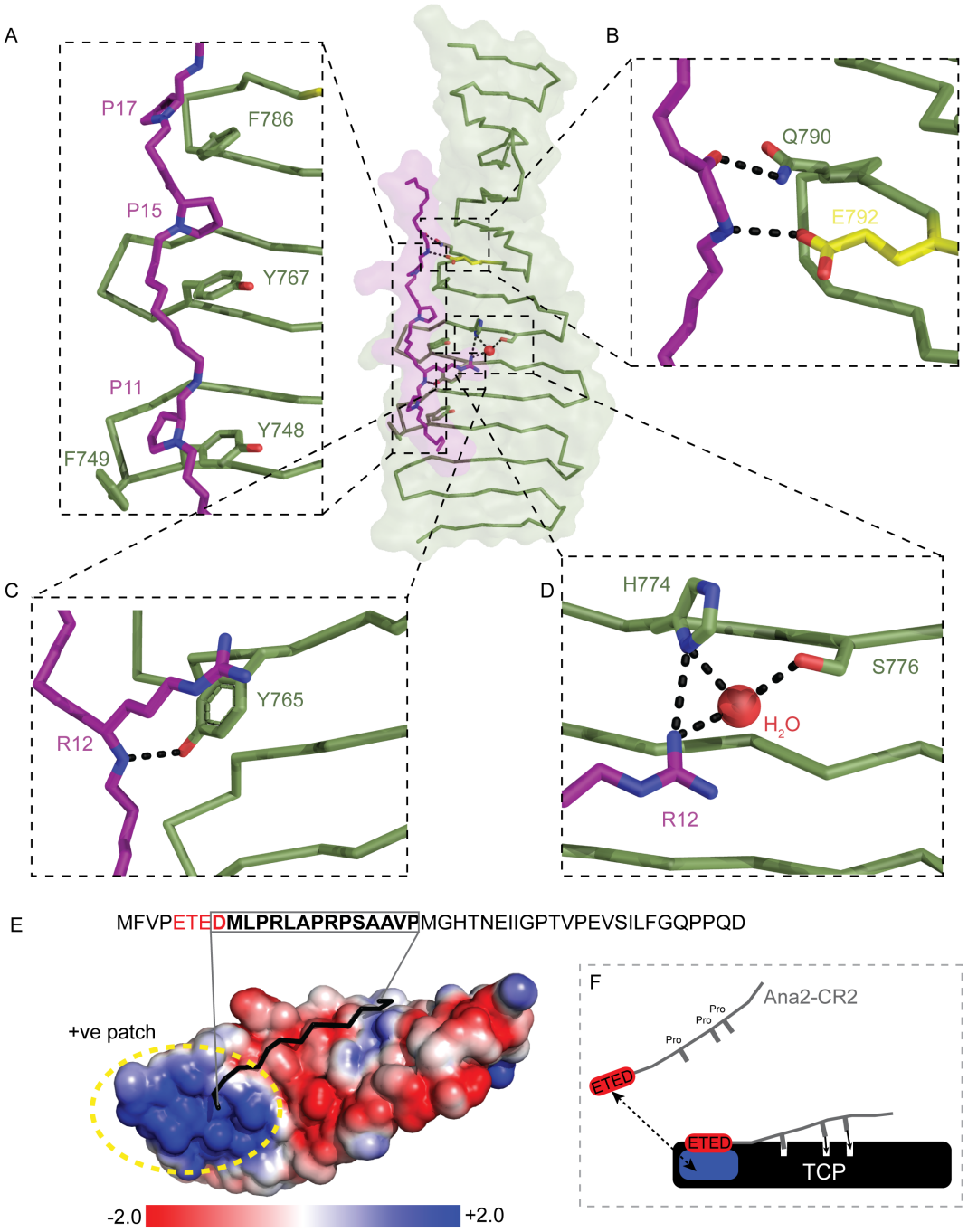


Figure 3.8.2.i: The SAS-4 TCP domain acts as a proline recognition domain, and binds a proline-rich region of the Ana2 CR2. Centre – the SAS-4 TCP domain in green ribbon and stick format inside a translucent green surface, and similarly, the Ana2 CR2 proline-rich motif, in purple. Hydrogen-bonds are shown as black dashed lines. (**A**) Proline rings of Ana2 stack against aromatic residues of the TCP domain. (**B**) Sidechains from the TCP domain H-bond with the mainchain of Ana2. (**C**) A cation-pi interaction forms between R12 of Ana2, and the aromatic ring of Y765 of the TCP domain. Additionally, an H-bond is formed between the Y765 hydroxl, and an Ana2 backbone amide. (**D**) A water-mediated H-bond network between R12 of Ana2 and residues of the TCP domain. (**E**) Surface charge analysis of the SAS-4 TCP domain, reveals a large basic zone towards the N-terminus (dashed yellow circle). The sequence of Ana2 that was included in the fusion construct is displayed overhead and residues that were visible in the structure are boxed and in bold. Just to the N-terminus of these residues is an acidic "ETED" motif. (**F**) It is possible that the Ana2 CR2 is recruited to the TCP domain via a long-range charge-based mechanism, that would also help to orientate the Ana2 CR2 before the "zipping up" of the proline-rich motif.

3.8.3: The SAS-4 : Ana2 interaction is highly conserved

Analysis of the MSA covering the Ana2 CR2 shows that residues seen to be interacting in this structure are conserved through to vertebrate species, and allow the definition of a core PRxxPxP motif (**Figure 3.8.3.i, A**). Analysis of the MSA covering the TCP domain similarly shows that interface residues are highly conserved with vertebrates (**Appendix 7.3,** also **Figure 3.9.1.i, A**). During the course of this work, my collaborator Dr Mark van Breugel also solved the structure of the orthologous STIL:CPAP complex from *D. rerio*. All chains of the *Drosophila* and *D. rerio* complexes superimpose well, with interspecies alignments giving an average RMSD of 1.2 ± 0.2 Å (average 118 ± 4 Cα pairs) (**Figure 3.8.3.i, B**).

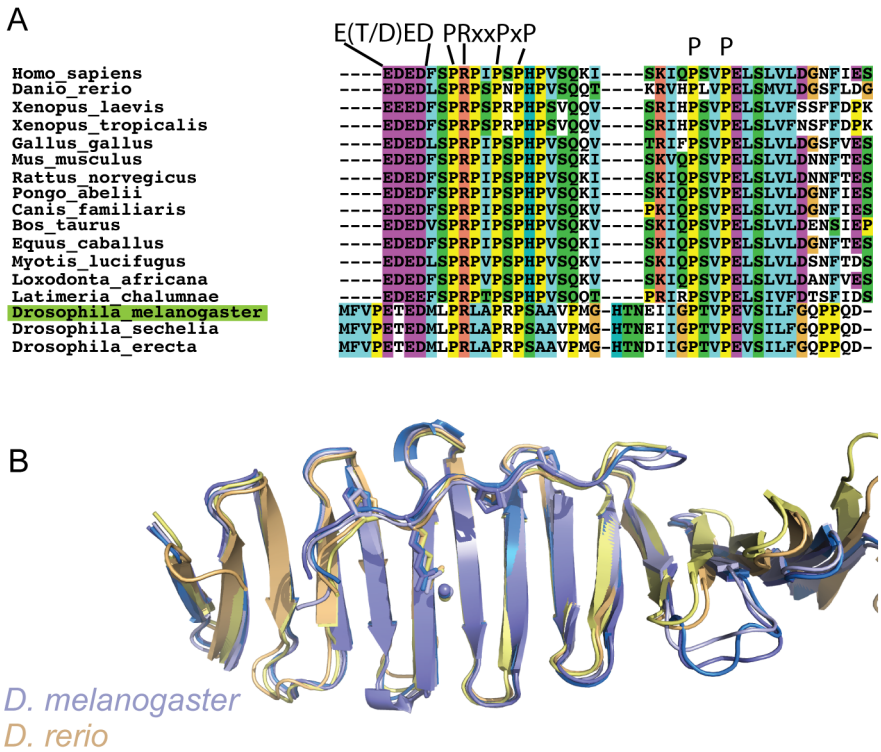


Figure 3.8.3.i: The Ana2:SAS-4 interaction is conserved through to vertebrates. (**A**) MSA analysis of the Ana2/STIL CR2 region. It can be seen that a PrxxPxP motif is conserved through to vertebrates. Additional downstream proline residues also appear to be conserved suggesting that the proline recognition mechanism may extend beyond what is visible in the crystal structures. An acidic motif is also conserved at the N-terminus of the CR2 domain. (**B**) All three chains from the *Drosophila* complex structure superimposed with both chains from the orthologous *D. rerio* complex structure, provided by my collaborator Dr Mark van Breugel. Interspecies alignments give an average RMSD of 1.2 ± 0.2 Å (average 118 ± 4 Cα pairs).

Both structures exhibited the same four interaction mechanisms, using conserved residues. These structural similarities confirm that the interaction between these two centriolar proteins represents a conserved step in centriole biogenesis, as well as demonstrating that Ana2 is a true orthologue of vertebrate STIL.

3.8.4: The crystal packing of the SAS-4 TCP domain causes fibrils to form

The crystal structure of the Ana2:SAS-4 complex contained three similar copies, which all exhibited the same interesting crystal packing. Each chain packed such that one translation of the unit cell along either the b, or c axes would allow continuation of a β-sheet throughout the crystal (**Figure 3.8.4.i, A,**), creating infinitely long fibrils. The crystal contact between chains in these fibrils mimics the natural extension of a β-meander, with conventional, antiparallel mainchain H-bonding between strands β15 of one chain, and β1 of the adjacent chain. The junction is further stabilised by polar sidechain-sidechain interactions (**Figure 3.8.4.i, C**). Due to this arrangement, it is not possible to say whether the Ana2 peptide fused to each TCP domain was initially bound to its own TCP domain, or to the next domain along (**Figure 3.8.4.i, B**).

In order for conventional, antiparallel β-sheet continuation to occur via a crystallographic translation of the unit cell, the repeat must take place after the sheet has twisted approximately 180˚. The TCP domain has a natural twist, turning 180˚ over ~15-16 strands. Therefore, this arrangement requires the predicted C-terminal strands (16-20) to be disordered in the crystal, such that the first strand of the next monomer, binds where the 16th strand would be predicted to bind. This packing may force a slight unwinding of the natural TCP domain twist, as in the *D. rerio* structures provided by my collaborator Dr Mark van Breugel - which do not exhibit such packing and are therefore not constrained - the twist appears to be slightly tighter, at ~13˚ per hairpin, as opposed to ~11˚ for the SAS-4 TCP domain. The *D. rerio* structures also have 19-20

strands, suggesting that the native structure of the sheet does indeed comprise all the predicted strands. These differences between the structures also explain why attempts at molecular replacement were especially difficult using the *D. rerio* TCP domain as a model.

This packing phenomenon is probably further facilitated by the polar sidechain contacts, but it is also interesting to note that β16 and β17 were predicted by Psipred to be particularly short strands covering only 3 and 4 residues respectively, and may therefore be more amenable to unfolding (**Appendix 7.3**). Overall, these features possibly allow crystal-packing restraints to override the natural inclination of the molecule to form a 20-strand β-sheet. Instead, the partial unwinding of each chain allows it to be locked stably into a repetitive structure in the crystal, which is further stabilised by other packing interactions along all crystallographic axes.

I suspect that this fibril formation phenomenon may well be a crystallisation artefact, simply providing a convenient and stable crystal contact, via a simple translation of the unit cell. This is important to consider, because since the publication of this study [Cottee et al., 2013] two further studies have revealed the structures of the TCP domain from *D. rerio* [Hatzopoulos et al., 2013] and *Drosophila* [Zheng et al., 2014b]. Interestingly, both exhibited similar β-sheet extension via a crystallographic translation of the unit cell. Hatzopolous et al. noticed that the periodicity of their repeat (79 Å) is close to that predicted for the ~8 nm repeats of tubulin dimers in microtubules, and so they proposed that the TCP domain might therefore act as a molecular ruler. However, the repeat in the Hatzopolous et al. crystals encompasses 17 strands rather than the 15 I observe in the *Drosophila* structure. By definition this means that the interfacing strands involved in the *Drosophila* and *D. rerio* fibrils are not equivalent (in fact both sets are entirely exclusive, with *Drosophila* fibrils using strands 1 and 15, and *D. rerio* fibrils using strands 3 and 19) (**Figure 3.8.4.i, D**).

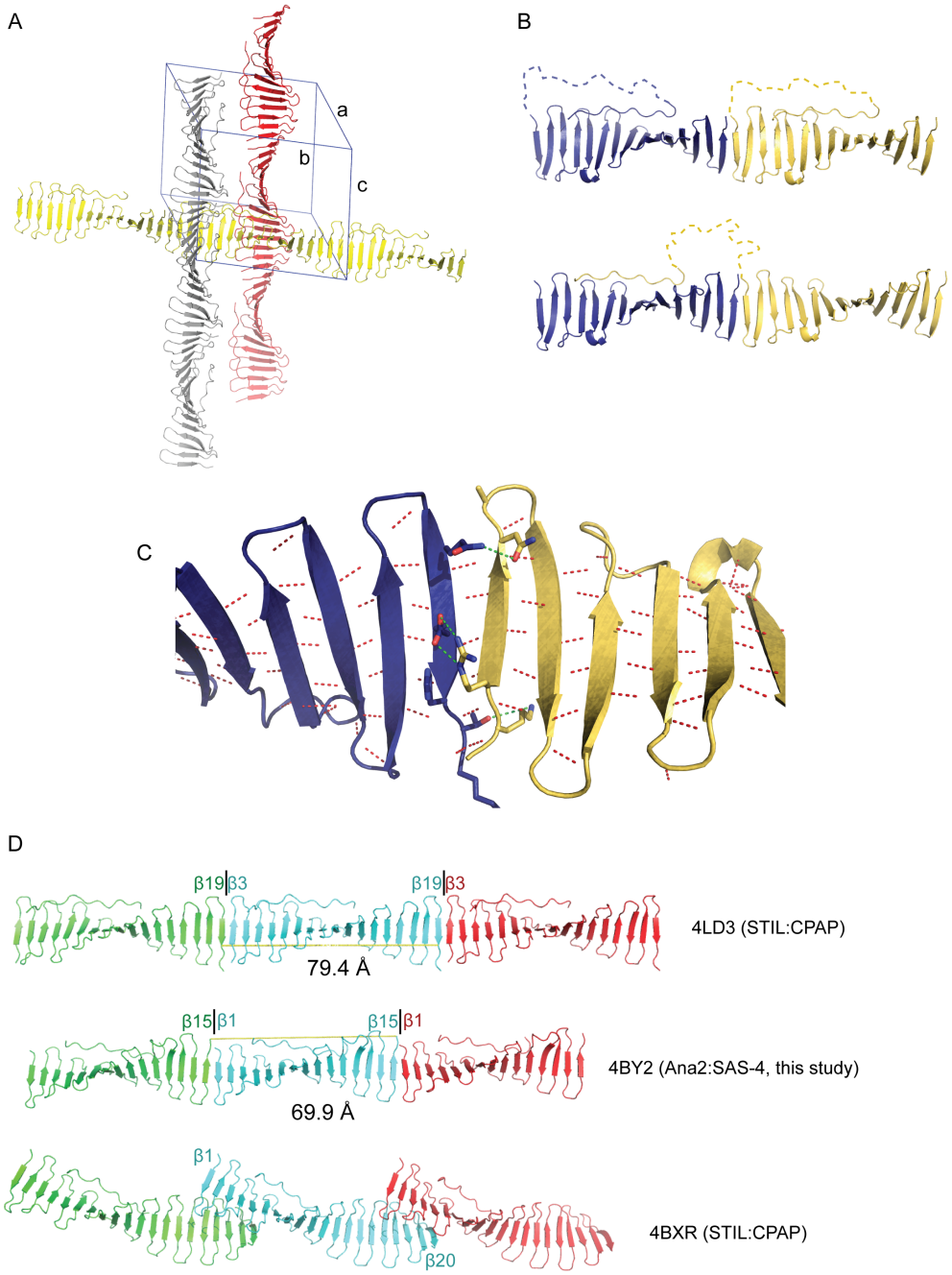


Figure 3.8.4.i: Crystal packing in the Ana2$^{1-47}$-SAS-4$^{700-901}$ structure. (**A**) All three chains form fibrils via translation of the unit cell along the *c* axis (chains A and C, grey and red cartoons) or the *b* axis (chain B, yellow cartoon). The depicted fibrils consist of three adjacent chains. Only the middle copy of each fibril is contained within the depicted unit cell, shown as a black box. (**B**) Two adjacent copies of chain B (blue and yellow cartoons) resulting from translation along the *b* axis. It is not known whether the Ana2 CR2 peptide region binds in a "cis" (above) or "trans" (below) fashion, as the linker is invisible. (**C**) Closeup of the inter-chain junction as in **B**. Red dashed lines indicate intra-mainchain H-bonds, highlighting the continuation of conventional anti-parallel β-sheet H-bonding. Green dashed lines show intrachain polar sidechain interactions that stabilise the junction, possibly explaining why it occurs specifically between strands β15:β1. (**D**) Comparison between crystal packing in 4LD3 [Hatzopolous et al., 2013], 4BY2 (The Ana2$^{1-47}$-SAS-4$^{700-901}$ structure) and 4BXR, the orthologous complex from *D. rerio* solved by Dr Mark van Breugel [Cottee et al., 2013]. 4LD3 and 4BY2 both form fibrils however the periodicity, and the strands involved in the junction are different.

These differences suggest that, if this fibril structure is physiologically relevant, its size and spacing is not conserved. The difference in strand number means that the *Drosophila* fibril repeat takes place over a shorter distance of ~70 Å, which could not form a repeating structure in register with the 8 nm tubulin repeats. Moreover, the *D. rerio* structures provided by Dr Mark van Breugel do not form fibrils and show the *D. rerio* TCP domain with its full 20 strands (**Figure 3.8.4.i, D, 4BXR**). These constructs were also longer at both the N and C terminus - perhaps this stabilised the complete β-sheet, and overrode its potential to form fibrils.

The non-conserved aspects of fibril formation therefore suggest that this phenomenon is an artefact of crystal-packing and/or construct design. However, the readiness of different TCP domains to form fibrils is an intriguing feature; whether such fibrils form *in vivo* remains to be determined.

3.9: Design of mutations to dissect the Ana2:SAS-4 interaction

In order to confirm that the crystallographic Ana2:SAS-4 interaction is physiologically relevant (particularly with respect to the involvement of the MCPH residue, E792V) it was necessary to design point mutations, or clusters thereof, and observe their effect on complex formation *in vitro* and *in vivo*.

3.9.1:Initial design of point mutant clusters

Prior to solving the structure of the Ana2:SAS-4 complex, only the structure of the *D. rerio* CPAP apo-TCP domain was known, provided by Dr Mark van Breugel. At this stage it was not known exactly which residues would comprise the Ana2:SAS-4 interface, and so the choice of point mutations was guided by sequence conservation. By mapping the CPAP/SAS-4 sequence conservation onto the TCP domain structure I observed that conserved residues mostly lay on one face of the β-sheet (**Figure 3.9.1.i, A**). For practicality, three clusters of mutations were initially designed to target localised sets of conserved residues. MC1 lay towards the N terminus, MC2 contained the MCPH residue E792 and lay near the centre, and MC3 lay towards the C-terminus. (**Figure 3.9.1.i, A, B)**. In light of the Ana2:SAS-4 complex crystal structure, it can now be seen that MC1 contains Y765, which makes extensive contacts with Ana2, and also K735, which contributes to the positively charged patch that may attract the CR2. Moreover, all three residues in MC2 are involved in the interaction, including E792, the MCPH residue. MC3 however, is much further away, and was not observed to participate in the Ana2:SAS-4 interface. (These residues are also highlighted in an alignment in **Appendix 7.3**). A construct that combined all 3 mutant clusters (MC123) was also designed. The clusters of point mutants were created by site-directed mutagenesis (QuikChange Lightning Multi kit, Agilent). All SAS-4 point mutations were to valine. Valine (like isoleucine) is favourable in β structures, as it is a β-branched residue. The added steric bulk close to the Cα serves to reduce peptide backbone mobility, and therefore

rigidifies beta strands [Pal and Chakrabarti, 2000]. Valine was therefore considered to be minimally disruptive to the fold of the domain. These mutation clusters were used for *in vivo* work, which is discussed in **Section 3.11**.

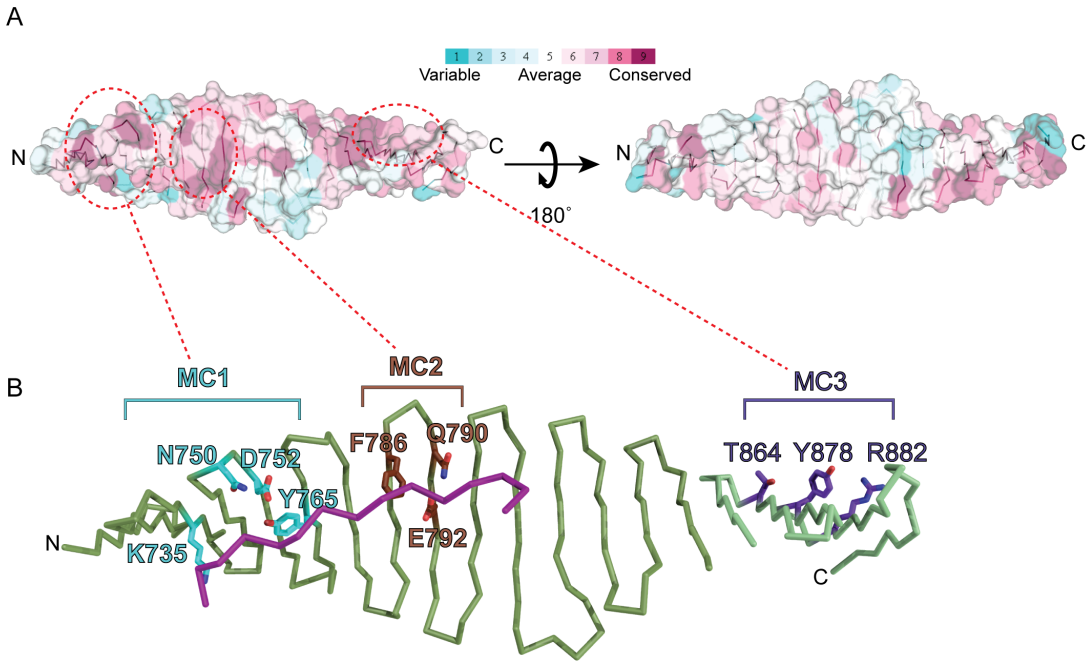


Figure 3.9.1.i: Design of SAS-4 point mutations. (**A**) Sequence conservation mapped onto the structure of the SAS-4 TCP domain, coloured according to the Consurf conservation score. Note that this is a composite structure, where the last five strands are modelled on the *D. rerio* CPAP TCP domain (due to these being disordered in the *Drosophila* structure). The SAS-4 TCP domain is shown in ribbon format within a semi-transparent surface representation. One face of the β-sheet (shown left) is more conserved than the other (shown right). Red dashed circles indicate the conserved patches targeted by mutation clusters MC1, MC2 and MC3. Targeted residues were substituted to valine. (**B**) The Ana2 (magenta ribbon) and SAS-4 TCP domain (dark green ribbon) complex. The five missing strands are modelled as in **A,** and are shown as a light green ribbon. Residues comprising MC1, MC2 and MC3 are shown as cyan, brown and dark purple sticks respectively. Note that Y765 and E792 were also targeted individually.

3.9.2: Design of single point mutants

Additional to the mutation clusters, an obvious target was also the MCPH mutation, E792V, in isolation. Moreover, upon solving the structure of the Ana2:SAS-4 complex, I was able to target additional individual residues with confidence. For this reason, a further SAS-4 mutant, Y765V was designed, Y765 being a residue involved in both a cation-pi, and H-bond interactions at the Ana2:SAS-4 interface (**Figure 3.8.2.i, C**). Both E792V and Y765V were used for *in vitro*

experiments discussed in **Section 3.10**. In light of the Ana2:SAS-4 structure, point mutations were also designed for the Ana2 CR2 (**Figure 3.9.2.i**), although these were only used for *in vivo* experiments which are discussed in **Section 3.11**. In the case of Ana2, all residues were substituted for alanine. Additionally, R16, a residue not involved at the interface, was also included as a control mutation.

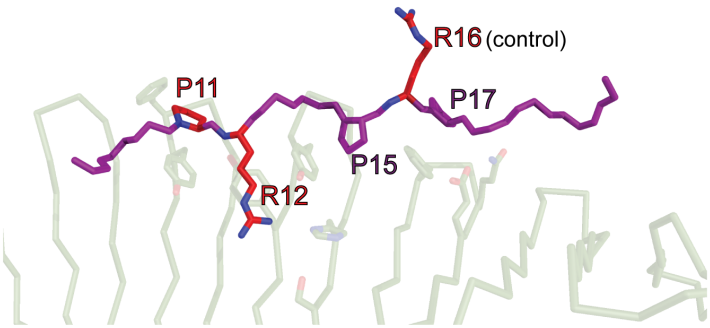


Figure 3.9.2.i: Design of Ana2 point mutations. The Ana2 peptide backbone is shown as magenta sticks, with the SAS-4 TCP domain shown behind in a transparent ribbon and stick format. Interfacing sidechains are shown as sticks. Residues targeted for mutation to alanine are shown as red sticks. R16 is not involved at the interface and is included as a negative control.

3.10: Attempts to characterise the Ana2:SAS-4 interaction *in vitro*

Previous pilot experiments (**Section 3.3**) suggested that the CR2 of Ana2 could interact with the SAS-4 TCP domain *in vitro*. Now that this interaction had been structurally characterised, I made further attempts to characterise the interaction using Surface Plasmon Resonance (SPR). The details of these attempts are described in **Appendix 7.5**. I was not able to calculate a Kd, however I was able to show that an MBP-fusion of SAS-4$^{700\text{-}901}$(WT) bound to Ana2$^{5\text{-}43}$ with a greater response than either of the equivalent MBP-SAS-4$^{700\text{-}901}$(E792V), or MBP-SAS-4$^{700\text{-}901}$(Y765V) mutants (**Figure 3.10.i**).

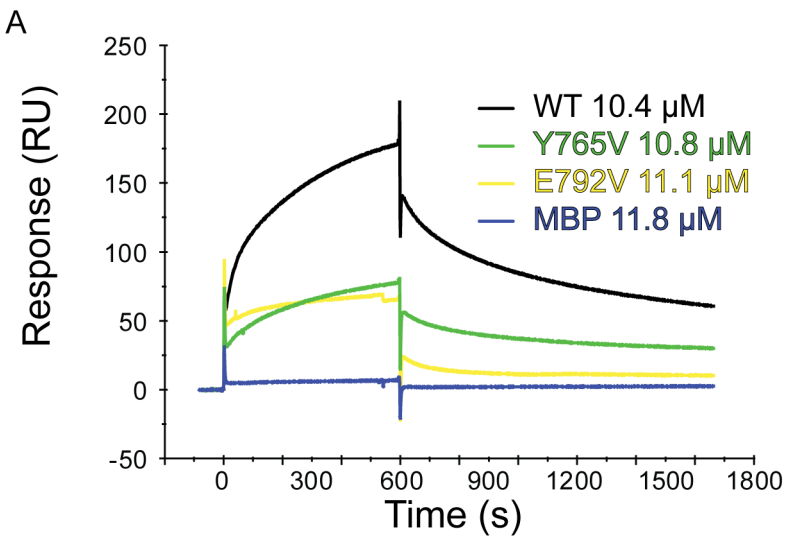


Figure 3.10.i: The wild-type SAS-4 TCP domain interacts more strongly with Ana2$^{5\text{-}43}$ than the Y765V and E792V mutants. In this Surface Plasmon Resonance (SPR) experiment, Ana2$^{5\text{-}43}$ was attached to the SPR chip surface. MBP-SAS-4$^{700\text{-}901}$ variants (WT, Y765V and E792V) as well as MBP alone, were flowed over the chip surface at ~equimolar concentrations. The wild-type MBP-SAS-4$^{700\text{-}901}$ binds Ana2$^{5\text{-}43}$ most strongly, as observed by the greater response. Consistent with the crystal structure, these point mutants appear to have a negative affect on binding.

3.11: The SAS-4:Ana2 interaction is essential for centriole duplication *in vivo*

An important part of this project was to assess the function of the Ana2:SAS-4 interaction interface *in vivo*. Because I had solved the structure of the *Drosophila* Ana2:SAS-4 interface, carrying out such an analysis was possible. Importantly, I wanted to assess the potential role of a mimic of the MCPH mutation (E792V in *Drosophila*), thus potentially simulating a human disease in *Drosophila*. Phenotypes associated with this mutation would be particularly interesting as a previous study suggested that it weakens, but does not abolish Ana2:SAS-4 binding [Tang et al., 2011]. Null mutants of *ana2* (*Ana2*^{169} and A*na2*^{719}, [Wang et al., 2011]) and *Sas4* (*sas4*^{S2214}, [Basto et al., 2006]) have the same null phenotypes of uncoordination and centriole loss, as previously described for S*as6* (**Chapter 2**). The aim of these experiments was to assay the ability of modified constructs, with an impaired Ana2:SAS-4 interface, to rescue these null phenotypes, thus establishing how essential this interface is for centriole assembly.

3.11.1: Design of transgenic constructs

C-terminally fused GFP constructs were designed for both SAS-4 and Ana2 (**Figure 3.11.1.i**). Full-length wild-type constructs were designed as well as constructs lacking either the SAS-4 TCP domain (SAS-4$^{\Delta TCP}$ (SAS-4$^{1-724}$)), or the Ana2 CR2 domain (Ana2$^{\Delta CR2}$ (Ana2$^{46-420}$)). Additionally, point mutated constructs or clusters thereof were designed as explained in **Section 3.9**. The constructs were cloned into a pUbq CT-GFP DEST vector, placing the fusions under the control of the Ubq promoter. Transgenic *Drosophila* lines were then generated by P-element mediated transformation. Injection of DNA and transformant screening was conducted by Genetic Services inc., or the Cambridge University Injection service. G1 transformants were genetically balanced by my collaborator, Dr. Nadine Muschalik, who subsequently performed all fly husbandry, specimen preparation, western blots and microscopy work.

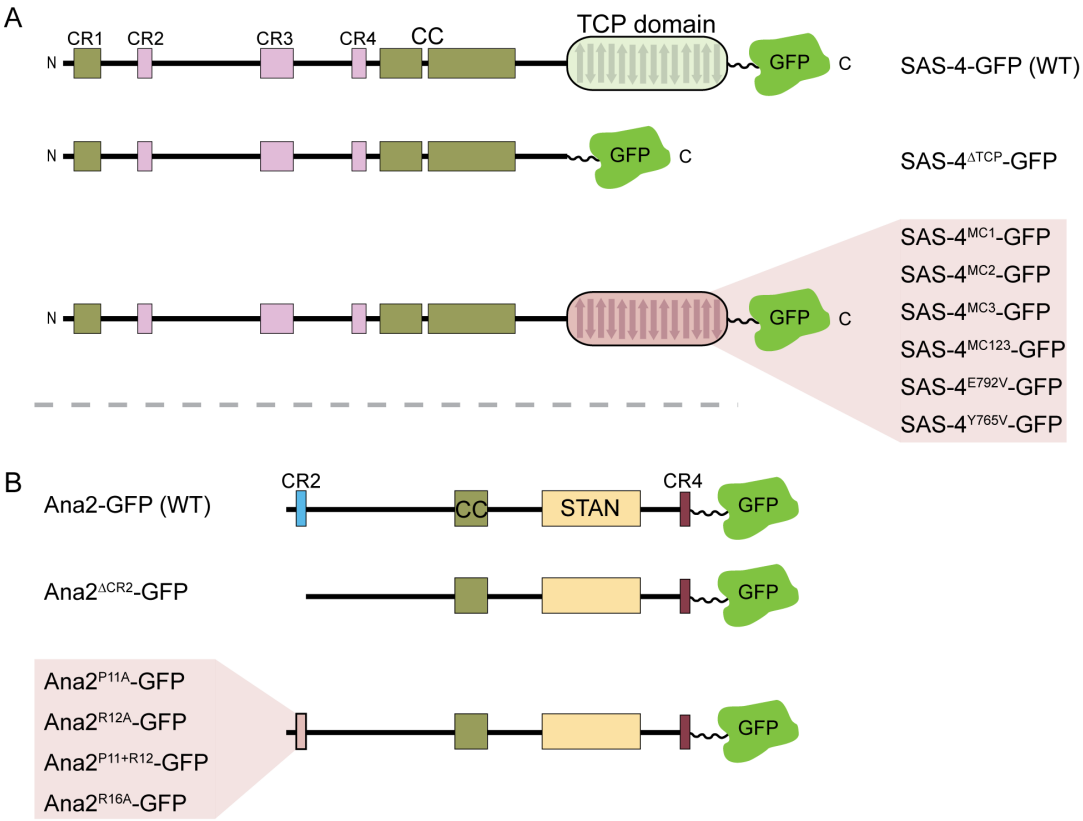


Figure 3.11.1.i: GFP-Fusion constructs for which transgenic *Drosophila* lines were made. (**A**) SAS-4-GFP constructs. A full-length WT SAS-4 construct was made as a control. To test the interaction with Ana2, SAS-4$^{\Delta TCP}$ was made (SAS-4$^{1-724}$). Additionally, full-length constructs with a mutated TCP domain were generated. Mutation clusters are detailed in **Figure 3.9.1.i**. (**B**) Ana2-GFP constructs. A full-length WT Ana2 construct was made as a control. To test the interaction with SAS-4, Ana2$^{\Delta CR2}$ was made (Ana2$^{46-420}$). Additionally, full-length constructs with a mutated CR2 domain were generated.

3.11.2: Ana2-GFP constructs can localise to centrioles

The Ana2 constructs were first assayed for their ability to localise to centrioles in live embryos, albeit in the presence of endogenous WT Ana2. Wild-type Ana2-GFP was visualised as pairs of tightly focussed foci, and so centriolar localisation was inferred from the presence of paired GFP foci. Interestingly, all Ana2-GFP mutant constructs were able to localise at similar levels to wild-type (**Figure 3.11.2.i**). This suggests that Ana2 can localise to the centriole independently of its interaction with SAS-4.

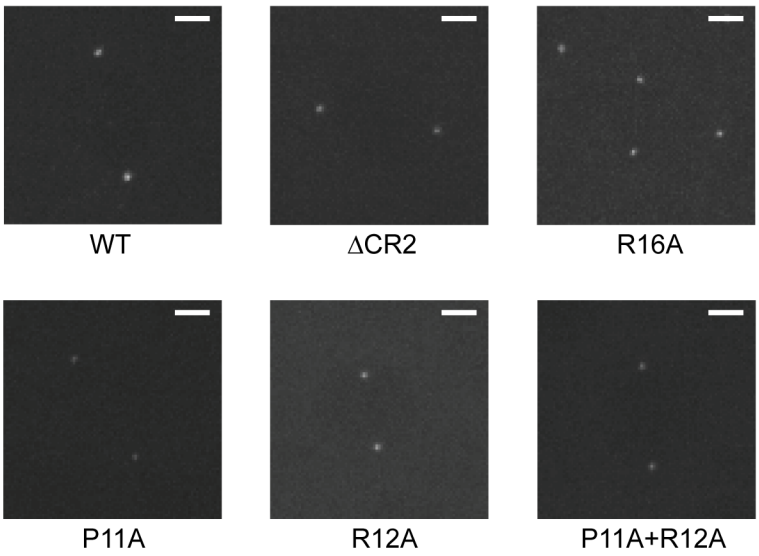


Figure 3.11.2.i: Localisation of Ana2-GFP constructs to centrioles in live embryos. Shown are representative images of paired GFP foci for each of the labelled constructs. Scale bars are 3 µm. Wild-type Ana2-GFP localised as a pair of bright focussed foci. Similar localisation was seen for all of the mutated constructs.

3.11.3: SAS-4-GFP can localise to centrioles, but localisation is much weaker when the interaction with Ana2 is compromised.

The SAS-4-GFP fusion constructs were also assayed for centriolar localisation in live embryos, in the presence of endogenous SAS-4. As expected, wild-type SAS-4-GFP could localise to centrioles, as judged by the presence of pairs of GFP foci. However, SAS-4$^{\Delta TCP}$-GFP, and all of the point-mutated constructs were much reduced in their localisation efficiency compared to wild-type (**Figure 3.11.3.i**). These results suggest that SAS-4 recruitment is significantly dependent on the TCP domain, and, more specifically, its interaction with Ana2, via the TCP domain. SAS-4 has a central CC, and so it is possible that residual localisation may be due to the association of the mutant protein with the endogenous WT SAS-4.

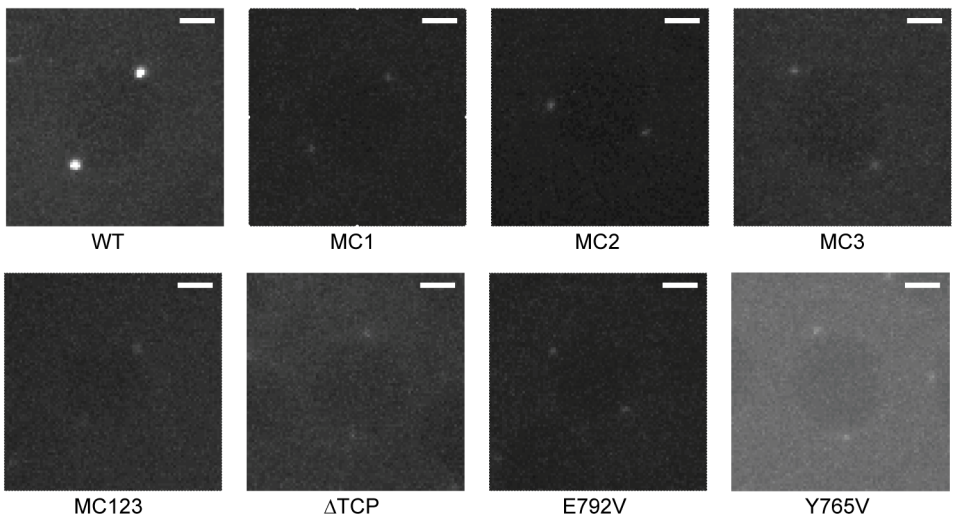
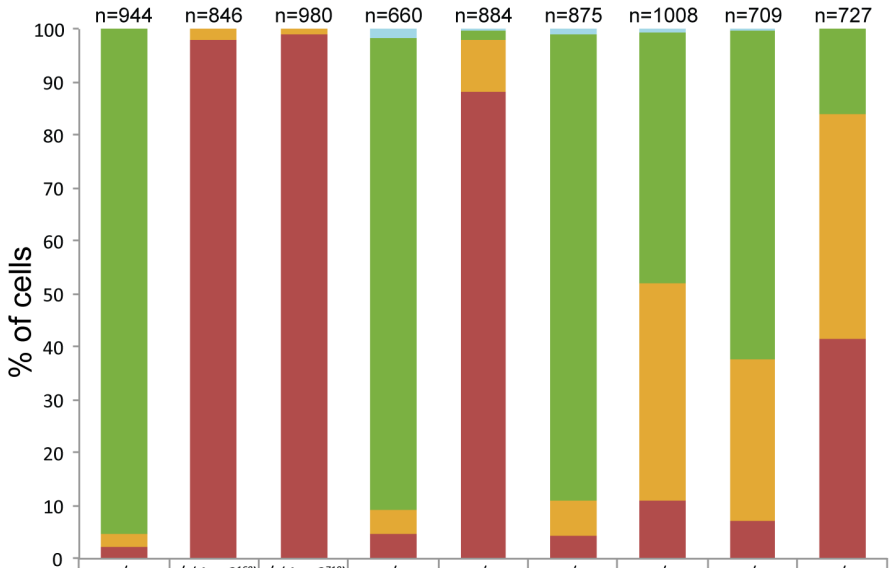


Figure 3.11.3.i: Localisation of SAS-4-GFP constructs to centrioles in live embryos. Shown are representative images of paired GFP foci for each of the labelled constructs. Scale bars are 3 µm. Wild-type SAS-4-GFP localised as a pair of bright focussed foci. All of the mutated constructs had very weak localisation in comparison.

3.11.4: The *Sas4* or *Ana2* null phenotypes can only be rescued by constructs that maintain the SAS-4:Ana2 interaction

The constructs were then crossed into their respective null genetic backgrounds, in order to assess their ability to rescue centrosome loss. A centrosome counting assay was performed, as described for the GFP-SAS-6 constructs in **Chapter 2**. Third instar larval brain cells were fixed and stained for the PCM marker, Cnn, and the outer centriole protein Asl. Foci staining for both proteins, in metaphase cells were counted as centrioles. All quantification was carried out blind.

In wild-type flies, ~95% of mitotic brains cells had two centrosomes. In the *Ana2*$^{-/-}$ mutants however, 98-99% of cells, had no centrosomes. Similarly, in the *Sas4*$^{-/-}$ flies, ~95% had no centrosomes. To summarise the rescue experiments, WT Ana2-GFP and WT SAS-4-GFP were both able to robustly rescue these phenotypes. However, Ana2 or SAS-4 mutant constructs could not robustly rescue these phenotypes where their point mutations were expected to disrupt the Ana2:SAS-4 interface (**Figures/Tables 3.11.4.i, and 3.11.4.ii**). In contrast, a control construct, Ana2^{R16A}-GFP (R16 was not involved at the binding interface in the structure) provided a wild-type level rescue. SAS-4^{MC3}-GFP also showed a robust rescue. Interestingly, the level of rescuing ability by each construct in these *in vivo* experiments roughly correlates with the expected severity of each mutation, as inferred from *in vitro* binding results carried out on the *D. rerio* homologues by Dr. Mark van Breugel. Using ITC, he was able to quantify the change in binding affinity conferred by various point mutations [Cottee et al., 2013]. Taken together, these *in vivo* results suggest that Ana2 is recruited upstream of SAS-4, and that the Ana2:SAS-4 interaction is absolutely essential for centriole duplication.



	+/+	-/-(*Ana2*^{169})	-/-(*Ana2*^{719})	-/- WT	-/- ΔCR2	-/- R16A	-/- P11A	-/- R12A	-/- P11A+R12A
3	0.0	0.0	0.0	1.7	0.1	0.9	0.5	0.1	0.0
2	95.2	0.0	0.0	89.2	1.9	88.2	47.5	62.2	16.1
1	2.6	1.9	0.8	4.2	9.8	6.5	41.0	30.6	42.5
0	2.1	98.1	99.2	4.8	88.1	4.3	11.0	7.1	41.4

Figure and Table 3.11.4.i: Quantification of centrosome numbers in third instar larval brains expressing Ana2-GFP transgenes. For different genotypes, the chart displays the proportion of brain cells in which the indicated number of centrosomes were counted. Cells from WT animals (*Ana2*$^{+/+}$) have predominantly two centrosomes per mitotic cell. Loss of *Ana2* (*Ana2*^{169} or *Ana2*^{719}) results in ~all cells having no centrosomes. This phenotype is rescued with the wild-type transgene, or the control mutant, R16A. Transgenes with the interface residues P11A or R12A, both show a less efficient rescue. Combination of these mutations results in an even less efficient rescue.



	+/+	-/-	-/- WT	-/- MC3	-/- MC2	-/- MC1	-/- MC123	-/- ΔTCP	-/- E792V	-/- Y765V
	Sas4 genotype									
	SAS-4 transgene									
3	0.0	0.0	0.8	0.3	0.3	0.0	0.1	0.0	0.4	0.1
2	95.2	0.6	96.3	81.9	14.1	3.1	1.8	1.0	54.8	56.3
1	2.6	4.24	2.5	13.3	42.2	22.2	8.0	3.8	31.1	32.1
0	2.1	95.2	0.4	4.5	43.5	74.7	90.1	95.1	13.7	11.5

Figure and Table 3.11.4.ii: Quantification of centrosome numbers in third instar larval brains expressing SAS-4-GFP transgenes. For different genotypes, the chart displays the proportion of brain cells in which the indicated number of centrosomes were counted. Cells from WT animals (*Sas4*$^{+/+}$) have predominantly two centrosomes per mitotic cell. Loss of *Sas4* (*Sas4*$^{-/-}$) results in ~all cells having no centrosomes. This phenotype is rescued with the wild-type transgene. Constructs harbouring mutation clusters rescued to differing extents, with MC1 rescuing the least efficiently, apart from the combined MC123. SAS-4$^{\Delta TCP}$ showed no rescue whatsoever. The E792V and Y765V point mutants showed a partial rescue.

3.12: Perspectives

In this Chapter, a bioinformatics approach was used to delineate the likely functional domains involved in the SAS-4:Ana2 interaction. Using a purified fusion protein that contained the relevant domains from both proteins, I was able to solve the crystal structure of the interaction interface. This structure is extremely interesting, as the TCP domain adopts a completely novel architecture - a non-globular, continuous β-sheet. The CR2 was predicted to be unstructured, and it is possible that it may fold upon encountering the TCP domain. Even in the complex, however, it did not adopt a canonical α/β secondary structure, but rather a polyproline helix. This interaction interface is also very interesting as it may be exemplary of other interactions in the centriole system, where few of the key proteins essential for centriole assembly are predicted to form, or even contain, globular domains. The crystal structure also illustrated the involvement of a conserved residue E792, which was directly involved at the SAS-4:Ana2 interface, and is mutated to valine in human MCPH. This study therefore provides the first structural insight linking centriole defects to a human disease.

Throughout this study I collaborated with Dr. Mark van Breugel, who performed similar studies using the *D. rerio* system. By working in separate model organisms we were able to show that the interaction was highly conserved across evolution. This is important, as centriole proteins often contain low levels of sequence identity, and indeed, it was not obvious to align and identify the Ana2 CR2 domain using MSA analysis. The fact that the core residues involved in this interaction are conserved therefore demonstrates its importance. Furthermore, the identical behaviour of STIL and Ana2 with regards to this interaction has enabled me to confirm that Ana2 is the true *Drosophila* orthologue of vertebrate STIL, as suggested by limited sequence and topological conservation in a previous study [Stevens et al., 2010b].

Interestingly, since the completion of this study, further structures of the SAS-4/CPAP TCP domain have been solved. The *D. rerio* TCP domain was again shown to form an identical complex with the CR2 of STIL, further confirming our findings [Hatzopoulos et al., 2013]. Additionally, the *Drosophila* SAS-4 TCP domain was crystallised in isolation, confirming that it does indeed exist as a solvent exposed β-sheet, away from its interaction with Ana2 [Zheng et al., 2014b]. As discussed in **Section 3.8.4**, TCP domain fibrils are observed in many of these crystal structures, however further work would be needed to determine whether these are of any significance.

My analysis of the *Drosophila* domains meant that the function of the interaction could be meaningfully dissected *in vivo*. Complete abrogation of the interaction interface of either protein resulted in the near-complete failure of centriole biogenesis. This demonstrates that the Ana2:SAS-4 interaction is an essential step in centriole biogenesis. However, milder mutations (such as the MCPH mutation, E792V) resulted only in an inefficiency to build centrioles. This has interesting implications for MCPH, as it suggests that it is an inefficiency to build centrioles - and therefore the reduced centriole number - that is likely to be the root cause of the abnormal brain development associated with this particular mutation. This may explain why many of the identified MCPH genes encode core centriole duplication factors. However, it remains unclear how inefficient centriole assembly leads to the development of small brains in humans.

The *in vivo* results also suggested, that despite impairment of either protein resulting in centriole loss, the proteins localised differently to centrioles in a wild-type background. Ana2 constructs, whether capable of the SAS-4 interaction or not, were able to localise to centrioles similarly to wild-type Ana2, suggesting that Ana2 is recruited to centrioles independently of its interaction with SAS-4. Mutated SAS-4 constructs however, were impaired in their localisation, suggesting

that SAS-4 is recruited to centrioles via Ana2. This result provides further evidence to studies in human cells, which suggest that HsSTIL is recruited to centrioles upstream of HsCPAP [Tang et al., 2011; Vulprecht et al., 2012]. However, the mechanisms underlying the centriolar recruitment of Ana2 remain poorly understood – a topic I attempt to address in **Chapters 4 and 5**.

Chapter 4: Structure/function analysis of Ana2 *in vivo*

4.0: Introduction

4.0.1: Ana2

As discussed in **Chapter 1**, *Drosophila* Ana2 is an essential centriole duplication protein that localises to the procentriole [Goshima et al., 2007; Stevens et al., 2010a]. Ana2 can drive *de novo* centriole formation when overexpressed and may be implicated with SAS-6 in the formation of the cartwheel [Stevens et al., 2010b]. Additionally, Ana2 has been shown to interact with another essential procentriole protein, SAS-4 [Cottee et al., 2013; Tang et al., 2011].

Ana2 belongs to the SAS-5/Ana2/STIL family of functional homologues. All three proteins share a central coiled-coil, and additionally contain a conserved C-terminal "STAN" domain. It has been shown in *C. elegans* that SAS-5 binds SAS-6 [Hilbert et al., 2013; Leidel et al., 2005; Qiao et al., 2012]. It is likely that the two proteins are recruited to the centriole together and that this is achieved via a direct interaction between ZYG-1 and SAS-6 [Lettman et al., 2013]. The case may not be so simple for Ana2 and STIL. Although Ana2 has been reported to weakly interact with SAS-6, this interaction has not been characterised in detail. No such interaction has been observed for STIL and SAS-6. In the case of STIL, one study has even proposed that STIL may be recruited upstream of SAS-6 [Vulprecht et al., 2012].

Although Ana2 interacts with SAS-4, Ana2 appears to be upstream of SAS-4 in the recruitment pathway (**Chapter 3**). The fact remains that the recruitment of Ana2 to the centriole remains

poorly understood. Additionally, the STAN domain has no known function, despite being an area of particular conservation.

4.0.2: Aims

- Define functional domains of Ana2.
- Analyse the ability of different Ana2 domains to localise to the centriole *in vivo*.
- Define the domain(s) necessary for centriolar localisation of Ana2.

4.0.3: Key findings

- Constructs lacking the STAN domain appear to localise less efficiently than wild-type Ana2.
- The STAN domain in isolation can weakly target to the centrosome.
- The self-association of Ana2 via the central coiled-coil is required for wild-type localisation.
- Constructs lacking both the CC and STAN domains fail to localise.

