STRUCTURAL AND BIOPHYSICAL ANALYSIS OF UBE3A AND PARTNER PROTEIN COMPLEXES

Thesis for the completion of a Doctorate of Philosophy at the University of Nottingham

Emma Cowan

Supervisors: Dr David Scott, Dr Stephen Carr, Dr Katie Cunnea, Dr Daniel Clare, Prof. Robert Layfield

Acknowledgements

I would like to thank everyone who has supported me throughout this project. Firstly to the MRC IMPACT DTP, the University of Nottingham, Diamond and eBIC, and the RCaH for providing the funding, equipment, and lab space necessary to carry out this work.

To my supervisors for your continued guidance and support throughout the project.

To Dave and Steve, thank you for teaching me how to do science. Thank you for always being open to chat about a weird result in the lab, or to discuss a new experiment idea, and especially for always checking my primers. This project would not have been possible, and this thesis would definitely not have been written, without your continued support and guidance.

To Katie, thank you for encouraging me to consider the real-world impact of the science as well as appreciating the experimental techniques. Thank you for sharing with me your passion for the work, and of course for teaching me everything I know about cryo-EM. I was always very grateful of your pep talks and support when the results weren't coming out as I hoped, and for your enthusiasm and encouragement when they were.

I would also like to thank all of my friends at the RCaH throughout the last few years for making the whole experience an enjoyable one. I don't think this project would have been possible, or certainly less enjoyable at least, without the many daily coffee breaks and lab chats.

Finally, I would like to thank my friends and family who have supported me through this degree despite having no idea what I do. Special thanks to my parents of course for raising me to appreciate science and for supporting and encouraging me to keep learning. The biggest thanks go to my patient boyfriend who has had to put up with me through the failed experiments, late evenings, and the entire writing process. Your support throughout this whole process has been invaluable. Final thanks to my cat, for keeping me company and keeping me sane through the writing stage.

Abstract

UBE3A is a human E3 ubiquitin ligase, responsible for the transfer of ubiquitin onto substrates to target them for degradation and other cellular processes. UBE3A has been implicated in a range of neurodevelopmental disorders, most notably Angelman Syndrome, but also Dup15Q syndrome, autism spectrum disorders, and schizotypies. Alongside this, it has also been shown to be involved in the oncogenesis of many different types of cancer, including prostate cancer, small-cell lung cancer, B-cell lymphomas, and most notably HPV-associated oropharyngeal cancer and cervical carcinomas. More recent studies have even shown a role of UBE3A in cardiomyopathies, Alzheimers, and the immune response to HIV, as well as identifying it as part of a diverse range of cellular signalling processes.

Despite the huge clinical significance of UBE3A, a full-length structure for the enzyme is not currently available. The catalytic HECT domain was solved by x-ray crystallography in 1999, but a large distance between the catalytic site and the binding site of the cognate E2 enzyme have raised more questions about the mechanism of UBE3A activity than answers. Various studies in the intervening years have revealed more insights into UBE3A's activity and interaction with cellular partners, but a full structure would enable a much better understanding of its mechanism. This is key for designing small molecules to counteract some of the effects of UBE3A mutations or oncogenic signalling disruptions, but also just for understanding how UBE3A works and how its loss may be compensated for in the majority of Angelman Syndrome cases.

In this work, I used a range of biophysical techniques, biochemical assays, and structural techniques in an attempt to characterise UBE3A and its interactions with partner proteins as completely as possible. Although many of the questions that I attempted to address remain unanswered, I was able to reconstitute a low-resolution cryo-EM map of isolated full-length UBE3A, even lower resolution models of UBE3A in complex with two of its binding partners, and a high resolution model of the RLD2 domain of the HERC2 protein. I was also able to demonstrate some of the properties of these interactions through SV-AUC, ITC, CD, and crosslinking and co-purification experiments. The different forms of UBE3A, both alone and in the presence of its binding partners, were subjected to *in vitro* assays in an attempt to determine the effects of complex formation of the ubiquitin ligase activity of UBE3A.