4.1: Bioinformatic analysis of Ana2

In order to design constructs to carry out a structure/function analysis, putative functional domains of Ana2 had to be defined. Using bioinformatic analysis on STIL and Ana2 sequences, domains were defined that were common to both protein families.

4.1.1: Ana2 and STIL contain several conserved features

In **Chapter 3**, bioinformatic analysis of STIL revealed a large N-terminal domain (CR1) that did not appear to be conserved through to *Drosophila*. Removal of this domain from STIL sequences allowed Ana2 and STIL sequences to be aligned in a 37 species multiple sequence alignment (MSA) using ClustalX [Larkin et al., 2007]. This allowed delineation of the SAS-4 binding region, CR2 (**Chapter 3**, **Figure 3.1.2.i**). Further analysis of this MSA revealed that two further regions with enriched conservation (CR3 and CR4) were shared between the two protein families (**Figure 4.1.1.i, A, B**). CR3 roughly covered the STAN domain identified by [Stevens et al., 2010b], although the inclusion of more sequences in the MSA allowed the boundaries of this domain to be refined to residues 316-383 of DmAna2. CR4 is at the extreme C-terminus of both proteins, and mapped to a region previously defined as the "Truncated in Microcephaly" (TIM) domain [Arquint et al., 2012]. This domain is largest in STIL proteins where it contains a KEN box degradation signal, essential for the degradation of STIL [Arquint and Nigg, 2014]. However there is no KEN signal in Ana2. Only the last ~20 residues of Ana2 aligned with the C-terminus of the STIL TIM domain, so in this study, this subsection is referred to as "CR4". Ana2 and STIL proteins have a predicted central coiled-coil domain, and so I identified the boundaries of the Ana2 CC using the COILs server [Lupas et al., 1991] (**Figure 4.1.1.i, C**). Secondary structural prediction by Psipred [Jones, 1999] indicated that Ana2 and STIL (from CR2 onwards) were low in secondary structural content, indicating that they may be intrinsically unfolded proteins. However, a helix was predicted, covering the predicted coiled-coil for both Ana2, and HsSTIL.

Further helices were predicted within the STAN domain, roughly consistent between both Ana2 and STIL. CR4 appears to be unstructured. The regions between these putative functional domains are generally unstructured and unconserved.

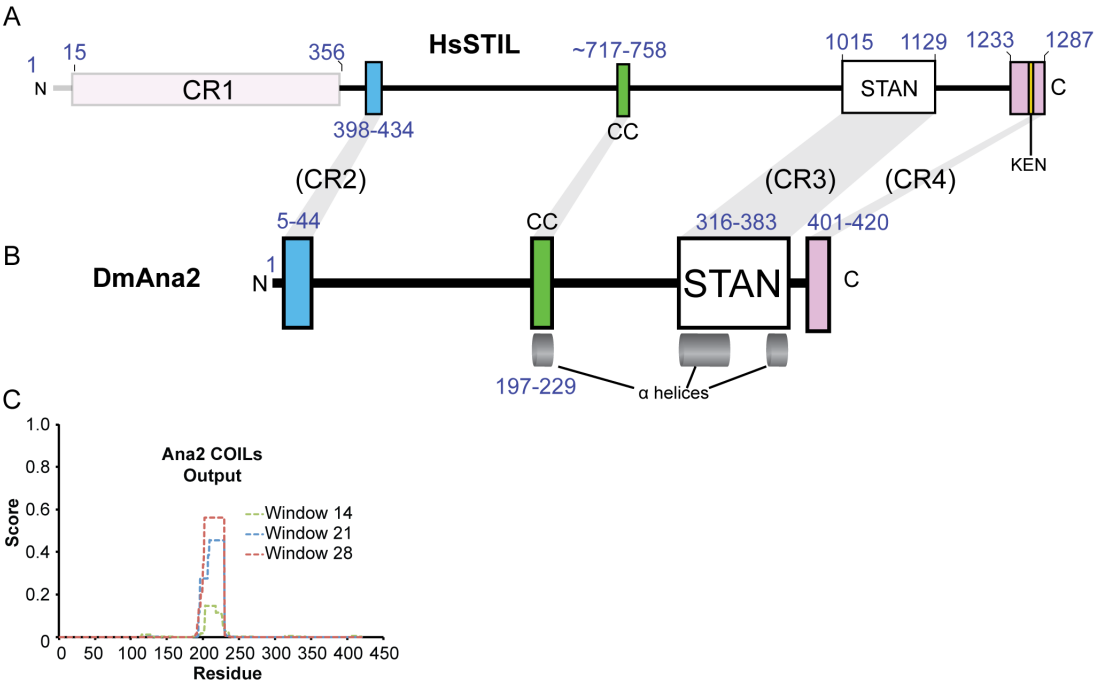


Figure 4.1.1.i: Summary of bioinformatic analysis of STIL and Ana2. (**A, B**) Schematics of HsSTIL **A**, and DmAna2, **B**. Note that DmAna2 is displayed on a larger scale. MSA analysis was used to delineate conserved regions (CRs) that were shared between these protein families. Equivalent domains are labelled and similarly coloured. CR2 is the region used to bind to the SAS-4 TCP domain (**Chapter 3**). The coiled-coil is defined by the correlation between a COILs server prediction for a coiled coil (**C**), and the Psipred prediction of a helix. The STAN domain was redefined from the (Stevens et al., 2010b) study to residues 316-383, based on MSA analysis. CR4 appeared to be more extensive in STIL and also contained a KEN degradation signal, apparently absent in Ana2. Psipred also predicted helices inside the STAN domain. Helices are depicted as grey cylinders. Similar helices were predicted for STIL.

4.2: Designing a structure/function analysis

4.2.1: An mRNA injection assay to assess centriolar localisation

Drosophila embryos provide an ideal system within which to monitor protein localisation to the centriole. During the syncytial stage of embryogenesis, nuclei undergo rapid mitosis, however, cytokinesis does not occur, and so the embryo is effectively a large multi-nucleate cell (the syncytial blastoderm). After several divisions, the nuclei migrate to the surface of the egg, allowing multiple nuclei to be visualised in the same plane, while they undergo further divisions. *In vitro* transcribed mRNAs, encoding GFP-fusions to a protein of interest, can be injected into the blastoderm. The GFP-fusion proteins are translated throughout the cytoplasm, and can be tracked using fluorescence microscopy. Because all the centrioles exist in a communal cytoplasm, the recombinant protein can be visualised at many centrioles/centrosomes at the same time, allowing quantitative results to be obtained in a relatively high-throughput manner, without the need for generation of recombinant flies (**Figure 4.2.1.i**).

I used this RNA injection approach to delineate the domains necessary for Ana2 localisation to the centriole. RNA constructs were designed that encoded various truncated forms of Ana2 fused to GFP. Starting with full length Ana2, domains were sequentially deleted from both the N and the C terminus in order to assess the effect on centriole localisation. Ideally this analysis would be performed in an Ana2 null background, but Ana2 is essential for centriole biogenesis so Ana2 null embryos arrest very early in development [Stevens et al., 2007]. Injections must therefore be carried out in a wild-type background. To verify centriolar localisation, injections were carried out in transgenic lines expressing the centrosomal marker, RFP-Cnn (**Figure 4.2.1.i, C**). Joanna Leveson, an undergraduate student I supervised, carried out the majority of the fly

husbandry, RNA microinjections, and microscopy work for these experiments. *Drosophila* embryos were prepared and injected as previously described [Oliveira et al., 2010].

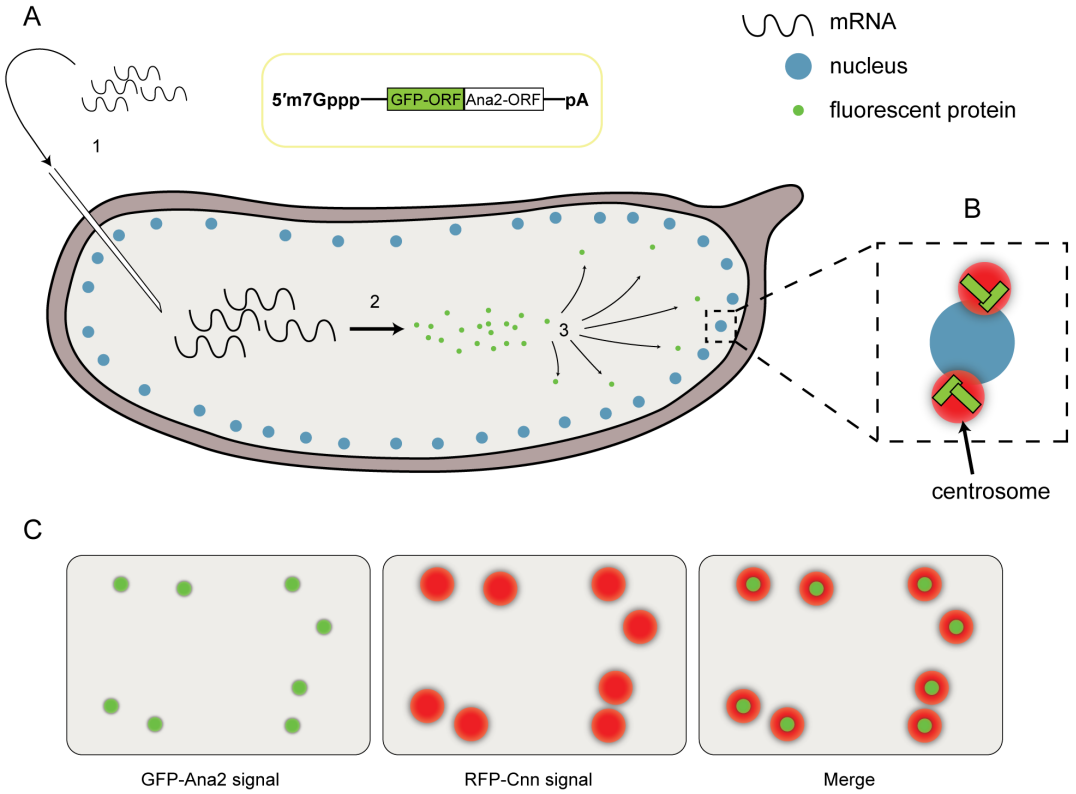


Figure 4.2.1.i: A schematic illustration of the mRNA injection assay. (**A**) Recombinant mRNA (yellow box) encoding GFP-Ana2 fusion constructs is (1) injected into the syncytial cytoplasm. The mRNA is translated *in vivo* (2) to produce fluorescent Ana2 molecules. These molecules are free to diffuse (3) throughout the cytoplasm, where they can localise to existing centrioles, or be incorporated into nascent centrioles. (**B**) Close up view showing how each nucleus is associated with a pair of centrosomes. At the centre of the centrosome lies a pair of centrioles. (**C**) Schematic of fluorescence signal visualised when fluorescent Ana2 localises at the centriole. GFP (Ana2) signal is seen as pairs of foci, each focus representing a pair of centrioles at the centre of a centrosome. (The individual centrioles are below the resolution of the microscope system used in these experiments). The GFP foci should co-localise with RFP-Cnn signal, which labels the PCM. If the protein is tightly localised to the centriole, the GFP foci should be seen as tight puncta within the PCM signal.

4.2.2: Construct design, cloning and RNA preparation

Ana2 constructs were designed as summarised in **Figure 4.2.2.i**. Constructs were PCR amplified from an Ana2 cDNA before cloning into a pDONRzeo gateway vector using attBP cloning, before transfer into either pRNA_NT-GFP_DEST or pRNA_CT-GFP_DEST using attLR cloning (**Appendix 7.1.2**). All constructs were designed to be in frame with GFP (encoded in the DEST

vector). Constructs for N-terminal GFP fusion had a STOP codon inserted at the end of the Ana2 ORF. Constructs not starting at the native N-terminus had a methionine inserted at the N-terminus. Internal deletions were carried out using site-directed mutagenesis (QuikChange II XL, Agilent).

RNA was obtained via *in vitro* transcription (**Appendix 7.1.6**), and its concentration normalised so that all constructs were injected at the same molar concentration. It should be noted, however, that the volume of injection, the volume of the cytoplasm, and the rate of translation and half-life of the mRNA/protein *in vivo* could not be controlled for, or assessed.

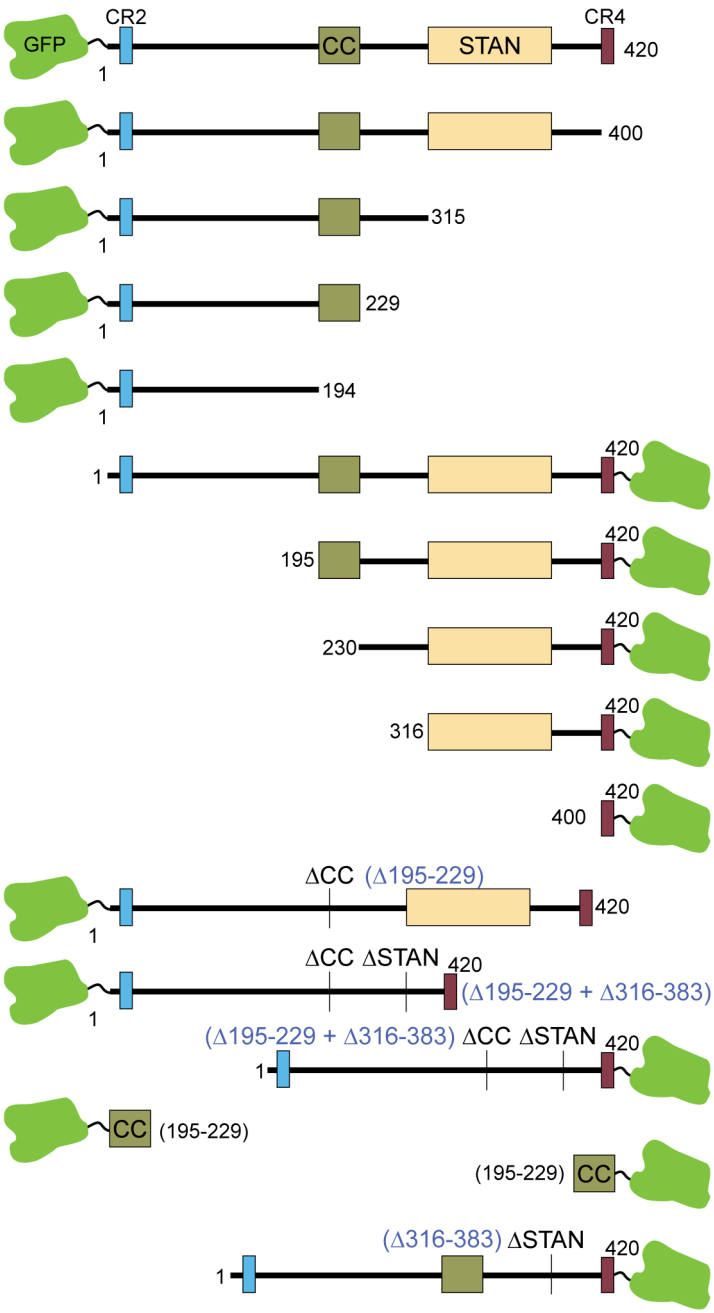


Figure 4.2.2.i: Ana2-GFP fusion constructs used during this study. Domains are labelled for the top construct, GFP-Ana2$^{1-420}$ (full-length). The amino acid termini for each Ana2 construct are labelled. Internally deleted domains are labelled and marked by vertical black lines. The exact amino acids deleted are labelled in blue.

Pilot experiments (not shown) were carried out using full-length, GFP-Ana2$^{1-420}$ as a positive control. It was determined that for RNA at 3.66 µM, 100 % of injected embryos displayed a strong localisation phenotype. However at 0.84 µM, 25% of embryos showed a "weak" localisation. Images were scored on a qualitative basis, using a single blind test

where the concentration of mRNA injected for a particular image was unknown. The quantification of injection results is discussed in greater detail in **section 4.3.1**. The concentration of 3.66 µM was therefore used as the benchmark for all further experiments, to enable as fair a comparison of results as possible.

4.3: Localisation of GFP-Ana2 fusion proteins to the centriole

4.3.1 Quantification of construct localisation

In order to effectively compare the injection results, results were quantified. Ideally, it would be desirable to quantify the fluorescence signal at each centriole/centrosome. Such an approach is not appropriate for this assay due to the uncontrollable variables such as the volume of injection, lifetime of RNA/protein, level of translation – all of which would influence the amount of signal detectable at centrioles/centrosomes. I decided therefore, to make a more qualitative judgement, categorising each embryo according to whether the centriole/centrosome staining was tight (indicating centriole localisation) or diffuse (indicating PCM localisation), and whether the signal was strong, weak or absent.

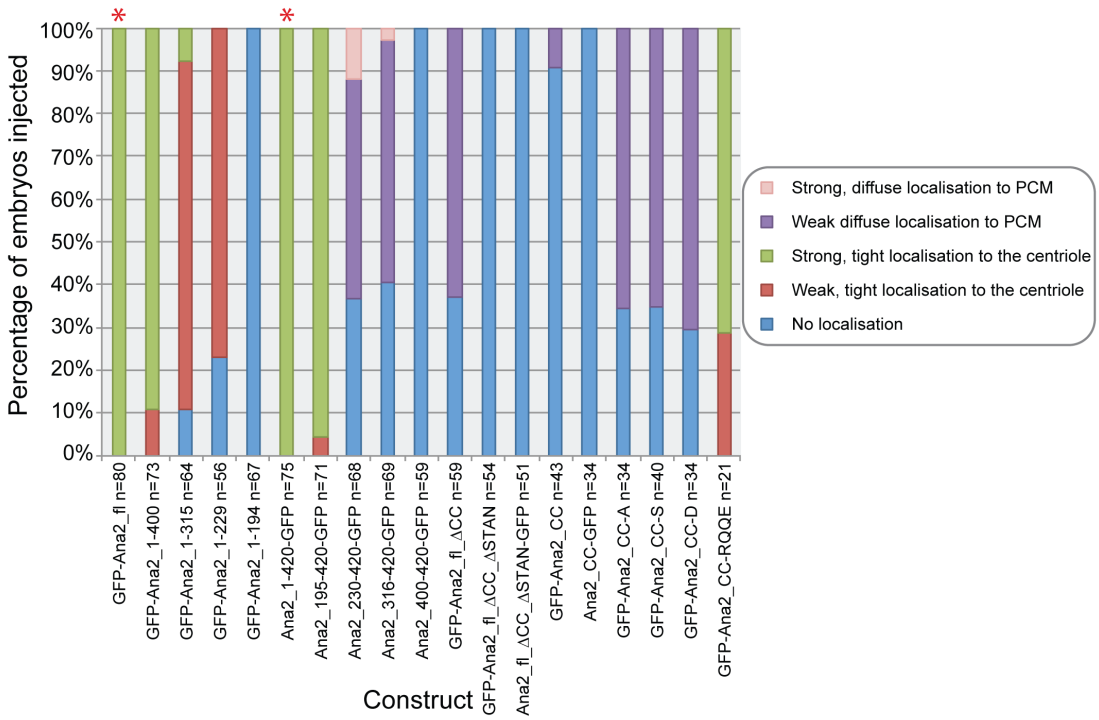


Figure 4.3.1.i: Quantitative comparison of the localisation seen for each construct. A blind, qualitative judgement was made for each embryo, placing it in one of 5 categories as described in the legend. For each construct, results are shown as the percentage of embryos that displayed a particular category of localisation. Red asterisks mark the full-length, wild-type constructs.

Importantly, this analysis was performed blind to avoid any bias from *a priori* expectation of the result. The percentage of embryos displaying each type of localisation could then be quantitatively compared. Results were usually very consistent between embryos injected with a given construct, suggesting this approach was valid and provides a useful way to compare these results. **Figure 4.3.1.i** shows these quantified results for all constructs.

Note that these quantification results will be re-displayed as each construct is discussed individually throughout this chapter.

4.3.2 Full-length GFP-Ana2 and Ana2-GFP localise strongly to centrioles

GFP-Ana2$^{1\text{-}420}$ and Ana2$^{1\text{-}420}$-GFP were both visible as bright, tightly focussed, paired foci (**Figure 4.1.2.i**). These foci colocalised with RFP-Cnn signal, which was much more diffuse, suggesting that the GFP-fusion proteins tightly localised to the centriole pairs, within the more diffuse PCM. 100% of analysed embryos showed this phenotype, so these constructs serve as robust controls, not only for the experimental approach, but also for using both the N and C terminal GFP-fusions, thus enabling meaningful results to be obtained from all constructs, fused at either terminus.

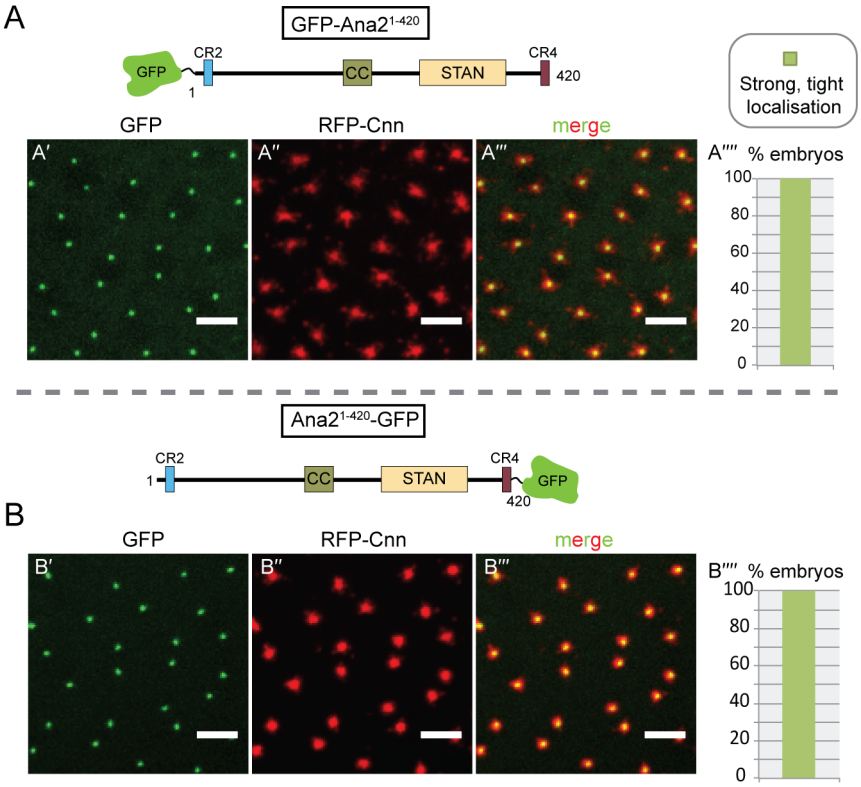


Figure 4.3.2.i: Localisation of full length Ana2 constructs. (**A**) GFP-Ana2$^{1-420}$ localises to strong, tight foci within the PCM, as labelled by RFP Cnn. (**B**) Ana2$^{1-420}$-GFP behaves identically. Representative images from a single embryo per construct are shown, scale bars are 5 µm. The quantification results for each construct are shown to the right. **(A⁗, B⁗)**.

4.3.3 Truncation of Ana2 domains from the C-terminus reveals that the CC and STAN domains are involved in centriolar localisation

A GFP-fusion protein lacking the CR4 domain (GFP-Ana2$^{1-400}$), was visible as bright, tightly focussed, paired foci **(Figure 4.3.3.i, A)** - very similar to the "wild-type" localisation shown by full-length Ana2$^{1-420}$-GFP. This result suggests that the CR4 domain is dispensable for centriole localisation. However, two GFP-fusion proteins lacking both the STAN and CR4 domains (GFP-Ana2$^{1-315}$, GFP-Ana2$^{1-229}$) did not show a wild-type localisation. Although tight foci were visualised, these foci were weak in the large majority of embryos visualised (**Figure 4.3.3.i, B**), indicating that the STAN domain is required for strong centriolar localisation. (Note that results

for GFP-Ana2$^{1-229}$ are displayed in **Appendix 7.6**). The residual localisation was attributable to the CC region. A GFP-fusion protein containing only the CR2 domain, and lacking the CC, STAN and CR4 domains (GFP-Ana2$^{1-194}$) failed to localise although it had been translated, as suggested by the detection of diffuse GFP signal throughout the cytoplasm (**Figure 4.3.3.i, C**).

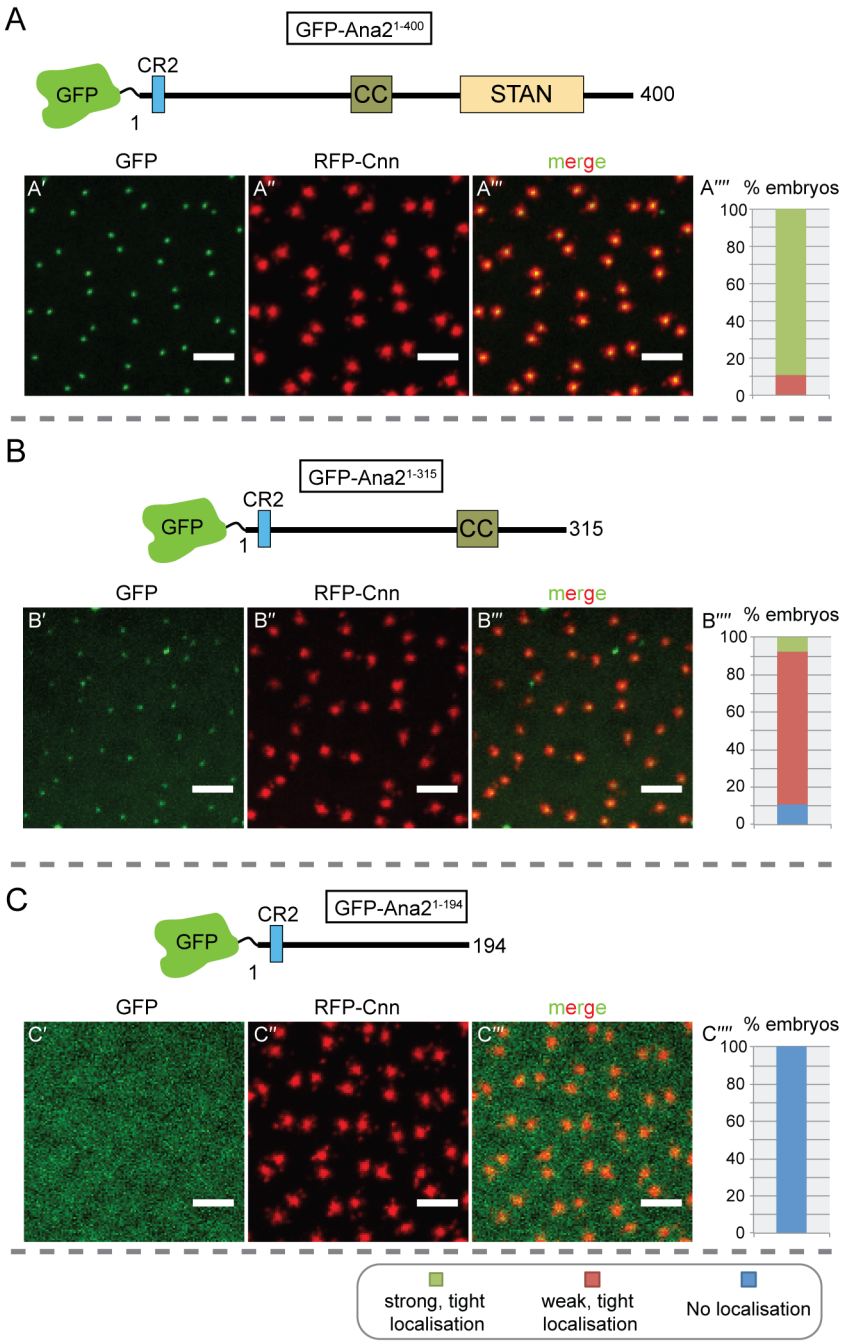


Figure 4.3.3.i: Localisation of C-terminally truncated GFP-Ana2 constructs. (**A**) GFP-Ana2$^{1-400}$ strongly and tightly localises to centrioles. (**B**) GFP-Ana2$^{1-315}$ localises weakly but tightly to centrioles. (**C**) GFP-Ana2$^{1-194}$ was spread diffusely and did not localise. Representative images from one embryo per construct are shown for the predominant manner of localisation. Scale bars are 5 µm. Quantification results are summarised in **(A⁗, B⁗, C⁗)**, coloured according to the legend underneath. Note that for presentation purposes, to show weak GFP signal, images have been adjusted to use the full scale of intensities.

This suggests that the CR2 does not target Ana2 to the centriole. Together, these results strongly suggest that in a wild-type background, the CC domain is essential for localisation the centriole, and although not essential, the STAN domain enhances the strength of localisation. Because the coiled-coil is a predicted self-association domain, it is possible that constructs containing a coiled-coil can localise with wild-type endogenous Ana2 in this assay.

4.3.4 Truncation of Ana2 domains from the N-terminus confirms that the CC and STAN domains are involved in centriolar localisation

A GFP-fusion protein lacking the CR2 domain (Ana2$^{194-420}$-GFP), was visible as bright, tightly focussed, paired foci **(Figure 4.3.4.i, A**) - very similar to the "wild-type" localisation shown by full-length Ana2$^{1-420}$-GFP. This confirms that the CR2 domain is not involved in localisation. GFP-fusion proteins lacking the CR2 and CC domains, but retaining the STAN and CR4 domains (Ana2$^{230-420}$-GFP, Ana2$^{316-420}$-GFP) had a distinctly non-wild-type localisation phenotype. In around 30-40% of embryos injected with these constructs, no centriole localisation was detected. However, in the remaining embryos, although much of the protein was diffuse in the cytoplasm, it was clearly enriched in regions that co-localised with the RFP-Cnn signal (**Figure 4.3.4.i, B**). (Note that results for Ana2$^{230-420}$-GFP are displayed in **Appendix 7.6**). As the CR4 domain appears to be dispensable for centriole localisation (**Section 4.3.3**), this result suggests that the STAN domain is able to weakly associate with the PCM. Indeed, a GFP-fusion protein containing only the CR4 domain (Ana2$^{400-420}$-GFP), was unable to localise (**Figure 4.3.4.i, C**), further suggesting that the CR4 domain does not target Ana2 to the centriole.

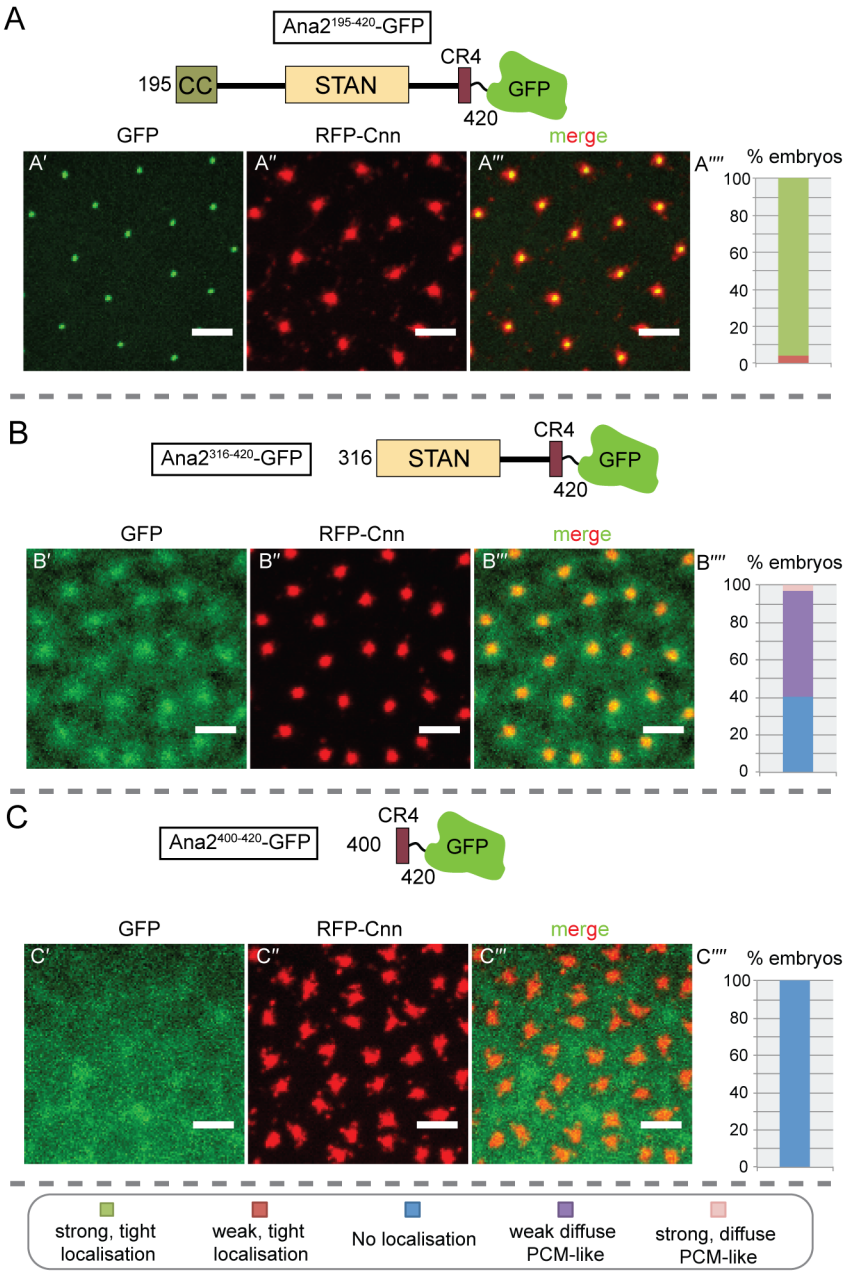


Figure 4.3.4.i: Localisation of N-terminally truncated Ana2-GFP constructs. (**A**) Ana2$^{195\text{-}420}$-GFP strongly and tightly localises to centrioles. (**B**) Ana2$^{316\text{-}420}$-GFP had a weak, diffuse, "PCM-like" localisation. (**C**) Ana2$^{400\text{-}420}$-GFP was spread diffusely and did not localise. Representative images from one embryo per construct are shown for the predominant manner of localisation. Scale bars are 5 µm. Quantification results are summarised in **(A⁗, B⁗, C⁗)**, coloured according to the legend underneath. Note that for presentation purposes, to show weak GFP signal, images have been adjusted to use the full scale of intensities.

4.3.5 Deletion of the CC and STAN domains from Ana2 confirms that they are responsible for localisation of Ana2 to the centriole

Results suggested that the CC was essential for tight centriolar localisation, and so residues 195-229 comprising the CC were cleanly deleted from Ana2 to create the GFP-fusion construct, GFP-Ana2$^{\Delta CC}$. In 37% of injected embryos, no localisation was evident for this construct. However, in the remaining embryos, although much of the protein was diffuse in the cytoplasm, it was clearly enriched in regions that co-localised with the RFP-Cnn signal (**Figure 4.3.5.i, A**). This weak, PCM-like localisation phenotype is identical to that observed in **section 4.3.4**. To prove that this weak PCM-like localisation was specifically attributable to the STAN domain, residues 316-383 comprising the STAN domain were cleanly deleted to create two fusion constructs, GFP-Ana2$^{\Delta CC\Delta STAN}$ and Ana2$^{\Delta CC\Delta STAN}$-GFP. In every observed embryo, both of these constructs failed to localise, and were instead spread diffusely throughout the cytoplasm (**Figure 4.3.5.i, B, C**). By cleanly deleting the CC and STAN domains, these results show that these domains are together responsible for wild-type localisation of Ana2 to the centriole.

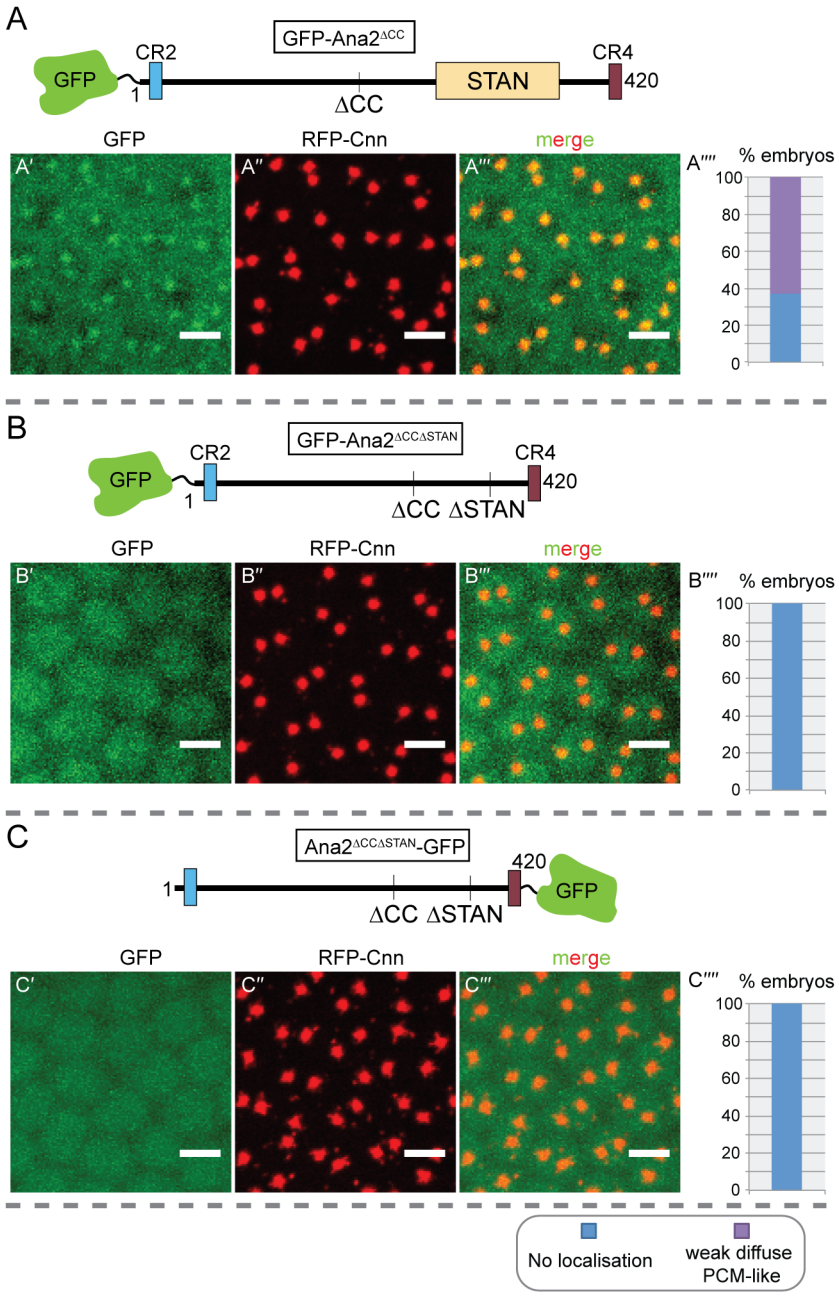


Figure 4.3.5.i: Localisation of GFP-Ana2 constructs the CC and STAN domains. (**A**) GFP- Ana2$^{\Delta CC}$ localises weakly and diffusely with the PCM. (**B, C**) GFP- Ana2$^{\Delta CC\Delta STAN}$ and Ana2$^{\Delta CC\Delta STAN}$-GFP were both spread diffusely and did not localise Representative images from one embryo per construct are shown for the predominant manner of localisation. Scale bars are 5 µm. Quantification results are summarised in **(A⁗, B⁗, C⁗)**, coloured according to the legend underneath. Note that for presentation purposes, to show weak GFP signal, images have been adjusted to use the full scale of intensities.

4.3.6: Fusion proteins containing only the CC domain do not localise to centrioles

In order to test whether the CC specifically targets Ana2 to the centriole, constructs were designed that contained only the CC, fused at either the N or C terminus, to GFP. These

constructs did not localise, however translation and folding of the proteins occurred, as evidenced by diffuse GFP signal (**Figure 4.3.6.i**). Although the CC is required for tight localisation, this result suggests that the CC itself does not target to the centriole. However, it is possible that a construct this small may not be able to function, due to steric hindrance from the close proximity of GFP. Unfortunately, as the termini of 195 or 229 have not been tested as functional fusion positions for GFP with a positive control, it is not possible to say whether this is the case. Therefore, it is still possible that the CC could be responsible for targeting Ana2 to the centriole.

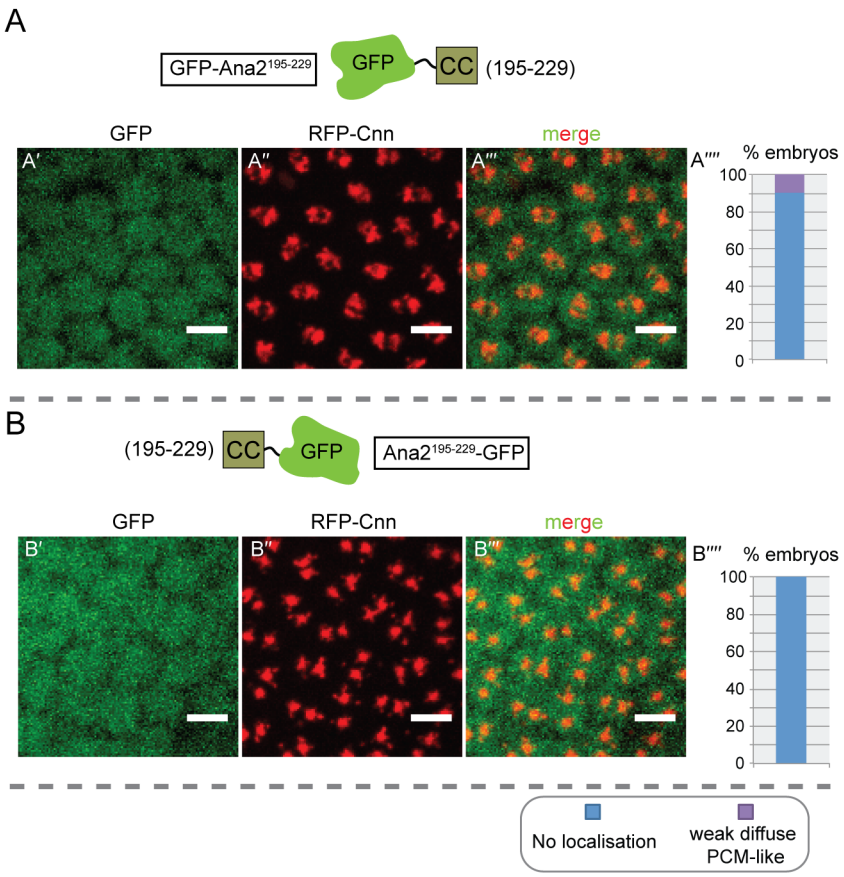


Figure 4.3.6.i: Localisation of GFP-Ana2 constructs containing only the CC domains (**A, B**) Both GFP-Ana2$^{195-229}$ and Ana2$^{195-229}$-GFP were spread diffusely and did not localise. Representative images from one embryo per construct are shown for the predominant manner of localisation. Scale bars are 5 µm. Quantification results are summarised in **(A⁗, B⁗)**, coloured according to the legend underneath. Note that for presentation purposes, to show weak GFP signal, images have been adjusted to use the full scale of intensities.

4.3.7: Point mutations of the Ana2 CC domain suggest that self-association of Ana2 is required for robust centriole localisation

Results so far suggested that the CC domain was essential for centriole localisation. Coiled-coils are a common dimerisation motif, and are present in many centriolar proteins (SAS-6, STIL/Ana2, SAS-4/CPAP, CEP135, CEP152/Asl). Dimeric coiled-coils are defined by a regular pattern of residues known as the heptad repeat. The supercoiling of α-helices allows them to adopt a pitch of 3.5 residues per turn, and so the heptad repeat is significant. Hydrophobic residues are usually found in the "**a**" and "**d**" positions of the **abcdefg** heptad repeat, and face the same side of the helix. The burying of these hydrophobic residues is the driving force behind coiled-coil formation.

In order to test whether the CC domain functions as a dimerisation domain I decided to model the CC as a dimer using my MSA analysis to assign the heptad register (**Figure 4.3.7.i, A**). 7 out of the 10 **a**/**d** position residues of the heptad were hydrophobic, suggesting that they might be able to drive dimerisation. Using this register, the sequence of the Ana2 CC was mapped onto the structure of a canonical CC to form a model (**Figure 4.3.7.i, B, C**). In order to disrupt dimerisation, and assess the affect on Ana2 localisation, I designed three RNA injection constructs, where the ten **a**/**d** residues were mutated to Ala (CC-A), Ser (CC-S) or Asp (CC-D). I also noticed that 4 of the exposed charged/polar residues in my model were highly conserved (**Figure 4.3.7.i, A, C**). In order to see whether these residues were important for localisation, a construct was designed that mutated these residues to Alanine (CC-RQQE).

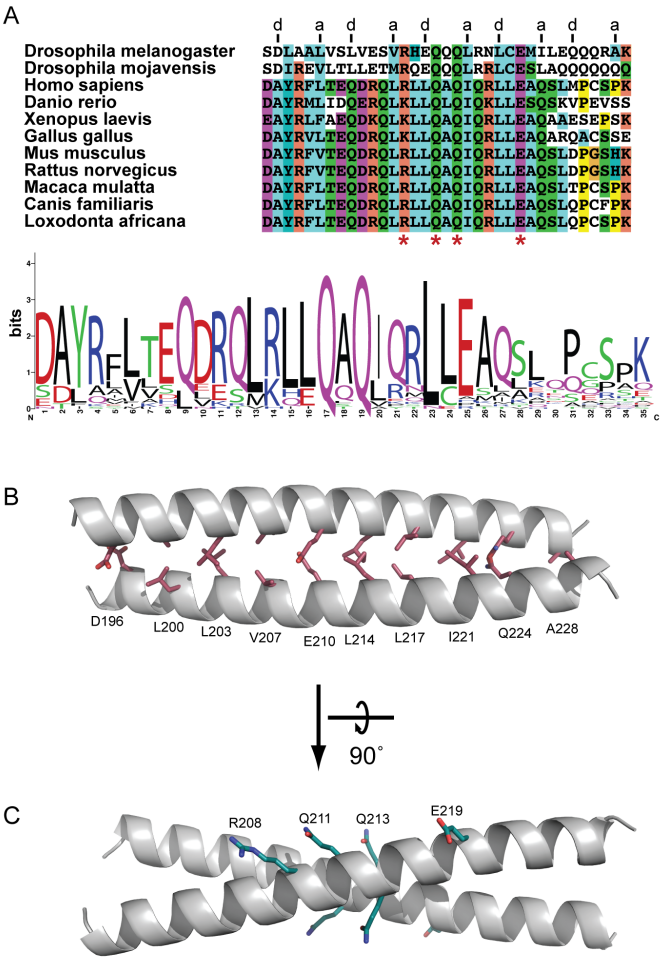


Figure 4.3.7.i: Design of Ana2 CC mutants. (**A**) MSA analysis of the CC region from STIL and Ana2 orthologues, coloured according to the ClustalX colour scheme. Note only 11 of the 37 species are shown for simplicity. The heptad register was manually assigned and is labelled above. Red asterisks mark the conserved non **a**/**d** residues. Below, a sequence logo of the CC region across these orthologues. (**B**) A structural model of the Ana2 CC as a canonical coiled-coil dimer. **a**/**d** residues are shown as magenta sticks, and all were point mutated to Ala, Ser or Asp in each of the CC-A, CC-S, or CC-D constructs respectively. (**C**) The same model, with the conserved R, Q, Q and E residues highlighted as cyan sticks. These were mutated to Ala in the CC-RQQE mutant.

When the CC-A/S/D constructs were injected into embryos, all three encoded GFP-fusion proteins localised in the weak and diffuse manner that is attributable to the STAN domain, consistent with a loss of function of the CC (**Figure 4.3.7.ii**). This result suggests that self-association via the coiled-coil is necessary for robust localisation of Ana2 to the centriole. The CC-RQQE mutant localised similarly to wild-type, although some embryos displayed slightly weaker localisation. This suggested that these residues play no essential role in localisation, although it is possible that the effect of these mutations is masked by association of the construct with endogenous wild-type protein.

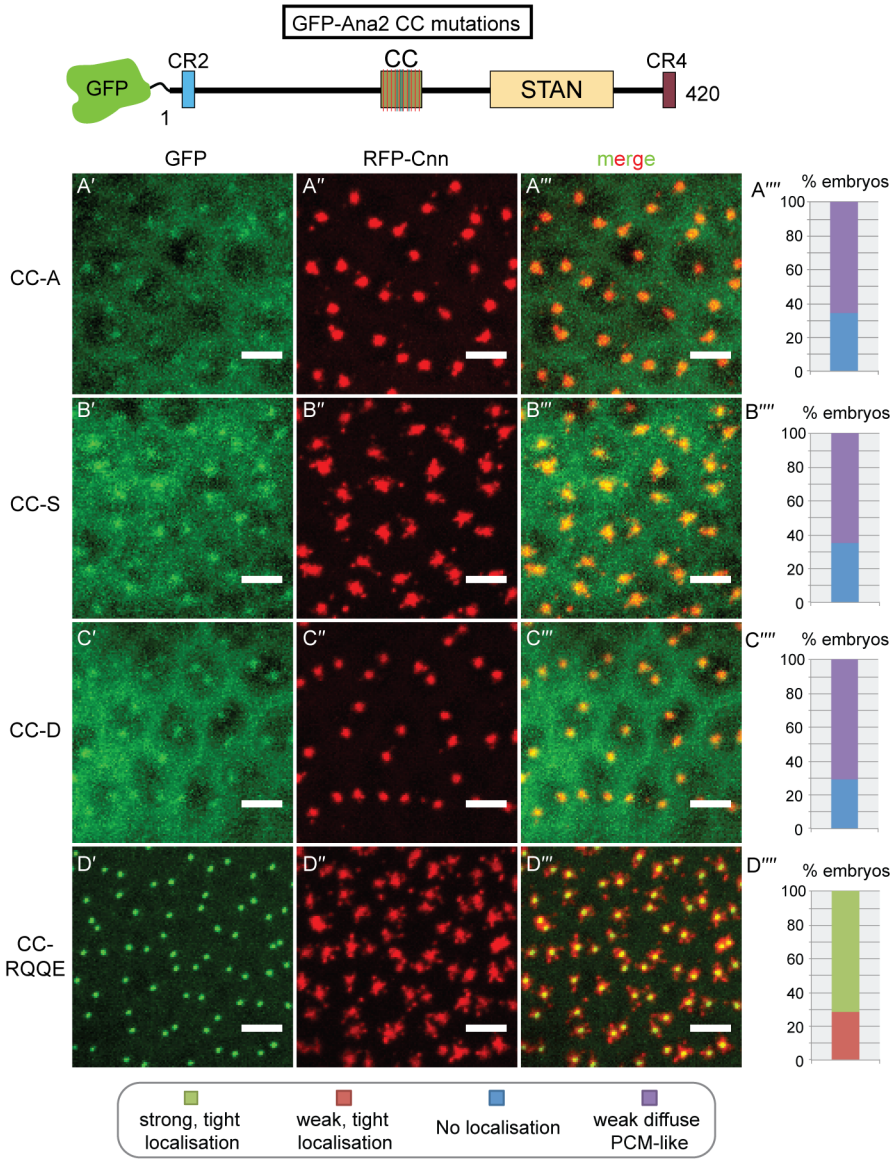


Figure 4.3.7.ii: Localisation of GFP-Ana2 constructs containing mutations in the CC domain. (**A, B, C**) GFP- Ana2 CC-A, CCS and CC-D mutants all behaved similarly. They localised weakly and diffusely with the PCM, consistent with loss of function of the CC. (**D**) The GFP- Ana2 CC-RQQE mutant localised similarly to wild-type, suggesting it retained CC functionality. Representative images from one embryo per construct are shown. Scale bars are 5 µm. Quantification results are summarised in (**A⁗, B⁗, C⁗, D⁗**), coloured according to the legend underneath. Note that for presentation purposes, to show weak GFP signal, images have been adjusted to use the full scale of intensities.