Table of Contents

Acknowledgements ***1***

Abstract ***2***

1 Introduction ***13***

1.1 Ubiquitination **13**

1.2 *UBE3A* Epigenetics and isoforms **15**

1.3 Neurodevelopmental Disorders **20**

1.3.1 Angelman Syndrome 20

1.3.2 Dup15q Syndrome 27

1.3.3 Autism Spectrum Disorders 29

1.3.4 Neuropsychiatric Disorders 29

1.4 Cancer **30**

1.4.1 Cervical Cancer 32

1.4.2 Oropharyngeal Cancer 34

1.4.3 Hepatocellular Carcinoma 35

1.4.4 Prostate Cancer 36

1.4.5 B-cell Lymphoma 37

1.4.6 Non-Small Cell Lung Cancer 37

1.5 Structures and Key Residues **38**

1.6 Project Aims **49**

2 Materials and Methods ***51***

2.1 Cells **51**

2.1.1 DNA Propogation 51

2.1.2 Protein Expression 51

2.1.3 Mammalian Cells 51

2.2 Plasmids **51**

2.2.1 UBE3A 51

2.2.2 UbcH7 52

2.2.3 E6 and p53 52

2.2.4 PSMD4 52

2.2.5 HERC2 53

2.2.6 RLD2 53

2.2.7 Ufrag 53

2.3 Materials **53**

2.3.1 Primers 53

2.3.2 DNA Sequencing 57

2.3.3 Gene Synthesis 57

2.3.4 Kits and Consumables 57

2.4 Cloning **60**

2.4.1 Transformations 60

2.4.2 PCR Mutagenesis 61

2.4.3 Restriction Enzyme Digests 62

2.4.4 InFusion Cloning 63

2.4.5 Gibson/HiFi Assembly 64

2.4.6 Colony PCR Screening 65

2.5 Protein Expression 65

2.5.1 Overexpression in *E. Coli* 65

2.5.2 Overexpression in Mammalian Cells 68

2.6 Protein Purification 68

2.6.1 Cell Lysis 68

2.6.2 HisTrap Purification 69

2.6.3 MBPTrap Purification 69

2.6.4 TEV/3C Digest 69

2.6.5 Reverse HisTrap Purification 69

2.6.6 Anion Exchange Chromatography 70

2.6.7 Size Exclusion Chromatography 70

2.6.8 Small-Scale Gravity Affinity Purifications 70

2.6.9 Strep-tag Gravity Purification 71

2.7 Protein Gel Electrophoresis 72

2.7.1 SDS-PAGE 72

2.7.2 Native PAGE 73

2.7.3 BN-PAGE 74

2.7.4 Horizontal Agarose Gel Electrophoresis 74

2.7.5 Vertical Agarose Gel Electrophoresis 75

2.8 Ubiquitination Assay 76

2.8.1 *In vitro* Assay 76

2.8.2 Western Blot 76

2.8.3 Densitometry Analysis 77

2.9 Biophysical Techniques 77

2.9.1 SV-AUC 77

2.9.2 ITC 79

2.9.3 Circular Dichroism 80

2.9.4 Thermal Melt Measurements 82

2.9.5 Glutaraldehyde Crosslinking 82

2.10 Electron Microscopy 83

2.10.1 Negative Stain Sample Preparation 83

2.10.2 Preparing Cryo-EM Grids 83

2.10.3 Clipping and Loading Grids 84

2.10.4 Screening Grids on the FEI Glacios 85

2.10.5 Data Collection on the Titan Krios 86

2.10.6 Data Processing 86

2.10.6.1 RELION 86

2.10.6.2 CryoSPARC 86

2.10.6.3 Motion Correction 87

2.10.6.4 Contrast Transfer Function (CTF) Estimation 87

2.10.6.5 Particle Picking 88

2.10.6.6 Particle Extraction 89

2.10.6.7 2D Classification 89

2.10.6.8 3D Classification 89

2.10.6.9 Ab initio Model Generation 89

2.10.7 UBE3A-only Pipeline 89

2.10.8 UBE3A+PSMD4 Pipeline 93

2.10.9 UBE3A+RLD2 Pipeline 96

2.11 X-Ray Crystallography 98
2.11.1 Hampton PCT 98
2.11.2 Setting Crystal Screens 99
2.11.3 Data Collection 100
2.11.4 Data Processing 101

3 Protein Expression and Purification 103

3.1 Expression of Proteins in a Bacterial vs Mammalian Cell System 103

3.2 Purification by Affinity Chromatography 104
3.2.1 UBE3A 104
3.2.2 UbcH7 106
3.2.3 PSMD4 109
3.2.4 RLD2 110
3.2.5 Ufrag 114

3.3 Purification by Ion Exchange Chromatography 115
3.3.1 UBE3A 115
3.3.2 PSMD4 116

3.4 Purification by Size Exclusion Chromatography 118
3.4.1 UBE3A 118
3.4.2 UbcH7 120
3.4.3 PSMD4 121
3.4.4 RLD2 123

3.5 UBE3A+E6+p53 124

3.6 HERC2 133
3.6.1 Cloning 134
3.6.2 Affinity Chromatography Purification and Optimisation of Gel Electrophoresis Detection 137
3.6.3 Identification through GFP Fluorescence 138

4 Biophysical Characterisation 140

4.1 Sedimentation Velocity Analytical Ultracentrifugation 140
4.1.1 UBE3A 140
4.1.2 UBE3A + PSMD4 141
4.1.3 UBE3A + RLD2 143

4.2 Isothermal Titration Calorimetry 144
4.2.1 UBE3A + PSMD4 144
4.2.2 UBE3A + RLD2 146
4.2.3 Ufrag + RLD2 148

4.3 Complex Formation Through Co-Purification 152
4.3.1 UBE3A+PSMD4 152
4.3.2 UBE3A+RLD2 154
4.3.3 Ufrag+RLD2 155

4.4 Circular Dichroism 159
4.4.1 UBE3A + RLD2 160

4.5 Nano Differential Scanning Fluorimetry 162
4.5.1 UBE3A+RLD2 162
4.5.2 UBE3A+PSMD4 163

4.5.3 UBE3A Buffer Optimisation 164
4.5.4 UBE3A Crosslinking 166
4.5.5 UBE3A+PSMD4 Crosslinking 173
4.5.6 UBE3A+RLD2 Crosslinking 175

4.6 Discussion 178

5 Biochemical Activity 182

5.1 *in vitro* Ubiquitination Assay 182

5.2 UBE3A Activity 182
5.2.1 UBE3A Autoubiquitination 182
5.2.2 UBE3A Activity in the Presence of RLD2 184
5.2.3 UBE3A Activity in the Presence of PSMD4 187
5.2.4 UBE3A Activity in the Presence of RLD2 and PSMD4 190

6 EM Sample Preparation and Optimisation 196

6.1 Negative Stain 196

6.2 Grid Types 198

6.3 Blotting Conditions 203

6.4 Sample Considerations 204

6.5 Screening Grids 205

7 Data Processing and Structures 207

7.1 UBE3A 207
7.1.1 Data Collection Considerations 207
7.1.2 Data Processing Considerations 210
7.1.3 UBE3A Structure 215

7.2 UBE3A+PSMD4 231
7.2.1 Data Collection Considerations 231
7.2.2 Data Processing Considerations 234
7.2.3 UBE3A+PSMD4 Structure 239

7.3 UBE3A+RLD2 242
7.3.1 Data Collection Considerations 242
7.3.2 Data Processing Considerations 242
7.3.3 UBE3A+RLD2 Structure 244

7.4 RLD2 248
7.4.1 Data Collection Considerations 248
7.4.2 RLD2 Structure 249
7.4.3 RLD2 Complex Crystals 264

8 Discussion 266

8.1 Oligomeric States of UBE3A 266

8.2 Substrate Binding Interfaces in UBE3A 269

8.3 Future Work 271

9 References 276

10 Appendices ***299***

10.1 Appendix 1 – UBE3A MSA **299**

10.2 Appendix 2 – pUBE3A UBE3A Sequence **300**

10.3 Appendix 3 – Disorder Predictions **301**

10.3.1 UBE3A 301

10.3.2 PSMD4 301

10.3.3 RLD2 301

10.4 Appendix 4 – RLDs MSA **302**

10.5 Appendix 5 – RobeTTA UBE3A Error Values **305**

10.6 Appendix 6 - Cryo-EM Theory **305**

10.6.1 Screening Grids on the Glacios 305

10.6.2 Data Collection on the Titan Krios 308

10.6.3 Data Processing 310

10.6.3.1 RELION 310

10.6.3.2 CryoSPARC 310

10.6.3.3 Motion Correction 311

10.6.3.4 CTF Estimation 313

10.6.3.5 Particle Picking 315

10.6.3.6 Particle Extraction 317

10.6.3.7 2D Classification 318

10.6.3.8 3D Classification 319

10.6.3.9 Ab initio Model Generation 321

10.6.3.10 Refinements and Postprocessing 321

10.7 Appendix 7 – PTM sites in UBE3A **323**

10.8 Appendix 8 – HERC2 Gel Electrophoresis **324**

Table of Figures

Figure 1: A ribbon representation of ubiquitin (1UBQ) with the residues involved in thioester bonds displayed in a ball-and-chain format 13

Figure 2: The three-step mechanism of ubiquitination by an E1, E2 and E3 enzyme. 15

Figure 3: The bipartite imprinting centre in the chromosome region 15q11.2. 17

Figure 4: Alternative splicing of the UBE3A gene 18

Figure 5: Genetic classes of Angelman syndrome 21

Figure 6: Six break points in the chromosome 15q11.2-13 region lead to different classes of deletions found in Angelman syndrome patients. 22

Figure 7: Some examples of downstream signalling processes regulated by UBE3A 24

Figure 8: Dendritic spines in AS neurons compared to healthy neurons 25

Figure 9: Genetic classes of Dup15q syndrome. 28

Figure 10: UBE3A is involved in several types of cancer, some are associated with a viral oncogene while others are triggered by changes in normal protein expression levels 31

Figure 11: An overview of the different identified regions within UBE3A, their locations within the amino acid sequence, cellular roles, and locations within the 3D protein. 40

Figure 12: The crystal structure of the UBE3A HECT domain in complex with the UbcH7 E2 enzyme (1C4Z). 41

Figure 13: The hydrophobic groove of the UBE3A small N-terminal subdomain, with the F63 residue of UbcH7 at the deepest point. 42

Figure 14: A surface representation of the UBE3A HECT domain with the key residues involved in the putative second E2~Ub binding site highlighted *43*
Figure 15: The proximal indexation model for polyubiquitin chain assembly catalysed by UBE3A *44*
Figure 16: A simulated demonstration of the proximal indexation model in action. *45*
Figure 17: The structure of the UBE3A AZUL domain. *46*
Figure 18: The ternary complex of UBE3A-E6-p53 solved by x-ray crystallography, showing the small LXXLL motif helix of UBE3A *47*
Figure 19: An overview of the current structural information available for UBE3A isoform 1. ..... *49*
Figure 20: The fragment of the UBE3A plasmid for which a sequence was provided. *51*
Figure 21: The pET15b-PSMD4 plasmid that was used to express a His-3C-PSMD4 construct in BL21 cells. *53*
Figure 22: A diagram representing the mechanism of the QuickChange mutagenesis protocol. *61*
Figure 23: A diagram demonstrating the overlap extension PCR process. *62*
Figure 24: Circularly polarised light is formed by a 90^{o} phase shift between the x and y components of the electric field *81*
Figure 25: A flowchart for the data processing steps involved in obtaining the UBE3A structure shown in chapter 7. ..... *92*
Figure 26: The FSC plot for the final model of UBE3A, generated in RELION. *93*
Figure 27: A flowchart for the data processing process involved in generating the final UBE3A+PSMD4 structure shown in chapter 7. ..... *95*
Figure 28: A flowchart for the data processing stages involved in generating the final UBE3A+RLD2 model shown in section 7.3.3. ..... *97*
Figure 29: The FSC plot for the final model of the UBE3A+RLD2 complex, generated in CryoSPARC *98*
Figure 30: Affinity chromatography of UBE3A using a 3-step HisTrap system *105*
Figure 31: The pETM41-UbcH7 plasmid used for expression of the UbcH7 protein with an N-terminal His tag, MBP tag, and TEV cleavage site. *107*
Figure 32: Affinity chromatography of UbcH7 using a 3-step MBPTrap and HisTrap system. *108*
Figure 33: Affinity chromatography of PSMD4 through a 3-step HisTrap system. *109*
Figure 34: The pETM11-RLD2 plasmid for expression of a His-TEV-RLD2 construct in E. coli cells. *110*
Figure 35: Affinity chromatography purification of His-MBP-RLD2 through a 4-step HisTrap plus MBPTrap system *111*
Figure 36: Affinity chromatography purification of His-RLD2. ..... *111*
Figure 37: Reverse HisTrap purification of RLD2 following small-scale TEV cleavage. *113*
Figure 38: The pETM40-Ufrag plasmid for expression of an MBP-TEV-Ufrag construct in E. coli cells. *114*
Figure 39: Affinity chromatography purification of MBP-Ufrag *115*
Figure 40: The anionic exchange purification stage of UBE3A purification *116*
Figure 41: Anionic Exchange of PSMD4 following affinity chromatography purification and SEC *117*
Figure 42: The SEC profile for UBE3A *119*
Figure 43: A) The SEC profile for UbcH7 ..... *120*
Figure 44: The SEC profile for PSMD4 following affinity chromatography *122*
Figure 45: The SEC profile for His-RLD2 following affinity chromatography purification *124*
Figure 46: The pACYCDuet1-E6-p53 plasmid used for co-expression of untagged E6 and p53 in BL21 or ArcticExpress cells. *125*
Figure 47: Initial purification of a co-expressed UBE3A-E6-p53 complex *126*
Figure 48: Strep resin purification of p53 co-expressed with E6. ..... *127*
Figure 49: The pCDFDuet1-E6-p53 plasmid for co-expression of untagged E6 and p53 in BL21 pLysS or Rosetta cells. *128*