4.4: Planning of experiments using Transgenic *Drosophila*

As discussed in **section 4.2.1**, mRNA injection experiments had to take place in the presence of endogenous wild-type Ana2. In order to avoid any effects from the presence of endogenous Ana2, I wanted to create transgenic *Drosophila* lines, and cross these into an *Ana2*$^{-/-}$ background.

4.4.1: Construct design

The mRNA injection assay suggested that the CC and STAN domains were together responsible for wild-type localisation in the presence of endogenous Ana2. I therefore designed four constructs to test the function of these domains in transgenic animals (**Figure 4.4.1.i**). As for Ana2-GFP constructs described in **Chapter 3**, these constructs were cloned into pUbq_CT-GFP (DEST) vectors, via Gateway cloning, to generate Ana2 constructs with GFP fused at the C-terminus. Purified plasmids were sent to Bestgene Inc, and the Cambridge University Injection service for generation of transgenic lines via P-element mediated transformation [Roberts, 1998]. Genetic mapping/balancing and fly husbandry are being performed by Dr Nadine Muschalik.

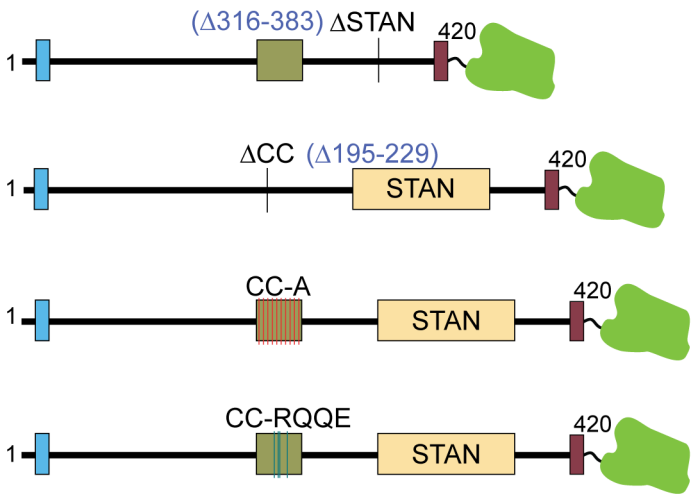


Figure 4.4.1.i: Constructs for which transgenic *Drosophila* lines have been generated.

4.4.2: Planned experiments

Experiments will be performed similarly to those discussed in **Chapters 2 and 3**. In order to confirm the RNA injection results, the ability of each construct to localise to centrioles in a wild-type background will be assayed. Importantly, the ability of each construct to rescue the uncoordination, and centriole loss phenotypes, associated with *Ana2*$^{-/-}$ mutants will be assessed. At the time of writing, these experiments are being performed by Dr Nadine Muschalik.

4.5: Perspectives

At the beginning of this study, little was understood about how Ana2 was recruited to the centriole, or what the roles of its domains were (other than the CR2 domain). Using RNA injection experiments with various mutated forms of Ana2 I have shown that both the coiled-coil and STAN domains of Ana2 are necessary for proper localisation to the centriole. More specifically, I disrupted the predicted self-association interface of the Ana2 CC to show that the function of the CC is likely to be a self-interaction domain. GFP-fusion proteins containing the STAN domain but lacking a CC displayed a weak, PCM-like localisation. However a construct lacking the CR2, and containing both the CC and STAN domains, had a wild-type localisation. This may suggest an avidity-based mechanism whereby the STAN domain targets Ana2 to the centriole, however it must be oligomerised via the CC in order to do so efficiently. This explanation relies on the residual localisation of the ΔSTAN construct (as in **Figure 4.3.3.i, B**) to be attributed to the CC or CR2 domains associating with endogenous Ana2.

However, there is an alternative explanation. It could not be solidly concluded that the CC itself does not target to the centriole. The fact that a ΔSTAN construct could still localise, tightly but weakly, to the centriole may suggest that the CC is responsible for initial centriolar targeting. This localisation is unlikely to be due to the CR2 domain, as this could be removed with ~no effect (**Figure 4.3.4.i, A**). Therefore, an alternative explanation is that the CC domain targets Ana2 to the centriole, and downstream actions provided by the STAN domain then stabilise this localisation. Indeed, the CC domain was the only common element in constructs that had a "tight" localisation. However, the unfortunate caveat of the RNA injection assay is that any construct with a CC could potentially oligomerise and localise with endogenous Ana2 protein. However, this was not the case for GFP-fusion proteins containing only the CC, as these failed to localise, possibly because the construct was sterically hindered.

Alternatively, it is not possible at this stage to rule out whether the Ana2 CC, rather than self-associating, actually associates with a different protein altogether.

Because of these complications, the results of transgenic rescue experiments will be particularly interesting. If constructs lacking functional CC or STAN domains cannot rescue centrosome loss, this will confirm the necessity of these domains. Interestingly, an existing report showed that the STAN domain was dispensable in Ana2 localisation to the centriole, and a ΔSTAN construct could rescue centriole loss in A*na2*$^{-/-}$ flies [Wang et al., 2011]. However discussion of this result in the [Wang et al., 2011] report is brief, and no quantification results are displayed, and so it will be interesting to see if my transgenic results will agree. Additionally, my experiments target very specific domain boundaries compared to the broader truncation approach adopted by [Wang et al., 2011]. Furthermore, my results suggest that the STAN domain, although perhaps not absolutely essential, still provides a significant stabilisation of Ana2 at the centriole.

Ana2 binds to SAS-4 via its CR2 domain, however, these results suggest that other Ana2 domains are responsible for its localisation to the centriole, which suggests a further interaction partner(s) must be present. Future experiments will no doubt focus on how Ana2 incorporates into the procentriole, specifically by identifying and characterising interactions with such putative interactions partners.

Chapter 5: *In vitro* characterisation of the Ana2 coiled-coil domain

5.0: Introduction

5.0.1: Insights from *in vivo* analyses

In vivo structure/function analysis (**Chapter 4**) had highlighted the importance of particular Ana2 domains in terms of centriolar localisation. This analysis highlighted that the STAN domain in isolation was able to weakly localise to the centriole or PCM, and also that a functional CC domain appeared to be a necessary feature for strong localisation. Due to its enriched conservation and ability to weakly localise, it was suspected that the STAN domain might bind to other centriole proteins. It was suspected that the coiled-coil may allow Ana2 to self-associate, and indeed, constructs with a mutated CC, or lacking a CC were strongly diminished in their localisation ability, suggesting that the self-association of Ana2 might be an important aspect of its function. Constructs lacking both domains failed to localise entirely.

5.0.2: Aims

- Recombinantly purify the STAN and CC domains in isolation.
- Assess the oligomerisation state of the Ana2 CC.
- Rationalise the effect of the CC mutants designed and discussed in **Chapter 4**.

5.0.3: Key findings

- The Ana2 central CC domain forms a stable, parallel 4-helix bundle that can also tether flanking domains.

- Point mutations of the CC described in **Chapter 4** render it monomeric.

- The CC region of HsSTIL oligomerises but its stoichiometry is unclear.

5.1: Characterisation of the STAN domain

5.1.1: Purification and characterisation of Ana2$^{236-420}$, containing the STAN + CR4 domains

Due to the STAN domain's role in localisation (discussed in **Chapter 4**) I wanted to purify and characterise it *in vitro*. During this study I was able to purify a C-terminal Ana2 fragment containing both the STAN and CR4 domains, however this fragment was not stable and appeared to degrade, resulting in a heterogeneous sample. In summary, I was able to obtain a small amount of sample enriched for the full-length fragment, and characterise it by circular dichroism (CD) spectroscopy. The fragment was largely unstructured, although the CD spectrum appeared to have some minor α-helical character, in keeping with the fairly confident prediction of helices in the STAN domain by Psipred (**Chapter 4, Section 4.1.1**). No further work was carried out with this fragment, and so its purification and CD characterisation are described in detail in **Appendix 7.7**.

5.2: Analysis of the central coiled-coil domain of Ana2

5.2.1: Design of a recombinant Ana2 coiled-coil construct

Following the RNA injection results (as detailed in **Chapter 4**) it was clear that the central CC domain was functionally important; the data could best be explained by the CC acting as a self-oligomerisation domain. I therefore wanted to characterise the behaviour of the Ana2 CC *in vitro* to test this possibility. Due to its relatively small size for a recombinant expression protein, I cloned Ana2$^{193-229}$ (containing the CC as predicted by the COILs server) into a custom vector designed to facilitate peptide expression. (The "pLip" vector, was constructed for Mark van Breugel by Mark Allen (LMB-MRC) (**Figure 5.2.1.i, A**). The plasmid uses the T7 system for bacterial expression, and fuses a TEV-cleavable and His-tagged Lipoyl domain from *Bacillus stearothermophilus* dihydrolipoamide acetyltransferase, at both the N- and C-terminus of the peptide of interest (**Figure 5.2.1.i, B**).

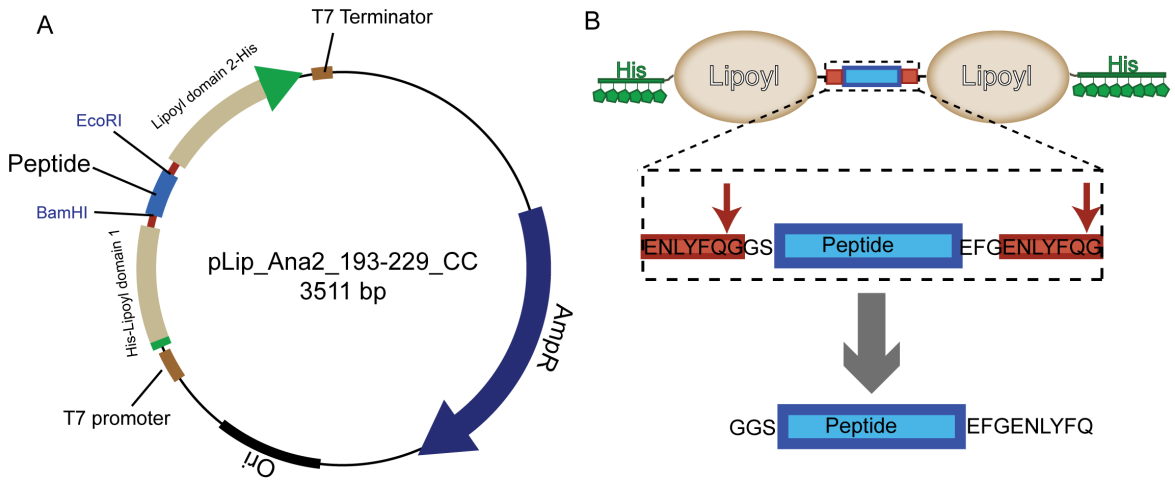


Figure 5.2.1.i: (**A**) The pLip plasmid, with the ORF for Ana2$^{193-229}$ cloned in-frame between the flanking Lipoyl domains. (**B**) The pLip system expression product, a peptide flanked by two cleavable His-tagged Lipoyl domains. Shown close-up is the peptide (blue) and surrounding sequence, including flanking TEV protease recognition sites (red arrows indicate position of cleavage). Upon cleavage with TEV protease (grey arrow), the peptide is liberated, albeit with some residual sequence from the cleavage sites.

The highly soluble Lipoyl domain acts as a vehicle for the expression of a peptide-containing fusion protein, (a diLipoyl-peptide) which, once purified, can be cleaved to isolate the peptide. Unlike for synthetically produced peptides, this system is amenable to mutagenesis.

Additionally, the highly negative charge of the Lipoyl domains usually overwhelms the contribution from the peptide, resulting in fusion proteins that have a predominantly acidic pI. This enables initial purification of different di-Lipoyl fusions under the same conditions. cDNA encoding Ana2$^{193\text{-}229}$ was amplified by a standard PCR protocol, and inserted into the pLip backbone between BamHI and EcoRI, via InFusion cloning (**Appendix 7.1.1**). (The fusion protein generated from this plasmid shall herein be denoted diLipoyl-Ana2_CC).

5.2.2: Expression and Purification of diLipoyl-Ana2_CC

The pLip_Ana2_193-229_CC plasmid was transformed into the *E. coli* expression strain, B834 (DE3). Recombinant protein was expressed and purified according to the standard procedure (**Appendix 7.1.4**) using Ni-NTA chromatography (**Figure 5.2.2.i**).

Amino Acids	Number Cys	MW	pI	ε_{280}	Abs 0.1%
252	3	27784.0	4.85	16960	0.611

Table 5.2.2.i Details of the diLipoyl-Ana2_CC-wt fusion protein, calculated using Protparam.

Typical yields of fusion protein were ~50mg per litre, but culture volume varied across different purifications. SDS-PAGE showed that Ni-NTA elution fractions were fairly pure, and exhibited a doublet band. It is known that the *B. stearothermophilus* Lipoyl domain can be partially lipoylated when expressed in *E. coli* cells [Dardel et al., 1990], possibly explaining the doublet. Peak fractions were concentrated using a 10 kDa MWCO centrifugal spin filter to ~3 ml and further purified by size exclusion chromatography (SEC) using an S75 16/60 column (GE Healthcare). The elution volume of ~0.41 CV was noted to be significantly earlier than expected for a fairly globular protein with a mass of ~28 kDa. (According to EMBL*, a 25 kDa globular protein elutes ~0.67 CV, and proteins eluting between ~0.42-0.46 CV have masses ranging from 150-670 kDa). The tandem Lipoyl domains of this construct create a "dumbbell" (**Figure 5.2.1.i, B**). Although this would likely increase the hydrodynamic radius compared with a more

* http://www.embl.de/pepcore/pepcore_services/protein_purification/chromatography/hiload16-60_superdex75/

uniformly shaped 28 kDa protein, this seems unlikely to account for this unexpectedly large apparent mass, suggesting that di-Lipoyl-Ana2_CC was likely oligomeric in solution.

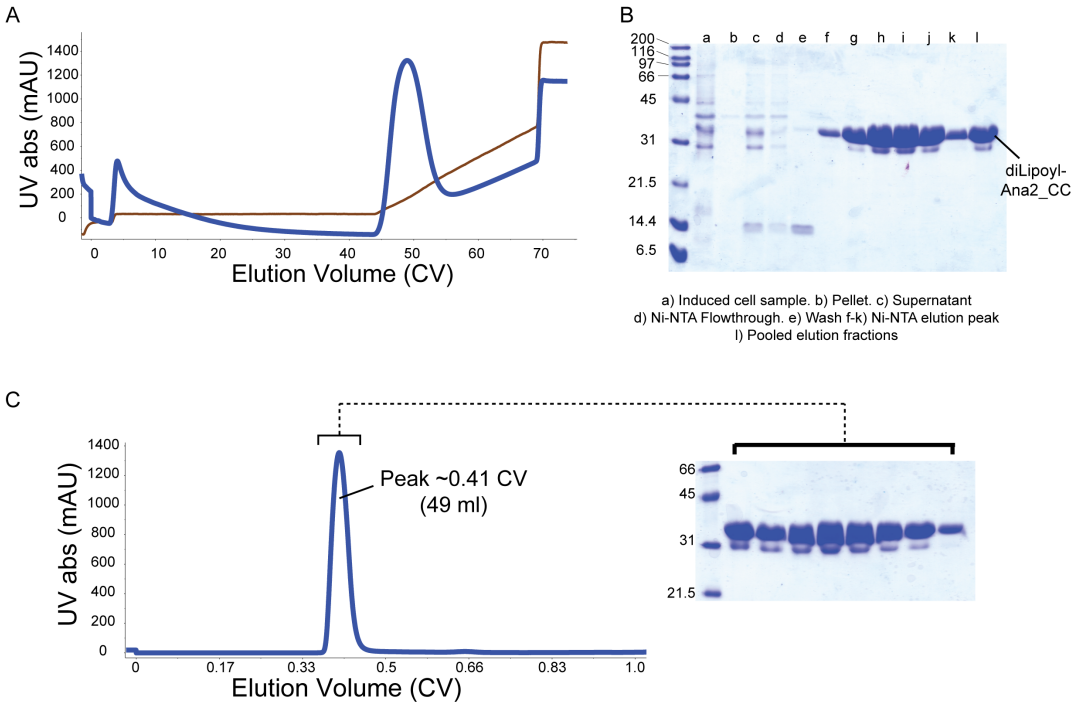


Figure 5.2.2.i: Purification of diLipoyl-Ana2_CC. (**A**) Elution of diLipoyl-Ana2_CC from Ni-NTA, using a linear imidazole gradient after a 40 CV 50 mM imidazole wash. Blue = UV absorption, brown = conductivity, effectively showing the concentration change of imidazole. (**B**) SDS-PAGE gel showing purification fractions. Elution fractions of diLipoyl-Ana2_CC appear fairly pure, although the protein appears to run as two bands, and slightly higher than its calculated mass. (**C**) SEC purification trace, blue = UV absorption, and SDS-PAGE analysis of peak fractions. Despite the single peak, the lower band can be seen to co-migrate. The doublet may be due to lipoylation of the Lipoyl domains [Dardel et al., 1990].

5.2.3: DiLipoyl-Ana2_CC is tetrameric in solution

In vivo structure/function analysis had suggested that the Ana2 coiled-coil might allow dimerisation (**Chapter 4**), and the purification suggested that diLipoyl-Ana2_CC was indeed oligomeric. SEC-MALS analysis was therefore carried out on diLipoyl-Ana2_CC in order to determine the stoichiometry of this construct over a range of concentrations (**Figure 5.2.3.i**). Samples were injected over an S200 10/30 column (GE Healthcare) using purification WB as the running buffer. Surprisingly, the MALS analysis revealed that the diLipoyl-Ana2_CC existed as a

stable species with a mass ranging from ~100-110 kDa, across an ~18-fold concentration range. This strongly suggests that this protein is behaving as a stable tetramer (the predicted mass of a diLipoyl-Ana2_CC tetramer is 111 kDa) As the Lipoyl domain is known to be monomeric [Dardel et al., 1993] the oligomeric activity of diLipoyl-Ana2_CC is most likely attributed to the Ana2_CC.

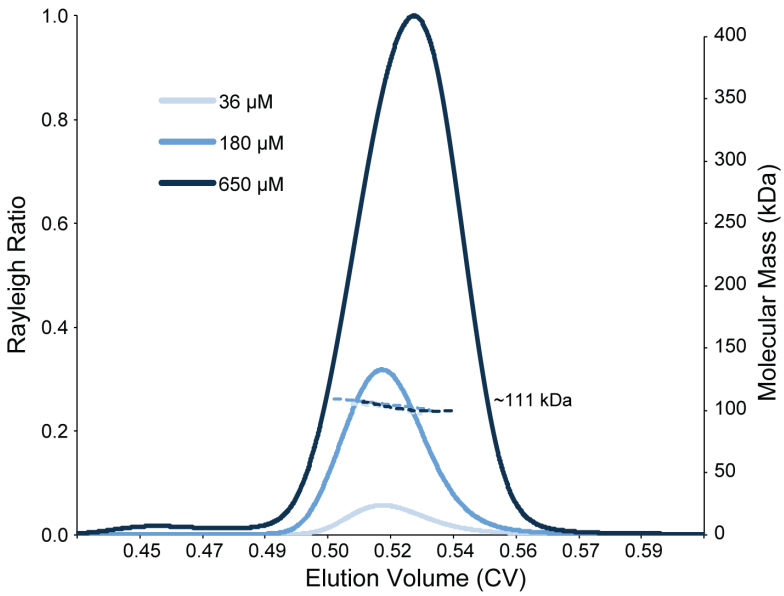


Figure 5.2.3.i: SEC-MALS analysis of diLipoyl-Ana2_CC (S200 1030 column). Solid lines show the relative Rayleigh ratio, dashed lines show the observed mass. Different concentrations are shown, coloured according to the legend. The calculated masses are consistent across all concentrations, and fall in the region of 100-110 kDa This suggests that diLipoyl-Ana2_CC forms a tetrameric species (calculated mass of 111 kDa).

5.2.4: Cleavage and Purification of Ana2_CC from the Lipoyl domains

Although it had been shown that the Ana2_CC could drive tetramerisation it was important to show whether it can do this in isolation, and whether steric hindrance from the Lipoyl domains may be limiting the extent to which it can oligomerise. I therefore devised a strategy to purify the Ana2_CC protein away from the Lipoyl domains.

Amino Acids	Number Cys	MW	pI	ε_{280}	Abs 0.1%
49	1	5636.3	5.13	1490	0.264

Table 5.2.4.i Details of the cleaved A2_CC protein calculated using Protparam.

Ni-NTA purified diLipoyl-Ana2_CC was cleaved by adding recombinant TEV protease at a ratio of 1 mg TEV : 20 mg fusion protein. This reaction (up to 40 ml) was dialysed against WB, with 3 mM DTT, using a 3 kDa MWCO dialysis membrane (Pierce). Initial incubation was at room temperature for 1-2 hours, followed by 4 ˚C overnight. Initial enrichment of Ana2_CC was

achieved by SEC (**Figure 5.2.4.i**). Tetrameric Ana2_CC would have a calculated mass of 22.5 kDa and would therefore be expected to elute between the TEV protein (~30 kDa) and the two Lipoyl domains (11.6 and 10.6 kDa, due to differences in attached linker sequence). Most important was the separation of the Lipoyl domains, which have a similar pI to Ana2_CC and would therefore be difficult to separate by ion-exchange chromatography (IEX). Optimal peak resolution for SEC was achieved using an S75 26/60 column and by concentrating the sample to ~2.5 ml (~25% of the maximum injection volume for this column). Separation in this manner was largely successful (**Figure 5.2.4.i, A**), however further polishing to remove residual TEV was achieved using IEX (**Figure 5.2.4.i, B**).

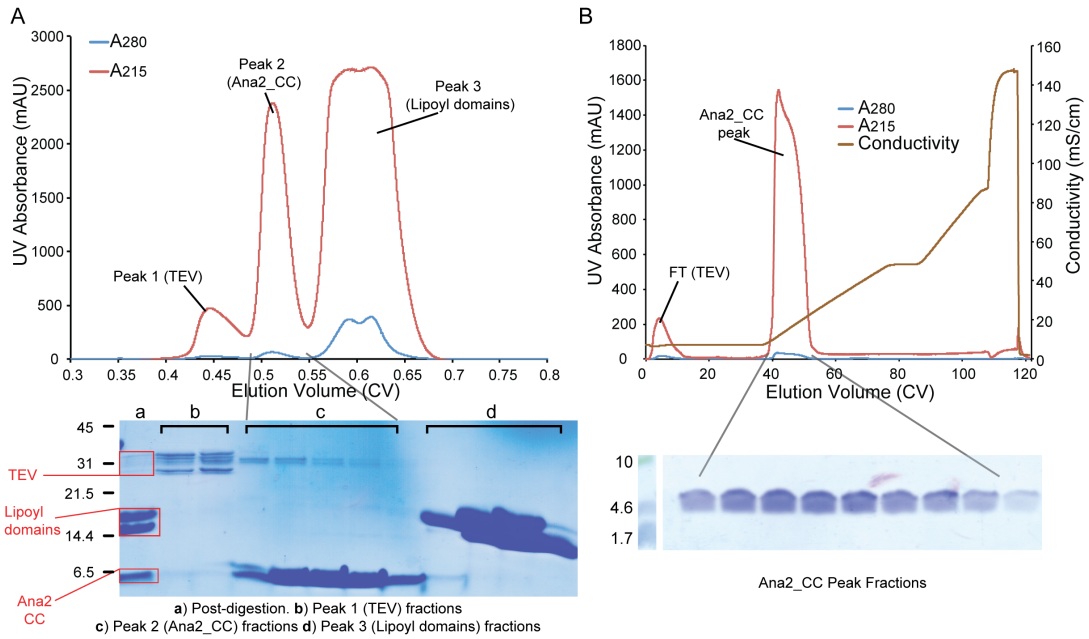


Figure 5.2.4.i: (**A**) SEC and 20% Tricine SDS-PAGE analysis of cleaved diLipoyl-Ana2_CC. Three main peaks were resolved on an S75 26/60 column, containing the TEV protease, Ana2_CC and the Lipoyl domains respectively, as indicated by SDS-PAGE analysis. Due to the relatively low ε_{280} of Ana2_CC, absorbance at both 280 (blue line) and 215 nm (red line) was used to trace protein elution. SDS-PAGE fraction "a" shows the sample post-cleavage. Red boxes indicate the bands for TEV, the Lipoyl domains, and Ana2_CC. Peak fractions are labelled "b, c and e". Ana2_CC fractions contain a small amount of residual TEV. (**B**) Anion exchange and 20% Tricine SDS-PAGE analysis of Ana2_CC. Again, both 280 and 215 nm were used to trace protein elution. TEV protease appears to flow through the column without binding, and was not detected in elution fractions. Ana2_CC eluted as a single broad peak. Despite its breadth, SDS-PAGE fractions of this peak showed that Ana2_CC appeared pure and homogenous.

Pooled SEC fractions were diluted until the sample was in a low ionic strength buffer (25 mM Tris pH 7.5, 50 mM NaCl). Anion exchange was carried out using two buffers, A (25 mM Tris pH 7.5, 1 mM DTT) and B (buffer A plus 2 M NaCl). A MonoQ 5/50 GL column (GE Healthcare) was equilibrated to 50 mM NaCl for 20 CV before loading the protein sample. TEV protease flowed through the column, which was then washed for a further 20 CV at 50 mM. A linear gradient of NaCl from 50-500 mM over 40 CV was then used to elute the column. Ana2_CC eluted as a broad peak with a mid-point of ~150 mM NaCl. No further protein eluted at higher ionic strength (**Figure 5.2.4.i, B**).

5.2.5: Isolated Ana2_CC forms a tight tetramer in solution

To determine the oligomeric state of purified Ana2_CC, SEC-MALS analysis was carried out (**Figure 5.2.5.i**). Ana2_CC was concentrated in a 10 kDa MWCO spin concentrator using A_{280} to assess concentration. A dilution series was prepared at equimolar concentrations to the diLipoyl-Ana2_CC analysed in **Section 5.2.3,** including a further concentration of 900 µM. Samples were injected over an S200 10/30 column (GE Healthcare) equilibrated in WB. Across all concentrations, the observed average mass was consistent with the calculated tetrameric mass of 22.5 kDa. During SEC, the sample is diluted as it diffuses over the column, and under these experimental conditions, the volumes of the peaks were ~10-fold greater than the volume injected. As the lowest analysed concentration was 36 µM, at this point I can only conclude that Ana2_CC forms a tetramer with a likely Kd below ~4 µM.

Further insight into the affinity of the interaction was gleaned via Electrospray-Ionisation Mass Spectrometry (ESI-MS, kindly performed by Dr David Staunton), which I carried out simply to confirm the identity of the Ana2_CC fragment. Preparation of samples for electrospray ionisation is a relatively harsh and usually denaturing process involving organic solvents,

acidification, dehydration and heat. It was remarkable therefore, to still observe signal for dimer, trimer and tetramer as well as for monomer (**Figure 5.2.5.i, B**).

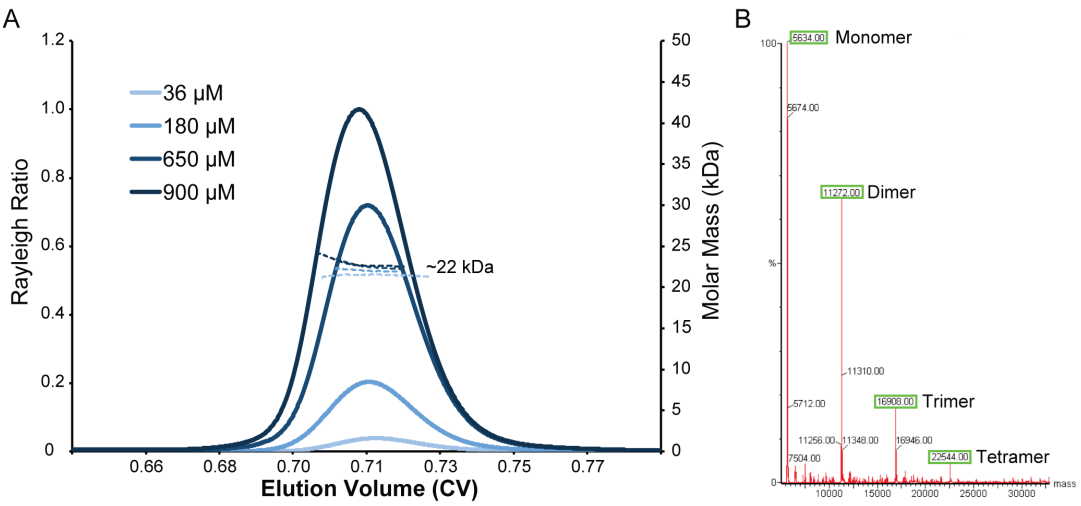


Figure 5.2.5.i: (**A**) SEC-MALS analysis of isolated Ana2_CC at various concentrations. Solid lines show the relative Rayleigh ratio, and dashed lines show the observed mass. Data for different concentrations are shown, coloured according to the legend. Ana2_CC eluted in a single peak for all tested concentrations. The average mass was consistent with a stable tetramer. (**B**) ESI-MS spectrum of Ana2_CC. Monomeric, dimeric, trimeric and tetrameric Ana2_CC are observed, with respective masses highlighted in green.

5.2.6: Ana2_CC is alpha-helical

In order to confirm the predicted secondary structure of Ana2_CC, a circular dichroism experiment was carried out (**Figure 5.2.6.i**). Ana2_CC was diluted to ~250 µg/ml and dialysed overnight at 4 ˚C against 10 mM $Na_xH_xPO_4$, pH 7.5. Initial spectra were averaged over four accumulations at room temperature between 190 and 260 nm, and normalised against a blank measurement taken with dialysis buffer. The spectrum clearly followed the characteristic features of an alpha-helical protein, showing positive signal between 190-200 nm, and negative signal between 200-260 nm, including a double dip at 208 and 222 nm. To assess the stability of this structure, a melting analysis was then performed by repeating experiments at 1 ˚C temperature increments between 10-80 ˚C, scanning between 200-240 nm. This showed that towards higher temperature, the alpha-helical characteristics were shifted and at 80 ˚C the

spectrum represented a mixture between alpha-helix and random-coil (**Figure 5.2.6.i, B**). Plotting the circular dichroism against temperature for readings at 222 nm (**Figure 5.2.6.i, C**), it was possible to see that there was little cooperativity in this transition, suggesting there were no sudden unfolding events. CD vs temperature plots at other wavelengths produced similar results. Interestingly the transition appeared to plateau after ~60 ˚C, and CD spectra from 60-80 ˚C were largely similar suggesting Ana2_CC had reached the extent of its thermal denaturation by ~60 ˚C. The sample was allowed to cool back to room temperature, and another reading was performed. The trace closely matched the pre-melt experiment, despite a small loss in signal presumably due to a small loss of material (**Figure 5.2.6.i, A**). These data suggest that the Ana2_CC structure is alpha-helical and is remarkably stable and may in fact not fully unfold.

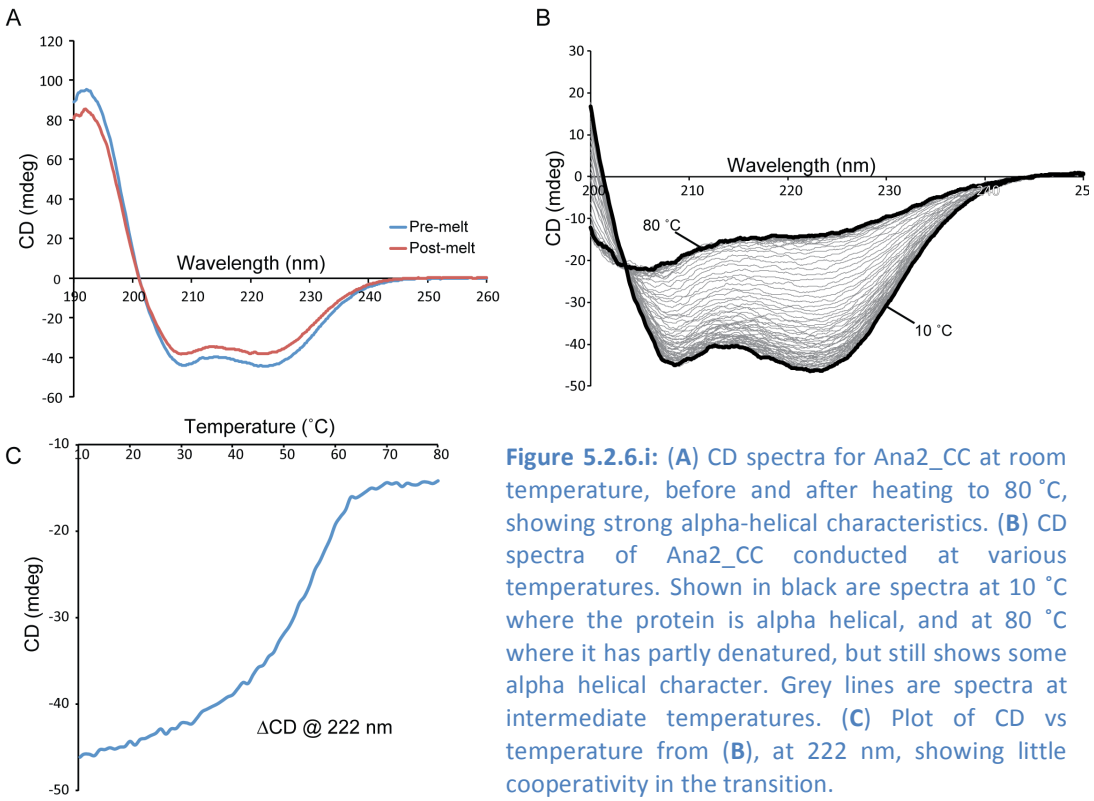


Figure 5.2.6.i: (**A**) CD spectra for Ana2_CC at room temperature, before and after heating to 80 ˚C, showing strong alpha-helical characteristics. (**B**) CD spectra of Ana2_CC conducted at various temperatures. Shown in black are spectra at 10 ˚C where the protein is alpha helical, and at 80 ˚C where it has partly denatured, but still shows some alpha helical character. Grey lines are spectra at intermediate temperatures. (**C**) Plot of CD vs temperature from (**B**), at 222 nm, showing little cooperativity in the transition.

5.2.7: Mutations within the CC domain modulate stoichiometry

Point mutations of Ana2_CC were described in **Chapter 4**, **Section 4.3.7** that were designed either to disrupt the oligomerisation of the CC (CC-A, CC-S, CC-D), or to target conserved and possibly solvent exposed residues within the predicted CC (CC-RQQE). Proteins containing the CC-A/S/D mutants all gave the same disrupted localisation phenotype in embryos as the ΔCC deletion (**Figures 4.3.7.ii** and **4.3.5.i** respectively), indicating that functionality of the CC had been abolished. The CC-RQQE mutant behaved similarly to wild-type suggesting these mutations retained essential CC functionality (at least in the presence of the WT protein) (**Figure 4.3.7.ii, D**). In order to assess how these mutations affected the ability of Ana2_CC to oligomerise, the mutant versions of Ana2_CC were cloned into the pLip vector. Large-scale recombinant expression and purification of diLipoyl-Ana2_CC mutants was carried out as for the wild-type protein, and a MALS experiment was performed. Wild-type diLipoyl-Ana2_CC was tetrameric even when harbouring the Lipoyl domains, and so analysis was performed on the diLipoyl fusion proteins (**Figure 5.2.7.i**). Samples were injected over an S200 10/30 column (GE Healthcare) equilibrated in WB. The mutant constructs were tested at 36 and 650 µM, the lower and upper concentrations previously tested for the wild-type protein. The CC-A, CC-S and CC-D mutants all eluted later compared to the wild-type protein, and their observed masses were consistent with their calculated monomeric masses. At 650 µM, CC-RQQE behaved similarly to the wild-type protein. It eluted at a similar position and with a similar profile, and its observed mass was consistent with being tetrameric. However at 36 µM, the peak eluted slightly later, and the observed average mass sloped steadily across the peak, ranging from ~92-56 kDa. The Kd of CC-RQQE appears to be weakened suggesting that one or more of the R208, Q211, Q213, or E219 residues may be involved in the self-association interface of Ana2_CC (see **Section 5.5.4**).

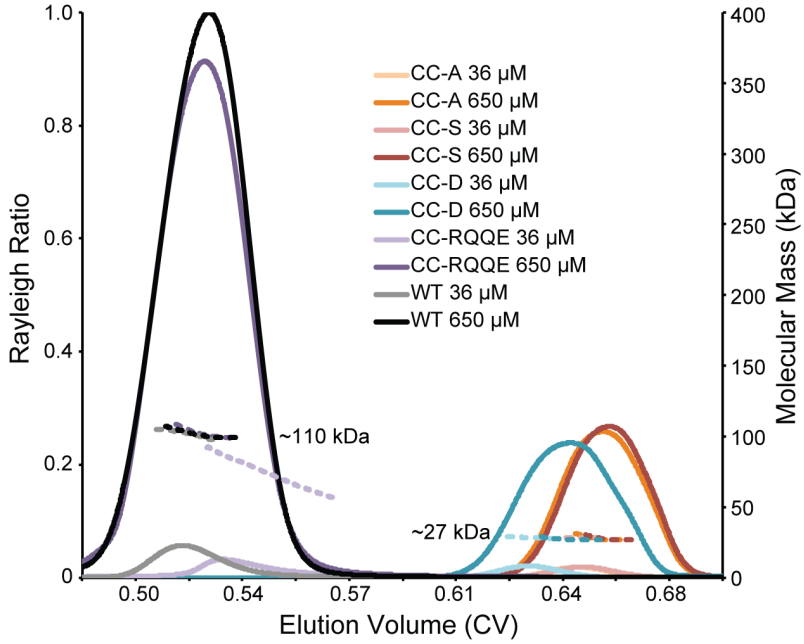


Figure 5.2.7.i: SEC-MALS analysis of diLipoyl-Ana2_CC variants compared to the WT experiment from **Section 5.2.3**. Solid lines show the relative Rayleigh ratio and dashed lines show the average observed mass, coloured as shown according to the legend. CC-A, CC-S and CC-D are monomeric at all concentrations. (~27 kDa is the average monomeric mass of the constructs) CC-RQQE is tetrameric at 650 µM, and has a range of stoichiometries at 36 µM. (110 kDa is the average tetrameric mass for all constructs)

5.2.8: Investigation into homologous systems

Ana2, STIL, and SAS-5 have all been implicated in cartwheel formation and are all predicted to contain a central CC domain that lies N-terminal to the STAN domain (**Figure 5.2.8.i**). I therefore wanted to investigate the oligomerisation behaviour of the CC domain from these Ana2 homologues, particularly as there is some evidence that suggests these species have different cartwheel structures [van Breugel et al., 2011, 2014; Hilbert et al., 2013; Kitagawa et al., 2011b]. I defined the central coiled-coil region from human HsSTIL (717-758), and *C. elegans* SAS-5 (131-173) based on sequence alignments, Uniprot database annotation, and analysis of the output from the COILs server (**Figure 5.2.8.i**). *E. coli*-optimised DNA sequences for each CC region were synthesised (Genewiz) and cloned into the pLip vector as for Ana2. diLipoyl fusions of the HsSTIL_CC and SAS-5_CC were expressed similarly to diLipoyl-Ana2_CC, but expression could only be obtained using C41 (DE3) cells. The proteins were then purified using Ni-NTA affinity,

and size exclusion chromatography. Because diLipoyl-Ana2_CC was able to oligomerise whilst still fused to the Lipoyl domains (**Section 5.2.3**), a MALS experiment was conducted on the diLipoyl fusions of STIL_CC and SAS-5_CC (**Figure 5.2.8.ii**).

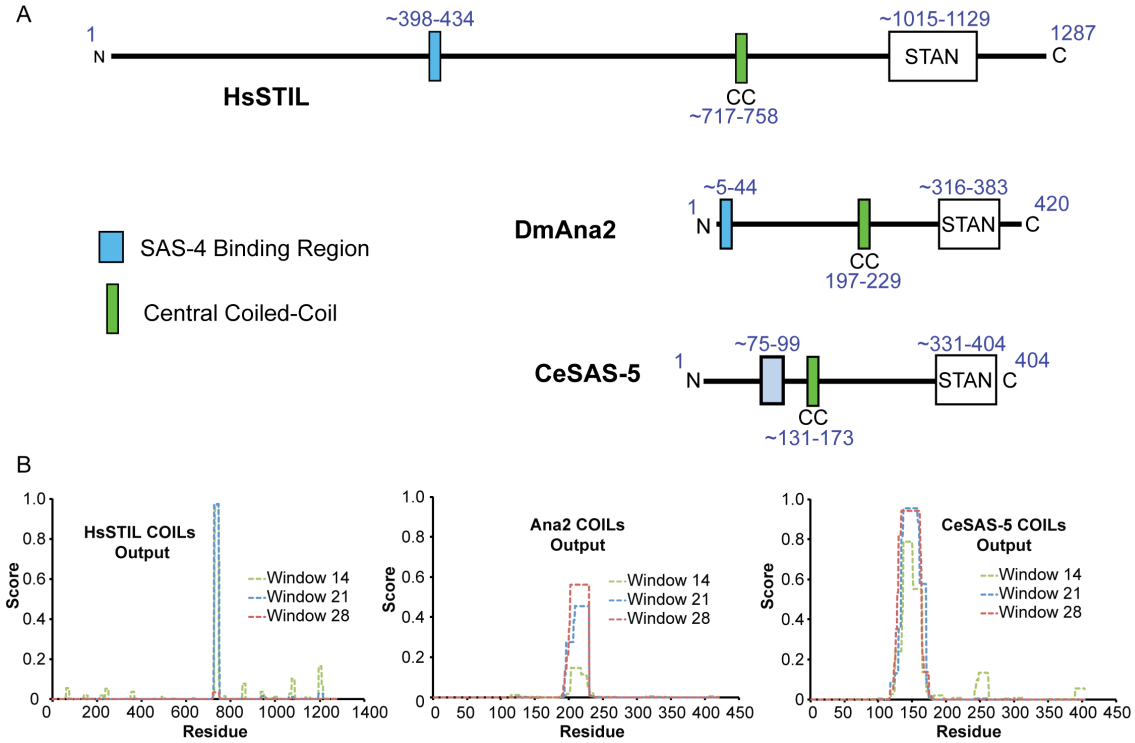


Figure 5.2.8.i: (**A**) Schematics of the STIL family of proteins, showing human STIL, *Drosophila* Ana2, and *C. elegans* SAS-5. The SAS-4 binding domain in SAS-5 is approximate, and based on an alignment given in [Hatzopolous et al., 2013]. (**B**) COILs server outputs showing the single predicted CC domain for each protein.

The proteins were concentrated to the same concentrations as those previously tested for diLipoyl-Ana2_CC. Samples were injected over an S200 10/30 column equilibrated in WB. Interestingly, at all concentrations, diLipoyl-SAS-5_CC eluted as a monomer, while the observed mass for diLipoyl-STIL_CC seemed to rise with each concentration tested – although it clearly did not adopt a stable tetrameric mass even at the highest concentration tested. The elution peaks of diLipoyl-STIL_CC had long tails, and the observed masses sloped, suggesting heterogeneity in the stoichiometry of the samples. At the higher concentration of 650 µM, the observed mass ranged from below to above trimeric. It is therefore possible that a tetramer could be the ultimate oligomer, but no firm conclusions can be drawn on this, as a stable species was not observed. It is also not possible to tell if the unstable oligomer was composed of a

tetramer/monomer mix, or if intermediate species were present. If STIL is tetrameric, it is curious that it has such a low self-affinity in comparison to the remarkably strong self-affinity exhibited by its orthologue Ana2. One possibility is that the design of these new constructs is not optimal for oligomerisation - there is very little sequence identity between the CC regions from the different species, making it difficult to precisely define the potential CC region. Either the termini of the constructs could be refined, or the Lipoyl domains cleaved off. Alternatively, perhaps in these other species, post-translational modifications are required for robust oligomerisation. Finally, it is possible that the Ana2 homologues simply do not form tetramers, and that this family of proteins can form different homo-oligomers (or doesn't form homo-oligomers at all) in other species. In summary, while it is intriguing that STIL and SAS-5 behave differently to Ana2, no firm conclusions can be drawn from these data without further investigation.

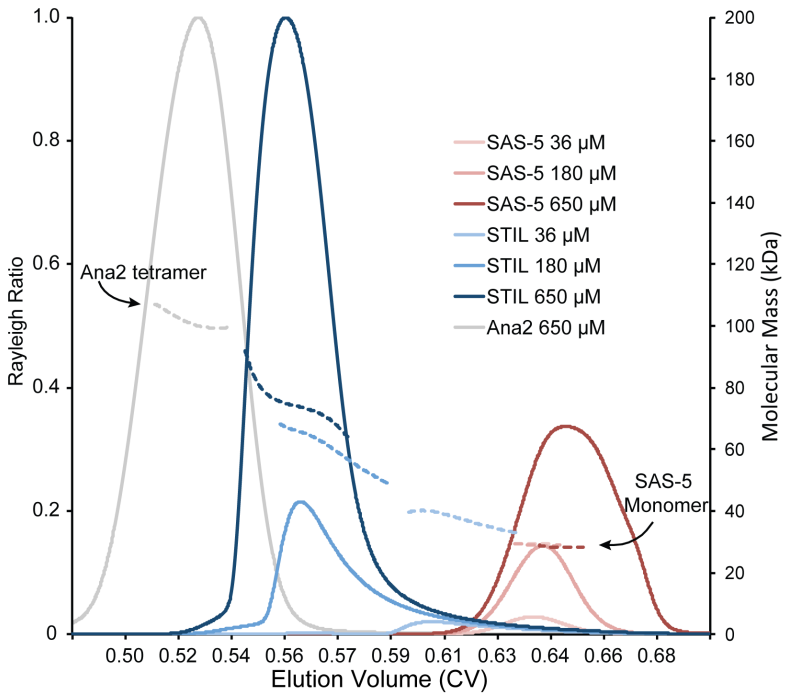


Figure 5.2.8.ii: MALS analysis of diLipoyl-STIL_CC and diLipoyl-SAS-5_CC. Solid lines show the relative Rayleigh ratio, and dashed lines show the observed average mass across each peak. Red lines show the data for SAS-5, and indicate a stable monomeric species. Blue lines show the data for STIL, which showed a range of stoichiometries. Shown for reference is the data for tetrameric diLipoyl-Ana2_CC in grey. Note that the monomeric mass for each construct is roughly equivalent.

5.3: Ana2_CC Crystallisation and Data Collection

The self-association of Ana2 is mediated by the central CC domain and is required for efficient localisation to centrioles *in vivo*. I therefore wanted to determine the structure of the CC domain at atomic resolution. Understanding the molecular determinants of this self-interaction will be important for rationalising Ana2's function *in vivo.* In particular it is of great interest to find out whether the interaction is parallel or anti-parallel, as this will have implications for the positioning of Ana2's N- and C-termini, and the subsequent arrangement of Ana2 binding partners, within the putative cartwheel structure. The structure may also rationalise the effect of Ana2 CC mutations and shed light on whether the CC binds to other proteins, or whether it functions solely as a homo-oligomerisation domain.

5.3.1: Crystallisation of Ana2_CC

Following ion-exchange chromatography, the Ana2_CC protein was highly purified, and behaved as a homogenous tetramer in solution, suggesting that this sample might be suitable for crystallisation studies. The sample was diluted into a crystallisation buffer of 20 mM Tris, 150 mM NaCl , 1 mM DTT, and then concentrated in a 10 kDa MWCO spin concentrator. The protein was highly soluble, and was concentrated to ~42 mg/ml (as judged by the A_{280}). Crystal trials were conducted using the sitting drop approach using 0.2 µl drops, set up with a Mosquito LCP robot (TTPLabtech) using various commercial screens. Phase separation and minor precipitative activity observed using the STURA/Macrosol screen suggested that 42 mg/ml was a suitable concentration for crystallisation. Remarkably, rod-like crystalline growth was observed within 10 minutes using the Morpheus screen, in conditions containing MPD, PEG 1K and PEG 3350. Other crystals grew overnight in the STURA/Macrosol screen, in 0.1M Na HEPES pH 7.5, 45% PEG 600 (**Figure 5.3.1.i**). In order to grow more crystals, and in anticipation of the possible need to grow Semet derivative crystals, an optimisation screen was set up based on the

STURA/Macrosol condition using a Dragonfly optimisation robot (TTPLabtech) (For exact details of crystallisation conditions, and the optimisation screen, see **Table 5.3.1.i** and **Appendix 7.9** respectively). Typically, crystals grew to their maximal size within 24-48 hours, and were fished after 24 hrs - 2 weeks. The presence of MPD or small PEGs at high concentration allowed all crystals to be cryoprotected by their own mother liquor. Crystals from which data were collected are denoted 1-5 (**Table 5.3.1.i**).