Figure 50: Attempted purification of the UBE3A-E6-p53 complex from separate expressions of His-UBE3A and the pCDF plasmid expressing E6 and p53. 129
Figure 51: The pACYCDuet1-E6 plasmid for expressing the untagged HPV16 E6 protein in BL21 cells. 130
Figure 52: The pCDFDuet1-p53 plasmid to enable co-expression of p53, E6, and UBE3A across three different plasmids. 131
Figure 53: Co-expression of UBE3A isoform 1, p53, and E6 from three different plasmids, separated with a gradient HisTrap purification 132
Figure 54: The coverage of the sequencing data for the HERC2 gene along with the placements of the designed primers for the gene. 136
Figure 55: Electrophoresis gels of different types demonstrating high molecular weight species within a HERC2 purification sample 137
Figure 56: Various electrophoresis gels showing high molecular weight fluorescent products within HERC2 purification samples 138
Figure 57: Fluorescence scans of a small scale HERC2 expression test. 139
Figure 58: The SV-AUC distribution of species within a UBE3A sample 141
Figure 59: The SV-AUC distribution of a complex of UBE3A and PSMD4 142
Figure 60: The SV-AUC distribution of a complex of UBE3A and RLD2 143
Figure 61: The ITC isotherm (upper) and integrated heat-per-injection plot for the interaction between UBE3A and PSMD4 145
Figure 62: The ITC isotherm (upper) and integrated heat-per-injection plot for the interaction between UBE3A and RLD2. 147
Figure 63: The ITC isotherm (upper) and integrated heat-per-injection plot for the interaction between MBP-Ufrag and His-RLD2 149
Figure 64: The ITC isotherm (upper) and integrated heat-per-injection plot for the interaction between MBP and His-RLD2 151
Figure 65: The association of UBE3A and PSMD4 visualised through SEC. 153
Figure 66: Size exclusion of a complex of UBE3A and RLD2 154
Figure 67: Purification of a complex of His-RLD2 and MBP-Ufrag through consecutive affinity chromatography purifications 156
Figure 68: A reverse MBPTrap purification of the cleaved RLD2+Ufrag complex 157
Figure 69: Size exclusion purification of a tagged RLD2+Ufrag complex 158
Figure 70: The interaction between UBE3A and RLD2 as studied using the circular dichroism technique 160
Figure 71: The fraction of each species that forms each secondary structure element 161
Figure 72: The thermal melt shift profile for a UBE3A+RLD2 complex compared to its individual constituents 163
Figure 73: The thermal melt shift profile for a UBE3A+PSMD4 complex compared to its individual constituents. 164
Figure 74: The SEC profiles of a UBE3A sample subjected to a range of buffers. 165
Figure 75: The thermal melt profile of UBE3A in HEPES buffers at different pH values 166
Figure 76: Initial visualisation of the UBE3A crosslinking test reactions through SDS-PAGE and a Coomassie-based dye 168
Figure 77: Thermal melt shift analysis of UBE3A after a crosslinking reaction 169
Figure 78: Thermal melt analysis of several small-scale crosslinking test reactions 170
Figure 79: A) The SEC profile of the weakly crosslinked 'UBE3A crosslink 1' sample 171
Figure 80: A) The SEC profile of the strongly crosslinked 'UBE3A crosslink 2' sample 172
Figure 81: A crosslinked sample of UBE3A+PSMD4 after SEC 174
Figure 82: A crosslinked sample of UBE3A+RLD2 after SEC 176
Figure 83: The in vitro autoubiquitination assay of UBE3A 183
Figure 84: The in vitro ubiquitination assay of UBE3A in the presence of RLD2 184
Figure 85: Densitometry analysis of the ubiquitination of UBE3A with and without RLD2 186
Figure 86: The in vitro ubiquitination of UBE3A in the presence of PSMD4. 188

Figure 87: in vitro ubiquitination activity of UBE3A in the presence of PSMD4..... 189
Figure 88: An in vitro ubiquitination assay involving UBE3A, RLD2, and PSMD4. 191
Figure 89: The in vitro ubiquitination assay in the presence of both RLD2 and PSMD4 193
Figure 90: The process of optimising a sample for data collection involves several steps before a protein structure can be generated 196
Figure 91: Negative stain images of UBE3A stained with a 1% uranyl acetate solution 197
Figure 92: The effects of different grid conditions on ice thickness and particle distribution . 199
Figure 93: Representative images of a cryo-EM sample on a 300 mesh copper QuantiFoil R1.2/1.3 grid with a hand-applied GrOx-DDM coating..... 200
Figure 94: Representative images of a cryo-EM sample on a 300 mesh copper QuantiFoil R1.2/1.3 grid without any extra coatings 201
Figure 95: A 400 mesh copper grid with a LaceyCarbon film and an extra layer of super fine continuous carbon film..... 202
Figure 96: A diagram of the different layers in a CMOD direct electron detector 209
Figure 97: The Laplace operator applied to a gradual edge. 212
Figure 98: The low-resolution model of full-length UBE3A..... 216
Figure 99: The AlphaFold model of UBE3A human isoform 2 218
Figure 100: A comparison of the AlphaFold model with the low resolution cryo-EM model of UBE3A..... 219
Figure 101: The Robetta deep learning predicted models for UBE3A isoform 1 220
Figure 102: A comparison of the Robetta model with the low resolution cryo-EM model of UBE3A..... 221
Figure 103: A comparison of the three models of UBE3A 222
Figure 104: The low resolution cryo-EM model of full-length UBE3A with the predicted models fitted in and coloured by domain regions 224
Figure 105: A comparison of the two predicted structures of UBE3A..... 227
Figure 106: A comparison of the AS-related missense mutations and the domain boundaries within the Robetta prediction and low-resolution EM model of UBE3A..... *229*
Figure 107: An Argand diagram showing the effect of the volta phase plate (VPP). 233
Figure 108: UBE3A+PSMD4 data collection on a 200 kV Glacios microscope equipped with a Falcon IV detector, dataset 1 235
Figure 109: UBE3A+PSMD4 data collection on a 200 kV Glacios microscope equipped with a Falcon IV detector, dataset 2 236
Figure 110: UBE3A+PSMD4 data collection on a 300 kV microscope with the phase plate 237
Figure 111: The low-resolution model of a UBE3A+PSMD4 sample solved by cryo-EM..... 240
Figure 112: A comparison of the low-resolution UBE3A+PSMD4 cryo-EM model with the predicted structures of UBE3A 241
Figure 113: An example of a micrograph picked using the blob-picker job in CryoSparc..... 243
Figure 114: The low-resolution cryo-EM model for a crosslinked complex of UBE3A and the RLD2 domain of HERC2 245
Figure 115: Model of the UBE3A and RLD2 monomers fitted into the low-resolution UBE3A+RLD2 cryo-EM map..... 246
Figure 116: The UBE3A and RLD2 models fitted into the cryo-EM map of the complex, with the UBE3A model coloured to show the location of known domains. 247
Figure 117: The crystal structure of RLD2 solved by X-ray crystallography..... 249
Figure 118: A multiple sequence alignment for the three RLD domains within the HERC2 protein. 251
Figure 119: A comparison of the three RLD domains of HERC2..... 252
Figure 120: A multiple sequence alignment for the RLD domains within all HERC proteins ... 254
Figure 121: Evolutionary analysis of the RLD sequences within HERC proteins..... 255
Figure 122: A comparison of the four known structures of RLD domains within HERC proteins. 257
Figure 123: Evolutionary analysis of the RLD domains within all human proteins..... 259

Figure 124: A structural comparison of all human RLD or RCC domains structures within the PDB. *261*
Figure 125: A ribbon diagram comparison of the RCC1 and HERC2 RLD2 structures *262*
Figure 126: RLD structures in complex with binding partners. *263*
Figure 127: Rearrangements of the Robetta predicted model of UBE3A show how movements around the flexible linker regions could allow bridging between the potential substrate binding regions and the HECT domain *270*
Figure 128 Full MSA of human and mouse UBE3A genes *299*
Figure 129: The sequence for the UBE3A gene within the pUBE3A plasmid provided by Dr. Martin Scheffner of the University of Konstanz. *300*
Figure 130: The disorder prediction for UBE3A predicted by IUPred2A *301*
Figure 131: The disorder prediction for UBE3A predicted by IUPred2A *301*
Figure 132: The disorder prediction for the RLD2 domain of HERC2 predicted by IUPred2A *301*
Figure 133: A multiple sequence alignment of all proteins containing an RLD domain *305*
Figure 134: The error estimate for the Robetta model of UBE3A isoform 1 ..... *305*
Figure 135: Motion correction of micrographs involves tracing the movement of particles across the frames of a micrograph and summarising it as a series of vectors *312*
Figure 136: The point spread function *313*
Figure 137: The effects of defocus values on the CTF *314*
Figure 138: The possible PTM sites in the UBE3A sequence identified by the PhosphoSite Plus server *323*
Figure 139: A small scale Talon purification of HERC2 run through PAGE under denaturing conditions *324*
Figure 140: Small scale purifications of HERC2 run through PAGE under native conditions *325*
Figure 141: A small scale purification trial of HERC2 subjected to a commercial blue-native PAGE gel *326*
Figure 142: Horizontal agarose gels for large protein visualisation *328*
Figure 143: Hand poured BioRad mini-protean tris-acetate gels with different acrylamide percentages *329*
Figure 144: 9% acrylamide large tris-acetate gels with HERC2 GFP-nanobody purification samples *330*
Figure 145: 6% acrylamide large tris-acetate gels with HERC2 GFP-nanobody purification samples *331*

Table of Tables

Table 1: A summary of the key interaction partners and targets of UBE3A in HPV-associated cervical cancer. *34*
Table 2: A summary of the key interaction partners and targets of UBE3A associated with HPV-associated oropharyngeal cancer *35*
Table 3: A summary of the key interaction partners and targets of UBE3A in HCV-associated hepatocellular carcinomas. *35*
Table 4: A summary of the interaction partners and targets of UBE3A in the development and progression of prostate cancer. *37*
Table 5: A summary of the key interaction partners and targets of UBE3A associated with B-cell lymphoma *37*
Table 6: A summary of the key interaction partners and targets of UBE3A associated with NSCLC progression. *38*
Table 7: A list of the primers used during this project. *57*

Table 8: An optimisation screen was attempted using small scale glutaraldehyde crosslinking reactions and a range of reaction conditions in order to determine the conditions that led to the most extensive UBE3A crosslinking without inducing aggregation of the sample. *167*
Table 9: The rate constants for each of the constituents of the UBE3A in vitro ubiquitination assays *193*
Table 10: The associated crystallography statistics for the RLD2 structure *250*

1 Introduction

1.1 Ubiquitination

UBE3A (also known as E6AP) is an enzyme involved in ubiquitination within human cells (Scheffner *et al.*, 1993). Ubiquitination is a form of post-translational modification that involves addition of ubiquitin, a 76 residue peptide, to proteins (Hershko and Ciechanover, 1992). Ubiquitination is predominantly known for targeting proteins for degradation, but it can also trigger a wide range of functions for target proteins, including cellular localisation, signal transduction, and even protein activation. (Hershko and Ciechanover, 1992; Komander *et al.*, 2009; Mukhopadhyay and Riezman, 2007). The seven lysine residues (K6, K11, K27, K29, K33, K48, K63) and the amino-terminal residue (M1) of ubiquitin can form the isopeptide or peptide bond that binds ubiquitin to its substrate (Fig. 1).

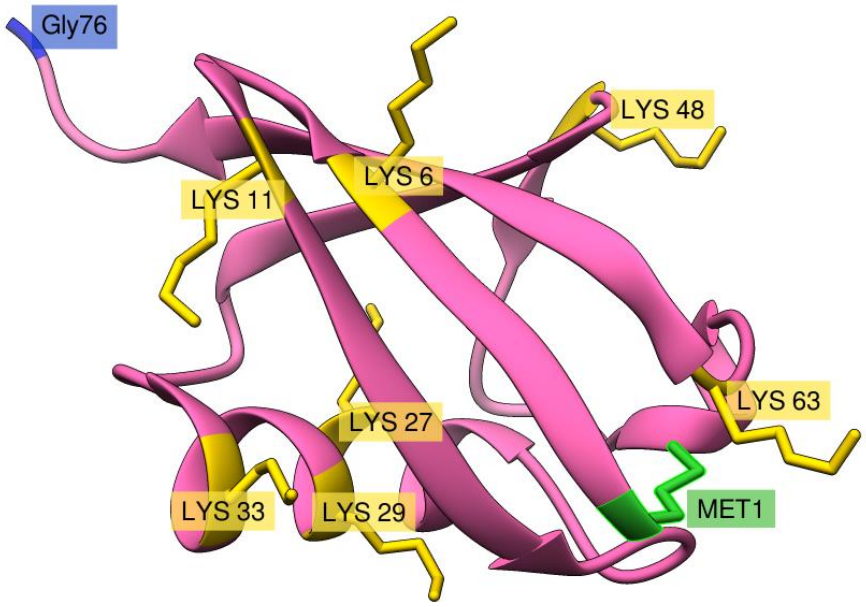


Figure 1: A ribbon representation of ubiquitin (1UBQ) with the residues involved in thioester bonds displayed in a ball-and-chain format. The lysine residues are highlighted in yellow, the N-terminal methionine residue in green, and the C-terminal glycine residue in blue.

The residue involved in formation of the ubiquitin bond, along with the number of ubiquitin units added, changes the shape of the ubiquitin tag, allowing the cell to recognise and respond differently to different forms of ubiquitination, leading to the range of cellular outcomes (Ye *et al.*, 2012). The most prevalent form of ubiquitination is K48, where several ubiquitin units are linked together through interactions between the K48 residue of one ubiquitin and the C-terminal glycine residue (G76) of the next, with the end of the chain bound to the substrate via an isopeptide bond between G76 of the ubiquitin unit and a lysine residue on the surface of the substrate protein (Hershko and Ciechanover, 1992; Ronchi *et al.*, 2017). This form of

ubiquitination marks a protein for recognition by the proteasome, a large protein complex responsible for the degradation of cellular proteins (Hersko and Ciechanover, 1992). Other types of ubiquitination have also been characterised including; K11 chains that are also linked to proteasomal degradation, linear M1 chains have been linked to NF-κB activation, K63 chains are implicated in NF-κB signalling, DNA repair and lysosomal targeting, and mono-ubiquitin tags are associated with protein interactions and cellular localisation (Woelk *et al.*, 2007; Akutsu *et al.*, 2016). However, other ubiquitin chain types are still not so well understood and could signal for an even wider variety of outcomes (Kliza and Husnjak, 2020). These chains include K6, K29, K33, K27 linkages, as well chains formed of more than one linkage type, both linear and branched (Woelk *et al.*, 2007).