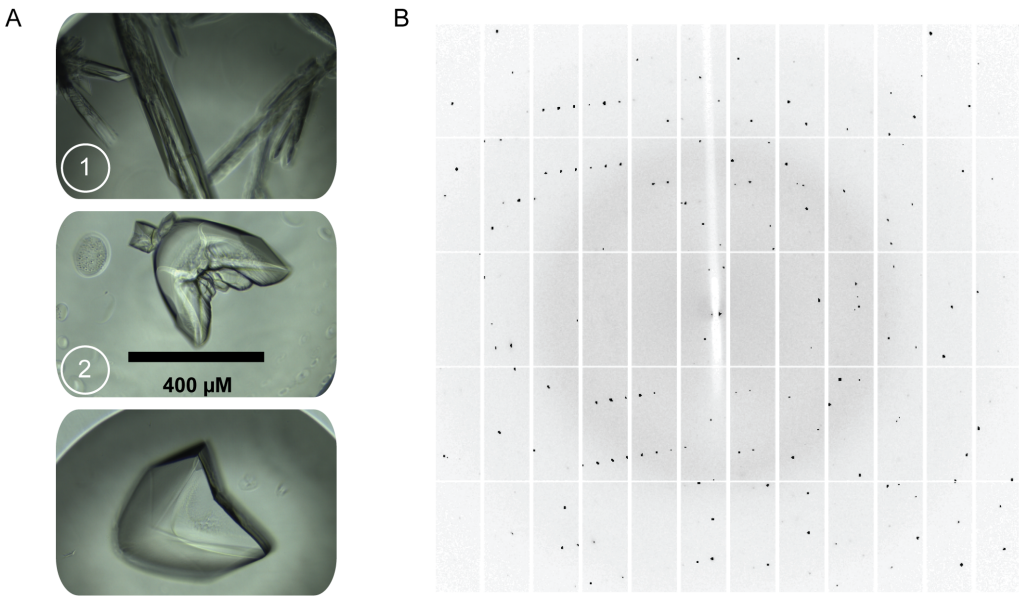


Figure 5.3.1.i: (**A**) Examples of Ana2_CC crystals. The top two panels show drops from which crystals 1 and 2 (**Table 5.3.1.i**) were harvested. All images are at the same scale. (**B**) Example diffraction image from the low-resolution pass of crystal 5 (**Table 5.3.2.i**). The detector was set to a 2.00 Å limit.

Crystal	Protein concentration (mg/ml)	Mother liquor	µl protein:µl Mother liquor	Temperature (˚C)
(1) 141_B8a_1	41.8	100 mM Na Hepes/MES mix, pH 7.5, 30 mM NaFl, 30 mM NaBr 30 mM NaI, 12.5 % MPD, 12.5% PEG 1000, 12.5% PEG 3350	0.15:0.05	21
(2) 143_E9a_1	42.7	100 mM Hepes (57.1 mM @ pH 7.2, 42.9 mM @ pH 8.2) 44% PEG 600	0.16:0.04	21
(3) 138_C8a_1	41.8	100 mM Na Hepes pH 7.5 45% PEG 600	0.15:0.05	21
(4) 143_F8a_1	42.7	100 mM Hepes (71.4 mM @ pH 7.2, 28.6 mM @ pH 8.2) 42% PEG 600	0.16:0.04	21
(5) 143_F8a_4	42.7	100 mM Hepes (71.4 mM @ pH 7.2, 28.6 mM @ pH 8.2) 42% PEG 600	0.16:0.04	21

Table **5.3.1.i:** Crystallisation conditions of Ana2_CC for crystals used in this study.

5.3.2: Data Collection and Processing of Ana2_CC

Data were initially collected at Diamond Light Source, I04-1. Despite crystal 1 having a visually different morphology to other crystals, all crystals indexed into the four-fold symmetric spacegroup I4, with similar cell dimensions, suggesting they were in fact the same crystal form. The four-fold symmetry of the spacegroup was intriguing considering that Ana2_CC forms a tetramer. Data were processed using Xia2 [Winter, 2010] in the 3daii mode, using XDS [Kabsch, 2010] to identify, index, and integrate spots, and AIMLESS [Evans and Murshudov, 2013] for scaling and merging. This produced datasets A2CCwt_1 and A2CCwt_2 as detailed in **Table 5.3.2.i**. The high-resolution shell data for A2CCwt_2 (1.09-1.07 Å) were still reliable, as shown from the CC½ of 0.965, and I/σI of 69.9, however, despite the high multiplicity of the dataset, this shell was only complete to 20.5%, as the detector could not move close enough to the crystal to efficiently collect higher resolution reflections.

Dataset	A2CCwt_1 (1.5Å)	A2CCwt_2 (1.07Å)	A2CCwt_3 (0.8Å)	A2CCwt_4 (Anom)
Crystal(s)	1	2/3(2sweep)	5(2sweep)	2/3(2sweep)/4
Beamline	Diamond I04-1	Diamond I04-1	Diamond I03	Diamond I04-1
Spacegroup	I4	I4	I4	I4
Wavelength	0.9200	0.9200	0.7293	0.9200
Unit cell dimensions (Å/°)	a = 33.79 b = 33.79 c = 74.47 α = 90.00 β = 90.00 γ = 90.00	a = 33.41 b = 33.41 c = 74.54 α = 90.00 β = 90.00 γ = 90.00	a = 33.27 b = 33.27 c = 74.28 α = 90.00 β = 90.00 γ = 90.00	a = 33.37 b = 33.37 c = 74.49 α = 90.00 β = 90.00 γ = 90.00
Resolution (Å)	23.89–1.50	30.49–1.07	30.36–0.80	37.24–1.06
Completeness (overall/inner/outer)	98.6/96.2/89.4	89.3/99.5/20.5	97.8/99.9/79.2	92.5/99.9/38.7
R_{merge} (overall/inner/outer)	0.064/0.060/0.551	0.214/0.360/0.161	0.038/0.034/0.547	0.092/0.144/0.116*
R_{pim} (overall/inner/outer)	0.027/0.023/0.370	0.048/0.092/0.095	0.013/0.011/0.312	0.019/0.031/0.089*
CC (½) (overall/inner/outer)	0.999/0.999/0.706	0.986/0.956/0.965	1.000/0.999/0.782	0.998/0.987/0.871
I/σI (overall/inner/outer)	18.0/39.9/2.2	378.4/498.2/69.9	21.5/71.3/2.2	55.0/80.3/9.8
Multiplicity (overall/inner/outer)	7.7/8.6/4.0	18.3/18.8/3.0	5.8/11.2/3.3	41.2/50.9/4.0
No. unique reflections	6,627	16,020	41,451	17,080
Anom Completeness	-	-	-	90.8/99.9/26.2
Anom Multiplicity	-	-	-	20.3/26.2/1.7
Anom Correlation	-	-	-	0.485/0.840/0.705**

Table 5.3.2.i: Processing statistics for Ana2_CC datasets used in this study. *Note that for the anomalous dataset, due to Bijvoet differences being searched for, R_{merge} and R_{pim} are presented for reflections within I+/I-. **Note that the high anom correlation in the outer shell is likely artefactual due to the incompleteness and low multiplicity measured for this shell. In reality anomalous correlations were fairly low beyond 3.0 Å.

A higher-resolution dataset to 0.8-Å (A2CC-wt_3, **Table 5.3.2.i**), was obtained by shooting crystal 5 at the tuneable beamline, Diamond I03. By using a shorter X-ray wavelength, diffraction occurred at smaller angles with respect to the incidence ray, allowing more spots to be collected by the detector. As well as a high-resolution sweep, a low-resolution sweep was also included to optimise the measurement of low-resolution reflections. This dataset was similarly processed using the XIA2 pipeline. One caveat of reducing the wavelength was an approximate four-fold reduction in photon flux compared to the standard I03 working wavelength (0.9763 Å). This resulted in relatively low intensity for the data, and, coupled with an outer shell CC½ of 0.782, it is likely that these crystals may well diffract useful data to a higher resolution than 0.80 Å. It is highly unusual for crystals of biological macromolecules to diffract to this resolution, and so a more optimal experimental setup might have been achieved using a beamline designed for small molecule crystallography. The spacegroup assignment of I4 was confident, as evident by the particularly low merging statistics, even for the highest-resolution shells (**Table 5.3.2.i**). To assess the possible number of Ana2_CC molecules in the asymmetric unit, the Matthews Coefficient [Matthews, 1968] was calculated, using a cell volume of 85027 $Å^3$ and a monomer mass of 5636 Da. This suggested it was only possible to have one copy of the Ana2_CC monomer in the asymmetric unit of the crystal, and even this would give a relatively low solvent content of 34.82%.

5.4: Solving The Crystal Structure of the Ana2 Central Coiled Coil Domain

Despite the lack of an obvious homology model, the Ana2_CC fragment was predicted to be a short helix of around 32 residues. It therefore seemed possible that a molecular replacement solution might be achieved using a simple helical model as there are fairly limited spatial arrangements that such a short fragment could take.

5.4.1: A Molecular Replacement Solution for the Ana2_CC structure

Analysis of 4-helical bundles in the PDB revealed that most had a left-handed supercoil, whether parallel or antiparallel. Left-handed supercoiling of helices is relatively common as it allows the natural 3.6 residue pitch of a helix to essentially adopt a 3.5 residue pitch with respect to the superhelical axis [Lupas and Gruber, 2005]. This ensures that particular positions in the helical heptad repeat can face in a consistent direction with respect to the supercoil axis. Right-handed four-helix bundles are more rare, and are usually not symmetrically arranged. Rather than coiling around each other, these helices are short, and simply cross each other in a way that could be described as having a right-handed topology. Due to these observations it was assumed Ana2_CC would form a left-handed supercoil, and so several helices from left handed four-helix bundles were trimmed to ~30 residues and used as MR search models, either with their native sequence, or mutated to polyalanine.

A successful model was found in PDB entry 1UO4 [Yadav et al., 2005], which is a GCN4 based, engineered four-helix antiparallel bundle with a left-handed supercoil. Molecular replacement was carried out in MOLREP [Vagin and Teplyakov, 2010] using chain A (residues 2-31) of 1UO4, and dataset A2CCwt_1. The top solution was obtained with a contrast score of 2.21. The solution did not produce any clashes, and the helix coiled around the four-fold symmetry axis to generate a parallel, left-handed tetrameric bundle. Solutions with similar packing were observed for

several MOLREP jobs using different search models, however this solution had the most favourable scores and was therefore pursued. The MOLREP score was by no means unambiguous, however, the search model was not an obvious homologue and so particularly high scores were not expected. To reduce bias from the inclusion of non-related sidechains, the placed search model was mutated to polyAla, and then refined against dataset A2CCwt_1 using REFMAC5 [Murshudov et al., 2011]. This produced a drop in the R_{work} and R_{free} values, and resulted in un-modelled helical density at the N terminus of the helix as well as density for sidechains, indicating that a valid solution had been found (**Figure 5.4.1.i**). Further alanine residues were modelled into the unfilled N-terminal helical density, and this new model was refined yielding an improved map with more obvious density for the missing sidechains. This allowed the register to be assigned, and so the polyAla model was mutated to the Ana2_CC sequence. After a further round of refinement, autobuilding using BUCCANEER [Cowtan, 2006] was carried out in order to reduce model bias.

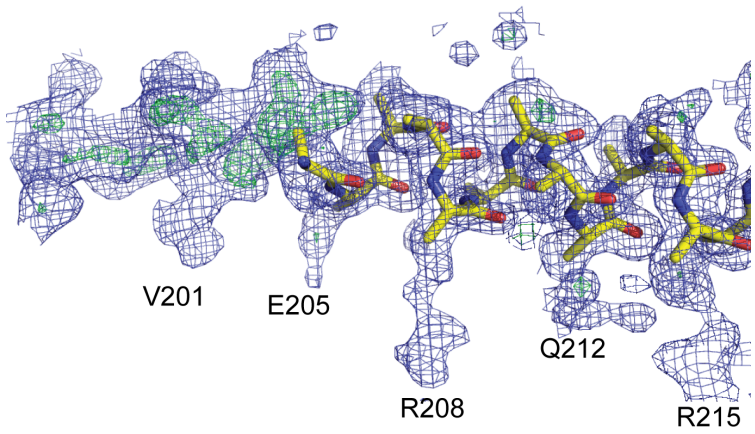


Figure 5.4.1.i: View of maps resulting from refinement of the initial MR solution, once it had been trimmed to polyAla. Blue density is the 2Fo-Fc map contoured at 1.0σ. Green density shows the Fo-Fc map at +3.0σ. Empty density (with positive residuals) is evident at the N-terminus. Pockets of empty density are also evident for the side chains (labelled), which should fill this space. Shown in yellow sticks is the polyAla model itself.

5.4.2 Refinement of Ana2_CC

Before refinement in the highest resolution dataset, the model was first refined in the A2CCwt_2 dataset (**Table 5.3.2.i**). This dataset was alternatively indexed compared to A2CCwt_1 (h, k, -l), requiring initial molecular replacement of the model into the new data. This now produced a

very unambiguous contrast score of 6.94, confirming that a correct solution had been developed. At this point in the study, the ultra high-resolution dataset, A2CCwt_3 was obtained (**Section 5.3.3**), and was indexed using A2CCwt_2 as a reference. Iterative refinement was continued with the A2CCwt_3 dataset, using Phenix.refine [Afonine et al., 2012], and REFMAC5 [Murshudov et al., 2011] with manual model building carried out in Coot [Emsley and Cowtan, 2004]. The resultant density was very detailed with electron clouds of individual nuclei clearly resolved (**Figure 5.4.2.i A**). To model this detail most accurately, the B-factors of all non-hydrogen atoms were modelled anisotropically. In many places, split conformations of side chains were obvious (**Figure 5.4.2.i B**). To accommodate these conformations, it was sometimes easiest to also split the main chain to accommodate the necessary Cα deviation. Hydrogens were modelled in their riding positions, as the resolution was not quite high enough to model individual positions, although they did influence the shape of the density around methyl and CH_2 groups, and in some cases it was possible to see the influence of particularly strong hydrogen bonds (**Figure 5.4.2.i C**). In these cases density could be seen to protrude along the δ^+-H bond, suggesting that electrons were especially populous along such bonds due to the electronegativity of the acceptor. Refinement was continued until the model adequately explained the dataset, as judged by the R_{work}/R_{free} values, which converged at 10.9/12.2.

The model was also validated for chemical feasibility using Molprobity. The Molprobity analysis proved less valuable than at lower resolution, as it relies largely on comparison between similar resolution structures, of which there are few. (At the time of writing, of 91,243 X-ray diffraction structures in the PDB, only 47 are solved to a resolution of 0.80 Å or better). It is recommended that the Molprobity score is lower than the resolution of the structure, however this system apparently breaks down at high resolution, as the Ana2_CC score of 1.47 is significantly higher than the resolution of 0.8 Å. However Ana2_CC was still placed in the 68th percentile of structures in its category.

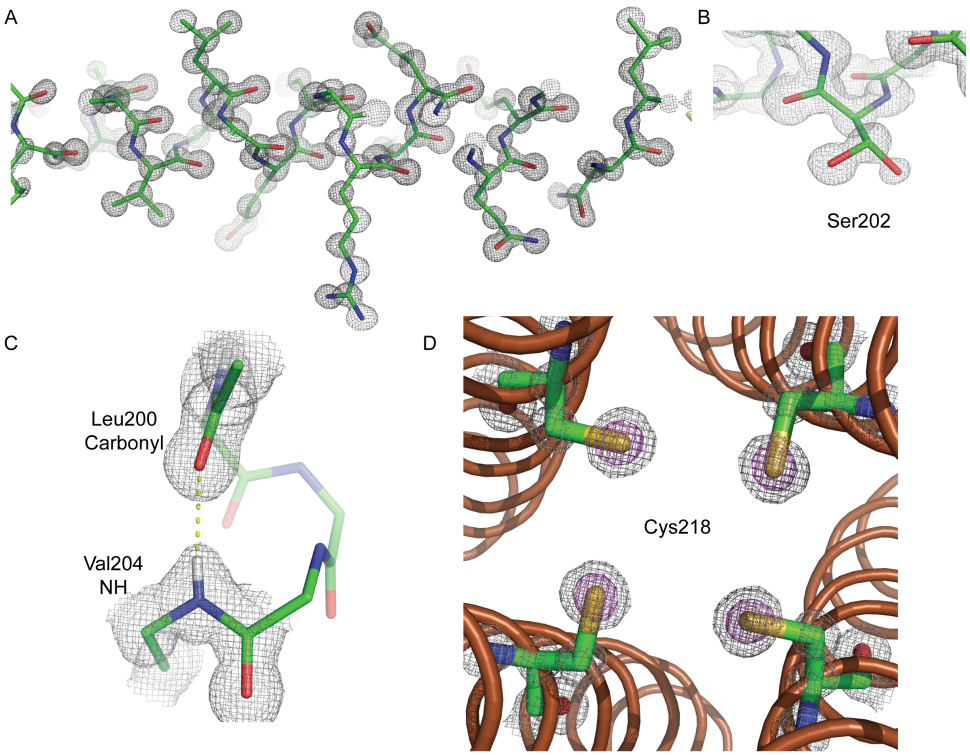


Figure 5.4.2.i: (**A**) Stick model of a region of the 0.8 Å Ana2_CC structure, showing representative electron density. The 2Fo-Fc map is shown in grey mesh, contoured at 1.8 σ and carved around protein atoms. (**B**) Close up view of serine 202, showing clear evidence of static disorder, modelled by fitting a dual conformation of the side chain. The 2Fo-Fc map is contoured at 1.0 σ. (**C**) Close-up view of an intra-helical hydrogen-bond between a mainchain carbonyl and amide. A dashed yellow line indicates the hydrogen bond. Electron enrichment across such bonds was visible as a hydrogen-oriented bump. The 2Fo-Fc map is contoured at 1.0σ, and selectively carved around the mainchain atoms in the foreground. (**D**) Close-up view showing anomalous signal located on the sulphur of cysteine 218. Shown is a view down the crystallographic four-fold *c* axis, around which the biologically relevant tetramer has been generated. The main chain is shown as a brown helical ribbon, and cysteine is shown in a green stick representation. The 2Fo-Fc map (calculated from the refinement of dataset A2CCwt_4(anom), 1.06 Å) is shown in grey mesh, carved around cysteine atoms, at 1.8 σ. The anomalous difference map (magenta mesh) is contoured at 4 σ and carved at 10 Å from the cysteines.

During refinement, an attempt was made to identify possible ordered chloride ions using anomalous signal. Chlorine and sulphur can be used to provide anomalous signal at certain wavelengths, however the longest wavelength for which I had data was 0.9200 Å, which would only provide extremely low anomalous signal. In summary, I was able to detect and process anomalous signal, and produce a phased anomalous difference map (the processing of this anomalous signal is described in detail in **Appendix 7.10**). However, the only significant peaks in this map corresponded to an ordered cysteine sulphur atom (**Figure 5.4.2.i, D**). This is

remarkable considering that at 0.9200 Å, the anomalous component, (f'') for sulphur is only 0.21 and is similar for chlorine (as calculated by CROSSEC [Cromer, 1983]). Thus, no chloride ions could be identified. Final refinement and validation statistics are presented in **Table 5.4.2.i**.

	A2CCwt
Resolution Range (Å)	30.36-0.80 (0.82-0.80)
R_{work}/R_{free} /% size test set	10.9/12.2/5.06% (21.0/20.2/4.84%)
No. Reflections working set	39346 (2338)
No. Reflections Test Set	2100 (119)
Number of atoms (non-H)	499
Waters	53
rmsd from ideal values: bond length (Å)/angles(˚)	0.022/1.982
Average B-factor	9.72
Correlation coefficient Fo-Fc/Fo-Fc free	0.985/0.980

Table 5.4.2.i: Processing and Refinement statistics for the Ana2_CC structure. Values in parentheses are for the highest resolution shell.

5.5: Analysis of the Ana2_CC structure

5.5.1: Ana2_CC is a parallel, symmetrical tetramer.

The Ana2_CC tetramer crystallised in the body-centred spacegroup I4, such that the tetramer CC axis aligned with the crystallographic four-fold axis. The asymmetric unit (ASU) therefore comprises a single monomer, and so, crystallographically speaking, the tetramer is built by a four-fold symmetry operation around the c axis. The lattice is built by the offset packing of tetramers (**Figure 5.5.1.i, A-C**), resulting in 8 copies of the asymmetric unit being present in each unit cell (**Figure 5.5.1.i, A-C**). Interfaces present in the crystal were analysed by the PISA server. The most significant interface present was the one between adjacent monomers in the tetramer, with a Complex Significance Score (CSS) of 0.954 and a predicted ΔG upon formation of -16.5 kcal/mol. Because the tetramer is rotationally symmetric, each monomer forms this interface on either side with neighbouring monomers. The only other interface that scored involved minor contacts between opposite monomers in the tetramer. Each monomer therefore makes contacts with all three other monomers in the tetramer. PISA predicted the resultant complex to be stable in solution, requiring an input of 63 kcal/mol to dissociate, and a solvation free energy gain of -91.6 kcal/mol. Each monomer buries 1944 $Å^2$ upon complex formation, which is 43% of the total solvent accessible area. Other contacts involved the packing of the tetramer into the lattice, and did not have a significant CSS.

It should be noted that the TEV protease recognition site, EFGENLYFQ was present at the C-terminus of the construct, and was ordered in the crystal. Although this sequence also adopted a helical conformation, the secondary structure assignment algorithm, DSSP [Kabsch and Sander, 1983], did not judge it to satisfy conventional alpha-helical geometry (**Figure 5.5.1.i, D**). This artefactual sequence is unlikely to significantly contribute to complex formation.

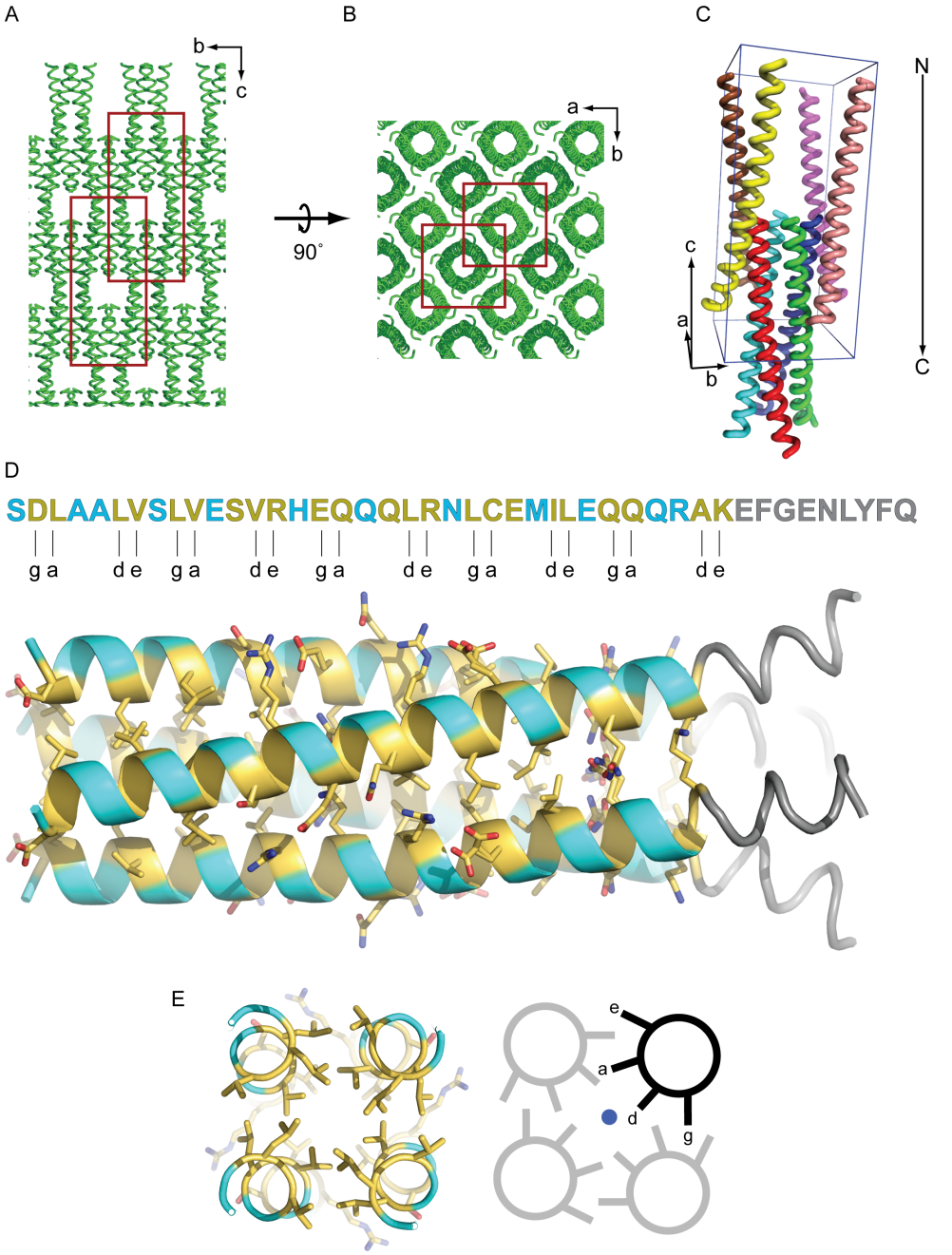


Figure 5.5.1.i: (**A, B**) Views of the Ana2_CC crystal lattice, along the *a* axis (**A**), and along the *c* axis (**B**). The Ana2_CC is shown in green ribbon format, and the unit cell is shown as a red box, twice in each view to show how in the I4 spacegroup, the unit cell overlaps with each adjacent cell due to the offset packing of the tetramer. (**C**) The contents of the unit cell, shown as a blue box, with the 8 copies of the ASU shown in different colours. (**D**) The Ana2_CC crystallographic tetramer in cartoon format, with 2˚ structure assigned by DSSP. Yellow residues in stick format are those judged to be ≥30% buried by PISA, while others are shown in blue. The TEV cleavage site is shown in grey. Shown above is the sequence, with the heptad register as assigned by the CC analysis program SOCKET. The tetrameric association is governed by the interactions between regularly spaced **g**, **a**, **d** and **e** residues, which were invariably ≥30% buried. (**E**) Left, a top down view (N-C) of the tetramer in ribbon format with **g/a/d/e** residues coloured in yellow. It can be seen that these residues face inwards with respect to the CC axis. A schematic of this is shown on the right, with the CC axis as a blue dot. At all points along the CC, the **g/a/d/e** residues have this distinct packing pattern.

First, its presence is not known to drive tetramerisation in other contexts. Second, the tetramer helices can be seen to splay apart in the crystal once the Ana2 sequence ends and the TEV site begins (**Figure 5.5.1.ii**). The distance was measured between Cα atoms of "e" position residues of the heptad repeat, across the diameter of the tetramer. For the "Y" residue of EFGENL**YF**Q, the Cα atoms were 17.3 Å apart, whereas the other "**e**" position residue Cαs of Ana2 were an average of 13.1±0.4 Å apart. Third, the EFGENLYFQ was deleted from the structure, and PISA analysis run again. The tetramer was still predicted to be a stable structure in solution, requiring 51 kcal/mol to dissociate and had a solvation free energy gain of -78.3 kcal/mol. Per residue, these energies are marginally higher than for the Ana2_CC + TEV site. The amount of buried surface area (1715.6 Å^2) was now 47% of each monomer. Subsequent analysis was therefore carried out with the non-biologically relevant TEV site omitted.

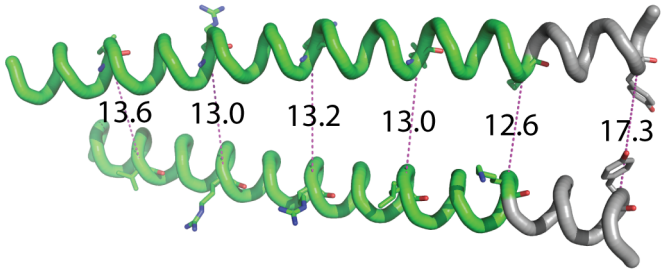


Figure 5.5.1.ii: Measurements (dashed lines) of the CC diameter along the helix. Measurements are between the Cα atoms of "**e**" position residues along the Ana2_CC, from opposite helices. Values given are in Å. The TEV cleavage site (shown in grey) does not drive tetramerisation and clearly splays away from the CC axis as demonstrated by the increased CC diameter. (The other two helices are omitted for clarity).

5.5.2: Ana2_CC tetramerisation is driven by polar, charged, and hydrophobic interactions

Upon the formation of the tetrameric interface, residues lying at the **g**, **a**, **d** and **e** positions of the heptad repeat were shown to be uniformly buried by PISA (**Figure 5.5.1.i, D**). Residues at position "**a**" in the tetrameric coiled coil frequently pack against the **d** and **g** residues of the neighbouring helix, and the **d** position residues against the **a** and **e** residues (**Figure 5.5.1.i, E**). This can lead to contexts where non-polar residues exhibit a hydrophobic knob-into-hole packing

arrangement. This was seen in the case of residues 200, 201, 203 and 204 (**Figure 5.5.2.i A**). However, the core of the tetramer is not entirely hydrophobic.

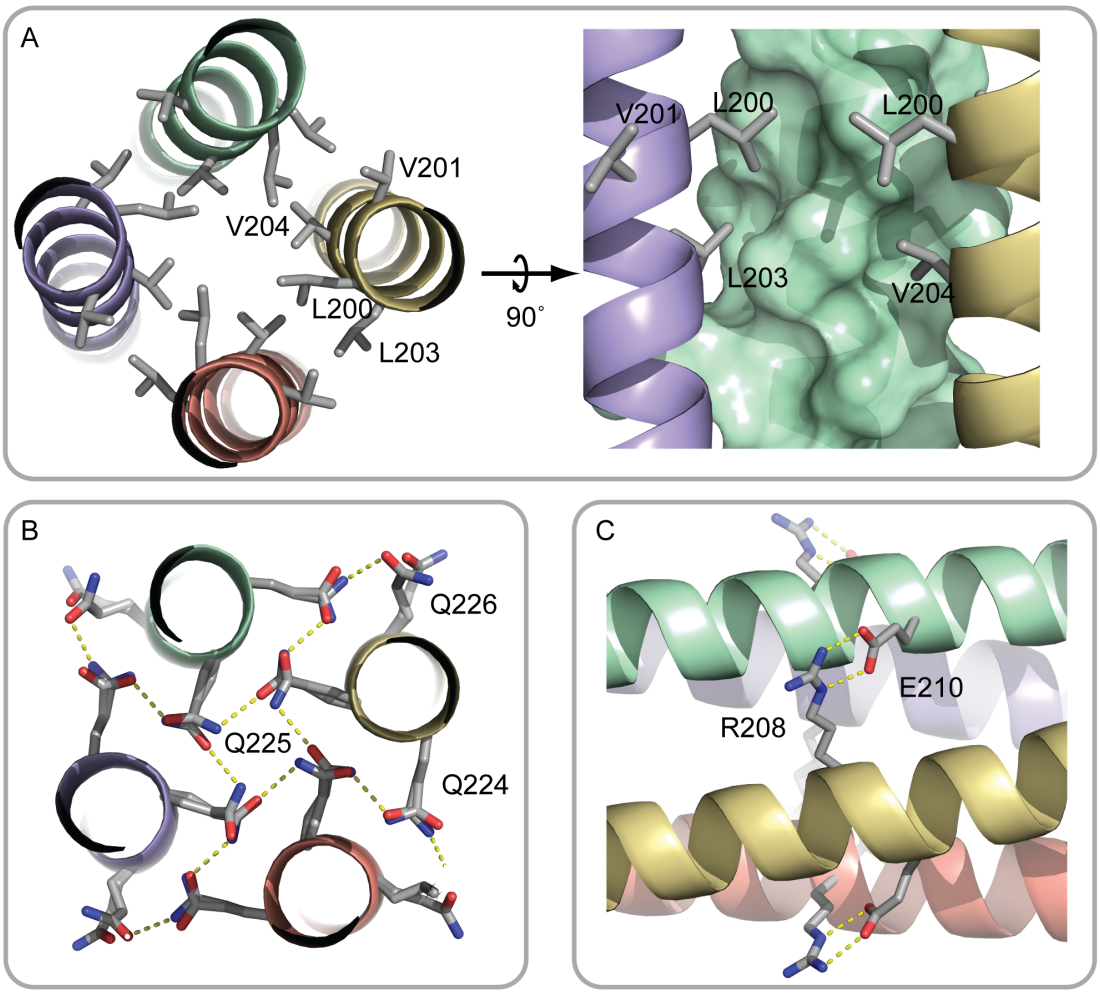


Figure 5.5.2.i: Views of the different interactions that hold the Ana2_CC tetramer together. (**A**) left, top down (N-C) view of a hydrophobic cluster. The labelled residues sit at the **g**, **a**, **d** and **e** positions of the heptad repeat, and pack closely forming a hydrophobic environment. Right, side on view of the same cluster, with one chain shown as a surface. The hydrophobic side-chains typically fit into "holes" which are hydrophobic pockets between side-chains of the adjacent monomer. (**B**) Top down view (N-C) of a QQQ triad which adopts positions **g**, **a** and **b** of the heptad. These polar sidechains form an inward facing hydrogen-bond network. Evidence could be seen for dual conformations of these sidechains, which can be seen to optimise the hydrogen bonding potential of the network. (**C**) Side-on view showing a salt bridge that is satisfied between adjacent chains.

The QQQ triad at residues 224, 225 and 226 (**g**, **a** and **b** positions) creates an internal hydrogen bond network (**Figure 5.5.2.i B**). Another QQQ triad exists in the helix, however it adopts a different position in the heptad register (**a**, **b** and **c**), and so the same hydrogen-bonding network is not formed here. However residue Q211 is also at the **a** position and faces inwards, this time coordinating a central sodium ion in a square-planar geometry (shown in **Figure 5.5.4.i, E**). (This

sodium could also be modelled by a water, as both species have ten electrons and therefore contribute similarly to scattering). The structure was analysed by the Check My Metal server [Zheng et al., 2014a], from which it was judged the coordinated species had consistent characteristics for a sodium ion). The tetramer is also stabilised by a cross-chain salt bridge between R208 and E210, which links adjacent chains. These residues adopt the peripheral **e** and **g** positions respectively, and so this interaction is much like one that would stabilise a canonical dimeric CC.

5.5.3: An analysis of CC conservation

The Ana2 central CC is tetrameric, which potentially has important biological implications for cartwheel biogenesis. Both *C. elegans* SAS-5 and vertebrate STIL proteins have a predicted CC, and it is therefore important to assess whether these are likely to share a similar structure and function. A multiple sequence alignment was carried out using Consurf [Glaser et al., 2003], aligning Ana2 *Drosophila* species orthologues with STIL orthologues, from CR2 onwards (**Figure 5.5.3.i, B**). It was not possible to convincingly align CeSAS-5 (not shown). Importantly, despite low overall sequence identity, the predicted CC regions of the STIL proteins aligned with the Ana2 CC. Conservation maps mainly to the centre of the Ana2 CC (**Figure 5.5.3.i, A**), although there is little correlation between conservation and buried residues. The alignment highlights a conserved core across the central ~2 heptad repeats of the CC. However this is slightly misleading, as the alignment is dominated by STIL orthologues, and it can be seen that although the STIL and Ana2 CCs aligned, their primary sequences are quite divergent. In Ana2, **g**, **a**, **d**, and **e** position residues are buried upon tetramerisation. The **g**, **a**, **d** and **e** positions were mapped onto the HsSTIL sequence to see if residues at these positions in STIL had the same characteristics as those in Ana2, and might therefore indicate a similar structure and function for the HsSTIL CC. Some of these residues were similar, although some were not.

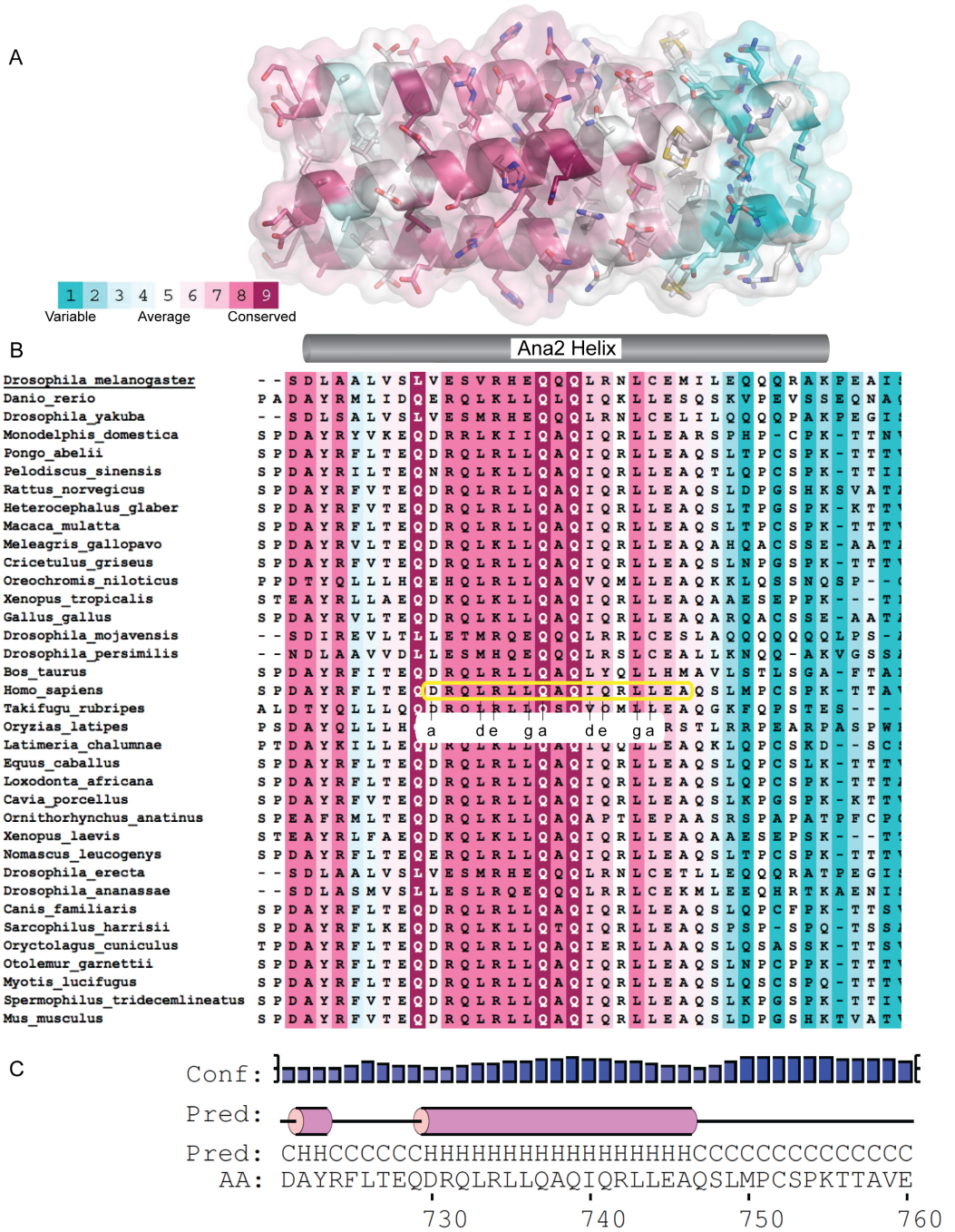


Figure 5.5.3.i: Analysis of CC conservation. (**A**) Ana2_CC tetramer (TEV site omitted) coloured according to the Consurf conservation scale. All side chains are shown in stick represention, within a semi-transparent molecular surface. (**B**) Multiple sequence alignment output from Consurf, which was able to align the predicted coiled-coil regions from vertebrate STIL proteins, with the CC regions from *Drosophila* species Ana2 proteins. Shown in a yellow box is the sequence from HsSTIL predicted to form a helix according to PSIPRED, the output of which is shown below in (**C**). This helix is predicted to be significantly shorter than the Ana2 CC helix. The **g**, **a**, **d** and **e** positions of the heptad repeat from the Ana2 CC are mapped onto the HsSTIL sequence. Some of these positions maintain residues with similar properties to those in Ana2, however some do not. It is therefore difficult to predict with confidence whether the predicted HsSTIL CC will have a similar structure/function to that from Ana2.

Because the mechanism of Ana2 tetramerisation is goverened by hydrophobic, polar and charged interactions, it is difficult to conclude the significance of a particular character of residue at a particular position. A further complication of this analysis is that it is quite difficult to accurately predict the boundaries of the STIL CC. The COILs server predicts a region ~717-758. Although PSIPRED does predict a helix in this region, the prediction is not particularly confident, is shorter, and is broken (**Figure 5.5.3.i, C**). What is clear however is that the CC region of Ana2/STIL proteins is particularly well conserved within each protein family, when compared to the rest of the protein. This indicates that the CC region of the protein is likely to be functionally important in all species. Although it can be seen from the results of **Section 5.2.8**, the HsSTIL CC is likely to confer multimerisation, absolute conservation of the structure and function of the HsSTIL CC is not easy to predict based on conservation or bioinformatics analysis. More experimental analysis will be needed to unambiguously answer this important question.

5.5.4: Rationalisation of Ana2_CC mutant constructs

In **Chapter 4**, **Section 4.3.7**, I described how mutant forms of the Ana2_CC were designed and tested for localisation *in vivo*, and also in **Section 5.2.7**, their multimerisation capacity *in vitro*. CC-A, CC-S and CC-D mutants were designed to disrupt self-association, by mutating predicted **a**/**d** position residues, based on a modelled dimeric CC. Fortunately, analysis of the Ana2_CC structure revealed that these residues are indeed involved in interface formation (**Figure 5.5.4.i A, B, C**), although according to the register assigned by the SOCKET server [Walshaw and Woolfson, 2001], they adopt the **d** and **g** positions, (**d** residues face inwards, and **g** residues are towards the periphery). Thus, mutations at these positions would drastically alter the tetramerisation interface, explaining why these mutant forms behave predominantly as monomers (**Section 5.2.7**). Interestingly, the RQQE mutation cluster was designed based on conserved residues that were predicted to be solvent exposed in a modelled dimeric structure

(**Figure 4.3.7.i, C**). Analysis of the tetramer structure reveals that this design was not optimal, as some of these residues are involved in the tetramer interface (**Figure 5.5.4.i A, D, E**). R208 is in an "**e**" position, and is involved in the salt bridge interaction (**Section 5.5.2**), while Q211 adopts an inward facing "**a**" position, and partakes in the coordination of a central sodium ion in the crystal; the presence of this ion may well be a crystallisation artefact, and *in vivo* this residue may simply interact via hydrogen bonding, similarly to Q225 (**Figure 5.5.2.i, B**). Thus, these two residues may be conserved because they are involved in the tetramerisation interface. The other two residues, Q213, and E219, are not involved in tetramerisation and are solvent exposed, with no obvious structural reason to explain their conservation. The involvement of R208 and Q211 in the tetramer interface presumably explains why alanine substitutions resulted in the weaker self-association for the RQQE protein, as seen by MALS (**Section 5.2.7**), although this disruption did not prevent localisation of the RQQE construct *in vivo* (**Figure 4.3.7.i, D**).

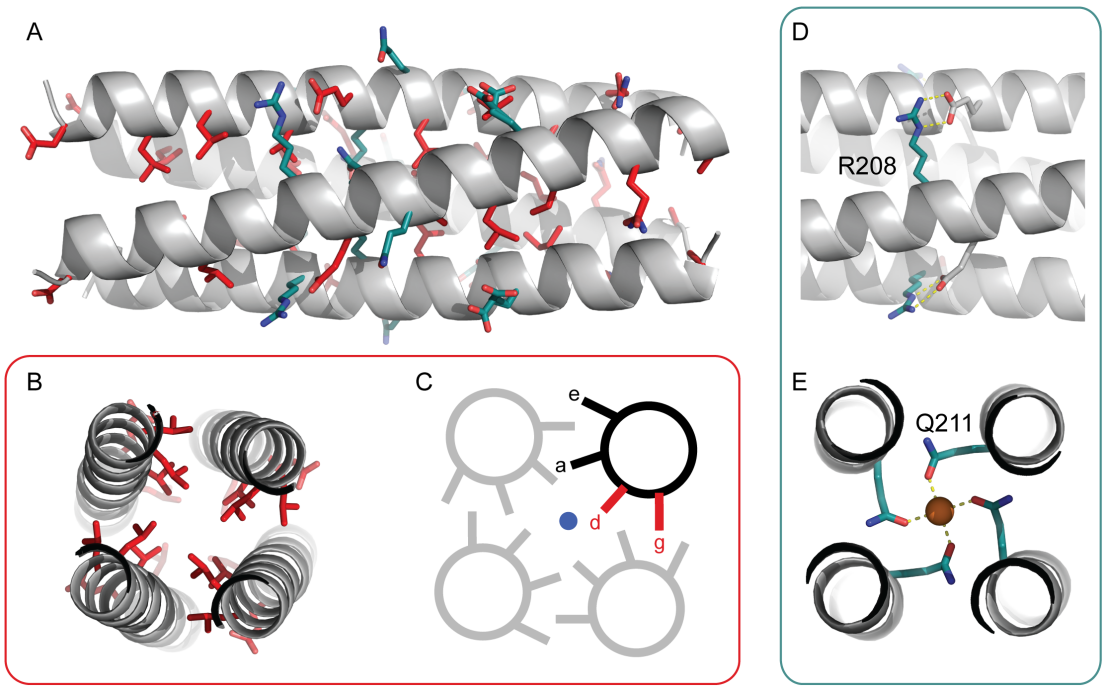


Figure 5.5.4.i: Mapping of mutated residues onto the Ana2_CC tetramer structure. (**A**) Backbone shown in grey cartoon format. Residues mutated in CC mutation clusters (CC-A, CC-S, CC-D) are shown as red sticks. CC-RQQE mutation cluster residues are shown as teal sticks. (**B**) Top down (N-C) view showing the CC mutated residues. As shown in the schematic in (C), these residues adopt the **d** and **g** positions out of the uniformly buried **g**, **a**, **d** and **e** positions. (The CC axis is shown as a blue dot). (**D**) R208, of the CC-RQQE cluster partakes in a cross-chain salt bridge. (**E**) Q211 of the CC-RQQE cluster faces inwards and coordinates a sodium ion (shown as an orange sphere).

5.5.5: Comparison of the Ana2_CC to other tetrameric coiled-coils

The parallel, homo-tetramerisation via a coiled coil domain that I have observed for Ana2_CC is quite unusual in biology. A model of the tetrameric Ana2_CC (TEV site omitted) was compared against the PDB using the DALI server [Holm and Rosenstrom, 2010] to find existing entries with similar tetrameric arrangements. Only a small proportion of hits were homo-tetrameric, parallel and had a left-handed supercoil. Other examples of this topology were found by manually searching the PDB and SCOP and are shown in **Figure 5.5.5.i**. Of all the examples found, there was quite a high variation in the twist and uniformity of the supercoil. As such, it was only possible to align the Ana2_CC tetramer with PDB entry 2O1J, (rotavirus enterotoxigenic protein NSP4, [Deepa et al., 2011]) which had a similar uniform twist. This superimposition gives an RMSD of 2.0 Å across 285 mainchain atom pairs. Other examples were more or less twisted, or had right handed or straight sections, or even kinks. Perhaps these variations explain why the DALI server missed many of these structures. It should also be noted that many examples of tetrameric CCs in the PDB were engineered or mutated proteins.

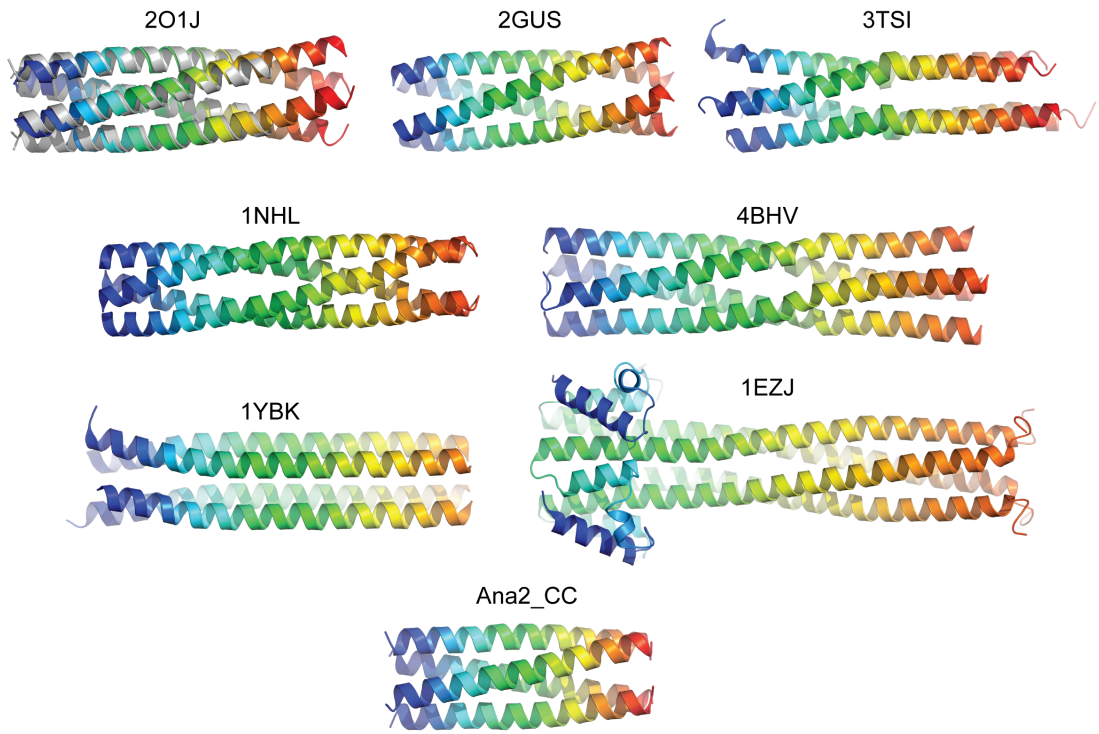


Figure 5.5.5.i: Comparison of various parallel tetrameric coiled-coils. All are shown in cartoon format, coloured N-C, blue to red. The closest match from the DALI server was 2O1J, onto which the Ana2_CC (TEV site omitted) could be superimposed with an RMSD of 2.0 Å. This superimposition is shown with the Ana2_CC tetramer structure as a grey cartoon.