Ubiquitination is a three-step process of activation, conjugation, and ligation, each catalysed by a different class of enzymes. The first step, activation, is catalysed by an E1 ubiquitin-activating enzyme, and is itself a two-step process. Firstly, the C-terminal carboxylate group of the ubiquitin monomer is acyl-adenylated, before transfer to an active-site cysteine residue, resulting in a thioester bond between the C-terminus of ubiquitin and the E1 sulfhydryl group (Fig. 2a). The second step, conjugation, is catalysed by an E2 ubiquitin-conjugating enzyme, which interacts with both the activated ubiquitin and the E1 enzyme to catalyse a transthioesterification reaction, transferring the ubiquitin from the E1 to the E2 active site cysteine (Fig. 2b). The final ligation step is catalysed by an E3 ubiquitin-ligase enzyme, of which UBE3A is an example (Scheffner *et al.*, 1993). The E3 enzyme transfers the ubiquitin to its target substrate, either by direct transfer from the E2 active site or via transthioesterification of the E3 active site depending on the type of E3 enzyme, completing the ubiquitination process (Pickart, 2001; Huang *et al.*, 1999; Scheffner *et al.*, 1995). The human genome encodes only 2 E1 enzymes, UBA1 and UBA6, around 40 E2 enzymes, and over 600 distinct E3 enzymes, many of which have several isoforms (Wang *et al*., 2017; Eletr and Kuhlman, 2007). The increased number of E3 enzymes allows for a higher degree of variation among them with at least 2 major subclasses of E3 enzymes identified: the HECT (Homology to E6AP C-Terminus) and RING (Really Interesting New Gene) ligases, each with a characteristic C-terminal motif and a distinct mechanism of catalysing the ligation reaction (Pickart, 2001). HECT ligases transfer the activated ubiquitin to a site within the HECT domain, before transfer to the substrate protein, while RING ligases bring the loaded E2 and the substrate within a close proximity to catalyse direct transfer of the ubiquitin from the E2 to the target protein (Fig. 2c) (Pickart, 2001). UBE3A is the founding example of a HECT-type enzyme, with the motif defined by a sequence similarity to the C-terminal sequence of the enzyme (Huibregste *et*

al., 1995).

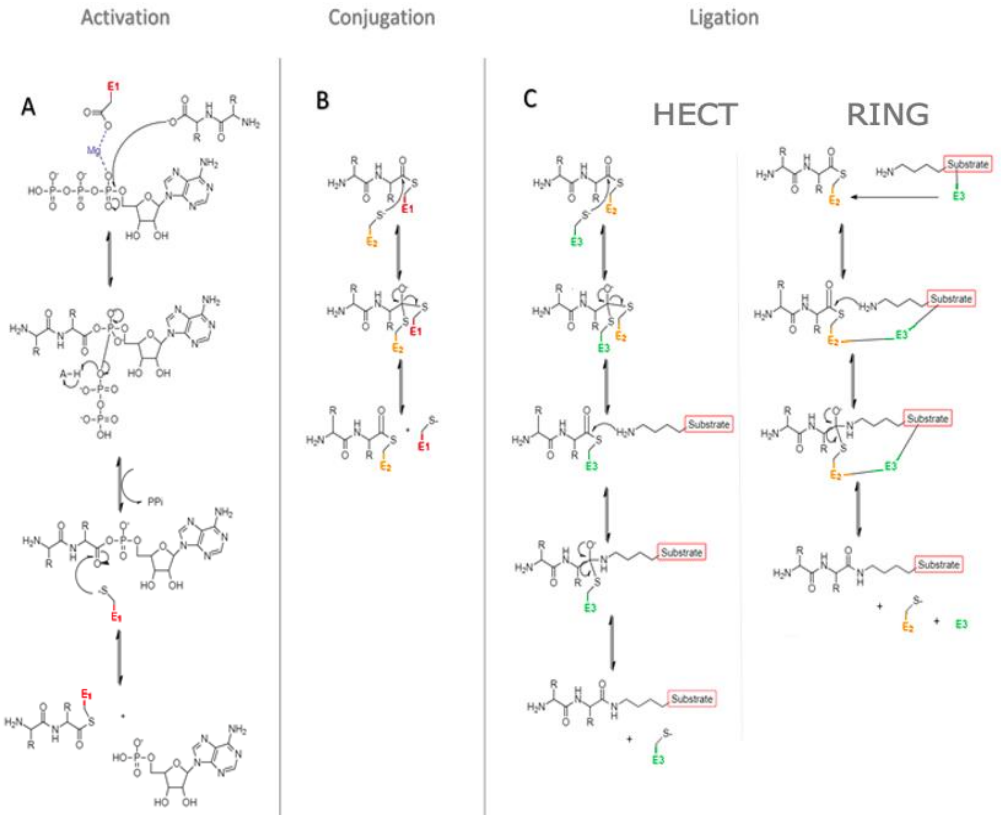


Figure 2: The three-step mechanism of ubiquitination by an E1, E2 and E3 enzyme. A) Activation of a ubiquitin moiety by an E1 enzyme, via acyl adenylation of the ubiquitin and subsequent conjugation. B) Conjugation of ubiquitin to the E2 enzyme. C) Ligation of the ubiquitin moiety to the substrate protein via one of two mechanisms. HECT ligases form a thioester intermediate with the ubiquitin before transfer to the substrate, while RING ligases bridge between the substrate and the E2 enzyme, bringing them into close proximity for the transthioestereification reaction to occur.

All ubiquitination processes occur via a combination of enzymes in the E1-E2-E3 cascade, and although each E3 enzyme has a preference for specific upstream E2 and E1 enzymes, there is a degree of redundancy and overlap in the system. A single E1 enzyme can interact with many E2 enzymes, and each E2 enzyme can react with several E3 enzymes, and each E3 enzyme can also interact with more than one E2 enzyme. This results in a complex system of cross-talk between ubiquitination cascades. UBE3A preferentially accepts ubiquitin from the UbcH7 E2 enzyme, but it can also function downstream of several UbcH5 isoforms and UbcH8, with the upstream E2 enzyme subtly altering the preferred lysines and substrate specificity of the ubiquitin ligase (David *et al.,* 2010; Eletr and Kuhlman, 2007; Huang *et al.,* 1999).