In total, only five biological examples of homotetrameric, parallel, left handed coiled coils were found (2O1J, 4BHV, 3TSI, 1EZJ and 1NHL) suggesting this sort of interaction is in fact highly unusual in biology. Interestingly, these were all from viral proteins, apart from 1NHL, which is a non-neuronal human SNARE complex protein, SNAP-23. SNAP-23 is thought to participate in a hetero-tetrameric complex, and the authors of 1NHL are unsure of the biological significance of the homo-tetramer [Freedman et al., 2003].

5.6 Perspectives

As I showed in **Chapter 4**, RNA injection experiments elucidated that both the coiled-coil and STAN domains of Ana2 are necessary for wild-type localisation to the centriole. *In vitro* analysis has confirmed that the Ana2 CC forms a parallel, tetrameric four-helix bundle. The tetramer appeared to have a very tight affinity; it was never observed in solution at a lower stoichiometry than tetrameric, and the tetramer was still detectable in a normally denaturing ESI-MS experiment. Indeed, four-helix bundles are thought to be especially stable forms of CC – the broad interface increases the number of possible interactions that stabilise the oligomer [Lupas and Gruber, 2005]. However, it would still be desirable to quantify the affinity of the Ana2 CC interaction. This might be difficult if the interaction is particularly tight, as it would require low concentrations to dissociate the protein, which might then be difficult to detect. However, as the observed interaction appears to be stable, this suggests that the Ana2 CC is homo-tetrameric in its native conformation. Indeed, it is likely that the local concentration of Ana2 once recruited to the centriole, would be particularly high, favouring oligomerisation. This is important to consider, as a previous study crystallised a homo-tetrameric coiled coil of SNAP23, involved in the vesicular transport SNARE complex (Freedman et al 2003). This protein is normally found in a hetero-tetrameric coiled-coil complex involving two other proteins, and so it was proposed that SNAP23 might form a tetrameric storage conformation when in isolation. This tetrameric result was a surprise to the authors, and was only observed *in crystallo*, however, I readily observed Ana2_CC as a tetramer *in vitro*.

I further showed that the CC mutations analysed in **Chapter 4**, abrogate tetramerisation, possibly explaining the observed loss of function of these constructs *in vivo*. This strongly suggests that the Ana2 CC also drives tetramerisation *in vivo*. It would however, be interesting to

confirm the oligomeric state of Ana2 *in vivo*. This was performed for HsSAS-6 using fluorescence correlation spectroscopy on the cytoplasm of live cells [Keller et al., 2014]. Using standards for GFP fluorescence, it is possible to monitor the passage of individual oligomers passing through an observation volume. However, it is questionable as to how accurate and informative this technique is at present, and it would only be able to monitor cytoplasmic oligomerisation - not that at the centriole, and it is the centriolar oligomerisation state that would hold the key to cartwheel assembly.

The tetrameric state of Ana2 was exciting, but not completely surprising. A very short biophysical study had previously reported that a recombinant fragment of *C. elegans* SAS-5 (residues 82-260) formed a tetramer in solution [Shimanovskaya et al., 2013]. The authors make no mention of the fact that this fragment contains the predicted coiled-coil, and there were no further efforts to delineate the determinant of tetramerisation. This result is interesting, because I purified a construct of *C. elegans* SAS-5 (residues 131-173), which is predicted to form the CC. In my MALS analysis, this fragment appeared to show no signs of self-association, which suggests perhaps that CeSAS-5 does not tetramerise by the same mechanism as Ana2. However, before I form hard conclusions on this, I would like to repeat the experiment with a redesigned construct, as the boundaries of the CeSAS-5 CC are by no means obvious. However, it is interesting that my STIL CC construct was able to oligomerise to some extent, although not as a stable tetramer. My study in **Chapter 3** confirmed that STIL and Ana2 share at least one orthologous assembly mechanism – the binding of SAS-4/CPAP, and so it is appealing to think that further functions might be maintained between both protein families. Again, no firm conclusions can yet be drawn from this result.

Very recently, a paper was published showing that a fragment of Ana2 – containing the CC – forms a tetramer, however - the authors' conclusion was that this was mediated by the Dynein Light Chain, LC8, (or Ctp as it is known in *Drosophila*) [Slevin et al., 2014]. LC8 is associated with cytoplasmic dynein, which is involved in spindle organisation. The authors provided structures of LC8 dimers, bound to 10-12-residue Ana2 peptide regions (Pep1 and Pep2) that flank the CC. The interactions were robust, with Kds of ~1 and 13 μM for Pep 1 and 2 respectively. They purified a fragment of Ana2, "Ana2M" that contained the CC plus the flanking Pep regions, however they could only purify it in the presence of LC8, with the resulting complex having a stoichiometry of $Ana2M_4$:$LC8_8$ by MALS. They did not report purification of the CC alone. It has been shown previously that Ana2 interacts with LC8, and that LC8 localises to the centrosome [Wang et al., 2011]. However, to my knowledge no studies have shown that LC8 is essential for centriole biogenesis. Considering the recent study by [Slevin et al., 2014], my results are especially important, as they show that tetramerisation is clearly driven by the CC alone. Although LC8 dimers may bind to regions flanking the CC, I suggest this has little to do with tetramerisation. Indeed – my CC mutants show that a non-oligomeric CC results in a complete loss of function *in vivo* (**Chapter 4**). Although two independent studies have now confirmed a link between Ana2 and LC8, it is hard to reconcile Ana2's likely localisation to the inner centriole lumen, with the DLC, which is not an integral centriole component, and therefore would not be expected to access the inner centriole to bind Ana2. Either that, or Ana2 has an additional role outside the centriole. Perhaps this is the reason the STAN domain on its own, localises to the PCM *in vivo*?

What is especially intriguing about the oligomerisation of Ana2 is that the convergence of SAS-6 spokes would cause the spoke region of SAS-6 to form a dimer of dimers, or even a tetramer. The co-localisation of both SAS-6 and Ana2 to the centriolar lumen, and this stoichiometric similarity is therefore appealing, especially as Ana2 has been implicated as a cartwheel regulating protein [Stevens et al., 2010a]. This result is also interesting with respect to a previous

far-Western blot experiment described in **Chapter 3, Section 3.3.1**. In this experiment, a fragment of Ana2, Ana2$^{1-228}$, was used that contained both the CR2, and the CC (and would therefore presumably be tetrameric). This fragment appeared to bind to the SAS-4 TCP domain, but also bound to a SAS-6 NT-CC dimer construct, SAS-6$^{1-216}$. As it was likely the CR2 domain of Ana2 that bound SAS-4, this might suggest that another region – possibly the CC, interacted with SAS-6. Although I have tried an experiment to see if the Ana2 CC interacts directly with a conserved section of SAS-6 CC (residues 313-432) (**Appendix 7.8**) no interaction was observed. It will be important to pursue whether the Ana2 CC might bind instead to SAS-6.$^{1-216}$. This result is also compatible with results obtained in another study, where in a Y2H experiment, DmSAS-6$^{1-210}$ interacted with full-length Ana2, and more specifically, a C-terminal portion of Ana2 (201-420) interacted weakly [Stevens et al., 2010b]. This Ana2 construct begins only part way through the CC, possibly compromising its tetramerisation, and therefore weakening the interaction. However, the only fragment of Ana2 common in both this experiment, and my far-Western experiment is the CC, suggesting that an interaction between the Ana2 CC and the NT of SAS-6 should be followed up.

Chapter 6: Discussion

6.1: Summary

In this thesis I have identified functional domains of three essential centriole assembly proteins, and have provided high-resolution structures that help to explain essential interfaces of these proteins in *Drosophila*. The biological implications of these interfaces were then explored *in vivo*. This has provided important information that will help our understanding of the spatio-temporal roles of these proteins in centriole assembly.

6.2: The temporal assembly pathway

My findings are largely in agreement with the pathway of centriole biogenesis proposed at the beginning of this thesis. Several steps have been confirmed, and I have been able to refine particular steps to a higher-resolution molecular understanding. The cartwheel is the first visible structure to be built in "canonical" centrioles [Allen, 1969; Anderson and Brenner, 1971; Cavalier-Smith, 1974; Guichard et al., 2010], and SAS-6 is one of the first proteins recruited to the centriole, after the centriolar kinase [Gönczy, 2012]. In worms the centriolar kinase can recruit SAS-6 independently of its kinase activity, and this also appears to be the case in flies (Alan Wainman, Janina Baumbach and Jordan Raff, unpublished observations). In agreement with studies published during the course of my thesis [van Breugel et al., 2011, 2014; Guichard et al., 2013; Kitagawa et al., 2011b], I have built a model showing that a plausible cartwheel structure can be assembled from *Drosophila* SAS-6 (**Chapter 2, Figure 2.7.2.i, C**).

Ana2/STIL is recruited to centrioles at approximately the same time, and possibly even before, SAS-6 [Vulprecht et al., 2012]. My studies have shown that Ana2 recruitment is dependent on both its central CC and STAN domains. Although I have a preliminary result suggesting that Ana2 may bind to SAS-6^{1-216} (**Chapter 3**, **section 3.3.1**), this has yet to be proven, and will be a focus of future experiments. It will also be necessary to further characterise the behaviour and role of the STAN domain. Although it has been shown that Ana2 interacts with the Dynein light chain, LC8, this behaviour is likely confined to *Drosophila* species [Slevin et al., 2014], and it remains to be shown whether LC8 is essential in centriole biogenesis. So far it has not been identified as a centriolar component in the genome-wide RNAi screens in worms or flies [Dobbelaere et al., 2008; Gönczy et al., 2000], although this is far from conclusive.

Cartwheel assembly is likely to depend on the actions of both Ana2 and SAS-6, and my studies have identified that both of these proteins self-assemble. In *Drosophila*, when Ana2 or SAS-6 are individually overexpressed in testes they form large aggregates, but these aggregates only form a cartwheel-like structure when the two proteins are co-overexpressed together, forming SAPs [Stevens et al., 2010a]. In future experiments I plan to collaborate with other members of the Raff lab to view these structures in the absence of either the SAS-6 N-N interface, or the Ana2 CC interface, both of which I have shown can be knocked out by point mutations. My prediction is that the self-assembly of both these proteins will be necessary to visualise these cartwheel-like structures. Preliminary results using my non-homo-oligomerising constructs suggest that the aggregates formed by either protein in isolation can no longer form (Nadine Muschalik, unpublished observations). It will also be interesting to understand why no physical interaction has been observed between STIL and SAS-6. In one *Drosophila* study, Sak was overexpressed with either of SAS-6 or Ana2, or both SAS-6 and Ana2. In all cases, levels of SAS-6, Ana2, or both were unusually enriched at centrioles [Stevens et al., 2010a]. Therefore, it may be that Ana2/STIL and SAS-6 are recruited by Sak/PLK4 independently, although simultaneously to the

procentriole. SAS-6 would then recruit CEP135, and Ana2/STIL would recruit SAS-4/CPAP. Both Cep135 and SAS-4 would be required to stabilise the cartwheel, and nucleate MTs at the periphery, as well as follow on to recruit Asl and the PCM.

Downstream of SAS-6 and Ana2, SAS-4 is recruited, and my studies have now shown that this is likely to occur via an interaction with the CR2 domain of Ana2. This interaction is conserved, and also rationalised a link to a particular MCPH mutation [Cottee et al., 2013]. Future studies will likely focus on how SAS-4 recruits downstream proteins such as Asl, and how it binds to MTs. Such studies are also likely to involve the solution of high-resolution structures.

Similarly, CEP135 has an important role in linking the cartwheel to MTs, [Lin et al., 2013] and I imagine similar structural studies will describe these interactions in the future. The linking of the central “spiral” of *C. elegans* to the MT singlets is likely to involve a different mechanism, as there is no CEP135 homologue in these animals; however, as the overall centriolar superstructure in this model organism is especially divergent, this may not be surprising.

6.3: Spatial organisation of centriole proteins

The structural study of many supramolecular assemblies involves a combination of high-resolution structures of protein/domain building blocks, and their placement into lower-resolution maps of the assembly as a whole. Such an approach has been carried in many systems such as the ribosome [Gao and Frank, 2005], the Type 3 Secretion System [Hodgkinson et al., 2009], GroEL [White et al., 1997] and many more. Recent advances have now provided such lower-resolution maps for the entire centriole, including the cartwheel, pinhead, and MT triplet blade (EMD-2329, EMD-2330) [Guichard et al., 2013]. The conserved structural knowledge of canonical SAS-6 orthologues, and their self-assembly, has allowed the convincing docking of the high-resolution SAS-6 structure into this low-resolution map.

However, aside from SAS-6, and the microtubules, we are currently limited in confidently building upon this model for several reasons. First, the EM map is of relatively low resolution (38 Å) in comparison to the size of the additional domains for which structures are available, namely the SAS-4:Ana2 complex, and the Ana2 CC. It is theoretically possible to create homology models of other domains to increase this catalogue of structures, however, centriole proteins are notoriously unstructured, or restricted to very basic structures such as segments of coiled-coils, and so this approach is not particularly feasible. For instance CEP135 is predicted to be almost entirely constructed from coiled-coil regions, and apart from the TCP domain, SAS-4 is predicted to be largely unstructured, or contain only coiled-coils. Additionally, the presence of such unstructured, or extended regions in these proteins also means there are many ways in which these different domains could be placed into the low-resolution map. As in the case of SAS-6, it would be desirable to model larger assemblies of these known domains and dock those into the map; however, our current knowledge of the superstructural assembly of these

domains is limited. Simply put – for most domains of centriolar proteins we could model anything and put them anywhere in the map.

Currently therefore, we can only speculate as to the location of the domains which have been structurally characterised. Due to the location of SAS-4/CPAP at the inner centriole wall [Lukinavičius et al., 2013; Mennella et al., 2012; Sonnen et al., 2012] and Ana2/STIL in the inner centriolar lumen [Arquint et al., 2012; Lukinavičius et al., 2013; Stevens et al., 2010a], it is likely that the TCP domain, bound to the N-terminus of Ana2 would be located at the centriolar wall, or somewhere in the pinhead region. However, there is plenty of density in the low-resolution map of the pinhead and areas close to the centriolar wall that could accommodate the SAS-4:Ana2 structure, so currently there is no further information to help us dock this structure. I will therefore limit my speculation to the role of the Ana2 CC, for which I have three suggestions as to how it could relate to the cartwheel model.

Before discussing these suggestions, it is important to consider the possible length of linker between the Ana2 CC and the CR2 domain or the STAN domain, as shown in **Figure 6.3.i**. Assuming these linkers are unstructured, I have based their maximal extended distance on a β-strand - the most extended linear conformation a series of residues can take (whilst obeying chemical restraints). Arranged in this way, a linker can stretch ~3.3 Å per residue, equating to a maximum CR2–CC distance of ~50 nm, and a maximum CC–STAN distance of ~29 nm (**Figure 6.3.i, B**).

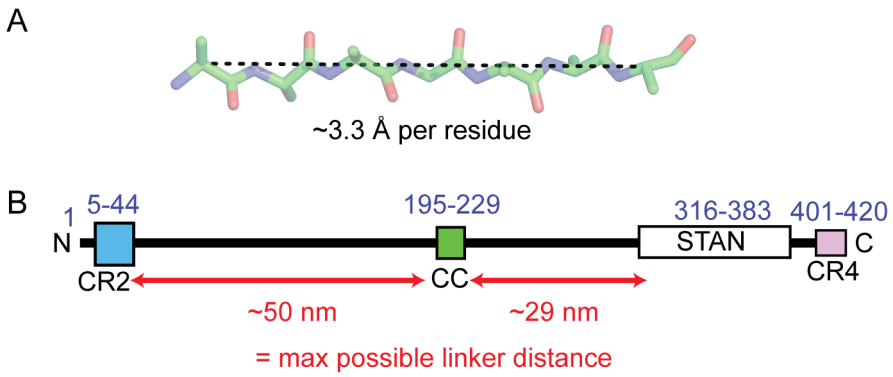


Figure 6.3.i: Maximum linker distances between the CR2 and STAN domains from the Ana2 CC. (**A**) A seven residue polyAla β-strand in stick format, where the residues span the ~maximum distance that is geometrically possible. In this arrangement, a linker could theoretically span ~3.3Å per residue. (**B**) A schematic representation of Ana2. According to the analysis in **A**, the linker regions between the CR2 to CC, and CC to STAN could span ~50 nm, and ~29 nm respectively.

6.3.1: The Ana2 CC tetramer as a hub separator

An interesting feature of the EM map is that despite any visible structural cue, each cartwheel hub is separated from the hub above and below it. The periodicity of the hubs is ~8 nm, however the hub itself is ~3.5-4 nm thick, resulting in a gap of ~4-5 nm between each hub. This gap is compatible with the ~5 nm length of the Ana2 CC (**Figure 6.3.1,i, C**), and so one possibility is that the Ana2 CC acts as a rod to separate the hubs (**Figure 6.3.1.i, A, B, D**). Indeed, coiled-coils often function as "molecular spacers" [Lupas and Gruber, 2005]. As discussed above, in a far western blot experiment I observed Ana2$^{1\text{-}228}$ (including the CC) to interact with SAS-6$^{1\text{-}216}$ (**section 3.3.1**). If this preliminary result is to be believed, this model would place the CC in close proximity to the SAS-6 N-terminus. Additionally, [Guichard et al., 2013] describe "cartwheel inner density" (CID) that resides inside the hub, close to the SAS-6 N-N interface, - approximately where I propose to dock the Ana2 CC (**Figure 6.3.1.i, B**). This density is the strongest in the map, and is therefore the only feature visible at high contour levels. However, the CID density exists mainly within the plane of the hub, and does not extend between the hubs (**Figure 6.3.1.i, C**). The caveat of this model is that one would expect such "inter-hub" rods to be a rigid and regular feature that would result in strong bridges of density between the hubs, and this is not the case

in the EM map. This model would also require the CR2 to have to stretch a long distance along the spoke in order to reach the proposed site of the SAS-4 TCP domain, in the pinhead. However, this may be possible, as the distance between the N-terminus of the Ana CC to the end of the SAS-6 spoke is less than 50 nm (**Figure 3.6.1.i, D, E**).

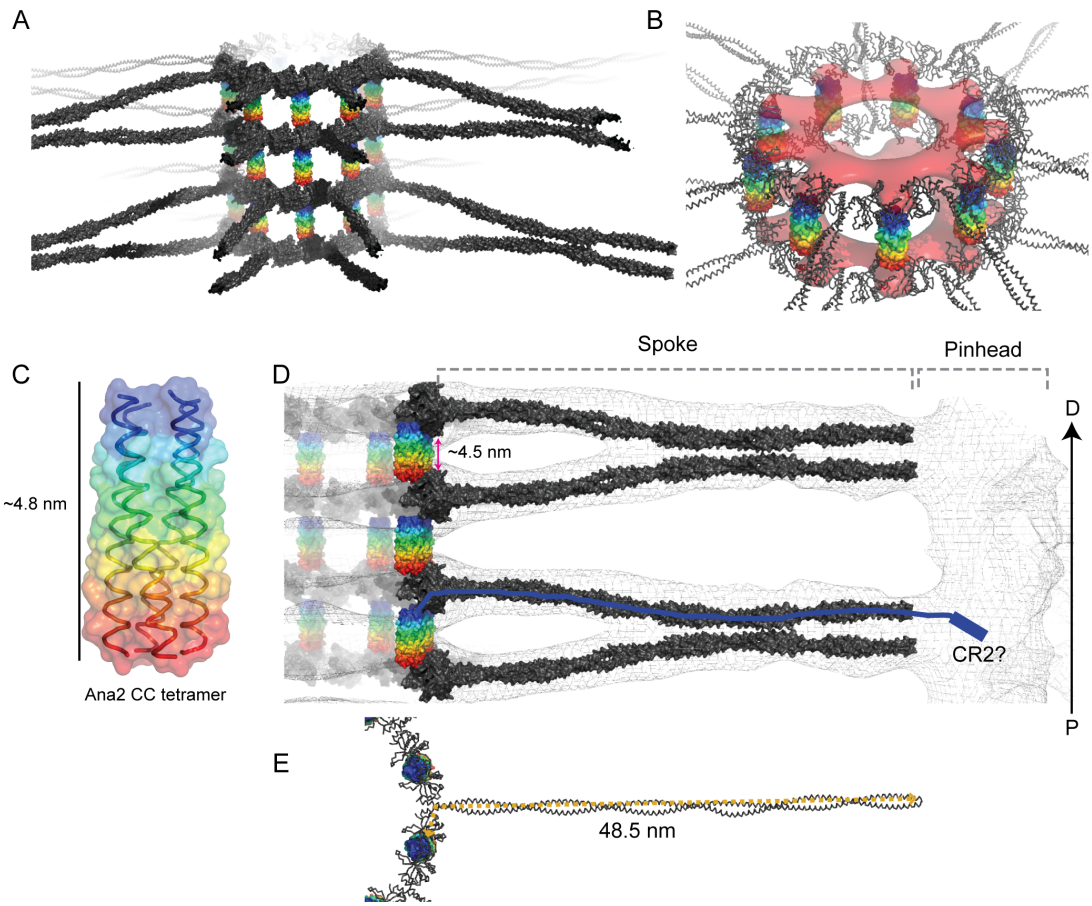


Figure 6.3.1.i: The Ana2 CC as a hub-separator. (**A**) View of the cartwheel model, with hubs separated by the Ana2 CC. SAS-6 is shown in black, either in spacefilling or ribbon format. The Ana2 CC tetramer is shown in spacefilling format, coloured blue to red, (N-C). (**B**) "The Cartwheel Inner Density" (CID) reported by Guichard et al., 2013. The density is shown as a translucent red surface, and at a high contour, where it is the only visible feature in the map. Note that fingers of density protrude over the top and bottom of the Ana2 CC, however, the CID is restricted mainly to within the plane of the hub. (**C**) The Ana2 CC is coloured as in **A**. The length of the tetramer is ~4.8 nm. (**D**) Longitudinal section through the cartwheel, inside the EM map, shown as grey mesh at a lower contour than in **B**, to reveal the spokes and pinhead. The inter-hub gap is ~4.5 nm. The N-terminal linker of Ana2 is shown as a blue line, travelling along a spoke in order that the CR2 domain can reach the pinhead – presumably the rough location of the SAS-4 TCP domain. (**E**) The distance the linker would have to span (dashed yellow line) is ~48.5 nm, which is just within the theoretical maximum length of this linker.

6.3.2: The Ana2 CC tetramer as a radial spoke crosslinker

A perhaps more attractive model is that Ana2 might stabilise the cartwheel symmetry by radially crosslinking the spokes (**Figure 6.3.2.i, A, C**). This would possibly explain features observed in electron micrographs of *Phlyctochytrium* centrioles [McNitt, 1974] (**Figure 6.3.2.i, B**). Although the distance of the crosslink is longer than the Ana2 CC, the linker need not be rigid to perform this role. In fact, its tetramerisation might also explain the convergence of the spokes at around these points. However, a major problem with this model is that the parallel nature of the Ana2 CC imposes an N-C polarity in the crosslinker, as shown in **Figure 6.3.2.i, A, C**. This would be hard to reconcile with the radial cartwheel symmetry, which displays no obvious clockwise or anticlockwise symmetry breaking features. However, it is possible that two antiparallel tetramers could arrange to make a bi-directional crosslinker (not shown). Either model would allow the CR2 to stretch towards the pinhead, and the STAN domain to have access to various regions of the SAS-6 spoke (**Figure 6.3.2.i, C**).

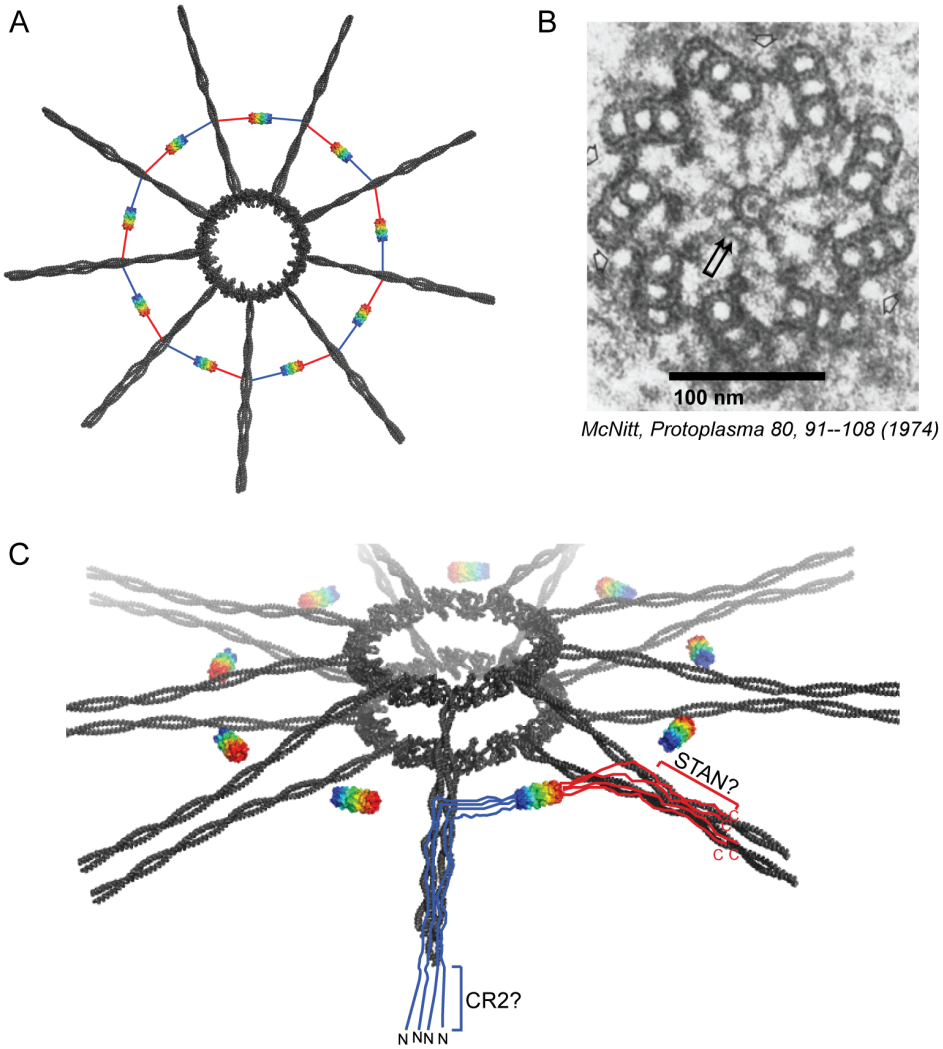


Figure 6.3.2.i: The Ana2 CC as a radial crosslinker. (**A**) Top-down view of the SAS-6 cartwheel, shown in black ribbon format. The Ana2 CC is coloured as in **Figure 6.3.1.i**, linker regions are shown as blue (NT) or red (CT) lines. (**B**) An example micrograph reported by McNitt, 1974, in which spoke crosslinkers were observed, indicated by the central double black arrows. (**C**) Angled view of **A**, with possible paths of the N- and C-terminal linkers shown as coloured lines. The CR2 domain could easily reach the pinhead in this model.

6.3.3: The Ana2 CC tetramer as a spoke-converger.

Alternatively, the tetramerisation function of the Ana2 CC may serve to tether the SAS-6 spokes at their convergence point (**Figure 6.3.3.i, A**). This model would justify the parallel nature of the tetramer, and could conveniently position all of the CR2 domains at the periphery of the centriole where they could bind to SAS-4. It could position the STAN domain towards the central hub where it may bind SAS-6; or alternatively, the STAN could also serve as a radial crosslinker, as shown in **Figure 6.3.3.i, A**), whereby the C-terminal linker and STAN domains would associate between radial spokes. This would be a more radially symmetrical crosslinking model compared to that proposed in **Section 6.3.2**.

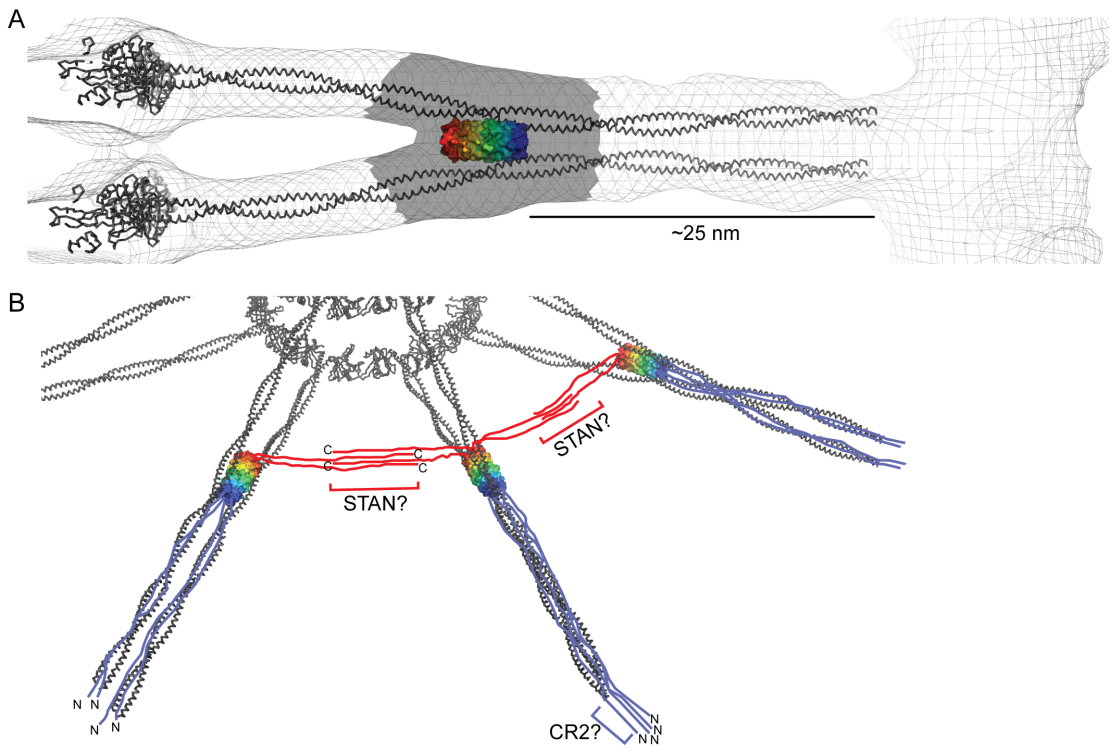


Figure 6.3.3.i: The Ana2 CC as a spoke-converger. (**A**) Side on view of the cartwheel, shown in black ribbon format, within the EM density, contoured to reveal both the spoke and pinhead density. The Ana2 CC is coloured as in **Figure 6.3.1.i**, with density around it highlighted as a shaded surface. The Ana2 CC is positioned at the convergence point between SAS-6 spokes, where there are four parallel helices. (**B**) Angled, top-down view. This model could also allow radial crosslinking, by the association of the STAN domains. The CR2 would simultaneously have easy access to the pinhead.

6.4: Conclusion

Although it is useful to think about these speculative models clearly more information will be required to allow the precise positioning of the Ana2 tetramer in the low-resolution EM density map. For instance, where are the Ana2 CR2 and STAN domains in relation to one another? One possible experiment would be to raise antibodies specifically against each of these domains and carry out immuno-gold labelling and observe their positions via electron microscopy. However, the resolution of this technique may not be high enough to discern the different locations of these domains if they are not well separated. Preliminary experiments have been carried out in the lab using various super-resolution light microscopy techniques to examine the localisation of GFP in N and C-terminal GFP-fusions of SAS-6 and Ana2. These experiments suggest that it may be possible to resolve the relative positions of the termini of these proteins in the cartwheel structure (Saroj Saurya and Alan Wainman, unpublished observations). Using 3D-SIM which has a theoretical resolution of ~100-120 nm, the N-terminus of SAS-6 is reproducibly located within the C-terminus of SAS-6, as predicted by the cartwheel model.

The function of the conserved STAN domain in Ana2/STIL proteins also remains to be elucidated. The most tempting possibility is that this region interacts directly with SAS-6, but it is unclear if, let alone how this interaction occurs *in vivo*. This information could lead to a more informed modelling attempt using the SAS-6 cartwheel model. Additionally, it will be important to understand whether STIL behaves similarly, and if not, why is this the case? Hopefully, the necessity of the STAN domain will soon be discerned from rescue experiments of *Ana2*$^{-/-}$ flies, with my designed Ana2$^{\Delta STAN}$-GFP construct.

Finally, it will be important to test whether the central CC domain present in all Ana2/STIL proteins forms a stable tetrameric structure in other species, as I have shown is the case with *Drosophila* Ana2.

How is the nine-fold symmetry of the centriole enforced to such a high standard? My overall conclusion is that the SAS-6 N-N and CC interfaces have been shown to greatly influence the morphology of the central cartwheel, and indeed, are likely to enforce a propensity towards nine-fold symmetry. However, despite the recent observation of SAS-6 crystallising as a nine-fold hub [van Breugel et al., 2014] it remains that SAS-6-only assemblies have readily been observed to form alternative stoichiometries *in vitro* [van Breugel et al., 2011; Kitagawa et al., 2011b]. Additionally, and perhaps most persuasively, SAS-6 null *C. reinhardtii* cells usually still form nine-fold symmetric centrioles [Nakazawa et al., 2007]. This fact has often been overlooked in the excitement over the potential of SAS-6 to form cartwheel-like structures on its own. I predict that the homo-oligomerisation properties of the Ana2/STIL family of proteins that I have identified here will also serve to ensure the efficient assembly of a nine-fold symmetric cartwheel. Due to the morphological differences of the *C. elegans* centriole, SAS-5 may not be a true orthologue of Ana2/STIL, however it likely serves an analogous role in *C. elegans* cartwheel/spiral organisation. The propensity towards nine-fold symmetry may then be further strengthened by downstream factors such as the angle of MT triplet attachment, and the A-C linker. Together, all of these factors combine to ensure the formation of the strict nine-fold symmetric centriole that is almost invariably observed *in vivo*.

7.0: Appendix

7.1: Standard Laboratory Procedures

7.1.1: Primer design, PCR amplification and ligation cloning

For non-gateway cloned vectors, primers were invariably designed to have ~15bp homology to the end-destination vector, and ~18 bp homology to the target sequence. This resulted in a PCR insert with flanking arms that were homologous to the insertion sites of the vector, allowing (if necessary/desired) the use of InFusion (Clontech, TaKaRa) cloning, which involves homologous recombination between the insert and vector. Note that primers and plasmids made as part of this study are included in separate Excel spreadsheets (**Supplementary CD**). For "pETM" series vectors, inserts were always cloned in using the NcoI and NotI sites. This necessitated the insertion of a glycine residue after the initiator methionine. Where the natural sequence had no methionine, one was inserted.

PCR amplification was carried out using Phusion polymerase master mix, with HF buffer (NEB).

Component	Volume (µl)
2 x Phusion HF master mix	25
Template (10 ng/µl)	1
Forward primer 10 µM	2
Reverse primer 10 µM	2
(optional) DMSO	1.5
aH_2O	To final 50 µl

Table 7.1.1.i: Components of a PCR reaction

Primer Tm values were calculated for the whole primer, according to the average of values calculated using the Promega Biomath Tm calculator*. The annealing temperature was usually 3 ˚C below the lowest primer Tm.

Step	Temperature (˚C)	Time (s)	Cycles
Initial Denaturation	98	25	1
Denaturation	98	10	35
Primer Annealing	3˚C below lowest primer Tm	15	
Extension	72	15-30 s/kb	
Final Extension	72	1 min/kb	1

Table 7.1.1.ii: PCR protocol

PCR inserts were analysed by 1 % agarose gel electrophoresis, and were processed using a Qiagen PCR cleanup kit. Vectors and inserts were restriction digested before ligation using conventional methods. Finalised plasmids were analysed by diagnostic digest and gel electrophoresis, and were sequenced across the insert (Source BioScience).

7.1.2: Gateway (attBP/attLR) Cloning

Gateway Cloning (Life Technologies) was used to construct plasmids used to generate transgenic Drosophila lines, (pUbq vectors) or for encoding mRNA (pRNA vectors). Primers were designed according to the manufacturer's protocol, and PCR amplification was carried out as in **Section 7.1.1**. PCR inserts were additionally purified by gel extraction, followed by gel purification (QIAquick Gel Extraction kit, Qiagen). BP and LR reactions were carried out according to the manufacturer's protocol, albeit using only half the volume of reagents.

* https://www.promega.com/techserv/tools/biomath/calc11.htm

7.1.3: Bacterial Transformation

20 µl chemically competent *E. coli* cells were incubated on ice for 30 min, with ~100 ng DNA, in a 15 ml Falcon tube to ensure thorough aeration after heatshock. Cells were heat-shocked for 90 sec at 42 ˚C using a water bath, before a 2 min incubation on ice. 200 µl 2xYT broth was added, and cells were incubated at 37 ˚C for 1 hour. In the case of transformation with an ampicillin resistant vector, this incubation could be shortened to 10 min. Cells were then plated onto agar plates containing appropriate antibiotic. For transformation of site-directed mutagenesis reactions, the manufacturers protocol was followed (QuikChange, Agilent) again, using 15 ml Falcon tubes.

7.1.4: Standard procedure for purifying soluble recombinant protein from the *E. coli* cytoplasm, and Ni-NTA purification details for all constructs.

Unless otherwise stated, the following describes the purification of proteins described in this thesis. Recombinant proteins were expressed using *E. coli* strain B834 (DE3) cells, which often proved to express most constructs (as assessed from SDS-PAGE analysis of small induced cultures). LB broth was usually used as the expression medium. The culture was inoculated either from a saturated overnight culture (~10 ml overnight added per litre), or from resuspension of cells from an agar plate (1/2 a plate per litre). Inoculated cultures of between 1-6 litres were grown at 37 ˚C, shaken at 180 rpm, until growth was mid-log phase, as assessed by an OD_{600} of ~0.6. Recombinant protein expression was induced by the addition of IPTG to 1 mM, and the cells were further cultured overnight at 21 ˚C. Cells were harvested by centrifugation at 4,000 x g for 20 min at 4 ˚C, resuspended in PBS (Dulbecco's, Sigma) and repelleted by centrifugation at 4,000 x g before freezing, and storage at -80 ˚C. Frozen bacterial cell pellets were thawed, and resuspended in Lysis buffer. This consisted of purification working buffer (WB) (as outlined in **Table 7.1.4**) to which had been added 1 mg/ml hen egg-white lysozyme (Sigma)

and ~400 U/ml Bovine pancreas DNaseI (Sigma). Additionally, one Protease Inhibitor Cocktail Tablet (PIT) (Roche, cOmplete EDTA-free) was added per 40 ml buffer. Typically, 5-10 ml Lysis buffer was used per gram of cell pellet, although this was modified where indicated. Typically, bacterial cell suspension was stirred for ~30 mins at 4 ˚C to ensure a homogenous resuspension. Cells were lysed via 3-5 passages through an Avestin Emulsiflex C5 homogeniser at 4 ˚C. The lysate was then clarified by centrifugation at 26,600 x g at 4˚C. Unless otherwise indicated, recombinant proteins were purified by Ni-NTA agarose affinity chromatography, using either a 1ml or 5ml Ni-NTA superflow cartridge (Qiagen), which was always pre-equilibrated in purification WB. The clarified supernatant was supplemented with imidazole to a particular loading concentration, in order to reduce non-specific binding. The cartridge was then washed for the indicated number of CVs, with WB, supplemented typically to 50 mM imidazole. The length of the wash step was influenced by the cleanliness of eluate from previous purification of that construct, or, when using an FPLC system, waiting for a stable UV readout. Elution of recombinant proteinin WB + imidazole was carried out either using a gradient of imidazole, or an isocratic elution, typically at 250-300 mM imidazole.

Construct/ Experiment	WB	Loading conc. [imidazole]	Ni-NTA	Wash [imidazole]	Elution [imidazole]
His-SAS-6, SAS-6-His, His-MBP-SAS-6, and His-GST-SAS-6	50 mM Na_xH_xPO4, 300 mM NaCl	(HT procedure) WB + 10 mM imidazole **Section 7.1.5**	Magnetic beads	HT procedure, **Section 7.1.5**	HT procedure, **Section 7.1.5**
His-SAS-6	50 mM Tris, pH 7.5, 150 mM NaCl, 1 mM DTT	25 mM	1 ml Cartridge	50 mM, 20 CV	Isocratic, 300 mM
SAS-6 fragments, (Section 2.1.3)	50 mM Tris, pH 7.5, 150 mM NaCl, 1 mM DTT	(HT procedure) WB + 10 mM imidazole **Section 7.1.5**	Magnetic beads	HT procedure, **Section 7.1.5**	HT procedure, **Section 7.1.5**
SAS-$6^{1-241/216/201}$ (WT/F143D)	50 mM Tris, pH 7.5, 150 mM NaCl, 1 mM DTT	25 mM	5 ml Cartridge	50 mM, 40 CV	Isocratic 250 mM
SAS-6_CC constructs (Section 2.8.2)	20 mM HEPES, pH 7.3, 200 mM NaCl	10 mM	0.5 ml Ni-NTA agarose resin	20 mM 10 CV	Isocratic 250 mM
SAS-$6^{313-432}$	50 mM $Na_xH_x(PO_4)$ pH 8.0, 200 mM NaCl	20 mM	5 ml Cartridge	50 mM, 20 CV	Isocratic 250 mM
SAS-$4^{725-901}$	50 mM Tris, pH 7.5, 150 mM NaCl	20 mM	1 ml Cartridge	50 mM, 40 CV	Isocratic 300 mM
SAS-$4^{707-901}$	50 mM Tris, pH 8.0, 150 mM NaCl	20 mM	5 ml Cartridge	50 mM, 40 CV	Gradient, 50-500 mM over 15 CVs
SAS-$4^{700-901}$	50 mM Tris, pH 8.5, 150 mM NaCl	20 mM	5 ml Cartridge	50 mM, 40 CV	Gradient, 50-500 mM over 30 CVs
$Ana2^{1-47}$-SAS-$4^{700-901}$	50 mM Tris, pH 8.5, 150 mM NaCl	20 mM	5 ml Cartridge	50 mM, 40 CV	Gradient, 50-500 mM over 30 CVs
SAS-$4^{700-901}$- $Ana2^{1-47}$	50 mM Tris, pH 8.5, 150 mM NaCl	20 mM	5 ml Cartridge	50 mM, 40 CV	Gradient, 50-500 mM over 30 CVs
diLipoyl-Ana2_CC	50 mM Tris, pH 7.5, 150 mM NaCl, 1 mM DTT	10 mM	5 ml Cartridge	50 mM, 40 CV	Gradient, 50-500 mM over 20 CVs
diLipoyl-Ana2 _CC-(A/S/D/RQQE)	50 mM Tris, pH 7.5, 150 mM NaCl, 1 mM DTT	10 mM	5 ml Cartridge	50 mM, 40 CV	Isocratic 250 mM
diLipoyl-STIL/SAS-5_CC	50 mM Tris, pH 7.5, 150 mM NaCl, 1 mM DTT	10 mM	5 ml Cartridge	50 mM, 40 CV	Isocratic 250 mM

Table 7.1.4.i: Various purification details for constructs purified in this thesis. For each construct, the purification working buffer (WB) is listed, along with details of purification via Ni-NTA chromatography. These include the concentration of imidazole ([imidazole]) used for loading/washing/elution, and also the type of (Qiagen) Ni-NTA product used, and the manner in which elution was achieved.

7.1.5: High-Throughput Ni-NTA affinity purification

Cells were typically grown in 5 ml cultures in 24 well high-volume plates (Flowgen) covered with a breathable membrane, and were inoculated with 100µl of overnight culture. Cultures were grown to an estimated OD_{600} of 0.6 (typically ~3-4 hours) before induction. Cells were harvested by centrifuging the plate at 4,000 x g for 20 min. Lysis buffer = as in **Section 7.1.4**, but with 2% Tween-20. 700 µl Lysis buffer was added to each cell pellet before lysis by incubation for 30 min on an orbital shaker, or until suspension was clear. Lysate was cleared by centrifugation at 4, 000 x g for 30 min, and supernatant was transferred to a 96 well plate (Greiner 655101). 20-30 µl magnetic Ni-NTA beads (Qiagen) was added to each well, and binding was carried out at RT for 30 min, whilst shaking. (Each well accommodated ~ 240 µl, and so this step was repeated multiple times where necessary). Beads were collected at the side of each well using a 96 well magnet, and supernatant discarded. (Where fewer samples were purified, 96 well plates could be swapped for 1.5 ml Eppendorf tubes, and a magnetic rack). Beads were washed in 200 µl wash buffer (working buffer + 20 mM imidazole + 0.05% Tween-20) by shaking for 5 min at RT. This was carried out 3 times. Beads were either eluted using (working buffer + 500 mM imidazole + 0.05% Tween-20), or were boiled in SDS-PAGE loading dye for direct analysis by SDS-PAGE.

7.1.6: In vitro transcription, and purification of recombinant mRNA

mRNA was obtained via *in vitro* transcription. The pRNA destination vector includes a T3 phage promoter upstream of the ORF, allowing RNA to be produced via an *in vitro* T3 transcription reaction (mMessage mMachine, Ambion, Life Technologies) according to the manufacturer's protocol, although reaction volumes were doubled in order to obtain mRNA at an effective concentration.

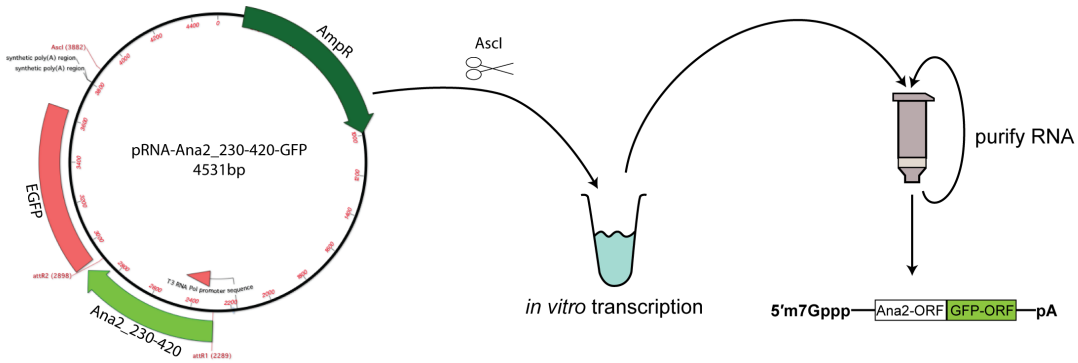


Figure 7.1.6.i: Recombinant mRNA production. On the left is a schematic of pRNA_Ana2_230-420-GFP plasmid as an example. pRNA plasmids are linearised and used as templates for *in vitro* transcription of RNA. The transcription reaction uses the T3 phage polymerase, and adds an m7G cap; the polyA tail is encoded in the pRNA template. RNA is then purified from the reaction mixture.

pRNA was linearised downstream of the polyA site using the AscI restriction enzyme, in order to provide a suitable linear substrate for transcription. Linearised vector DNA was purified by either precipitation, or using a PCR cleanup kit (Qiagen), and assessed for complete digestion using DNA gel electrophoresis. Transcribed mRNA was purified using an RNeasy Minelute cleanup kit (Qiagen) (**Figure 7.1.6.i**). In order to obtain sufficiently high concentrations, eluted RNA was re-applied to the elution spin column, and re-eluted. RNA was then analysed by gel electrophoresis to assess integrity of the product. The concentration was measured via the A_{230}, using a Nanodrop 1000 (Thermo Scientific) in "ssRNA" mode. The concentration was normalised to 3.66µM according to the theoretical length of each RNA, before storage at -80 ˚C until use.

7.1.7: Multi Angle Light Scattering (MALS)

MALS takes advantage of the angular dependence of light scattering by differently sized objects. The proteins in solution are smaller than the wavelength of incident light and so they scatter light according to Rayleigh scattering. Smaller objects scatter light isotropically, however, the larger the object, the less light is scattered perpendicular to the incident beam. MALS therefore monitors scattering from various angles to assess the size and shape of the molecule. MALS also analyses the refractive index of the sample in order to assess its polarisability, in order to relate this to the intensity of the scattered light. Together this information allows the calculation of the absolute mass of the scatterer. MALS is therefore typically used to assess the oligomerisation of a sample in solution over a range of concentrations.

In this thesis MALS analysis was carried out in-line after size exclusion chromatography, in order to separate discrete oligomeric species prior to mass determination. Columns used are detailed in the results. In all cases, a 100 µl sample size was injected over the column. Concentrations of injected samples are detailed throughout the results. A Dawn Heleos-II, and Optilab-rex monitor (Wyatt) respectively measured the light scattering and refractive index. Data were analysed using ASTRA software (Wyatt) assuming a dn/dc value of 0.186 ml/g. MALS experiments were kindly run, and their results processed by Dr. Steven Johnson.

7.2: X-ray crystallography practice

7.2.1: Data Collection

X-ray crystallography centres on the observation of "reflections" of incident X-rays, which are diffracted from crystallographic lattice planes according to Bragg's law. In this thesis, all data were collected at a synchrotron radiation source using the rotation method, under cryo-conditions (100 ˚K). Crystals were mounted onto a goniometer and rotated about a fixed axis such that each reflection from each plane could be measured. Images are taken over a small oscillation range (typically 0.1-1.0˚) such that a complete dataset can be obtained without the overlapping of different spots. Indexing software can then automatically assign to each spot its Miller index. Indexing is initially carried out on a subset of scouting images, in order to deduce the lattice type. From this it is possible to create a strategy to collect a complete dataset, containing all of the possible unique reflections within a given resolution range, whilst limiting the exposure to X-rays and therefore radiation damage to the sample, if necessary. In particular, as data collection represents the last truly experimental step in crystallography, it is desirable to collect as complete a dataset as possible – i.e. measure all of the available reflections to as high a resolution as the crystal will diffract, and also to obtain an appropriate redundancy in the observations – the multiplicity.

$$\text{Completeness} = \frac{\text{Number of unique reflections observed}}{\text{Total possible unique reflections}}$$

$$\text{Multiplicity} = \frac{\text{Total number of observed reflections}}{\text{Number of unique reflections observed}}$$

7.2.2: Data Processing

Once a dataset has been collected, each reflection must be indexed to assign it to a particular spot on the reciprocal lattice (each corresponding to a particular set of lattice planes). In defining

the reciprocal lattice, the dimensions of the real-space unit cell are also determined. The intensity of each spot contains the diffracted information of the entire ASU content and so the intensity of each spot must then be determined. This step in known as integration, because the detector measures the intensity of each spot as a 3D Gaussian, and so the volume under this peak must be determined.

A processed dataset contains an intensity associated with each unique reflection that was observed. However, many of the observed reflections are multiple observations of a unique reflection. Additionally some reflections will be partially recorded across more than one image, and must be combined. This is especially the case when data is collected using small oscillations or “fine-slicing” which is beneficial to avoid the merging of adjacent spots, and increases the accuracy of the measurement of the reflection. In order to combine these reflections, it is first necessary to apply a scaling factor. Equivalent reflections should have an equal intensity, however due to various aspects such as beamline fluctuations, or detector settings this is not always the case, so these factors are corrected for. Finally, the scaled partial and multiply observed equivalent reflections are combined into a final dataset, a process known as merging.

Because multiple observations of each reflection are recorded in a diffraction experiment, the comparison of equivalent observations allows the calculation of several metrics that can describe the precision of the data. Analysis of these metrics can inform about the quality of the data as a function of resolution, or between diffraction images, or rotation ranges of the crystal. Throughout the thesis I report some of these metrics and so I will discuss them here. Note that due to the large number of unique reflections in a dataset, sets of reflections are usually grouped into similar resolution ranges or shells/bins, within which the errors should be on a similar magnitude. The metrics are therefore reported as an average for all the reflections in a

particular shell. Throughout the thesis I report the values averaged over the whole dataset, the lowest resolution (inner) shell, and the highest resolution (outer) shell.

$$\mathrm{R_{merge}} = \frac{\Sigma_{hkl}\Sigma_i|\boldsymbol{I_i} - \langle \boldsymbol{I_{hkl}} \rangle|}{\Sigma_{hkl}\Sigma_i\langle \boldsymbol{I_{hkl}} \rangle}$$

$$\mathrm{R_{pim}} = \frac{\Sigma_{hkl}\sqrt{\frac{1}{n-1}}\Sigma_i|\boldsymbol{I_i} - \langle \boldsymbol{I_{hkl}} \rangle|}{\Sigma_{hkl}\Sigma_i\langle \boldsymbol{I_{hkl}} \rangle}$$

R_{merge} describes the difference between intensities of individual observations compared to the mean for equivalent observations, expressed as a fraction of the mean. It is therefore a measure of the consistency within measurements of unique reflections. Although this value is usually reported, it has a main shortfall. Because these errors are summed, increased multiplicity (which should improve the precision of the dataset) causes the R_{merge} to rise. Simply – higher multiplicity datasets appear worse than low multiplicity datasets. The alternative metric R_{pim} takes data redundancy into account by down-weighting the contribution to the sum error for redundant observations (highlighted in grey). While R_{merge} describes the overall error, R_{pim} more accurately describes the precision of the averaged measurements.

Another data quality indicator is the CC½ value. In this case, unmerged reflections are randomly split into two half datasets, between which the intensities of equivalent reflections can be correlated, with the score expressed as a fraction of perfect correlation. This method should be completely independent of redundancy.

It is also important to consider the mean signal to noise ratio I/σI. A general rule of thumb is that data is unreliable when this value drops below ~2, however the other metrics should also be considered when making this judgement.

Higher-resolution reflections are relatively weaker and so have a lower signal to noise ratio, eventually resulting in statistically less consistent measurements for the outer shells. Analysis of these quality indicator metrics can therefore ultimately inform as to whether the inclusion of data beyond a particular resolution is useful in the dataset. However, they can also be used to decide whether certain images (batches) should be discarded from a dataset as reflection intensities in these batches may significantly deviate from equivalent reflections in other images.

Finally, the intensities of the measured spots are only proportional to the actual amplitude of the reflected X-rays:

$$|F_{hkl}| = \sqrt{\frac{I}{LP}}$$

The final step in processing is therefore to estimate these amplitudes (F) from the intensities (I). At this stage, corrections are also made for each unique reflection, based on geometry dependent Lorentz and Polarisation (LP) factors that are inherent in the experimental setup.

7.2.3: The phase problem

The final, processed dataset is a list of each unique reflection in terms of its position in the reciprocal lattice (Miller index), and its associated amplitude, which are the only measurable components in a diffraction experiment. However, this information describes only part of the "structure factor". The structure factor, F_{hkl} is the Fourier transform of the electron density of the real-space crystal lattice – this density being the ultimate goal for an X-ray diffraction experiment. The structure factor can also be expressed in terms of the atomic scattering components of the crystal lattice, however for simplicity, it is defined below in terms of electron density:

$$\mathrm{F_{hkl}} = \mathrm{v} \iiint_{0}^{1} \rho\,(x, y, z)\; e^{2\pi i(hx+ky+lz)} dx\; dy\; dz$$

However, as shown below, the full structure factor, for each separate reflection, comprises two components - the aforementioned amplitude, but also the phase (φ) of the reflection, which cannot be measured. This is the phase problem.

$$\mathrm{F_{hkl}} = |\mathrm{F_{hkl}}|\; e^{i\phi_{hkl}}$$

In order to produce an electron density map and ultimately build a structure, the phases must be "solved". This can be done via several techniques. I will summarise the techniques used in this thesis below.

7.2.4: Phase Solution via Molecular Replacement (MR)

This method was used for the solution of my SAS-6, and Ana2 CC structures. If we have a model of our asymmetric unit, and know how to arrange it in the lattice, then we can calculate the phases *in silico*, and use them to create the electron density map. Molecular replacement therefore works on the basis of creating a surrogate "search" model of the protein. Search models are often based on homologous protein structures with ≥30% sequence identity, however this is merely a guideline to suggest adequate structural similarity. In practise, any model can be used, as long as it adequately describes as much of the scattering from the real protein as possible. The phases and map can then be calculated from this model, using the experimentally observed amplitudes. However, this can only be done if the model is positioned in the unit cell – and therefore forms an identical crystal lattice - in the same way as the protein under study. The first step is to therefore align the model with your protein in real space – but – whereas the model is in real space, your observed data describing your protein lattice is in

reciprocal space. Both data forms can however be reconciled in the same form: the Patterson map.

The intensity of a reflection is dependent on the distances between in-plane atoms. Therefore, by setting all the phases to zero, the Fourier transform of only the intensity information gives rise to a map that describes only the interatomic vectors. This is the Patterson map, and can be calculated from both the model, and the observed intensities (**Figure 7.2.4.i**). The tails of each vector are positioned at the origin, where there is a large “origin” peak. At the head of each vector is a small peak with a magnitude proportional to the atomic mass of the atoms involved in that vector. The distance from the origin to the peak is equal to the interatomic vector. The map is also centrosymmetric, because each vector that maps atom *x* to atom *y*, has a corresponding opposite vector, mapping atom *y* to atom *x*.

In Molecular Replacement, a “search“ Patterson is calculated from the search model, and is aligned with the experimental Patterson in two steps, rotation, and then translation (**Figure 7.2.4.i**). Because the interatomic vectors describe real-space, these same rotation and translation operations will also place the search model in the correct place in the unit cell.

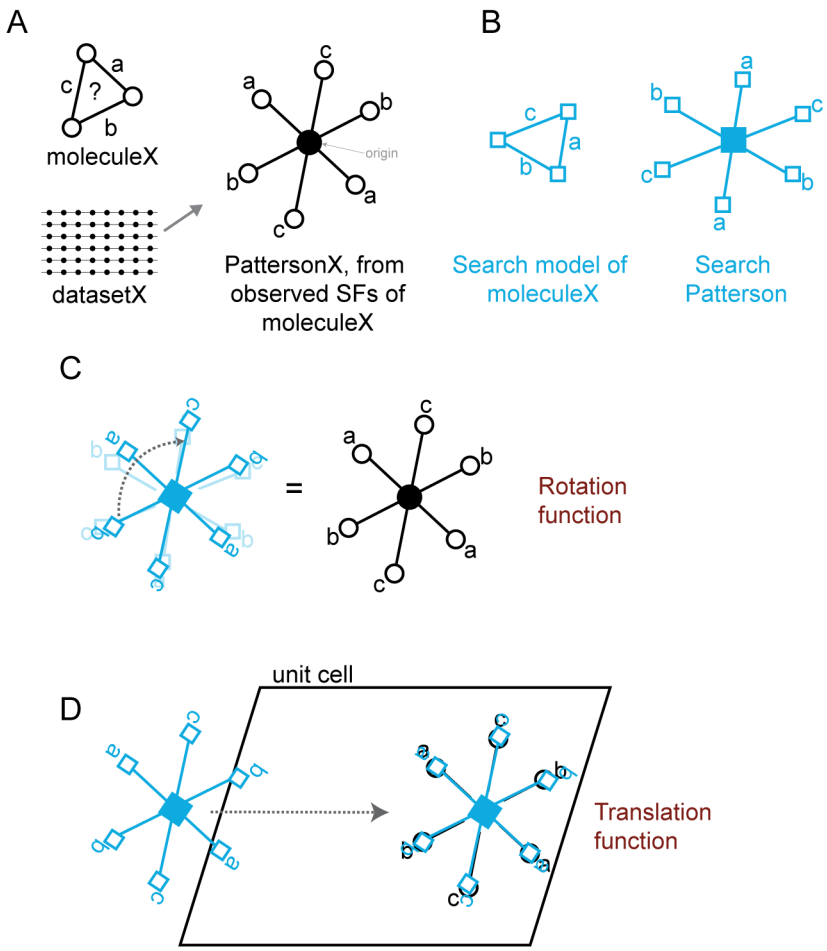


Figure 7.2.4.i: The Patterson function and molecular replacement. (**A**) Left, A simple 3-atom molecule "X" for which we have no crystal structure, and below, "datasetX" collected from moleculeX. Interatomic vectors are labelled a, b and c. Right, PattersonX, calculated from the Fourier transform of datasetX, without phases. (**B**) Left, a search model of moleculeX, and right, Patterson calculated from this model. (**C**) The Rotation Function, where the Search Patterson is rotated to match the orientation of PattersonX. (**D**) The Translation Function, where the correctly orientated Search Patterson is translated into the same position in the unit cell as PattersonX. Phase information can now be calculated from this solution.

MR programs therefore first perform a rotation function (RF). Once the correct rotation is deduced (or several, well scoring rotations), the translation function (TF) is performed. Although this process is largely automated, solutions output by the programs must be scrutinised by the end user according to certain metrics. In this thesis, I used two programs, PHASER [McCoy et al., 2007] and MOLREP [Vagin and Teplyakov, 2010]. PHASER uses a Log-likelihood-gain (LLG) function, which describes how much better the observed data are explained by the search model compared to a random atom model. It then scores the significance of this based on the Z score, which is a measure of how many standard deviations the solution's LLG is above the LLG

for randomly orientated models. The authors suggest a final TFZ score of ≥8 is a definite solution, however this should also be a standout solution[†]. In MOLREP, solutions are given a score based on the correlation coefficient between the observed intensities, and those generated by the solution, combined with a packing function, which describes any steric overlap resulting from the packing. This score is then given a "contrast" which describes how unique it is compared to the mean score of other solutions. The authors suggest a contrast of ≥2.5 describes a definite solution[‡]. Ultimately, any correct solution should improve during refinement, described in **Section 7.2.6**.

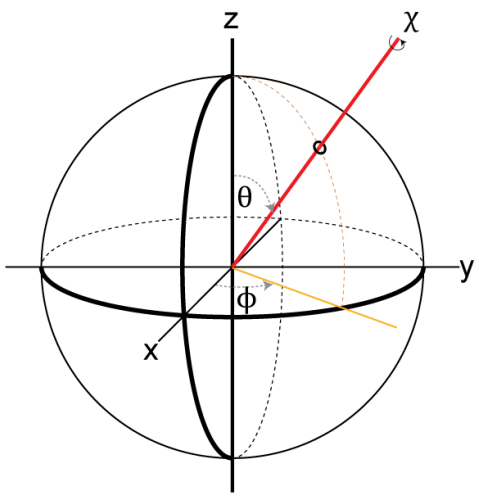


Figure 7.2.4.ii: The construction of the self-rotation function in spherical polar angles. The Patterson is rotated χ degrees about an axis, that is first defined by θ degrees from the z axis, then φ degrees from the x axis.

The Patterson can also give information about the arrangement of molecules in the unit cell, which is useful for guiding MR jobs. For this, a "self-rotation function" (SRF) is performed, whereby the Patterson map is rotated and superimposed with respect to itself, around a particular axis. In the presence of crystallographic, and non-crystallographic symmetry (NCS), the Patterson map should overlap with itself when rotated about a crystallographic or NCS symmetry axis, resulting in peaks in the SRF. The construction of the SRF is explained in **Figure 7.2.4.ii**, with the output typically viewed as a stereographic projection.

[†] http://www.phaser.cimr.cam.ac.uk/index.php/Molecular_Replacement

[‡] http://ccp4wiki.org/~ccp4wiki/wiki/index.php?title=Molecular_replacement_with_MolRep

7.2.5: Single or Multi-wavelength Anomalous Diffraction (SAD/MAD)

This technique was used to solve the structure of the Ana2:SAS-4 complex, and takes advantage of the different ways certain atoms can interact with X-ray energy. With normal protein atoms, each atom contributes equally to the phase shift experienced by X-rays when they are scattered by atoms in a protein crystal. Each reflection, (h, k, l) in a dataset is related to another at (-h, -k, -l) due to reflection from opposite sides of the same lattice plane. The dataset therefore has an inherent centrosymmetry. These reflections have the same amplitude, but have inverse phases, and this phenomenon is known as Friedel's law (**Figure 7.2.5.i, A**). The related reflections are known as as Friedel pairs, one denoted I+, the other I-.

However, it is possible to derivatise protein crystals with heavy, "anomalously" scattering atoms, which have electronic transition states corresponding to particular X-ray wavelengths. In this case, "anomalous" scattering can occur, where the X-ray can interact with the electronic oscillations of the anomalous scatterer. Normally the electronic property of the atomic scattering component can be described as f^0. In the case of anomalous scattering, this property is split into real and imaginary components, described by;

$$f^0 = f' + if''$$

where f' represents the real scattering component, while if'' describes the anomalous component. The result is that the reflection experiences a 90˚ phase lag. The phase contributions of the anomalous atoms to the I+ reflection, and the I- reflection are now no longer equivalent, resulting in slight amplitude and phase differences between the Friedel mates. Friedel's law is therefore broken, and this is most conveniently represented as an Argand diagram (**Figure 7.2.5.i**).

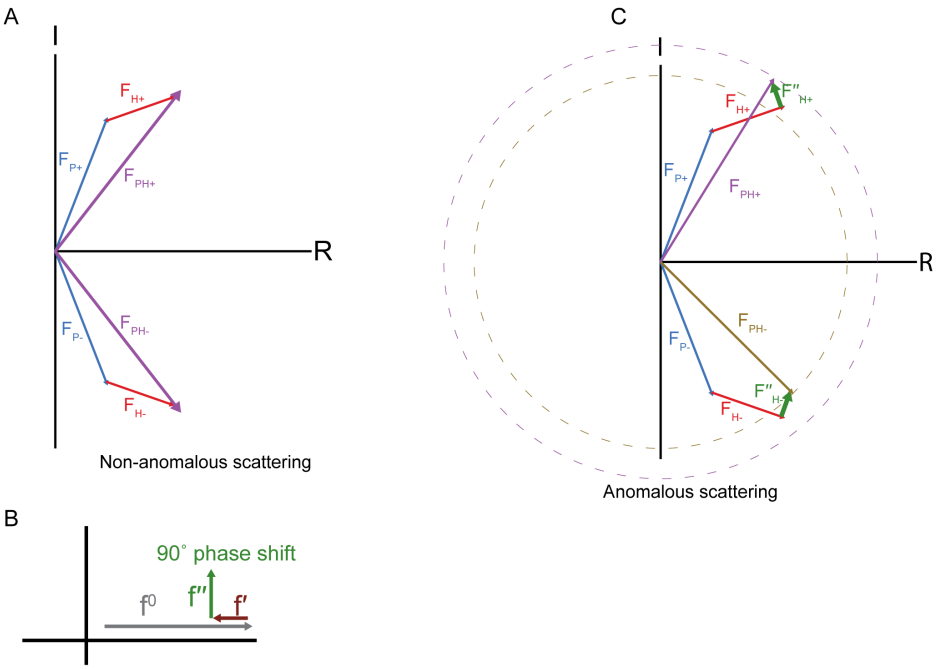


Figure 7.2.5.i: Breaking Friedel's law. (**A**) Argand diagram describing the structure factors of a pair of Friedel reflections from a heavy atom derivative. The amplitude and phase are the real and imaginary components respectively. Each reflection comprises scattering from all the protein atoms (F_P) and the heavy atoms (F_H), which are summed to form the final reflection (F_{PH}). The Friedel mates (F_{PH+} and F_{PH-}) have equal amplitude and opposite phase. (**B**) Under anomalous scattering conditions, the scattering from anomalous atoms takes on an anomalous component, f'', which contributes a 90˚ phase shift. (**C**) The anomalous scattering component causes the Friedel mates to no longer have equivalent phase or amplitude. Note that although the phase and magnitude of all contributions is shown, the only measurable components are the amplitudes of FPH+ and FPH-, which are illustrated by the dashed purple and gold circles, respectively.

Although the phase cannot be measured, the difference in intensity between Friedel pairs can be measured, known as the anomalous difference. As only the anomalous atoms contribute to this difference, this anomalous signal can be analysed separately from the rest of the dataset. Using Patterson methods, the locations of the heavy atoms can be located to define a heavy atom substructure. Once the locations of the heavy atoms are known, the structure factors of the heavy atoms can be calculated, and used to deduce the phase for each reflection. This process can also be represented as Argand diagrams, known as Harker constructions (**Figure 7.2.5.ii, A, B**). In the case of one anomalous dataset, there are two phase solutions for each reflection. This is the case in a single wavelength anomalous diffraction (SAD) experiment. This

ambiguity is best resolved by measuring an additional dataset at an alternative wavelength, which will have a different value for f' and f'' (**Figure 7.2.5.ii, C**). Once the phases have been estimated for each reflection, the full structure factor is known, and an electron density map corresponding to the entire ASU can be calculated (experimental density), into which the structure can be built.

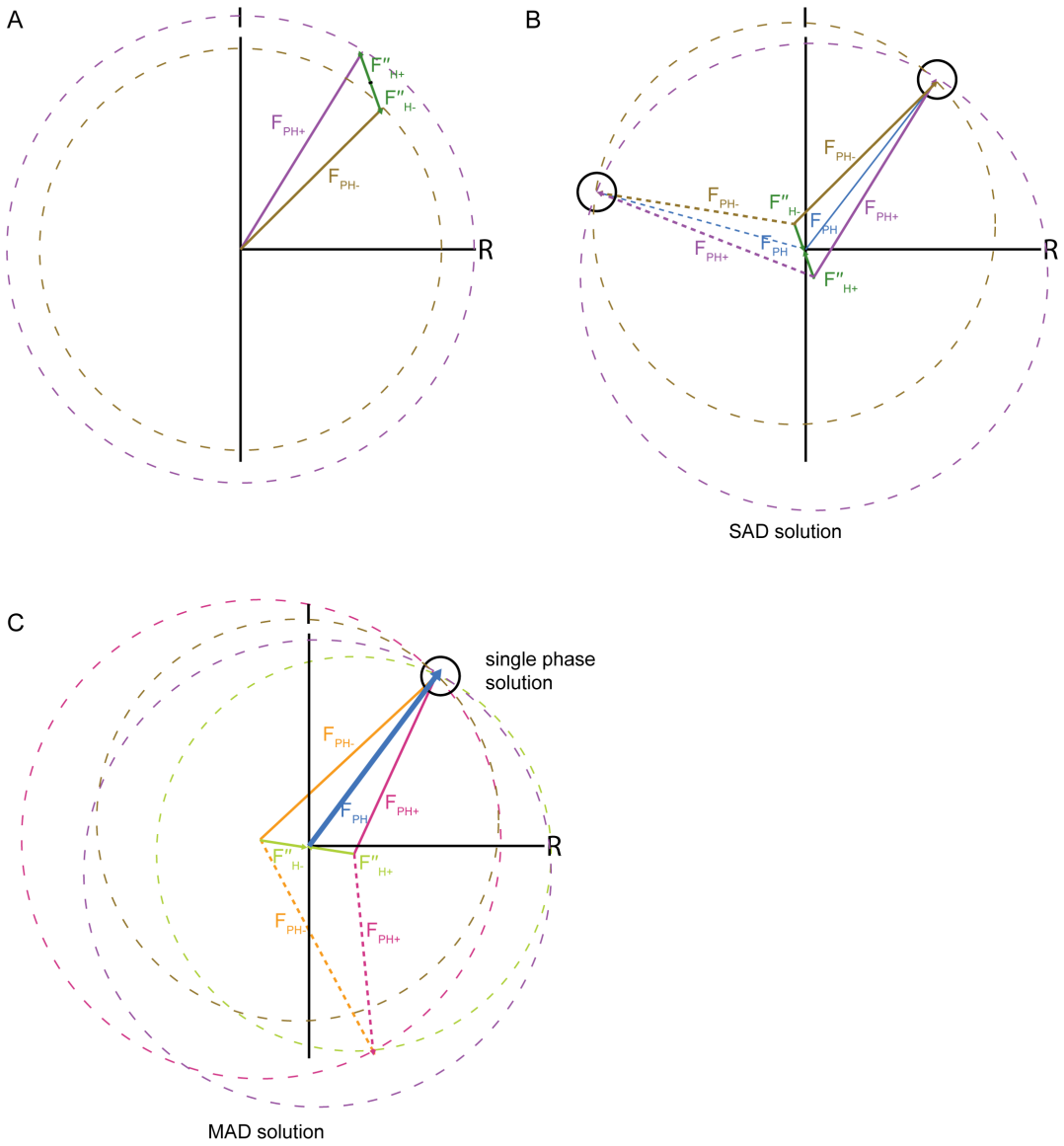


Figure 7.2.5.ii: Phase solution via SAD and MAD. (**A**) Information available in a SAD experiment, after solution of the HA substructure. The magnitudes of F_{PH+} and F_{PH-} have been measured (dashed circles) and also now, the phased, anomalous components of the structure factors, (F''_{H+} and F''_{H-}) have been calculated from the known coordinates of the heavy atom substructure. (**B**) The magnitudes of F_{PH+} and F_{PH-} can now be re-plotted from the ends of the F''_{H+} and F''_{H-} vectors. This causes them to converge at two positions, either of which could explain a phase solution for the non-anomalous structure factor, (F_{PH}). (**C**) In a MAD experiment, this process is repeated with a second anomalous wavelength with a different F'' contribution. By combining this new information with the original SAD solution, the answer can be narrowed to a single phase solution.

7.2.6: Refinement

During refinement, various parameters of the structure can be adjusted. These include atomic positions, the B-factors or "atomic displacement parameters", and the occupancy, (if for instance a particular residue is seen to adopt multiple conformations, it must only fractionally occupy each conformation). These parameters will all affect the atomic scattering components of the structure factor, and so are modified in order to best describe the observed diffraction data. During refinement, phases are recalculated from the improved structure, which is now a more accurate source of phase information than the original phase solution. A new electron density map can then be calculated using these phases, and the original observed structure factor amplitudes. Phase improvement leads to a more accurate map, into which the structure can be more accurately modelled, and then re-refined, and so refinement is an iterative procedure. Most maps are not of high enough detail to refine against without the use of "restraints". These are *a priori* assumptions that can be made about the general chemistry of the protein, such as bond lengths and angles. Additionally, where there are multiple copies of a molecule in the unit cell, assuming they should be similar, NCS restraints can be used to aid refinement by comparing the different copies.

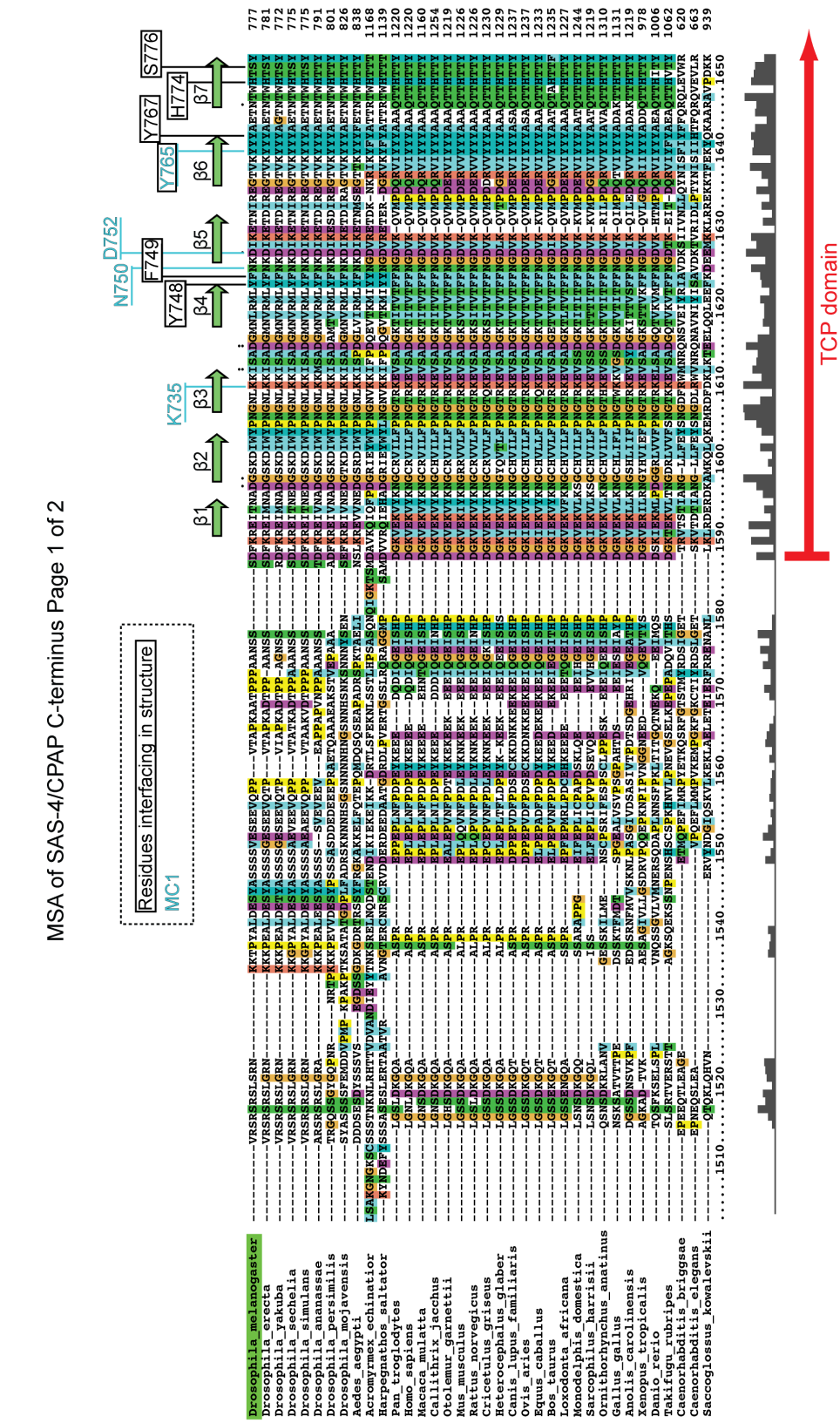
The progress of refinement is monitored by the R factor, which describes the difference between the experimentally observed amplitudes, and those calculated from the model:

$$R = \frac{\sum_{hkl} ||\boldsymbol{F_{obs}}| - |\boldsymbol{F_{calc}}||}{\sum_{hkl} |\boldsymbol{F_{obs}}|}$$

However, this system is prone to model bias, as phases from the model are used to calculate the map into which the model is built. This can result in a model that is overly modelled to reduce the R factor without being more accurate. In order to assess this bias, ~5% of all reflections in the dataset are not included in the refinement process (the "test set" or "R free"

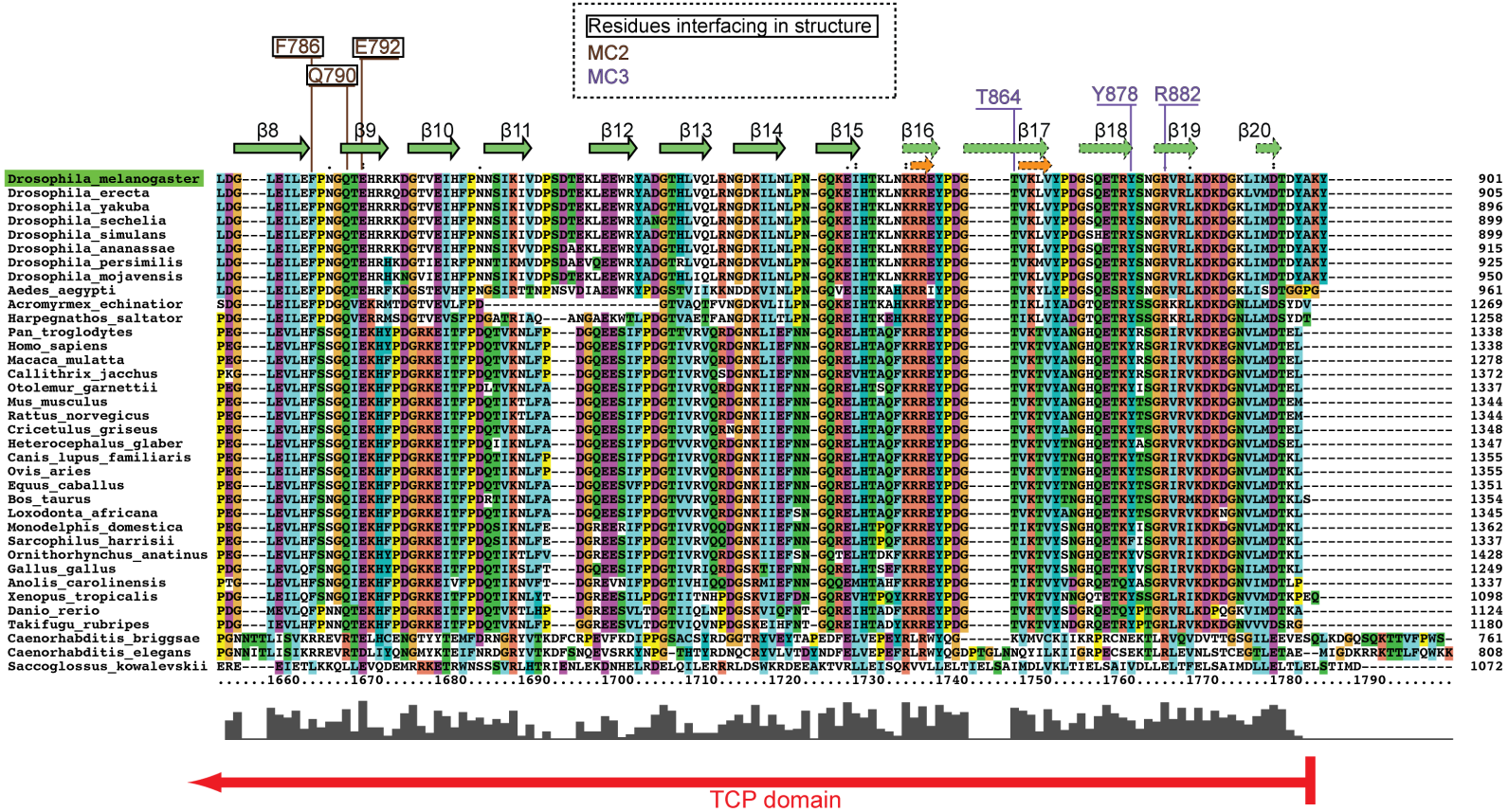
set). If the model is truly accurate, it will be able to predict these missing reflections, and again this is measured with the "R_{free}". The remaining reflections that are used for refinement are known as the working set. Throughout this thesis I therefore quote two values, the R_{work}, and the R_{free}.

7.3: MSA of SAS-4 orthologue TCP domains



1st part of TCP domain. Solid green sheets were determined experimentally in this study.
Black boxed residues are those seen to be interfacing with Ana2 in the structure presented in this study.
Cyan underlined residues are those mutated in mutation cluster 1 (MC1).

MSA of SAS-4/CPAP C-terminus Page 2 of 2



2nd part of TCP domain. Solid green sheets were determined experimentally in this study.
Dashed green arrows represent beta sheets predicted based on homology to *D.rerio*
CPAP. The orange arrows represent particularly short strands predicted by psipred.
Black boxed residues are those seen to be interfacing with Ana2
in the structure presented in this study. Brown underlined residues are those mutated in mutation cluster 2 (MC2).
Purple underlined residues are those mutated in mutation cluster 3 (MC3)

7.4: Purification of the Ana2 N-terminus + Coiled coil (Ana2$^{1-228}$)

Unsuccessful attempts were made to purify Ana2_1-122, and Ana2_1-177, however it was possible to purify Ana2$^{1-228}$, covering the CR2, to the end of the CC (**Figure + Table 7.4.i**). pETM-14_Ana2_1-228 was transformed into B834 (DE3) cells. Cells were grown in TB at 37 ˚C until an OD_{600} > 0.6. Recombinant protein expression was induced using 1 mM IPTG, with overnight growth at 21 ˚C. Cells were harvested by centrifugation, and then washed in PBS before being pelleted and frozen at -80 ˚C until purification.

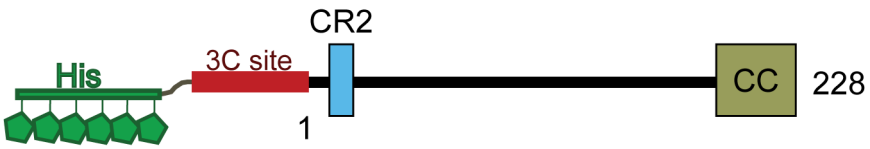


Figure 7.4.i: Schematic of His-Ana2$^{1-228}$

Amino Acids	Number Cys	MW	pI	ε_{280}	Abs 0.1%
248	4	27410.7	6.59	12950	0.473

Table 7.4.i Details of His-Ana2$^{1-228}$ calculated using Protparam.

Purification working buffer (WB) was 50 mM Tris, pH 8.0, 300 mM NaCl, 1 mM βME). Cells were lysed according to the standard protocol (**Section 7.1.4**). The supernatant was supplemented with WB + 1M imidazole, to 20 mM imidazole before loading onto a 1 ml Ni-NTA cartridge, pre-equilibrated in WB + 20 mM imidazole using a peristaltic pump. The Ni-NTA cartridge was transferred to an FPLC system for washing and elution (**Figure 7.4.ii**). The column was washed with WB + 50 mM imidazole for 25 CV. Ana2$^{1-228}$ had a peculiar affinity for Ni-NTA, as shown in (**Figure 7.4.ii**), with full-length and fairly pure protein eluting across the entirety of a 50 mM – 1 M imidazole gradient as oppose to eluting in a discrete peak. In subsequent purification, elution was carried out using a 1 M imidazole step. Upon dialysis into fresh working buffer, the protein crashed out of solution almost immediately and couldn't be retrieved. It could not be deduced whether the imidazole provided stability through high ionic strength (hence the inclusion of 300 mM NaCl in the WB) or through a particular interaction specific to imidazole. Due to the high Ni-NTA affinity and tendency to aggregate in this manner, it is possible that

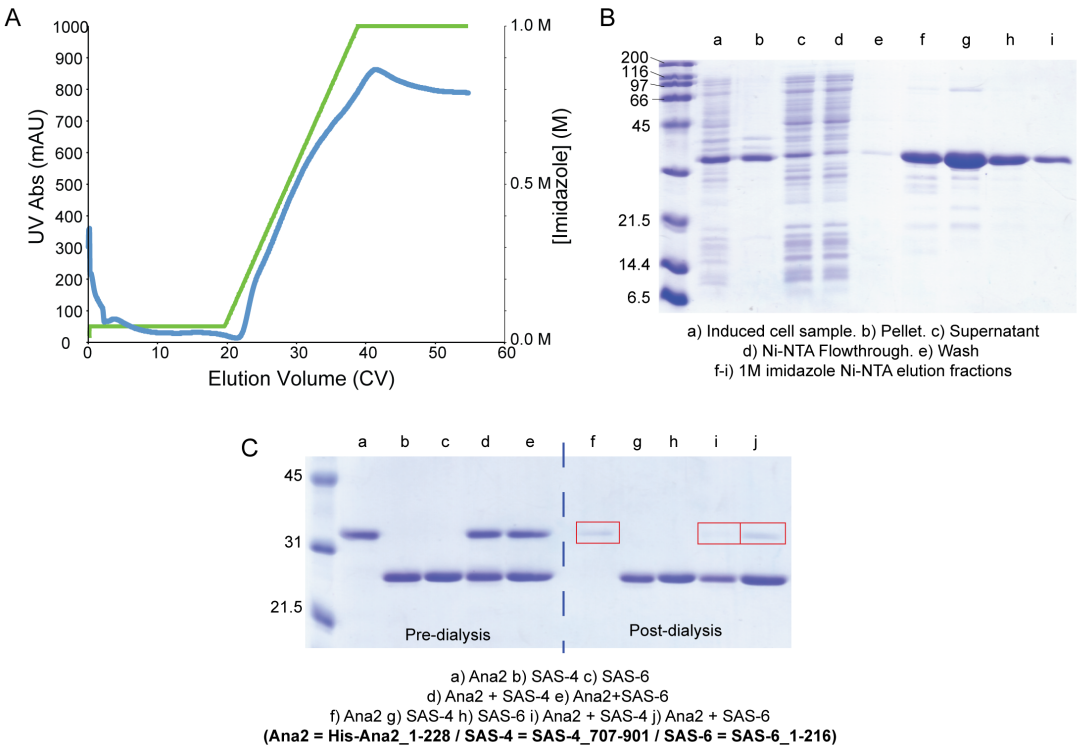


Figure 7.4.ii: Purification of His-Ana2$^{1-228}$. (**A**) Elution of His-Ana2$^{1-228}$ from Ni-NTA, using a linear imidazole gradient (green). Blue = UV absorption. (**B**) 15% SDS-PAGE gel showing purification fractions obtained using a static 1 M imidazole elution step. Elution fractions were typically fairly pure and contained full-length protein. (**C**) 15% SDS-PAGE analysis of an attempt to stabilise Ana2 in solution using SAS-4 and SAS-6. Ana2 was only soluble in the presence of Imidazole. It can be seen that once imidazole has been removed via dialysis, bands representing Ana2$^{1-228}$ (in the red boxes) are greatly diminished. Therefore the presence of SAS-4 or SAS-6 does not appear to increase the stability of Ana2$^{1-228}$.

Ana2$^{1-228}$ self-associates. Further experiments were conducted in the presence of imidazole. As Ana2$^{1-228}$ was already fairly pure, no further purification steps were taken. An attempt was made to see if SAS-6$^{1-216}$ or the SAS-4 TCP domain (residues 707-901) could stabilise the Ana2 in solution. SAS-6 or SAS-4 were mixed with the Ana2 in equimolar ratios, in WB, 500 mM imidazole. Imidazole was then dialysed away against fresh WB. SDS-PAGE (**Figure 7.4.ii, C**) was used to assess the presence of Ana2 before and after dialysis. In all cases, Ana2 was largely depleted from solution after dialysis, suggesting SAS-4 and SAS-6 did not stabilise the construct. Further work on this construct was abandoned, although it was used in a far-western blot experiment described in **Chapter 3, Section 3.3.1.**

7.5: Attempts to characterise the Ana2:SAS-4 interaction *in vitro*

Following the solution of the structure of the Ana2:SAS-4 complex, attempts were made to deduce the Kd of this interaction. Attempts were ultimately unsuccessful via ITC and SPR, however it was possible to show that the WT SAS-4 TCP domain bound the Ana2 CR2 with a greater response than SAS-4 TCP domain mutants.

7.5.1: Generation of MBP-fusion proteins

The SAS-4$^{700\text{-}901}$ construct was sparingly soluble and purifications produced poor yields – typically one purification would generate enough protein for 1-2 ITC experiments, which was not sufficient considering that this technique needed optimisation. In an attempt to increase both the yield and solubility of the construct, it was fused to MBP, by cloning the codon optimised SAS-4$^{700\text{-}901}$ cDNA into the pETM-44 bacterial expression vector (**Figure 7.5.1.i, A**). The MBP also allows for affinity purification using dextrin sepharose. The MCPH mutation, E792V, and separately the Y765V mutation, were introduced to the pETM-44_MBP-SAS-4$^{700\text{-}901}$ vector using site-directed mutagenesis (QuikChange II XL, Agilent), according to the manufacturer's protocol. As a control, MBP itself was generated by insertion of two STOP codons, TAATAA, immediately before the SAS-4 open reading frame such that the resultant protein contained all elements of the construct, bar only the TCP domain. Details of these constructs are outlined in **Table 7.5.1.i**. Plasmids encoding these constructs were transformed into *E. coli* strain B834 (DE3) cells, and

Construct	Amino Acids	Number Cys	MW (Da)	pI	ε_{280}	Abs 0.1%
MBP-SAS-4^{WT}	591	0	66131.7	5.95	100730	1.524
MBP-SAS-4^{E792V}	591	0	66101.7	6.03	100730	1.525
MBP-SAS-4^{Y765V}	591	0	66067.6	5.95	99240	1.503
MBP	387	0	42672.5	5.49	66350	1.556

Table 7.5.1.i Details of MBP-SAS-4$^{700\text{-}901}$ fusion proteins and MBP calculated using Protparam.

grown in 1 L LB culture at 37 ˚C, until an OD_{600} of ~0.6 was reached. Protein expression was induced via the addition of 1 mM IPTG at 21 ˚C, overnight. Cells were harvested by centrifugation, washed in PBS and frozen until purification. All MBP-SAS-$4^{700-901}$ proteins (and MBP alone) were purified in the same manner, using a working buffer (WB) of 50 mM $Na_xH_xPO_4$, pH 7.4, 150 mM NaCl. Cells were lysed using the standard procedure (**Appendix 7.1.4**). Clarified supernatant was loaded onto a 5 ml MBPtrap (GE Healthcare) equilibrated in WB, at 5 ml/min using a peristaltic pump. The column was washed with 10 CVs of WB, before elution using WB + 10 mM maltose. All protein usually eluted within 4-5 ml. SDS-PGE analysis indicated that elution fractions were very pure, containing only fusion protein, and free MBP (**Figure 7.5.1.i, C**). Elution fractions containing the fusion protein were then further purified by SEC, using an S200 16/60 column (GE Healthcare), and WB as the running buffer. Invariably there was a void peak, containing (presumably aggregated) MBP-SAS-$4^{700-901}$, a large peak containing MBP-SAS-$4^{700-901}$, and an ensuing small peak or shoulder containing free MBP (**Figure 7.5.1.i, D**). Pooled fractions containing fusion protein were snap frozen in liquid nitrogen before storage at -80 ˚C. Typical yields were ~30 mg protein per litre of bacterial culture, although this differed accordingly with each protein's solubility limit. During concentration of these proteins, it was noticed that MBP-SAS-$4^{700-901}$(WT) had a solubility limit of ~5.5 mg/ml (~85 µM). In molar terms, this is about equivalent to the solubility limit of the SAS-$4^{700-901}$ construct in isolation. As such, MBP fusion appears only to increase purification yield, and not the solubility of the construct, which seemed to be dominated by the TCP domain. However, the MBP-fusion appeared to be more stable, with less material being lost during processes such as protein concentration.

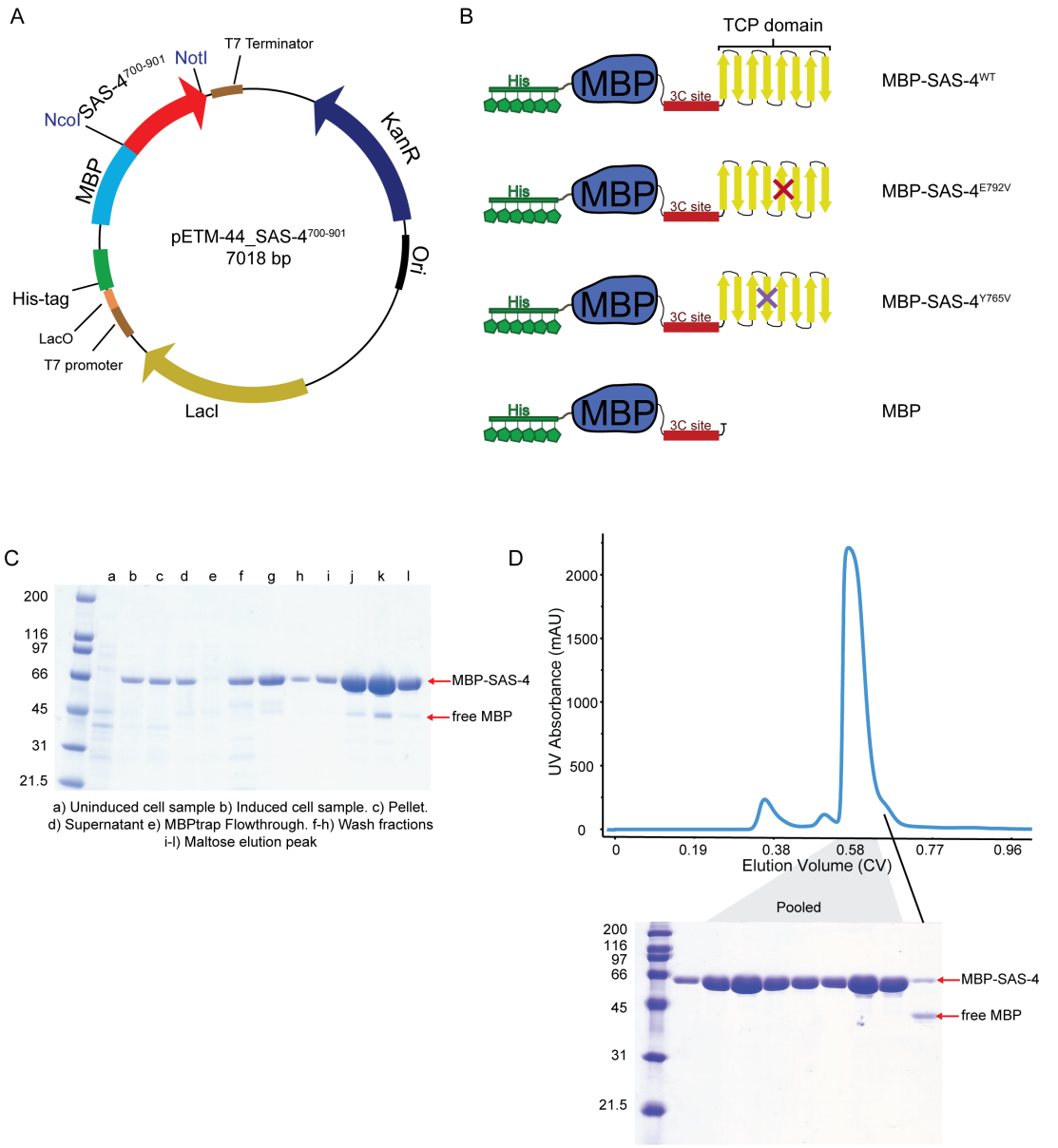


Figure 7.5.1.i: Design and purification of MBP-SAS-4$^{700-901}$ fusion proteins. (**A**) The pETM-44 expression vector, with cDNA encoding SAS-4$^{700-901}$ cloned in-frame after MBP. (**B**) Schematics of the MBP fusion proteins produced from the pETM-44 plasmids. (**C**) SDS-PAGE analysis of a purification of MBP-SAS-4^{WT}. This is representative of all the fusion proteins. (**D**) SEC trace from an S200 16/60 column and subsequent SDS-PAGE analysis, showing a main peak containing MBP-SAS-4.

7.5.2: Production of non-tagged mutant TCP domains for circular dichroism analysis

If the point mutations disrupt the fold of the TCP domain, then it would be impossible to conclude the significance of any observed loss of binding in the point mutants. Therefore, it was important to confirm that the mutant TCP domains could still fold similarly to the wild-type protein. Circular dichroism spectroscopy was therefore carried out on the mutant proteins in order to analyse if their spectra deviated from that of the wild-type protein. These mutants included MC1, MC2, MC3, MC123, E792V and Y765V. Obviously, it was necessary to purify the SAS-4 TCP domains in isolation in order to carry out these experiments. Although it was possible to make MBP-fusion constructs and cleave the MBP, it was not possible to separate it via reverse Ni-NTA, SEC or IEX methods (not shown). To isolate the SAS-4 TCP domains alone, it was easiest to purify the His-tagged versions of the constructs, cleave the His-tag, and purify by SEC chromatography as described for wild-type SAS-4$^{700\text{-}901}$ in **Chapter 3 Section 3.2.2**. The details of cleaved TCP domain constructs are shown in **Table 7.5.2.i**. Aside from the pH of the WB, all mutant proteins were purified as described for SAS-4$^{700\text{-}901}$(WT). It was not possible to purify large amounts of these proteins, and in all cases, the large majority of the protein was contained in the insoluble fraction, as was the case for SAS-4$^{700\text{-}901}$(WT). In the case of SAS-4$^{700\text{-}901}$(MC1) and SAS-4$^{700\text{-}901}$(MC123) it was not possible to purify any soluble protein, and so CD spectroscopy could not be carried out on these two mutant constructs.