1.2 *UBE3A* Epigenetics and isoforms

UBE3A is the gene that encodes the UBE3A protein in humans. *UBE3A* appears to have a complicated evolutionary history; it has orthologs in organisms that

pre-date the nervous system, yet it has been lost in several higher order lineages, and across some nematode lineages, including *C. elegans* (Sato, 2017; Grau-Bové *et al.,* 2013). An interesting feature of *UBE3A* evolution is the emergence of a paternal imprinting mechanism that occurs after the emergence of viviparous mammals (Rapkins *et al.,* 2006).

Genomic imprinting refers to a parent of origin specific regulation of gene expression. *UBE3A* is paternally imprinted in neurons, which means that in neuronal cells only the maternal copy of the gene is expressed (Knoll *et al.,* 1989; Rougeulle *et al.,* 1997). Analysis of both mRNA and protein levels suggest that *UBE3A* is expressed in significant levels in most tissue types of the human body (Uhlen *et al.,* 2015; Tissue expression of UBE3A - The Human Protein Atlas, 2020; Sirois *et al.*, 2020), and in all cells except neurons it is expressed biallelically (Rougeulle *et al.,* 1997; Yamasaki *et al.,* 2003). The reason for this paternal imprinting solely in neurons is not understood. It has been suggested that it functions to allow tighter regulation of UBE3A levels in the brain as its dysregulation leads to a variety of disorders, but studies have shown that the maternal expression increases concordantly with decreasing paternal expression of the gene to maintain protein levels despite the paternal imprinting (Hillman *et al.,* 2017). One suggestion is that the imprinting acts to regulate the expression of different UBE3A isoforms, rather than total levels of UBE3A (Lopez *et al.,* 2019), although how this would occur is not explained.

The silencing of paternal *UBE3A* in neurons is controlled by a series of epigenetic marks around a bipartite imprinting centre in the chromosomal region 15q11.2, roughly 1 Mb upstream of the *UBE3A* gene (LaSalle *et al.,* 2015; Lalande and Calciano, 2007; Fig. 3). The bipartite nature of this imprinting centre refers to two separate conditions caused by genes within and around this area; Prader-Willi Syndrome (PWS) is caused by loss of the maternally imprinted genes *MKRN3, MAGEL2, NDN, PWRN1, NPAP1,* and *SNRPN*, while Angelman Syndrome (AS) is caused by defects in the paternal imprinting of *UBE3A*, and in some cases other genes downstream as well (Knoll *et al.,* 1989; LaSalle *et al.,* 2015; Lalande and Calciano, 2007). The 15q11.2 region encodes the *SNRPN-SNORD* locus, which contains the *SNRPN* (Small Nuclear Ribonucleoprotein-associated Protein N) gene, the SNORD116 and SNORD115 loci, which encode several tandem repeats of snoRNA (small nucleolar RNA) molecules *SNORD116* (snoRNA region D116) and *SNORD115* (snoRNA region D115) respectively, as well as the *UBE3A* gene (Fig. 3) (LaSalle *et al.,* 2015; Lalande and Calciano, 2007).

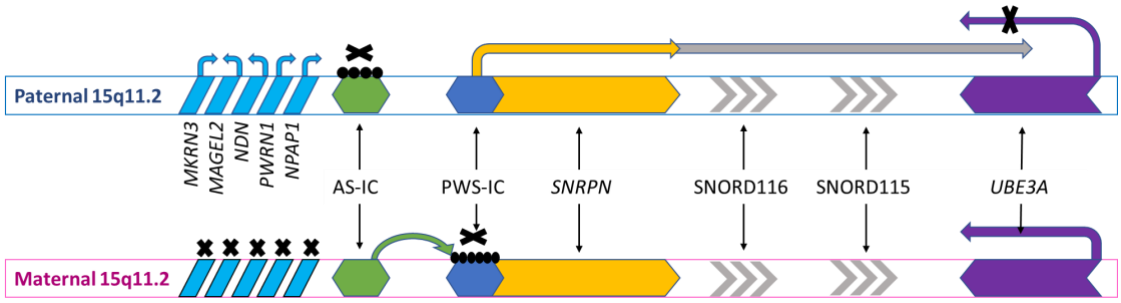


Figure 3: The bipartite imprinting centre in the chromosome region 15q11.2. The coloured blocks represent different DNA elements, with the upstream maternally imprinted genes shown in light blue, the AS-IC in green, the PWS-IC in dark blue, the SNRPN gene in yellow, the SNORD116 and SNORD115 loci in grey, and the UBE3A gene in purple. Black circles represent methylated sites on the DNA, black crosses represent inactivated regions. The purple arrow represents transcription of the UBE3A gene, the yellow arrow represents transcription of the SNRPN gene in non-neuronal cells, and the grey arrow represents transcription of SNRPN and the UBE3A-ATS in mature neurons. The blue arrows represent transcription of maternally imprinted genes in neuronal cells.

The key differences in maternal and paternal copies of this 15q11.2 region are epigenetic markers on two imprinting centres, regions upstream of the coding region that do not encode any gene products but control the expression of the downstream areas. Oocyte-specific transcription of noncoding exons upstream of the maternal AS-IC protect the maternal copy from methylation during early development, while the paternal AS-IC is silenced by methyl marks. The un-methylated maternal AS-IC triggers methylation of the maternal PWS-IC via an unknown mechanism, resulting in active transcription of the downstream *SNRPN-SNORD* locus on the paternal allele, while the maternal copy is silenced (Shemer *et al.,* 2000; LaSalle *et al.,* 2015; Lalande and Calciano, 2007). In all cell types, activation of the PWS-IC triggers translation of the downstream *SNRPN* gene, but in an environment specific to post-natal neuronal cells, the paternal allele is influenced to transcribe past the end of the *SRNPN* gene, through the SNORD116 and SNORD115 loci, and through the *UBE3A* gene in an anti-sense direction. The product of this elongated transcription is a long piece of mRNA termed the *UBE3A* antisense transcript (*UBE3A-ATS*) (Rougeulle *et al.,* 1998; Meng *et al.,* 2012; Fig. 3). This prevents meaningful expression of the paternal *UBE3A* gene through the collision model, where polymerases transcribing the gene in both directions will collide and be knocked from the DNA, preventing full transcription of the gene (Rougeulle *et al.,* 1998; LaSalle *et al.,* 2015; Lalande and Calciano, 2007).

Aside from its interesting epigenetic regulation, another level of complexity is added to the *UBE3A* gene via alternative splicing of the transcript, resulting in

three different isoforms of the UBE3A protein (Fig. 4) (Yamamoto *et al.,* 1997).