SAS-4$^{700\text{-}901}$ mutant	Amino Acids	MW (Da)	pI	ε_{280}	Abs 0.1%	WB pH	Possible to purify?
WT	206	23631.4	7.98	34380	1.455	8.5	Yes
E792V	206	23601.4	8.55	34380	1.457	7.5	Yes
Y765V	206	23567.3	7.99	32890	1.396	8.5	Yes
MC1	206	23507.3	8.00	32890	1.399	8.5	No
MC2	206	23524.3	8.55	34380	1.461	7.5	Yes
MC3	206	23508.3	7.15	32890	1.399	8.5	Yes
MC123	206	23277.2	8.00	31400	1.349	8.5	No

Table 7.5.2.i Details of His-tag-cleaved SAS-4$^{700\text{-}901}$ WT and mutants calculated using protparam. No constructs contained any cysteines. Note that the calculated pI of these constructs is very close to physiological pH, which may account for their low yield and solubility and in some cases, inability to purify any protein altogether. In each case, the working buffer (WB) was 50 mM Tris, 150 mM NaCl. The pH of WB used for each construct is indicated, and was chosen based on the pI, as well as ensuring that histidines would be sufficiently charged for Ni-NTA affinity chromatography.

Purified SAS-4$^{700-901}$ constructs were diluted to ~100 µg/ml and dialysed overnight using a 10 kDa MWCO dialysis membrane (Pierce), into 10 mM $Na_xH_xPO_4$, with a pH matching the respective WB, outlined in **Table 7.5.2.i**.). CD Spectra were measured at room temperature on either a Jasco j-720, or a Jasco j-815 machine, and were averaged over 4 accumulations, normalised against a blank measurement taken with dialysis buffer. The spectra for the mutant domains are compared to the wild-type spectra (from **Chapter 3 Section 3.2.4**) in **Figure 7.5.2.i**, and can be seen to be highly similar, with all spectra following a predominantly beta-type profile. These results suggest that these single point mutations still allow the domain to fold in a wild-type manner. Any observed loss of binding in affinity measurement techniques must therefore result from the specific loss of these residues as oppose to misfolding. It must be borne in mind that

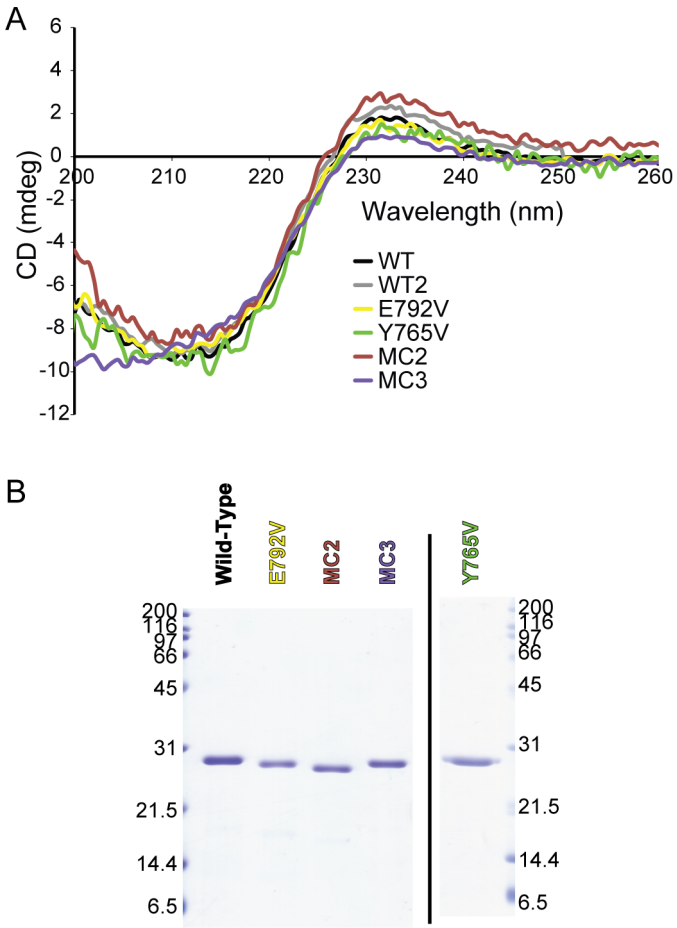


Figure 7.5.2.i: Circular Dichroism analysis of SAS-4$^{700-901}$ point mutants. (**A**) Buffer subtracted CD traces of SAS-4$^{700-901}$ (WT) and the indicated mutants, all indicating a predominantly β type fold. To correct for concentration differences, an arbitrary correction factor is applied to some traces in order to allow comparison between curves. Two separate wild-type preps are shown. (**B**) Coomassie-stained SDS-PAGE gels showing the purified recombinant proteins.

the inability to produce MC1 or MC123 may suggest that these mutants did not fold properly, however this is not possible to conclude.

7.5.3: Surface Plasmon Resonance (SPR) experiments.

Because SAS-4 and Ana2 did not appear to be amenable to ITC experimentation (**Chapter 3 Section 3.3.2**) I tried to use an alternative technique, Surface Plasmon Resonance (SPR), to characterise the interaction between the SAS-4 TCP domain and the Ana2 CR2 domain. This technique requires the coupling a protein (bait) to the surface of an SPR sensor chip, and measuring the binding response when a ligand (prey) is flowed through a flow cell, over the chip surface. The response is proportional to the mass of material that has bound the bait. SPR experiments were conducted using the MBP-SAS-$4^{700-901}$ fusion proteins described in **Section 7.5.1**, and synthetically generated peptides spanning Ana2 residues 5-43. A Biacore 3000 machine was used, with CM5 sensor chips (Biacore). These chips have four available flow cells, usually three experimental (Fc1, Fc2, Fc3) and a reference or baseline flow cell (Fc4), to which no bait (or inert bait) is attached. The prey can be flowed through all flow cells in series. The chip surface is coated with a dextran matrix to which bait can be attached. All experiments were carried out in HBS-EP buffer (GE Healthcare), pH 7.4.

Initially, MBP-SAS-$4^{700-901}$ proteins were immobilised directly onto the sensor chip surface via primary amine coupling (**Figure 7.5.3.i, A**). This has the caveat that where there are multiple primary amines (lysine residues, and N terminal NH_2 group), the bait can be attached to the chip in multiple conformations, some of which may obscure binding sites. MBP-SAS-$4^{700-901}$(WT), MBP-SAS-$4^{700-901}$(E792V) and MBP-SAS-$4^{700-901}$(Y765V) were immobilised to Fc1, 2 and 3 respectively, while MBP alone was immobilised to Fc4 as a control. Approximately 6000 RU of bait was attached for each protein, however it was noticed that the MBP-SAS-$4^{700-901}$ proteins coupled much more efficiently than did MBP alone, for which only ~2000 RU could be attached. This suggests that the TCP domain (which contains 20 lysine residues, distributed on either face of the β-sheet) couples much more efficiently to the chip surface than MBP. Coupled in this way,

the TCP domain may not be well presented to lumen of the flow cell, and access to the peptide would be limited (**Figure 7.5.3.i, A**). Indeed, upon flowing over Ana2$^{5-43}$, the response recorded was very weak and so this approach was abandoned.

In order to present the TCP domain more efficiently to the lumen of the flow cell, the next approach sought to couple the MBP-SAS-4$^{700-901}$ proteins indirectly. Rabbit Polyclonal αMBP antibodies (courtesy of whoever made DTACC Abs) were first coupled to the chip surface of all flow cells via primary amine coupling, reaching a total of ~5500-9000 RU of antibody per flow cell. The MBP-SAS-4 fusion proteins were then immobilised by binding to the coupled antibody (MBP-SAS-4$^{700-901}$(WT), MBP-SAS-4$^{700-901}$(E792V), MBP-SAS-4$^{700-901}$(Y765V) and MBP only, were immobilised to Fc1, 2, 3 and 4 respectively) (**Figure 7.5.3.i, B**). Unfortunately, only ~500-2000 RU of each MBP fusion protein could be coupled in this manner, suggesting that the antibody was not effectively immobilising the MBP fusion proteins, and thus limited the number of available binding sites on the chip surface. It is possible that only a small proportion of antibody was appropriately orientated to bind to MBP. Again, flowing over the peptide resulted in only a very weak response, and so this approach was abandoned.

The molecular weight of the Ana2$^{5-43}$ peptide is relatively small at 4172 Da. This inherently limits the response that can be seen from its binding to the chip surface, and so would require a large number of binding sites on the chip to see a significant response. Additionally, it is difficult to measure the concentration of the peptide, as it does not absorb at 280 nm. It was therefore decided that it would be easier to couple the Ana2 peptide as bait, and measure the binding of the comparatively larger MBP-SAS-4$^{700-901}$ proteins, the concentrations of which could also be more accurately measured, using the A_{280}.

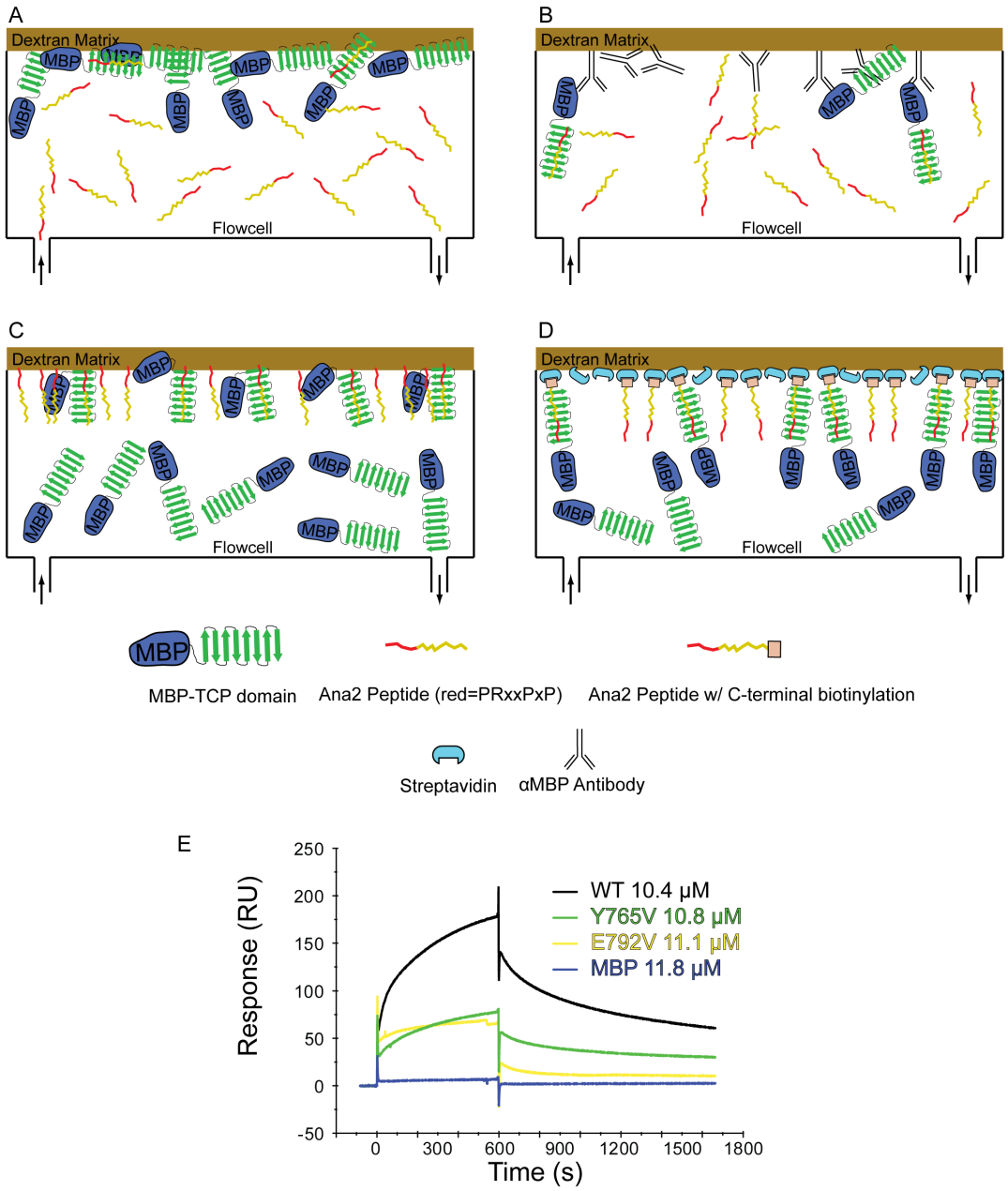


Figure 7.5.3.i: SPR analysis of the Ana2:SAS-4 interaction. (**A-D**) Different experimental approaches attempted. Molecules in these representations are labelled in the legend below. (**A**) Direct coupling of MBP-SAS-$4^{700-901}$ as bait, and $Ana2^{5-43}$ as prey. It is likely that most MBP-SAS-4700-901 molecules coupled via the SAS-$4^{700-901}$ part of the molecule. (**B**) Indirect coupling of MBP-SAS-$4^{700-901}$ as bait via binding to immobilised αMBP antibodies. This method was inefficient, possibly due to the orientation of the antibodies on the chip surface. (**C**) Direct coupling of $Ana2^{5-43}$ as bait via the N-terminus. The core binding motif (red) was bound in close proximity to the chip surface. (**D**). Indirect coupling of a biotinylated version of $Ana2^{5-43}$ to immobilised streptavidin. (**E**) Background subtracted SPR response traces from approach **C**, using ~equimolar amounts of MBP-SAS-4 variants. The wild-type protein clearly has a higher binding response than either the MCPH mutant, E792V, or the Y765V mutant. The MBP alone has a negligible signal in comparison.

Initially, the peptide was coupled directly using primary amine chemistry. As Ana2$^{5-43}$ contains no lysine residues, there is only one primary amine, at the N-terminus, which means the peptide would couple in a uniform orientation. Unfortunately, the N-terminus of the peptide contains the conserved binding element, ETEDPRxxPxP, which would be in close proximity to the chip surface, and may therefore be obscured from binding. Binding would also require the N-terminus of the TCP domain to come into close proximity to the chip surface – difficult, because MBP is fused at this terminus (**Figure 7.5.3.i, C**). Ana2$^{5-43}$ was coupled to Fc1 (~3600 RU) Fc2 (~1050 RU) and Fc3 (~490 RU) in order to trial how many binding sites would be required on the chip to observe a robust binding response. Fc4 was left blank as a control. MBP-SAS-4 proteins could then be flowed over each channel at a range of concentrations. Using this approach, the binding response was much stronger than the previous attempts. However, there were difficulties with dissociation of the MBP-SAS-4$^{700-901}$ proteins (not shown), which meant that repeat experiments were difficult to interpret. A new chip was generated and a short experiment was carried out, injecting MBP-SAS-4$^{700-901(}$WT), MBP-SAS-4$^{700-901}$(E792V) and MBP-SAS-4$^{700-901}$(Y765V) and MBP over the chip at ~equimolar amounts, in order to directly compare the magnitudes of binding response from the different proteins. This approach showed quite clearly that the wild-type protein gave a higher response than either of the mutants, and the MBP control gave essentially no response (**Figure 7.5.3.i, E,** as shown in **Chapter 3 Figure 3.10.i**).

In a final approach, Ana2$^{5-43}$ was synthesised (Peptide Protein Research Ltd. UK), that had a C-terminal biotinylation. Using Primary amine coupling, a CM5 chip was coated with streptavidin to ~ 4000-6000 RU across all Fcs. Biotinylated Ana2$^{5-43}$ was then coupled indirectly, via binding of the biotin to the immobilised streptavidin, to Fc1, 2 and 3 (~900 RU for each, to obtain triplicate data). No peptide was coupled to Fc4, to enable its use as a reference channel. In this approach, the peptide orientation was reversed, allowing much better access for binding, due to the exposure of the N-terminus of the peptide to the flow cell lumen, and a correspondingly

convenient geometry for binding by MBP-SAS-4$^{700-901}$ (**Figure 7.5.3.i, D**). Initially, binding of MBP-SAS-4$^{700-901}$(WT) was analysed, by injecting it over the chip at a range of concentrations, from 0.625-80 μM. This approach enabled the binding responses at each concentration to reach equilibrium, however again, there were issues with the dissociation of MBP-SAS-4 from the chip, meaning that data over repeat experiments were difficult to process (not shown).

Together, the ITC and SPR results presented in this thesis suggested that it would not be trivial to characterise the affinity of the *Drosophila* Ana2:SAS-4 interaction. The Ana2 CR2 is highly hydrophobic, and the TCP domain appears to be inherently unstable in the absence of the CR2. Although desirable to show in the *Drosophila* system, the *in vitro* binding of the complex had been investigated in the more amenable *D. rerio* system by my collaborator, and as the system appeared to be highly orthologous, I abandoned further attempts at characterising the interaction in vitro. As the molecular details of the interaction were known from the crystal structure, the interaction was dissected *in vivo* (**Chapter 3 Section 3.11**).

7.6: Localisation of GFP-Ana2$^{1-229}$ and Ana2$^{230-420}$-GFP in RNA injected embryos

These results were described in **Chapter 4**, **Section 4.3**, although these particular images were not shown. GFP-Ana2$^{1-229}$ behaved similarly to GFP-Ana2$^{1-315}$. Ana2$^{230-420}$-GFP behaved similarly to Ana2$^{316-420}$-GFP.

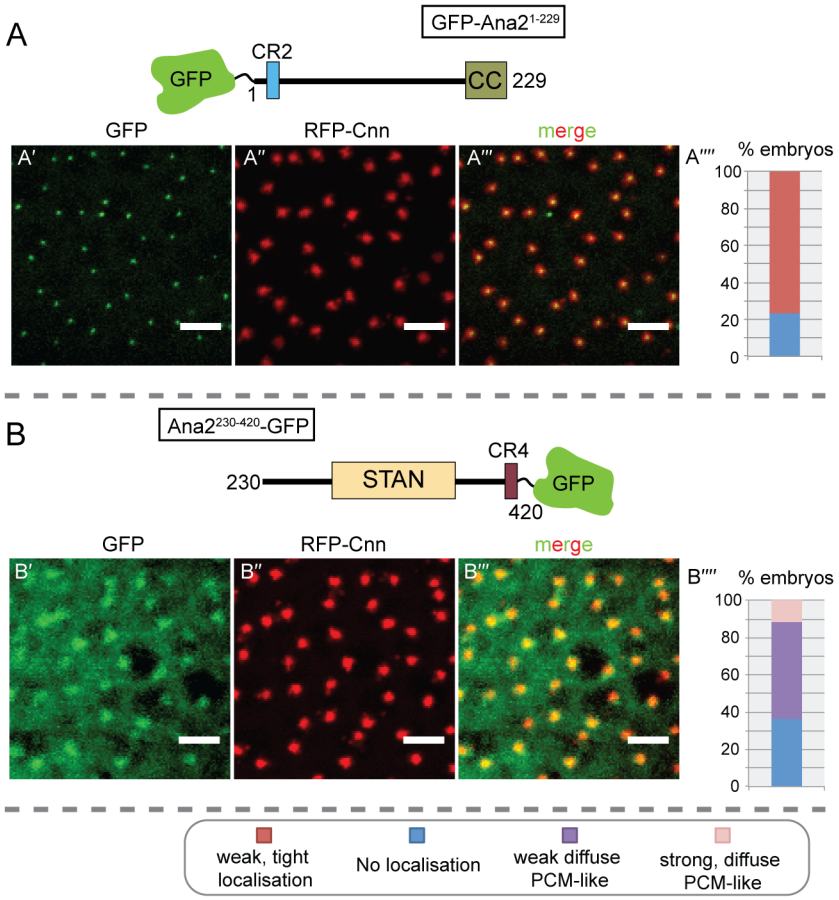


Figure 7.6.i: Localisation of GFP-Ana2$^{1-229}$ and Ana2$^{230-420}$-GFP in RNA-injected embryos. (**A**) GFP-Ana2$^{1-229}$ localised tightly, but weakly to the centriole. A representative image is shown for embryos where localisation was evident. (**B**) Ana2$^{230-420}$-GFP had a STAN attributable, weak and diffuse, PCM-like localisation, for which a representative image is shown. Note that the image had been adjusted to see the weak GFP signal. To the right in either panel is a bar chart detailing the percentage of embryos where particular type of localisation were observed, coloured according to the legend below. Scale bars are 5 µm.

7.7: Characterisation of the Ana2 C-terminal domains

7.7.1: Purification of Ana2$^{236-420}$ containing the STAN + CR4

Ana2$^{236-420}$, containing the conserved STAN and CR4 domains, was PCR amplified from *E. coli* codon-optimised Ana2 cDNA (GeneArt), and inserted into the pETM-14 vector via InFusion cloning (Clontech, **Section 7.1.1**). This plasmid would produce an N-terminal, 3C-cleavable His-tagged construct (**Figure + Table 7.7.1.i**).

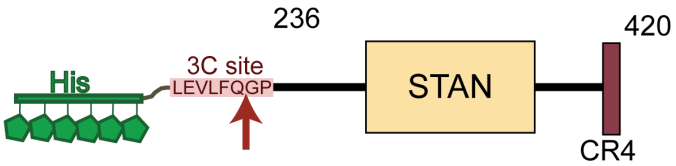


Figure 7.7.1.i: Schematic of His-Ana2$^{236-420}$, showing the 3C cleavage site.

Amino Acids	Number Cys	MW	pI	ε_{280}	Abs 0.1%
189	2	20972	8.56	8940	0.426

Table 7.7.1.i Details of the cleaved Ana2$^{236-420}$ protein calculated using Protparam.

pETM-14_Ana2$^{236-420}$ was transformed into *E. coli* B834 (DE3) cells, which were grown in TB at 37 ˚C until the OD_{600} was ~0.6. Protein production was induced overnight at 21 ˚C using 1 mM IPTG. Cells were harvested, and washed in PBS before being pelleted and stored at -80 ˚C. . The calculated pI of the His-tagged construct is fairly basic at 8.52. A working buffer (50 mM Tris, pH 7.5, 150 mM NaCl, 1 mM DTT) pH of 7.5 was chosen in the hope that the target would be charged, whilst ensuring the pH was not too low for Ni-NTA purification. Cells were lysed according to the standard protocol (**Appendix 7.1.4**). The supernatant was supplemented to 20 mM imidazole before loading onto a 1 ml Ni-NTA cartridge, pre-equilibrated in WB + 20 mM imidazole. The Ni-NTA cartridge was transferred to an FPLC system for washing and elution (**Figure 7.7.1.ii, A**). The column was washed with WB + 50 mM imidazole for 40 CV before elution with a gradient of imidazole, 50 mM to 300 mM over 30 CVs. His-Ana2$^{236-420}$ eluted in a discrete peak, and although the predominant species was full-length protein, gel fractions

(**Figure 7.7.1.ii, B**) indicated some contaminants or degradation products were present. To cleave the His-tag, 3C was added at a ratio of 1 mg 3C : 10 mg target and the reaction was dialysed overnight against fresh WB. The reaction was concentrated to 5ml using a 10 kDa MWCO centrifugal spin filter, and spun to remove a small amount of precipitate, before loading onto an S75 16/60 column (GE Healthcare) in order to perform SEC using WB for the running buffer (**Figure 7.7.1.iii, A**).

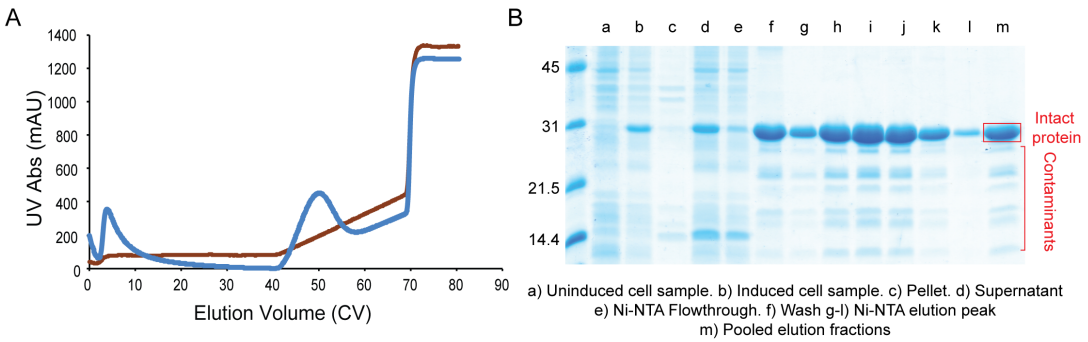


Figure 7.7.1.ii: Purification of Ana2$^{236-420}$. (A) Elution of His-Ana2$^{236-420}$ from Ni-NTA, using a linear imidazole gradient. Blue = UV absorption, brown = conductivity, effectively showing the concentration change of imidazole. (B) 15% SDS-PAGE gel showing purification fractions. Elution fractions are predominantly full length protein, however contaminants or degradation products are present in the sample.

Ana2$^{236-420}$ eluted as a broad peak with multiple shoulders, perhaps explicable by the fact that the protein is predicted to be predominantly unfolded, or by the presence of degradation products. SDS-PAGE analysis showed His-tag cleavage was complete, but separation from 3C protease was not entirely successful. Degradation products were only partially separated. In order to enrich the sample for full-length protein, IEX chromatography was used. SEC fractions were dialysed into a low ionic strength buffer (25 mM MES pH 6.3, 50 mM NaCl). Under these conditions the Ana2$^{236-420}$ is positively charged and will bind to a cation exchange resin. This pH is close to the pI of GST-3C, and so it was unlikely that 3C would bind the column. Cation exchange was carried out using two buffers, A (25 mM MES pH 6.3, 1 mM DTT) and B (buffer A plus 2 M NaCl). A MonoS 5/50 GL column (GE Healthcare) was equilibrated to 50 mM NaCl for 20 CV before loading the protein sample. Some protein flowed through the column, which was then washed for a further 20 CV at 50 mM NaCl. A linear gradient of NaCl from 50-500 mM over 40 CV

was then used for elution. Enriched full-length Ana2$^{236-420}$ eluted as a sharp peak with a mid-point of ~108 mM NaCl. Various contaminants eluted in other small peaks. SDS-PAGE analysis showed that enrichment was successful, although there was still some slight contamination. Electrospray-ionisation Mass Spectrometry analysis confirmed the dominant species was indeed Ana2$^{236-420}$; although another smaller species was present. (**Figure 7.7.1.iii).**

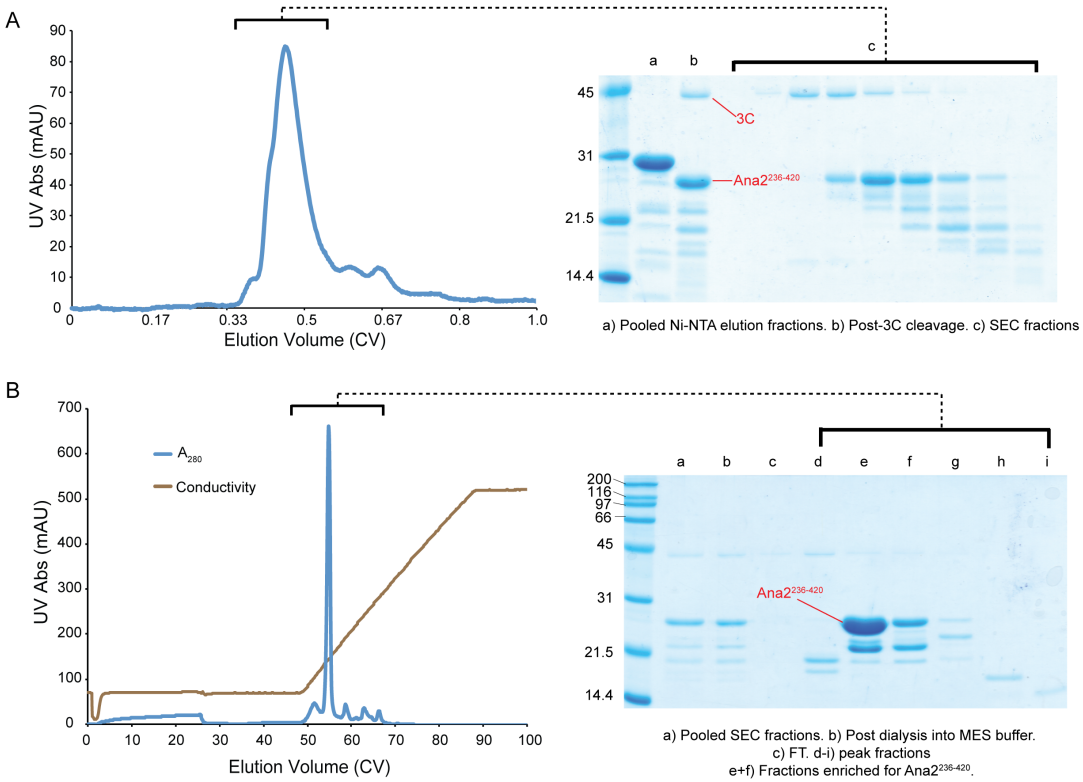


Figure 7.7.1.iii: SEC and IEX of Ana2$^{236-420}$. (**A**) S75 16/60 elution profile and 15% SDS-PAGE analysis of cleaved Ana2$^{236-420}$, showing incomplete separation of Ana2$^{236-420}$. (B) IEX chromatography 15% SDS-PAGE gel showing purification fractions. Fractions e + f were most enriched for full-length Ana2$^{236-420}$ and were pooled for further study.

7.7.2: Circular Dichroism analysis of Ana2$^{236-420}$

IEX Purified Ana2$^{236-420}$ was diluted to ~400 µg/ml and dialysed at 4 ˚C using a Pierce Mini Dialysis Device (MWCO 10 kDa) against 10 mM $Na_xH_xPO_4$, pH 7.0. An average spectrum over four accumulations was measured at room temperature, between 190 and 260 nm, and normalised against a blank measurement taken with dialysis buffer. The minimum trough in the spectrum

was ~200 nm, and the trace was also negative between 210 and 230 nm. Canonical random-coil spectra minima are usually <200 nm, and are positive from 210 – 230 nm. These features suggest that Ana2$^{236-420}$ is largely unstructured, but it is possible that there is some alpha helical influence on the spectrum. It is therefore possible that Ana2$^{236-420}$ has some level of secondary structure, consistent with the PSIPRED prediction of helices in the STAN domain (discussed in **Chapter 4**). Because this construct was heterogeneous and largely unstructured, further study was abandoned. A new construct has since been designed containing only the STAN domain, however this has yet to be tested.

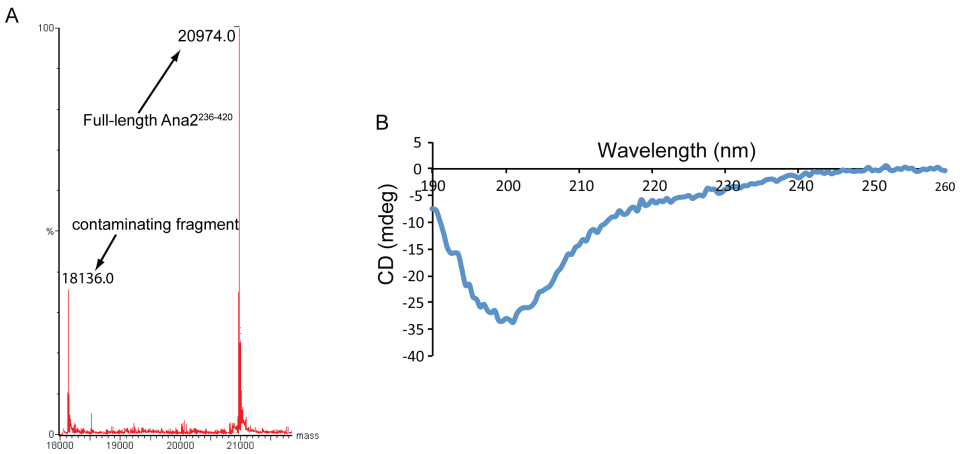


Figure 7.7.2.i: MS and CD of Ana2$^{236-420}$. (**A**) Mass Spectrometry analysis of enriched Ana2$^{236-420}$. The main peak had a mass of 20974 Da, only 2 Da from the calculated mass of the full-length construct. (The lesser peak is presumably the main degradation product visible by SDS-PAGE analysis in **Figure 7.7.1.iii, B**). (**B**) Background subtracted circular dichroism spectrum of Ana2$^{236-420}$. The spectrum is most similar to that of a random coil, however it may show some alpha helical influence.

7.8: Trial MST experiment between the Ana2 coiled-coil and SAS-6$^{313-432}$

My observation that Ana2_CC forms a highly stable tetramer may have important biological implications. The converged region of SAS-6 is also likely to be a parallel tetramer, or dimer of dimers. This stoichiometric similarity, combined with these two proteins' colocalisation to the centriolar lumen may not be a coincindence. I was therefore interested to see whether Ana2_CC can interact with a conserved region of SAS-6, likely to lie in the converged region of the cartwheel spoke (**Chapter 2**, **Section 2.7.3**). A preliminary Microscale Thermophoresis (MST) experiment was therefore carried out to investigate this. Purified SAS-6$^{313-432}$-GFP (**Chapter 2**, **Section 2.8.3**) and Ana2_CC (**Chapter 5**, **Section 5.2.4**) were dialysed into 50 mM Tris pH 7.5, 150 mM NaCl, 1 mM DTT, 0.05% Tween-20, overnight at 4 ˚C. A 1:2 dilution series of Ana2_CC was prepared, from 7000 µM, 3500 µM, 1750 µM etc. to 0.42 µM. 20 µl of each Ana2_CC dilution was added to 20 µl aliquots of 0.07 µM SAS-6$^{313-432}$-GFP (a concentration that was found to give a suitable GFP fluorescence signal for MST, of ~ 1000 AU). These samples were then loaded into standard MST capillaries (NantoTemper) via capillary action. MST scans were carried out using 30-second laser pulses at 20% and 40% laser power, with a 20% LED illumination using a NanoTemper Monolith NT.115. The MST timetraces (**Figure 7.8.1.i, A**) indicated smooth diffusion was taking place via negative thermophoresis and no aggregation was evident. Analysis of the normalised thermophoresis and T-jump (**Figure 7.8.1.i, B**) indicated that there was no obvious difference in thermophoretic behaviour of SAS-6$^{313-432}$-GFP with increasing Ana2_CC concentration. This suggests there is no interaction. In the case of an interaction, one would expect a difference in thermophoretic behaviour between bound and unbound species that is distinct from the noise of the baseline. Interestingly, analysis of the baseline GFP fluorescence signal (before MST is carried out) showed that despite equimolar amounts of SAS-6-GFP being present in each sample, fluorescence increased correlatively with the concentration of Ana2_CC (**Figure 7.6.1.i, C**). This could also be seen in the cap scan (not shown). This is unlikely be due to pipetting error, because SAS-6-GFP was aliquotted independently of the preparation of Ana2_CC

dilutions. Additionally, the change in fluorescence is progressive, and quite dramatic. (There is >50% increase in SAS-$6^{313-432}$-GFP fluorescence comparing the 0 µM Ana2_CC capillary to the 3500 µM Ana2_CC capillary). It is known that binding in close proximity to a fluorophore can affect fluorescence, and so it is therefore possible that this effect is due to binding. However, one would expect that such a binding event as an Ana2 tetramer to a SAS-6 molecule might drastically alter the thermophoretic behaviour. It is also possible that Ana2_CC may interact with the capillary, or even the eppendorf tubes used to mix the two proteins, resulting in an artefactual enrichment of SAS-6-GFP in the capillary lumen. These data are therefore ambiguous, and so in order to follow up this question, it might be wise to opt for a different interaction method such as SPR or ITC.

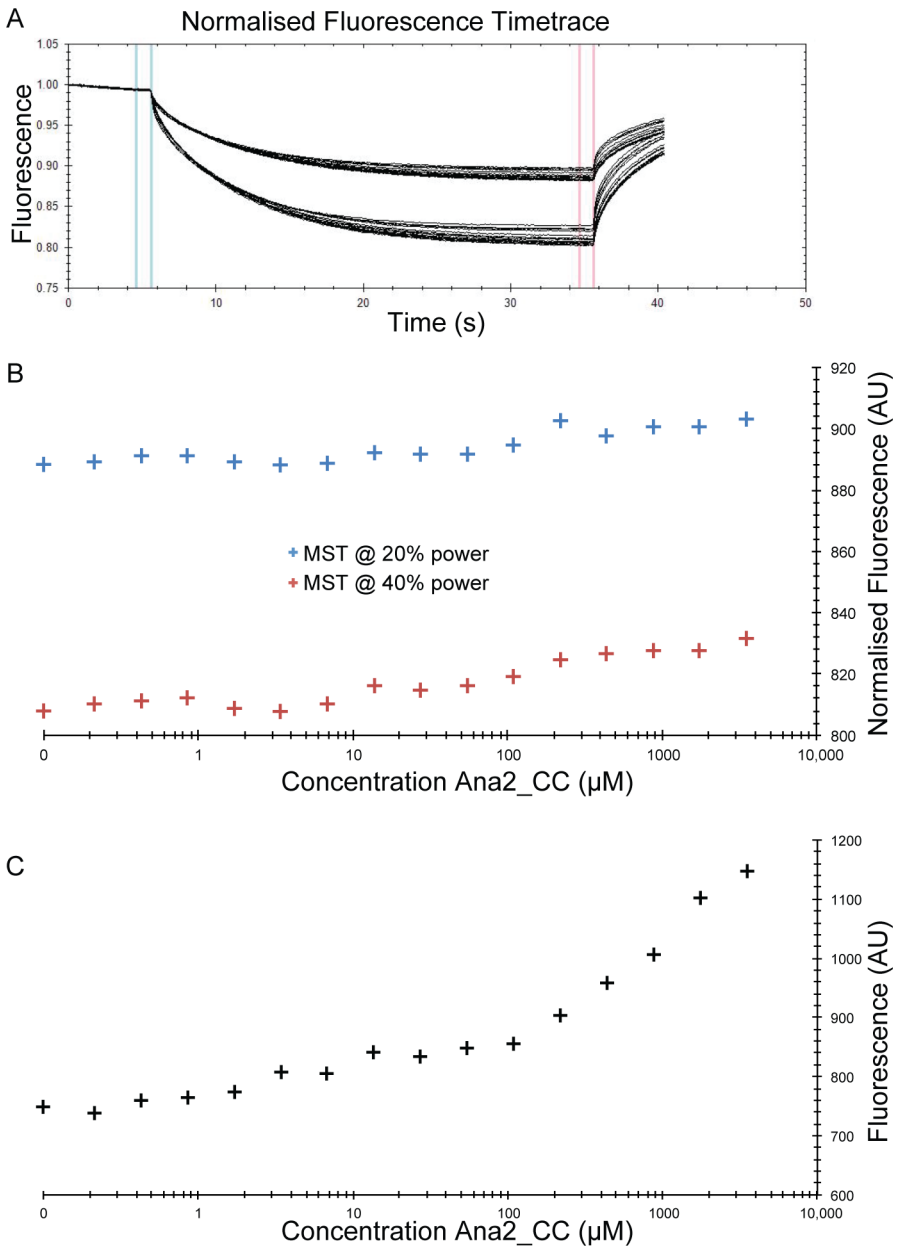


Figure 7.8.1.i: MST to investigate whether Ana2_CC interacts with a conserved region of the SAS-6 spoke (SAS-6$^{313-432}$-GFP). (**A**) MST timetraces, normalised to start at 1.0. Blue lines indicate where "cold" fluorescence measurements were taken. Red lines indicate "hot" fluorescence after thermophoresis has reached equilibrium. The traces are smooth and continuous, indicating no aggregation is occurring. (**B**) Thermophoresis + T-jump fluorescence signal vs Ana2_CC concentration, at two different MST powers. There appears to be no significant change in thermophoresis with ligand concentration, indicating no complex formation. One would expect complex formation to cause an obvious shift in fluorescence. (**C**) "Cold" fluorescence signal vs Ana2_CC concentration before the MST measurement. The presence of Ana2_CC is clearly affecting fluorescence in some manner (>50% difference in SAS-6-GFP fluorescence between 0 μM Ana2, and 3500 μM Ana2).

7.9: Optimisation screen for Ana2_CC crystallisation

In **Chapter 5 Section 5.3.1**, an optimisation screen was produced in order to obtain crystals of the Ana2 CC construct, and the details are displayed here. The pH was screened between 7.2 and 8.2 via the mixing of two stocks of HEPES, at pH 7.2 and 8.2 respectively. The PEG550 MME concentration was optimised in one half of the screen, while the PEG600 concentration was optimised in the other half.

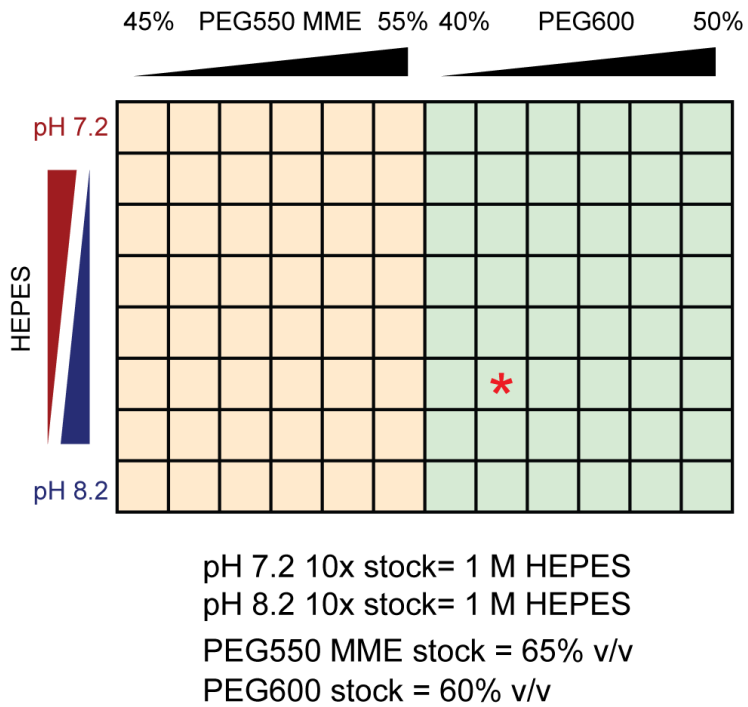


Figure 7.9.i: Optimisation screen for the crystallisation of Ana2 CC. pH was varied between 7.2 and 8.2 along the y axis. The final concentration of HEPES was 100 mM, although the exact pH in each screen in unknown. PEG550 MME and PEG600 concentrations were varied evenly between the indicated values. The red asterisk indicates the condition in which crystals 4 and 5 (**Chapter 5**, **Table 5.3.1**) were grown, however the screen produced large, high diffraction quality crystals in many drops.

7.10: Processing and refinement of Ana2_CC anomalous signal

In **Chapter 5, Section 5.4.2**, I described how I searched for possible chloride ions in the Ana2 CC structure, using anomalous signal processed from some datasets. I discuss the processing of this data here.

In keeping with the high-resolution of the structure, many ordered waters were observed, often with high coordination and low B-factors. The strength of density at these sites suggested that some of these waters might in fact be chloride ions. The correct replacement of these atoms would more accurately explain the data and further improve the maps, and so I wanted to see if I could find evidence of chlorine, using anomalous signal. Chlorine and sulphur are able to provide anomalous signal in x-ray diffraction experiments and are used for S/Cl site finding, to aid SAD phasing. Although the optimal resonant frequency for S/Cl k-shell absorption lies at >4 Å, the X-ray absorption edge is not sharp, and anomalous signal (albeit weak) can still be observed at smaller wavelengths. An attempt was therefore made to see if anomalous signal could be found at the highest wavelengths for which Ana2_CC diffraction data had been collected. If useful anomalous signal could be found, it could potentially be used to unambiguously identify chloride ions in the anomalous difference density maps. 1.54 Å is typically used to search for S/Cl anomalous signal, however the longest wavelength for which I had data was 0.9200 Å. Four sweeps of three crystals were indexed and processed using XIA2, mode 3daii, searching for consistent Bijvoet pair amplitude differences. This approach generated a dataset that was complete to 1.06 Å (Dataset A2CCwt_4(anom), **Chapter 5**, **Table 5.3.2.i**). Despite the non-optimal wavelength, Analysis by AIMLESS [Evans and Murshudov, 2013] suggested that the anomalous signal was significant, as judged by the RMS correlation ratio. The resolution bins covering 37.24-3.87 Å were most significant, and anomalous correlation beyond this resolution was low, although still considered by AIMLESS to be just significant. Perhaps the

relatively high intensity and multiplicity of the data meant that the anomalous differences – even though weak – were precisely measured, resulting in these significant correlations. The anomalous dataset, A2CCwt_4(anom) was then reindexed to match the 0.8 Å dataset A2CCwt_3, and adopt its FreeR set reflections. In order to generate an anomalous difference map, phases were first obtained by refining the A2CCwt model against dataset A2CCwt_4(anom) using REFMAC5. The resultant phase information (PHWT and FOM) from the refined "A2CCwt_4(anom).mtz" was then copied back into dataset A2CCwt_4(anom) using CCP4_CAD, to ensure that experimental amplitudes were used for map generation, thereby limiting bias. A phased anomalous difference map was then generated using CCP4_FFT, feeding in the phase information (PHWT, FOM), and the anomalous amplitude information (DANO, SIGDANO). The only peak that stood out over the noise had a height of 11.33σ and was located over the cysteine sulphur (F**igure 5.4.2.i, D**). This particularly ordered cysteine (γ-sulphur B-factor = 5.56) was located near to the crystallographic four-fold c axis, and faces towards the centre of the tetramer. As no other significant peaks were found in the anomalous map, it was not possible to locate ordered chloride ions, and as such, none were placed in the model.

References

Abrahams, J.P., and Leslie, A.G.W. (1996). Methods used in the structure determination of bovine mitochondrial F1 ATPase. Acta Crystallogr. Sect. D *52*, 30–42.

Afonine, P.V., Grosse-Kunstleve, R.W., Echols, N., Headd, J.J., Moriarty, N.W., Mustyakimov, M., Terwilliger, T.C., Urzhumtsev, A., Zwart, P.H., and Adams, P.D. (2012). Towards automated crystallographic structure refinement with *phenix.refine*. Acta Crystallogr. D Biol. Crystallogr. *68*, 352–367.

Afzelius, B. (2004). Cilia-related diseases. J. Pathol. *204*, 470–477.

Allen, R.D. (1969). The Morphogenesis of Basal Bodies and Accessory Structures of the Cortex of the Ciliated Protozoan Tetrahymena Pyriformis. J. Cell Biol. *40*, 716–733.

Alves-Cruzeiro, J.M. da C., Nogales-Cadenas, R., and Pascual-Montano, A.D. (2013). CentrosomeDB: a new generation of the centrosomal proteins database for Human and Drosophila melanogaster. Nucleic Acids Res. gkt1126.

Alvey, P. (1986). Do adult centrioles contain cartwheels and lie at right angles to each other? Cell Biol. Int. Rep. *10*, 589–598.

Andersen, J.S., Wilkinson, C.J., Mayor, T., Mortensen, P., Nigg, E.A., and Mann, M. (2003). Proteomic characterization of the human centrosome by protein correlation profiling. Nature *426*, 570–574.

Anderson, R.G.W., and Brenner, R.M. (1971). The Formation of Basal Bodies (centrioles) in the Rhesus Monkey Oviduct. J. Cell Biol. *50*, 10–34.

Aplan, P.D., Lombardi, D.P., and Kirsch, I.R. (1991). Structural characterization of SIL, a gene frequently disrupted in T-cell acute lymphoblastic leukemia. Mol. Cell. Biol. *11*, 5462–5469.

Arquint, C., and Nigg, E.A. (2014). STIL Microcephaly Mutations Interfere with APC/C-Mediated Degradation and Cause Centriole Amplification. Curr. Biol. *24*, 351–360.

Arquint, C., Sonnen, K.F., Stierhof, Y.-D., and Nigg, E.A. (2012). Cell-cycle-regulated expression of STIL controls centriole number in human cells. J. Cell Sci.

Awad, S., Al-Dosari, M.S., Al-Yacoub, N., Colak, D., Salih, M.A., Alkuraya, F.S., and Poizat, C. (2013). Mutation in PHC1 implicates chromatin remodeling in primary microcephaly pathogenesis. Hum. Mol. Genet. *22*, 2200–2213.

Azimzadeh, J., and Marshall, W.F. (2010). Building the Centriole. Curr. Biol. *20*, R816–R825.

Balestra, F.R., Strnad, P., Flückiger, I., and Gönczy, P. (2013). Discovering Regulators of Centriole Biogenesis through siRNA-Based Functional Genomics in Human Cells. Dev. Cell *25*, 555–571.

Barr, A.R., Kilmartin, J.V., and Gergely, F. (2010). CDK5RAP2 functions in centrosome to spindle pole attachment and DNA damage response. J. Cell Biol. *189*, 23–39.

Basto, R., Lau, J., Vinogradova, T., Gardiol, A., Woods, C.G., Khodjakov, A., and Raff, J.W. (2006). Flies without Centrioles. Cell *125*, 1375–1386.

Basto, R., Brunk, K., Vinadogrova, T., Peel, N., Franz, A., Khodjakov, A., and Raff, J.W. (2008). Centrosome Amplification Can Initiate Tumorigenesis in Flies. Cell *133*, 1032–1042.

Bazzi, H., and Anderson, K.V. (2014). Acentriolar mitosis activates a p53-dependent apoptosis pathway in the mouse embryo. Proc. Natl. Acad. Sci. *111*, E1491–E1500.

Beisson, J., and Wright, M. (2003). Basal body/centriole assembly and continuity. Curr. Opin. Cell Biol. *15*, 96–104.

Bettencourt-Dias, M., and Glover, D.M. (2007). Centrosome biogenesis and function: centrosomics brings new understanding. Nat Rev Mol Cell Biol *8*, 451–463.

Bettencourt-Dias, M., Rodrigues-Martins, A., Carpenter, L., Riparbelli, M., Lehmann, L., Gatt, M.K., Carmo, N., Balloux, F., Callaini, G., and Glover, D.M. (2005). SAK/PLK4 is required for centriole duplication and flagella development. Curr. Biol. CB *15*, 2199–2207.

Bettencourt-Dias, M., Hildebrandt, F., Pellman, D., Woods, G., and Godinho, S.A. (2011). Centrosomes and cilia in human disease. Trends Genet. *27*, 307–315.

Bilgüvar, K., Öztürk, A.K., Louvi, A., Kwan, K.Y., Choi, M., Tatlı, B., Yalnızoğlu, D., Tüysüz, B., Çağlayan, A.O., Gökben, S., et al. (2010). Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. Nature *467*, 207–210.

Blachon, S., Gopalakrishnan, J., Omori, Y., Polyanovsky, A., Church, A., Nicastro, D., Malicki, J., and Avidor-Reiss, T. (2008). Drosophila asterless and Vertebrate Cep152 Are Orthologs Essential for Centriole Duplication. Genetics *180*, 2081–2094.

Bobinnec, Y., Khodjakov, A., Mir, L.M., Rieder, C.L., Eddé, B., and Bornens, M. (1998). Centriole Disassembly In Vivo and Its Effect on Centrosome Structure and Function in Vertebrate Cells. J. Cell Biol. *143*, 1575–1589.

Bonaccorsi, S., Giansanti, M.G., and Gatti, M. (1998). Spindle Self-organization and Cytokinesis During Male Meiosis in asterless Mutants of Drosophila melanogaster. J. Cell Biol. *142*, 751–761.

Bond, J., Roberts, E., Mochida, G.H., Hampshire, D.J., Scott, S., Askham, J.M., Springell, K., Mahadevan, M., Crow, Y.J., Markham, A.F., et al. (2002). ASPM is a major determinant of cerebral cortical size. Nat. Genet. *32*, 316–320.

Bond, J., Roberts, E., Springell, K., Lizarraga, S., Scott, S., Higgins, J., Hampshire, D.J., Morrison, E.E., Leal, G.F., Silva, E.O., et al. (2005). A centrosomal mechanism involving CDK5RAP2 and CENPJ controls brain size. Nat. Genet. *37*, 353–355.

Bornens, M., Paintrand, M., Berges, J., Marty, M.-C., and Karsenti, E. (1987). Structural and chemical characterization of isolated centrosomes. Cell Motil. Cytoskeleton *8*, 238–249.

Bourke, E., Dodson, H., Merdes, A., Cuffe, L., Zachos, G., Walker, M., Gillespie, D., and Morrison, C.G. (2007). DNA damage induces Chk1-dependent centrosome amplification. EMBO Rep. *8*, 603–609.

Boveri, T. (2008). Concerning the Origin of Malignant Tumours by Theodor Boveri. Translated and annotated by Henry Harris. J. Cell Sci. *121*, 1–84.

Van Breugel, M., Hirono, M., Andreeva, A., Yanagisawa, H., Yamaguchi, S., Nakazawa, Y., Morgner, N., Petrovich, M., Ebong, I.-O., Robinson, C.V., et al. (2011). Structures of SAS-6 Suggest Its Organization in Centrioles. Science.

Van Breugel, M., Wilcken, R., McLaughlin, S.H., Rutherford, T.J., and Johnson, C.M. (2014). Structure of the SAS-6 cartwheel hub from Leishmania major. eLife *3*, e01812–e01812.

Bricogne, G., Vonrhein, C., Flensburg, C., Schiltz, M., and Paciorek, W. (2003). Generation, representation and flow of phase information in structure determination: recent developments in and around *SHARP* 2.0. Acta Crystallogr. D Biol. Crystallogr. *59*, 2023–2030.

Brito, D.A., Gouveia, S.M., and Bettencourt-Dias, M. (2012). Deconstructing the centriole: structure and number control. Curr. Opin. Cell Biol. *24*, 4–13.

Brown, N.J., Marjanović, M., Lüders, J., Stracker, T.H., and Costanzo, V. (2013). Cep63 and Cep152 Cooperate to Ensure Centriole Duplication. PLoS ONE *8*, e69986.

Carvalho-Santos, Z., Machado, P., Branco, P., Tavares-Cadete, F., Rodrigues-Martins, A., Pereira-Leal, J.B., and Bettencourt-Dias, M. (2010). Stepwise evolution of the centriole-assembly pathway. J Cell Sci *123*, 1414–1426.

Carvalho-Santos, Z., Azimzadeh, J., Pereira-Leal, J.B., and Bettencourt-Dias, M. (2011). Tracing the origins of centrioles, cilia, and flagella. J. Cell Biol. *194*, 165–175.

Carvalho-Santos, Z., Machado, P., Alvarez-Martins, I., Gouveia, S.M., Jana, S.C., Duarte, P., Amado, T., Branco, P., Freitas, M.C., Silva, S.T.N., et al. (2012). BLD10/CEP135 Is a Microtubule-Associated Protein that Controls the Formation of the Flagellum Central Microtubule Pair. Dev. Cell *23*, 412–424.

Cavalier-Smith, T. (1974). Basal Body and Flagellar Development During the Vegetative Cell Cycle and the Sexual Cycle of Chlamydomonas Reinhardii. J Cell Sci *16*, 529–556.

Chang, J., Cizmecioglu, O., Hoffmann, I., and Rhee, K. (2010). PLK2 phosphorylation is critical for CPAP function in procentriole formation during the centrosome cycle. EMBO J. *29*, 2395–2406.

Chavali, P.L., Pütz, M., and Gergely, F. (2014). Small organelle, big responsibility: the role of centrosomes in development and disease. Philos. Trans. R. Soc. Lond. B Biol. Sci. *369*, 20130468.

Cizmecioglu, O., Arnold, M., Bahtz, R., Settele, F., Ehret, L., Haselmann-Weiß, U., Antony, C., and Hoffmann, I. (2010). Cep152 Acts as a Scaffold for Recruitment of Plk4 and CPAP to the Centrosome. J. Cell Biol. *191*, 731–739.

Conduit, P.T., Richens, J.H., Wainman, A., Holder, J., Vicente, C.C., Pratt, M.B., Dix, C.I., Novak, Z.A., Dobbie, I.M., Schermelleh, L., et al. (2014). A molecular mechanism of mitotic centrosome assembly in Drosophila. eLife e03399.

Corbit, K.C., Aanstad, P., Singla, V., Norman, A.R., Stainier, D.Y.R., and Reiter, J.F. (2005). Vertebrate Smoothened functions at the primary cilium. Nature *437*, 1018–1021.

Cottee, M.A., Raff, J.W., Lea, S.M., and Roque, H. (2011). SAS-6 oligomerization: the key to the centriole? Nat Chem Biol *7*, 650–653.

Cottee, M.A., Muschalik, N., Wong, Y.L., Johnson, C.M., Johnson, S., Andreeva, A., Oegema, K., Lea, S.M., Raff, J.W., and van Breugel, M. (2013). Crystal structures of the CPAP/STIL complex reveal its role in centriole assembly and human microcephaly. eLife *2*.

Cowtan, K. (2006). The Buccaneer software for automated model building. 1. Tracing protein chains. Acta Crystallogr. Sect. D *62*, 1002–1011.

Cowtan, K. (2010). Recent developments in classical density modification. Acta Crystallogr. Sect. D *66*, 470–478.

Cromer, D.T. (1983). Calculation of anomalous scattering factors at arbitrary wavelengths. J. Appl. Crystallogr. *16*, 437–437.

Culver, B.P., Meehl, J.B., Giddings, T.H., and Winey, M. (2009). The Two SAS-6 Homologs in Tetrahymena thermophila Have Distinct Functions in Basal Body Assembly. Mol. Biol. Cell *20*, 1865–1877.

Cunha-Ferreira, I., Bento, I., Pimenta-Marques, A., Jana, S.C., Lince-Faria, M., Duarte, P., Borrego-Pinto, J., Gilberto, S., Amado, T., Brito, D., et al. (2013). Regulation of Autophosphorylation Controls PLK4 Self-Destruction and Centriole Number. Curr. Biol. *23*, 2245–2254.

Dammermann, A., Müller-Reichert, T., Pelletier, L., Habermann, B., Desai, A., and Oegema, K. (2004). Centriole Assembly Requires Both Centriolar and Pericentriolar Material Proteins. Dev. Cell *7*, 815–829.

Dammermann, A., Maddox, P.S., Desai, A., and Oegema, K. (2008). SAS-4 is recruited to a dynamic structure in newly forming centrioles that is stabilized by the γ-tubulin–mediated addition of centriolar microtubules. J. Cell Biol. *180*, 771–785.

Dardel, F., Packman, L.C., and Perham, R.N. (1990). Expression in Escherichia coli of a sub-gene encoding the lipoyl domain of the pyruvate dehydrogenase complex of Bacillus stearothermophilus. FEBS Lett. *264*, 206–210.

Dardel, F., Davis, A.L., Laue, E.D., and Perham, R.N. (1993). Three-dimensional Structure of the Lipoyl domain from Bacillus stearothermophilus Pyruvate Dehydrogenase Multienzyme Complex. J. Mol. Biol. *229*, 1037–1048.

Darvish, H., Esmaeeli-Nieh, S., Monajemi, G.B., Mohseni, M., Ghasemi-Firouzabadi, S., Abedini, S.S., Bahman, I., Jamali, P., Azimi, S., Mojahedi, F., et al. (2010). A clinical and molecular genetic study of 112 Iranian families with primary microcephaly. J. Med. Genet. *47*, 823–828.

D'Assoro, A.B., Barrett, S.L., Folk, C., Negron, V.C., Boeneman, K., Busby, R., Whitehead, C., Stivala, F., Lingle, W.L., and Salisbury, J.L. (2002). Amplified Centrosomes in Breast Cancer: A Potential Indicator of Tumor Aggressiveness. Breast Cancer Res. Treat. *75*, 25–34.

Deepa, R., Durga Rao, C., and Suguna, K. (2011). PDB entry 2O1J (To Be Published).

Delattre, M., Leidel, S., Wani, K., Baumer, K., Bamat, J., Schnabel, H., Feichtinger, R., Schnabel, R., and Gonczy, P. (2004). Centriolar SAS-5 is required for centrosome duplication in C. elegans. Nat Cell Biol *6*, 656–664.

Delattre, M., Canard, C., and Gönczy, P. (2006). Sequential Protein Recruitment in C. elegans Centriole Formation. Curr. Biol. *16*, 1844–1849.

Dix, C.I., and Raff, J.W. (2007). Drosophila Spd-2 recruits PCM to the sperm centriole, but is dispensable for centriole duplication. Curr. Biol. CB *17*, 1759–1764.

Dobbelaere, J., Josué, F., Suijkerbuijk, S., Baum, B., Tapon, N., and Raff, J. (2008). A Genome-Wide RNAi Screen to Dissect Centriole Duplication and Centrosome Maturation in Drosophila. PLoS Biol. *6*.

Al-Dosari, M.S., Shaheen, R., Colak, D., and Alkuraya, F.S. (2010). Novel CENPJ mutation causes Seckel syndrome. J. Med. Genet. *47*, 411–414.

Doxsey, S., McCollum, D., and Theurkauf, W. (2005a). Centrosomes in Cellular Regulation. Annu. Rev. Cell Dev. Biol. *21*, 411–434.

Doxsey, S., Zimmerman, W., and Mikule, K. (2005b). Centrosome control of the cell cycle. Trends Cell Biol. *15*, 303–311.

Dzhindzhev, N.S., Yu, Q.D., Weiskopf, K., Tzolovsky, G., Cunha-Ferreira, I., Riparbelli, M., Rodrigues-Martins, A., Bettencourt-Dias, M., Callaini, G., and Glover, D.M. (2010). Asterless is a scaffold for the onset of centriole assembly. Nature *467*, 714–718.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. *60*, 2126–2132.

Evans, P. (2005). Scaling and assessment of data quality. Acta Crystallogr. D Biol. Crystallogr. *62*, 72–82.

Evans, P.R. (1997). Scaling of MAD Data. In Proceedings of the CCP4 Study Weekend, January 1997, (Warrington: CCLRC Daresbury Laboratory), pp. 97–102.

Evans, P.R., and Murshudov, G.N. (2013). How good are my data and what is the resolution? Acta Crystallogr. D Biol. Crystallogr. *69*, 1204–1214.

Fong, C.S., Kim, M., Yang, T.T., Liao, J.-C., and Tsou, M.-F.B. (2014). SAS-6 Assembly Templated by the Lumen of Cartwheel-less Centrioles Precedes Centriole Duplication. Dev. Cell *30*, 238–245.

Freedman, S.J., Song, H.K., Xu, Y., Sun, Z.-Y.J., and Eck, M.J. (2003). Homotetrameric Structure of the SNAP-23 N-terminal Coiled-coil Domain. J. Biol. Chem. *278*, 13462–13467.

Ganem, N.J., Godinho, S.A., and Pellman, D. (2009). A mechanism linking extra centrosomes to chromosomal instability. Nature *460*, 278–282.

Gao, H., and Frank, J. (2005). Molding Atomic Structures into Intermediate- Resolution Cryo-EM Density Maps of Ribosomal Complexes Using Real-Space Refinement. Structure *13*, 401–406.

Genin, A., Desir, J., Lambert, N., Biervliet, M., Aa, N.V.D., Pierquin, G., Killian, A., Tosi, M., Urbina, M., Lefort, A., et al. (2012). Kinetochore KMN network gene CASC5 mutated in primary microcephaly. Hum. Mol. Genet. *21*, 5306–5317.

Giansanti, M.G., Bucciarelli, E., Bonaccorsi, S., and Gatti, M. (2008). Drosophila SPD-2 Is an Essential Centriole Component Required for PCM Recruitment and Astral-Microtubule Nucleation. Curr. Biol. *18*, 303–309.

Gibbons, I.R., and Grimstone, A.V. (1960). On Flagellar Structure in Certain Flagellates. J. Biophys. Biochem. Cytol. *7*, 697–716.

Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. (2003). ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information. Bioinformatics *19*, 163–164.

Goetz, S.C., and Anderson, K.V. (2010). The primary cilium: a signalling centre during vertebrate development. Nat. Rev. Genet. *11*, 331–344.

Gomez-Ferreria, M.A., Rath, U., Buster, D.W., Chanda, S.K., Caldwell, J.S., Rines, D.R., and Sharp, D.J. (2007). Human Cep192 Is Required for Mitotic Centrosome and Spindle Assembly. Curr. Biol. *17*, 1960–1966.

Gönczy, P. (2012). Towards a molecular architecture of centriole assembly. Nat. Rev. Mol. Cell Biol. *13*, 425–435.

Gönczy, P., Echeverri, C., Oegema, K., Coulson, A., Jones, S.J.M., Copley, R.R., Duperon, J., Oegema, J., Brehm, M., Cassin, E., et al. (2000). Functional genomic analysis of cell division in C. elegans using RNAi of genes on chromosome III. Nature *408*, 331–336.

Gopalakrishnan, J., Guichard, P., Smith, A.H., Schwarz, H., Agard, D.A., Marco, S., and Avidor-Reiss, T. (2010). Self-assembling SAS-6 Multimer Is a Core Centriole Building Block. J. Biol. Chem. *285*, 8759–8770.

Goshima, G., Wollman, R., Goodwin, S.S., Zhang, N., Scholey, J.M., Vale, R.D., and Stuurman, N. (2007). Genes Required for Mitotic Spindle Assembly in Drosophila S2 Cells. Science *316*, 417–421.

Guernsey, D.L., Jiang, H., Hussin, J., Arnold, M., Bouyakdan, K., Perry, S., Babineau-Sturk, T., Beis, J., Dumas, N., Evans, S.C., et al. (2010). Mutations in Centrosomal Protein CEP152 in Primary Microcephaly Families Linked to MCPH4. Am. J. Hum. Genet. *87*, 40–51.

Guichard, P., Chretien, D., Marco, S., and Tassin, A.-M. (2010). Procentriole assembly revealed by cryo-electron tomography. EMBO J *29*, 1565–1572.

Guichard, P., Desfosses, A., Maheshwari, A., Hachet, V., Dietrich, C., Brune, A., Ishikawa, T., Sachse, C., and Gönczy, P. (2012). Cartwheel Architecture of Trichonympha Basal Body. Science.

Guichard, P., Hachet, V., Majubu, N., Neves, A., Demurtas, D., Olieric, N., Fluckiger, I., Yamada, A., Kihara, K., Nishida, Y., et al. (2013). Native Architecture of the Centriole Proximal Region Reveals Features Underlying Its 9-Fold Radial Symmetry. Curr. Biol.

Gul, A., Hassan, M.J., Hussain, S., Raza, S.I., Chishti, M.S., and Ahmad, W. (2006). A novel deletion mutation in CENPJ gene in a Pakistani family with autosomal recessive primary microcephaly. J. Hum. Genet. *51*, 760–764.

Habedanck, R., Stierhof, Y.-D., Wilkinson, C.J., and Nigg, E.A. (2005). The Polo kinase Plk4 functions in centriole duplication. Nat Cell Biol *7*, 1140–1146.

Hatch, E.M., Kulukian, A., Holland, A.J., Cleveland, D.W., and Stearns, T. (2010). Cep152 interacts with Plk4 and is required for centriole duplication. J. Cell Biol. *191*, 721–729.

Hatzopoulos, G.N., Erat, M.C., Cutts, E., Rogala, K.B., Slater, L.M., Stansfeld, P.J., and Vakonakis, I. (2013). Structural Analysis of the G-Box Domain of the Microcephaly Protein CPAP Suggests a Role in Centriole Architecture. Structure *21*, 2069–2077.

Hilbert, M., Erat, M.C., Hachet, V., Guichard, P., Blank, I.D., Flückiger, I., Slater, L., Lowe, E.D., Hatzopoulos, G.N., Steinmetz, M.O., et al. (2013). Caenorhabditis elegans centriolar protein SAS-6 forms a spiral that is consistent with imparting a ninefold symmetry. Proc. Natl. Acad. Sci.

Hiraki, M., Nakazawa, Y., Kamiya, R., and Hirono, M. (2007). Bld10p Constitutes the Cartwheel-Spoke Tip and Stabilizes the 9-Fold Symmetry of the Centriole. Curr. Biol. *17*, 1778–1783.

Hirono, M. (2014). Cartwheel assembly. Philos. Trans. R. Soc. B Biol. Sci. *369*, 20130458.

Hodges, M.E., Scheumann, N., Wickstead, B., Langdale, J.A., and Gull, K. (2010). Reconstructing the evolutionary history of the centriole from protein components. J. Cell Sci. *123*, 1407–1413.

Hodgkinson, J.L., Horsley, A., Stabat, D., Simon, M., Johnson, S., da Fonseca, P.C.A., Morris, E.P., Wall, J.S., Lea, S.M., and Blocker, A.J. (2009). Three-dimensional reconstruction of the Shigella T3SS transmembrane regions reveals 12-fold symmetry and novel features throughout. Nat Struct Mol Biol *16*, 477–485.

Holland, A.J., Lan, W., Niessen, S., Hoover, H., and Cleveland, D.W. (2010). Polo-like kinase 4 kinase activity limits centrosome overduplication by autoregulating its own stability. J. Cell Biol. *188*, 191–198.

Holm, L., and Rosenstrom, P. (2010). Dali server: conservation mapping in 3D. Nucleic Acids Res. *38*, W545–W549.

Howell, P.L., and Smith, G.D. (1992). Identification of heavy-atom derivatives by normal probability methods. J. Appl. Crystallogr. *25*, 81–86.

Hsu, W.-B., Hung, L.-Y., Tang, C.-J.C., Su, C.-L., Chang, Y., and Tang, T.K. (2008). Functional characterization of the microtubule-binding and -destabilizing domains of CPAP and d-SAS-4. Exp. Cell Res. *314*, 2591–2602.

Huangfu, D., Liu, A., Rakeman, A.S., Murcia, N.S., Niswander, L., and Anderson, K.V. (2003). Hedgehog signalling in the mouse requires intraflagellar transport proteins. Nature *426*, 83–87.

Hung, L.-Y., Tang, C.-J.C., and Tang, T.K. (2000). Protein 4.1 R-135 Interacts with a Novel Centrosomal Protein (CPAP) Which Is Associated with the γ-Tubulin Complex. Mol. Cell. Biol. *20*, 7813–7825.

Hung, L.-Y., Chen, H.-L., Chang, C.-W., Li, B.-R., and Tang, T.K. (2004). Identification of a Novel Microtubule-destabilizing Motif in CPAP That Binds to Tubulin Heterodimers and Inhibits Microtubule Assembly. Mol. Biol. Cell *15*, 2697–2706.

Hussain, M.S., Baig, S.M., Neumann, S., Nürnberg, G., Farooq, M., Ahmad, I., Alef, T., Hennies, H.C., Technau, M., Altmüller, J., et al. (2012). A Truncating Mutation of CEP135 Causes Primary Microcephaly and Disturbed Centrosomal Function. Am. J. Hum. Genet. *90*, 871–878.

Hussain, M.S., Baig, S.M., Neumann, S., Peche, V.S., Szczepanski, S., Nürnberg, G., Tariq, M., Jameel, M., Khan, T.N., Fatima, A., et al. (2013). CDK6 associates with the centrosome during mitosis and is mutated in a large Pakistani family with primary microcephaly. Hum. Mol. Genet. *22*, 5199–5214.

Ibañez-Tallon, I., Heintz, N., and Omran, H. (2003). To beat or not to beat: roles of cilia in development and disease. Hum. Mol. Genet. *12*, R27–R35.

Ishikawa, H., and Marshall, W.F. (2011). Ciliogenesis: building the cell's antenna. Nat. Rev. Mol. Cell Biol. *12*, 222–234.

Jana, S.C., Marteil, G., and Bettencourt-Dias, M. (2014). Mapping molecules to structure: unveiling secrets of centriole and cilia assembly with near-atomic resolution. Curr. Opin. Cell Biol. *26*, 96–106.

Jerka-Dziadosz, M., Gogendeau, D., Klotz, C., Cohen, J., Beisson, J., and Koll, F. (2010). Basal body duplication in paramecium: The key role of Bld10 in assembly and stability of the Cartwheel. Cytoskeleton *67*, 161–171.

Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. *292*, 195–202.

Kabsch, W. (2010). XDS. Acta Crystallogr. D Biol. Crystallogr. *66*, 125–132.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577–2637.

Kalay, E., Yigit, G., Aslan, Y., Brown, K.E., Pohl, E., Bicknell, L.S., Kayserili, H., Li, Y., Tüysüz, B., Nürnberg, G., et al. (2011). CEP152 is a genome maintenance protein disrupted in Seckel syndrome. Nat. Genet. *43*, 23–26.

Keller, D., Orpinell, M., Olivier, N., Wachsmuth, M., Mahen, R., Wyss, R., Hachet, V., Ellenberg, J., Manley, S., and Gönczy, P. (2014). Mechanisms of HsSAS-6 assembly promoting centriole formation in human cells. J. Cell Biol. *204*, 697–712.

Keller, L.C., Romijn, E.P., Zamora, I., Yates III, J.R., and Marshall, W.F. (2005). Proteomic Analysis of Isolated Chlamydomonas Centrioles Reveals Orthologs of Ciliary-Disease Genes. Curr. Biol. *15*, 1090–1098.

Kemp, C.A., Kopish, K.R., Zipperlen, P., Ahringer, J., and O'Connell, K.F. (2004). Centrosome Maturation and Duplication in C. elegans Require the Coiled-Coil Protein SPD-2. Dev. Cell *6*, 511–523.

Kernan, M.J. (2007). Mechanotransduction and auditory transduction in Drosophila. Pflüg. Arch. - Eur. J. Physiol. *454*, 703–720.

Khan, M.A., Rupp, V.M., Orpinell, M., Hussain, M.S., Altmüller, J., Steinmetz, M.O., Enzinger, C., Thiele, H., Höhne, W., Nürnberg, G., et al. (2014). A missense mutation in the PISA domain of HsSAS-6 causes autosomal recessive primary microcephaly in a large consanguineous Pakistani family. Hum. Mol. Genet. ddu318.

Kilburn, C.L., Pearson, C.G., Romijn, E.P., Meehl, J.B., Giddings, T.H., Culver, B.P., Yates, J.R., and Winey, M. (2007). New Tetrahymena basal body protein components identify basal body domain structure. J. Cell Biol. *178*, 905–912.

Kirkham, M., Müller-Reichert, T., Oegema, K., Grill, S., and Hyman, A.A. (2003). SAS-4 Is a C. elegans Centriolar Protein that Controls Centrosome Size. Cell *112*, 575–587.

Kitagawa, D., Busso, C., Flückiger, I., and Gönczy, P. (2009). Phosphorylation of SAS-6 by ZYG-1 Is Critical for Centriole Formation in C. elegans Embryos. Dev. Cell *17*, 900–907.

Kitagawa, D., Flückiger, I., Polanowska, J., Keller, D., Reboul, J., and Gönczy, P. (2011a). PP2A Phosphatase Acts upon SAS-5 to Ensure Centriole Formation in C. elegans Embryos. Dev. Cell *20*, 550–562.

Kitagawa, D., Vakonakis, I., Olieric, N., Hilbert, M., Keller, D., Olieric, V., Bortfeld, M., Erat, M.C., Flückiger, I., Gönczy, P., et al. (2011b). Structural Basis of the 9-Fold Symmetry of Centrioles. Cell *144*, 364–375.

Kitagawa, D., Kohlmaier, G., Keller, D., Strnad, P., Balestra, F.R., Flückiger, I., and Gönczy, P. (2011c). Spindle positioning in human cells relies on proper centriole formation and on the microcephaly proteins CPAP and STIL. J. Cell Sci.

Kleylein-Sohn, J., Westendorf, J., Le Clech, M., Habedanck, R., Stierhof, Y.-D., and Nigg, E.A. (2007). Plk4-Induced Centriole Biogenesis in Human Cells. Dev. Cell *13*, 190–202.

Kohlmaier, G., Lončarek, J., Meng, X., McEwen, B.F., Mogensen, M.M., Spektor, A., Dynlacht, B.D., Khodjakov, A., and Gönczy, P. (2009). Overly Long Centrioles and Defective Cell Division upon Excess of the SAS-4-Related Protein CPAP. Curr. Biol. *19*, 1012–1018.

Kumar, A., Girimaji, S.C., Duvvari, M.R., and Blanton, S.H. (2009). Mutations in STIL, Encoding a Pericentriolar and Centrosomal Protein, Cause Primary Microcephaly. Am. J. Hum. Genet. *84*, 286–290.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947–2948.

Leal, G.F., Roberts, E., Silva, E.O., Costa, S.M.R., Hampshire, D.J., and Woods, C.G. (2003). A novel locus for autosomal recessive primary microcephaly (MCPH6) maps to 13q12.2. J. Med. Genet. *40*, 540–542.

Leidel, S., and Gönczy, P. (2003). SAS-4 Is Essential for Centrosome Duplication in C. elegans and Is Recruited to Daughter Centrioles Once per Cell Cycle. Dev. Cell *4*, 431–439.

Leidel, S., Delattre, M., Cerutti, L., Baumer, K., and Gonczy, P. (2005). SAS-6 defines a protein family required for centrosome duplication in C. elegans and in human cells. Nat Cell Biol *7*, 115–125.

Leslie, A.G.W., and Powell, H.R. (2007). Evolving Methods for Macromolecular Crystallography - The Structural Path to the Understanding.

Lettman, M.M., Wong, Y.L., Viscardi, V., Niessen, S., Chen, S., Shiau, A.K., Zhou, H., Desai, A., and Oegema, K. (2013). Direct Binding of SAS-6 to ZYG-1 Recruits SAS-6 to the Mother Centriole for Cartwheel Assembly. Dev. Cell *25*, 284–298.

Li, J.B., Gerdes, J.M., Haycraft, C.J., Fan, Y., Teslovich, T.M., May-Simera, H., Li, H., Blacque, O.E., Li, L., Leitch, C.C., et al. (2004). Comparative Genomics Identifies a Flagellar and Basal Body Proteome that Includes the BBS5 Human Disease Gene. Cell *117*, 541–552.

Li, S., Fernandez, J.-J., Marshall, W.F., and Agard, D.A. (2012). Three-dimensional structure of basal body triplet revealed by electron cryo-tomography. EMBO J. *31*, 552–562.

Lin, Y.-C., Chang, C.-W., Hsu, W.-B., Tang, C.-J.C., Lin, Y.-N., Chou, E.-J., Wu, C.-T., and Tang, T.K. (2013). Human microcephaly protein CEP135 binds to hSAS-6 and CPAP, and is required for centriole assembly. EMBO J. *32*, 1141–1154.

Linck, R.W., and Stephens, R.E. (2007). Functional protofilament numbering of ciliary, flagellar, and centriolar microtubules. Cell Motil. Cytoskeleton *64*, 489–495.

Lingle, W.L., Barrett, S.L., Negron, V.C., D'Assoro, A.B., Boeneman, K., Liu, W., Whitehead, C.M., Reynolds, C., and Salisbury, J.L. (2002). Centrosome amplification drives chromosomal instability in breast tumor development. Proc. Natl. Acad. Sci. U. S. A. *99*, 1978–1983.

Loncarek, J., and Khodjakov, A. (2009). Ab ovo or de novo? Mechanisms of centriole duplication. Mol. Cells *27*, 135–142.

Loncarek, J., Hergert, P., Magidson, V., and Khodjakov, A. (2008). Control of daughter centriole formation by the pericentriolar material. Nat. Cell Biol. *10*, 322–328.

Lucas, E.P., and Raff, J.W. (2007). Maintaining the proper connection between the centrioles and the pericentriolar matrix requires Drosophila centrosomin. J. Cell Biol. *178*, 725–732.

Lukinavičius, G., Lavogina, D., Orpinell, M., Umezawa, K., Reymond, L., Garin, N., Gönczy, P., and Johnsson, K. (2013). Selective Chemical Crosslinking Reveals a Cep57-Cep63-Cep152 Centrosomal Complex. Curr. Biol. *23*, 265–270.

Lupas, A.N., and Gruber, M. (2005). The Structure of α-Helical Coiled Coils. In Advances in Protein Chemistry, and J.M.S. David A. D. Parry, ed. (Academic Press), pp. 37–38.

Lupas, A., Van Dyke, M., and Stock, J. (1991). Predicting coiled coils from protein sequences. Science *252*, 1162–1164.

Ma, H.T., and Poon, R.Y.C. (2011). How protein kinases co-ordinate mitosis in animal cells. Biochem. J. *435*, 17–31.

Marthiens, V., Rujano, M.A., Pennetier, C., Tessier, S., Paul-Gilloteaux, P., and Basto, R. (2013). Centrosome amplification causes microcephaly. Nat. Cell Biol. *15*, 731–740.

Matsuura, K., Lefebvre, P.A., Kamiya, R., and Hirono, M. (2004). Bld10p, a novel protein essential for basal body assembly in Chlamydomonas. J. Cell Biol. *165*, 663–671.

Matthews, B.. (1968). Solvent content of protein crystals. J. Mol. Biol. *33*, 491–497.

McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., and Read, R.J. (2007). Phaser crystallographic software. J. Appl. Crystallogr. *40*, 658–674.

McIntyre, R.E., Lakshminarasimhan Chavali, P., Ismail, O., Carragher, D.M., Sanchez-Andrade, G., Forment, J.V., Fu, B., Del Castillo Velasco-Herrera, M., Edwards, A., van der Weyden, L., et al. (2012). Disruption of Mouse Cenpj, a Regulator of Centriole Biogenesis, Phenocopies Seckel Syndrome. PLoS Genet *8*, e1003022.

McNitt, R. (1974). Centriole ultrastructure and its possible role in microtubule formation in an aquatic fungus. Protoplasma *80*, 91–108.

Megraw, T.L., Sharkey, J.T., and Nowakowski, R.S. (2011). Cdk5rap2 exposes the centrosomal root of microcephaly syndromes. Trends Cell Biol. *21*, 470–480.

Mennella, V., Keszthelyi, B., McDonald, K.L., Chhun, B., Kan, F., Rogers, G.C., Huang, B., and Agard, D.A. (2012). Subdiffraction-resolution fluorescence microscopy reveals a domain of the centrosome critical for pericentriolar material organization. Nat. Cell Biol. *14*, 1159–1168.

Mottier-Pavie, V., and Megraw, T.L. (2009). Drosophila Bld10 Is a Centriolar Protein That Regulates Centriole, Basal Body, and Motile Cilium Assembly. Mol. Biol. Cell *20*, 2605–2614.

Murshudov, G.N., Skubak, P., Lebedev, A.A., Pannu, N.S., Steiner, R.A., Nicholls, R.A., Winn, M.D., Long, F., and Vagin, A.A. (2011). REFMAC5 for the refinement of macromolecular crystal structures. Acta Crystallogr. D Biol. Crystallogr. *67*, 355–367.

Nakazawa, Y., Hiraki, M., Kamiya, R., and Hirono, M. (2007). SAS-6 is a Cartwheel Protein that Establishes the 9-Fold Symmetry of the Centriole. Curr. Biol. *17*, 2169–2174.

Nicholas, A.K., Khurshid, M., Désir, J., Carvalho, O.P., Cox, J.J., Thornton, G., Kausar, R., Ansar, M., Ahmad, W., Verloes, A., et al. (2010). WDR62 is associated with the spindle pole and is mutated in human microcephaly. Nat. Genet. *42*, 1010–1014.

Nigg, E.A., and Raff, J.W. (2009). Centrioles, Centrosomes, and Cilia in Health and Disease. Cell *139*, 663–678.

Nigg, E.A., and Stearns, T. (2011). The centrosome cycle: Centriole biogenesis, duplication and inherent asymmetries. Nat. Cell Biol. *13*, 1154–1160.

Nigg, E.A., Čajánek, L., and Arquint, C. (2014). The centrosome duplication cycle in health and disease. FEBS Lett. *588*, 2366–2372.

Novak, Z.A., Conduit, P.T., Wainman, A., and Raff, J.W. (2014). Asterless Licenses Daughter Centrioles to Duplicate for the First Time in Drosophila Embryos. Curr. Biol. *24*, 1276–1282.

O'Connell, K.F., Caron, C., Kopish, K.R., Hurd, D.D., Kemphues, K.J., Li, Y., and White, J.G. (2001). The C. elegans zyg-1 Gene Encodes a Regulator of Centrosome Duplication with Distinct Maternal and Paternal Roles in the Embryo. Cell *105*, 547–558.

O'Driscoll, M., Jackson, A., and Jeggo, P.A. (2006). Microcephalin: A Causal Link Between Impaired Damage Response Signalling and Microcephaly. Cell Cycle *5*, 2339–2344.

Ohta, T., Essner, R., Ryu, J.-H., Palazzo, R.E., Uetake, Y., and Kuriyama, R. (2002). Characterization of Cep135, a Novel Coiled-Coil Centrosomal Protein Involved in Microtubule Organization in Mammalian Cells. J. Cell Biol. *156*, 87–100.

Oliveira, R.A., Hamilton, R.S., Pauli, A., Davis, I., and Nasmyth, K. (2010). Cohesin cleavage and Cdk inhibition trigger formation of daughter nuclei. Nat. Cell Biol. *12*, 185–192.

Pal, D., and Chakrabarti, P. (2000). β-Sheet propensity and its correlation withparameters based on conformation. Acta Crystallogr. D Biol. Crystallogr. *56*, 589–594.

Park, S.-Y., Park, J.-E., Kim, T.-S., Kim, J.H., Kwak, M.-J., Ku, B., Tian, L., Murugan, R.N., Ahn, M., Komiya, S., et al. (2014). Molecular basis for unidirectional scaffold switching of human Plk4 in centriole biogenesis. Nat. Struct. Mol. Biol. *21*, 696–703.

Peel, N., Stevens, N.R., Basto, R., and Raff, J.W. (2007). Overexpressing Centriole-Replication Proteins In Vivo Induces Centriole Overduplication and De Novo Formation. Curr. Biol. *17*, 834–843.

Pelletier, L., Özlü, N., Hannak, E., Cowan, C., Habermann, B., Ruer, M., Müller-Reichert, T., and Hyman, A.A. (2004). The Caenorhabditis elegans Centrosomal Protein SPD-2 Is Required for both Pericentriolar Material Recruitment and Centriole Duplication. Curr. Biol. *14*, 863–873.

Pelletier, L., O/'Toole, E., Schwager, A., Hyman, A.A., and Muller-Reichert, T. (2006). Centriole assembly in Caenorhabditis elegans. Nature *444*, 619–623.

Pfaff, K.L., Straub, C.T., Chiang, K., Bear, D.M., Zhou, Y., and Zon, L.I. (2007). The Zebra fish cassiopeia Mutant Reveals that SIL Is Required for Mitotic Spindle Organization. Mol. Cell. Biol. *27*, 5887–5897.

Pihan, G.A., Purohit, A., Wallace, J., Knecht, H., Woda, B., Quesenberry, P., and Doxsey, S.J. (1998). Centrosome Defects and Genetic Instability in Malignant Tumors. Cancer Res. *58*, 3974–3985.

Pihan, G.A., Wallace, J., Zhou, Y., and Doxsey, S.J. (2003). Centrosome Abnormalities and Chromosome Instability Occur Together in Pre-invasive Carcinomas. Cancer Res. *63*, 1398–1404.

Puklowski, A., Homsi, Y., Keller, D., May, M., Chauhan, S., Kossatz, U., Grunwald, V., Kubicka, S., Pich, A., Manns, M.P., et al. (2011). The SCF-FBXW5 E3-ubiquitin ligase is regulated by PLK4 and targets HsSAS-6 to control centrosome duplication. Nat Cell Biol *13*, 1004–1009.

Qiao, R., Cabral, G., Lettman, M.M., Dammermann, A., and Dong, G. (2012). SAS-6 coiled-coil structure and interaction with SAS-5 suggest a regulatory mechanism in C. elegans centriole assembly. EMBO J.

Ramamurthy, V., and Cayouette, M. (2009). Development and disease of the photoreceptor cilium. Clin. Genet. *76*, 137–145.

Roberts, D.B. (1998). Drosophila: a practical approach (IRL Press at Oxford University Press).

Rodrigues-Martins, A., Riparbelli, M., Callaini, G., Glover, D.M., and Bettencourt-Dias, M. (2007a). Revisiting the Role of the Mother Centriole in Centriole Biogenesis. Science *316*, 1046–1050.

Rodrigues-Martins, A., Bettencourt-Dias, M., Riparbelli, M., Ferreira, C., Ferreira, I., Callaini, G., and Glover, D.M. (2007b). DSAS-6 Organizes a Tube-like Centriole Precursor, and Its Absence Suggests Modularity in Centriole Assembly. Curr. Biol. *17*, 1465–1472.

Roque, H., Wainman, A., Richens, J., Kozyrska, K., Franz, A., and Raff, J.W. (2012). Drosophila Cep135/Bld10 maintains proper centriole structure but is dispensable for cartwheel formation. J. Cell Sci. *125*, 5881–5886.

Roy, S. (2012). Cilia and Hedgehog: When and how was their marriage solemnized? Differentiation *83*, S43–S48.

Schmidt, T.I., Kleylein-Sohn, J., Westendorf, J., Le Clech, M., Lavoie, S.B., Stierhof, Y.-D., and Nigg, E.A. (2009). Control of Centriole Length by CPAP and CP110. Curr. Biol. *19*, 1005–1011.

Schuh, M., and Ellenberg, J. (2007). Self-Organization of MTOCs Replaces Centrosome Function during Acentrosomal Spindle Assembly in Live Mouse Oocytes. Cell *130*, 484–498.

Sharp, D.J., McDonald, K.L., Brown, H.M., Matthies, H.J., Walczak, C., Vale, R.D., Mitchison, T.J., and Scholey, J.M. (1999). The Bipolar Kinesin, KLP61F, Cross-links Microtubules within Interpolar Microtubule Bundles of Drosophila Embryonic Mitotic Spindles. J. Cell Biol. *144*, 125–138.

Sheldrick, G.M. (2007). A short history of SHELX. Acta Crystallogr. A *64*, 112–122.

Shimada, M., and Komatsu, K. (2009). Emerging Connection Between Centrosome and DNA Repair Machinery. J. Radiat. Res. (Tokyo) *50*, 295–301.

Shimanovskaya, E., Qiao, R., Lesigang, J., and Dong, G. (2013). The SAS-5 N-terminal domain is a tetramer, with implications for centriole assembly in C. elegans. Worm *2*, e25214.

Shimanovskaya, E., Viscardi, V., Lesigang, J., Lettman, M.M., Qiao, R., Svergun, D.I., Round, A., Oegema, K., and Dong, G. (2014). Structure of the C. elegans ZYG-1 Cryptic Polo Box Suggests a Conserved Mechanism for Centriolar Docking of Plk4 Kinases. Structure *22*, 1090–1104.

Siller, K.H., and Doe, C.Q. (2009). Spindle orientation during asymmetric cell division. Nat. Cell Biol. *11*, 365–374.

Sillibourne, J.E., Tack, F., Vloemans, N., Boeckx, A., Thambirajah, S., Bonnet, P., Ramaekers, F.C.S., Bornens, M., and Grand-Perret, T. (2010). Autophosphorylation of Polo-like Kinase 4 and Its Role in Centriole Duplication. Mol. Biol. Cell *21*, 547–561.

Sir, J.-H., Barr, A.R., Nicholas, A.K., Carvalho, O.P., Khurshid, M., Sossick, A., Reichelt, S., D'Santos, C., Woods, C.G., and Gergely, F. (2011). A primary microcephaly protein complex forms a ring around parental centrioles. Nat. Genet. *43*, 1147–1153.

Slevin, L.K., Nye, J., Pinkerton, D.C., Buster, D.W., Rogers, G.C., and Slep, K.C. (2012). The Structure of the Plk4 Cryptic Polo Box Reveals Two Tandem Polo Boxes Required for Centriole Duplication. Structure.

Slevin, L.K., Romes, E.M., Dandulakis, M.G., and Slep, K.C. (2014). The Mechanism of Dynein Light Chain LC8-mediated Oligomerization of the Ana2 Centriole Duplication Factor. J. Biol. Chem. *289*, 20727–20739.

Smart, O.S., Womack, T.O., Flensburg, C., Keller, P., Paciorek, W., Sharff, A., Vonrhein, C., and Bricogne, G. (2012). Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER. Acta Crystallogr. D Biol. Crystallogr. *68*, 368–380.

Sonnen, K.F., Schermelleh, L., Leonhardt, H., and Nigg, E.A. (2012). 3D-structured illumination microscopy provides novel insight into architecture of human centrosomes. Biol. Open.

Sonnen, K.F., Gabryjonczyk, A.-M., Anselm, E., Stierhof, Y.-D., and Nigg, E.A. (2013). Human Cep192 and Cep152 cooperate in Plk4 recruitment and centriole duplication. J. Cell Sci. *126*, 3223–3233.

Stein, N. (2008). CHAINSAW: a program for mutating pdb files used as templates in molecular replacement. J. Appl. Crystallogr. *41*, 641–643.

Stevens, N.R., Raposo, A.A.S.F., Basto, R., St Johnston, D., and Raff, J.W. (2007). From Stem Cell to Embryo without Centrioles. Curr. Biol. *17*, 1498–1503.

Stevens, N.R., Dobbelaere, J., Wainman, A., Gergely, F., and Raff, J.W. (2009). Ana3 is a conserved protein required for the structural integrity of centrioles and basal bodies. J. Cell Biol. *187*, 355–363.

Stevens, N.R., Roque, H., and Raff, J.W. (2010a). DSas-6 and Ana2 Coassemble into Tubules to Promote Centriole Duplication and Engagement. Dev. Cell *19*, 913–919.

Stevens, N.R., Dobbelaere, J., Brunk, K., Franz, A., and Raff, J.W. (2010b). Drosophila Ana2 is a conserved centriole duplication factor. J. Cell Biol. *188*, 313–323.

Strnad, P., and Gönczy, P. (2008). Mechanisms of procentriole formation. Trends Cell Biol. *18*, 389–396.

Strnad, P., Leidel, S., Vinogradova, T., Euteneuer, U., Khodjakov, A., and Gönczy, P. (2007). Regulated HsSAS-6 Levels Ensure Formation of a Single Procentriole per Centriole during the Centrosome Duplication Cycle. Dev. Cell *13*, 203–213.

Stura, E.A., and Wilson, I.A. (1991). Applications of the streak seeding technique in protein crystallization. J. Cryst. Growth *110*, 270–282.

Suhre, K., and Sanejouand, Y.-H. (2004a). ElNémo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. Nucleic Acids Res. *32*, W610–W614.

Suhre, K., and Sanejouand, Y.-H. (2004b). On the potential of normal-mode analysis for solving difficult molecular-replacement problems. Acta Crystallogr. D Biol. Crystallogr. *60*, 796–799.

Sun, T., and Hevner, R.F. (2014). Growth and folding of the mammalian cerebral cortex: from molecules to malformations. Nat. Rev. Neurosci. *15*, 217–232.

Tang, C.-J.C., Fu, R.-H., Wu, K.-S., Hsu, W.-B., and Tang, T.K. (2009). CPAP is a cell-cycle regulated protein that controls centriole length. Nat. Cell Biol. *11*, 825–831.

Tang, C.-J.C., Lin, S.-Y., Hsu, W.-B., Lin, Y.-N., Wu, C.-T., Lin, Y.-C., Chang, C.-W., Wu, K.-S., and Tang, T.K. (2011). The human microcephaly protein STIL interacts with CPAP and is required for procentriole formation. EMBO J. *30*, 4790–4804.

Tavosanis, G., Llamazares, S., Goulielmos, G., and Gonzalez, C. (1997). Essential role for γ-tubulin in the acentriolar female meiotic spindle of Drosophila. EMBO J. *16*, 1809–1819.

Thornton, G.K., and Woods, C.G. (2009). Primary microcephaly: do all roads lead to Rome? Trends Genet. *25*, 501–510.

Tsou, M.-F.B., Wang, W.-J., George, K.A., Uryu, K., Stearns, T., and Jallepalli, P.V. (2009). Polo kinase and separase regulate the mitotic licensing of centriole duplication in human cells. Dev. Cell *17*, 344–354.

Vagin, A., and Teplyakov, A. (2010). Molecular replacement with MOLREP. Acta Crystallogr. Sect. D *66*, 22–25.

Valente, E.M., Rosti, R.O., Gibbs, E., and Gleeson, J.G. (2014). Primary cilia in neurodevelopmental disorders. Nat. Rev. Neurol. *10*, 27–36.

Varmark, H., Llamazares, S., Rebollo, E., Lange, B., Reina, J., Schwarz, H., and Gonzalez, C. (2007). Asterless Is a Centriolar Protein Required for Centrosome Function and Embryo Development in Drosophila. Curr. Biol. *17*, 1735–1745.

Vonrhein, C., Blanc, E., Roversi, P., and Bricogne, G. (2007). Automated structure solution with autoSHARP. Methods Mol. Biol. Clifton NJ *364*, 215–230.

Vulprecht, J., David, A., Tibelius, A., Castiel, A., Konotop, G., Liu, F., Bestvater, F., Raab, M.S., Zentgraf, H., Izraeli, S., et al. (2012). STIL is required for centriole duplication in human cells. J. Cell Sci.

Walshaw, J., and Woolfson, D.N. (2001). SOCKET: a program for identifying and analysing coiled-coil motifs within protein structures. J. Mol. Biol. *307*, 1427–1450.

Wang, C., Li, S., Januschke, J., Rossi, F., Izumi, Y., Garcia-Alvarez, G., Gwee, S.S.L., Soon, S.B., Sidhu, H.K., Yu, F., et al. (2011). An Ana2/Ctp/Mud Complex Regulates Spindle Orientation in Drosophila Neuroblasts. Dev. Cell *21*, 520–533.

Wang, X., Tsai, J.-W., Imai, J.H., Lian, W.-N., Vallee, R.B., and Shi, S.-H. (2009). Asymmetric centrosome inheritance maintains neural progenitors in neocortex. Nature *461*, 947–955.

White, H.E., Chen, S., Roseman, A.M., Yifrach, O., Horovitz, A., and Saibil, H.R. (1997). Structural basis of allosteric changes in the GroEL mutant Arg197-->Ala. Nat. Struct. Biol. *4*, 690–694.

Winey, M., and O'Toole, E. (2014). Centriole structure. Philos. Trans. R. Soc. B Biol. Sci. *369*, 20130457.

Winter, G. (2010). xia2: an expert system for macromolecular crystallography data reduction. J. Appl. Crystallogr. *43*, 186–190.

Woods, C.G., Bond, J., and Enard, W. (2005). Autosomal Recessive Primary Microcephaly (MCPH): A Review of Clinical, Molecular, and Evolutionary Findings. Am. J. Hum. Genet. *76*, 717–728.

Yabe, T., Ge, X., and Pelegri, F. (2007). The zebrafish maternal-effect gene cellular atoll encodes the centriolar component sas-6 and defects in its paternal function promote whole genome duplication. Dev. Biol. *312*, 44–60.

Yadav, M.K., Redman, J.E., Leman, L.J., Alvarez-Gutiérrez, J.M., Zhang, Y., Stout, C.D., and Ghadiri, M.R. (2005). Structure-Based Engineering of Internal Cavities in Coiled-Coil Peptides†,‡. Biochemistry (Mosc.) *44*, 9723–9732.

Yang, Y.J., Baltus, A.E., Mathew, R.S., Murphy, E.A., Evrony, G.D., Gonzalez, D.M., Wang, E.P., Marshall-Walker, C.A., Barry, B.J., Murn, J., et al. (2012). Microcephaly Gene Links Trithorax and REST/NRSF to Control Neural Stem Cell Proliferation and Differentiation. Cell *151*, 1097–1112.

Yu, T.W., Mochida, G.H., Tischfield, D.J., Sgaier, S.K., Flores-Sarnat, L., Sergi, C.M., Topçu, M., McDonald, M.T., Barry, B.J., Felie, J.M., et al. (2010). Mutations in WDR62, encoding a centrosome-associated protein, cause microcephaly with simplified gyri and abnormal cortical architecture. Nat. Genet. *42*, 1015–1020.

Zacharias, D.A., Violin, J.D., Newton, A.C., and Tsien, R.Y. (2002). Partitioning of lipid-modified monomeric GFPs into membrane microdomains of live cells. Science *296*, 913–916.

Zheng, H., Chordia, M.D., Cooper, D.R., Chruszcz, M., Müller, P., Sheldrick, G.M., and Minor, W. (2014a). Validation of metal-binding sites in macromolecular structures with the CheckMyMetal web server. Nat. Protoc. *9*, 156–170.

Zheng, X., Gooi, L.M., Wason, A., Gabriel, E., Mehrjardi, N.Z., Yang, Q., Zhang, X., Debec, A., Basiri, M.L., Avidor-Reiss, T., et al. (2014b). Conserved TCP domain of Sas-4/CPAP is essential for pericentriolar material tethering during centrosome biogenesis. Proc. Natl. Acad. Sci. 201317535.

Zhu, F., Lawo, S., Bird, A., Pinchev, D., Ralph, A., Richter, C., Müller-Reichert, T., Kittler, R., Hyman, A.A., and Pelletier, L. (2008). The Mammalian SPD-2 Ortholog Cep192 Regulates Centrosome Biogenesis. Curr. Biol. *18*, 136–141.