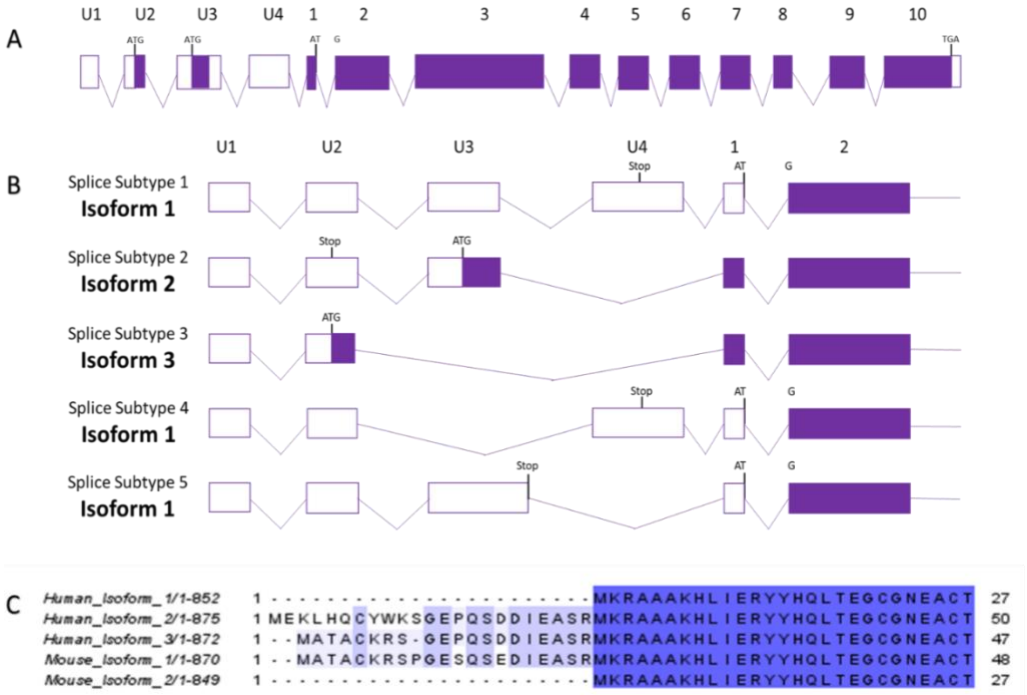


Figure 4: Alternative splicing of the UBE3A gene. A) The exons that make up the human UBE3A gene. U1-U4 are untranslated exons while 1-10 are translated exons (labelled from isoform 1). The potential start sites for the different isoforms are labelled as 'ATG', exons involved in each isoform are coloured in purple, and the stop site for all isoforms is labelled as 'TGA' in exon 10. B) The alternative splicing products of UBE3A. There are 5 different splicing products, but the placement of start and stop codons leads to only 3 protein isoforms after translation. C) Multiple Sequence Alignment of the three human isoforms and the two accepted mouse isoforms of UBE3A, showing how the different splicing patterns change the N-terminal sequence of the gene. The regions correlating to the first 50 residues of human isoform 2 are shown. Sequences are highlighted in shades of blue correlating to a percentage sequence identity between the five sequences. Residues that are present in a higher percentage of aligned sequences are shown highlighted in darker shades of blue, while residues that are present in only one of the aligned sequences are shown with a white background.

In humans, these three isoforms differ only in their extreme N-terminus, where isoform 2 is the longest sequence, isoform 1 is identical to isoform 2 apart from a loss of the first 23 amino acids, and isoform 3 is similar to isoform 2 except that the first 10 residues of isoform 2 are replaced with 8 different amino acids (Yamamoto *et al.,* 1997; UniProt accession Q05086; Appendix 1). Isoform 2 has been determined as the canonical sequence for the gene (The UniProt Consortium, 2019) but it is likely that isoform 1, the shortest isoform, is the most predominant in human cells (Sirois *et al.,* 2020). Much of what I know about *UBE3A* comes from work in mice (*Mus muscularis*), where the mouse gene is commonly annotated as *Ube3a*, and it

comprises three isoforms (The UniProt Consortium, 2019; UniProt accession O08759), although the classification of mouse isoform 3 as a full isoform has recently been called into question (Avagliano Trezza *et al.,* 2019). Although the mouse and human genes are not 100% identical, they are very similar (Fig. 4c), with human isoform 1 sharing a sequence identity of 97% with mouse isoform 2 (Appendix 1), and human isoform 3 shares a 95.7% sequence identity with mouse isoform 1 (Appendix 1). Human isoform 2 does not have a direct equivalent in the mouse proteome, although it is more similar to isoform 1 than isoform 2 (Fig. 4c). Mouse isoform 3 also has no direct equivalent, but as mouse isoform 3 is identical to mouse isoform 2 until the C-terminus, where it lacks the entire HECT domain, its physiological relevance has come into doubt (Avagliano Trezza *et al.,* 2019).

Research in mice suggests that Ube3a is present in both the nucleus and cytoplasm of cells. It is thought that the Ube3a protein binds to another protein called PSMD4 (also known as Rpn10 and S5a) for transport into the nucleus, where the shorter mouse isoform 2 is retained while the extra N-terminal region on mouse isoform 1 causes it to return to the cytoplasm (Avagliano Trezza *et al.,* 2019). It has also been shown that mouse isoform 2 is the most abundant form of the protein, and the most involved in the pathogenesis of an AS phenotype (Avagliano and Trezza *et al.,* 2019). This correlates to the human isoform 1 being the most predominant and physiologically important isoform, despite UniProt's assignment of isoform 2 as the canonical sequence (UniProt accession Q05086). However, a recent study in human cells suggests that all 3 human isoforms of UBE3A are present across both the cytoplasm and the nucleus (Sirois *et al.,* 2020). This study identifies isoform 1 as the most prevalent, making up 84-88% of all UBE3A in the cell, however, the distribution of isoforms 2 and 3 amongst the remaining 12-16% is not specified. This study suggests that there is not one nuclear and one cytoplasmic isoform, as postulated by studies in mice, but rather that all isoforms are spread across both sub-cellular locations, suggesting that the difference between the isoforms may instead be in their substrate specificity or level of ubiquitin ligase activity, regardless of cellular location (Sirois *et al.,* 2020). The observation that UBE3A is present in both the nucleus and cytoplasm, but is more concentrated in the nucleus, has also been observed in post-mortem human brain cells (Burette *et al.,* 2018).

Angelman syndrome is a pathological state caused by loss of UBE3A in neurons (Kishino *et al.,* 1997; LaSalle *et al.,* 2015; Tan and Bird, 2016). The disorder has been well characterised in terms of an observable phenotype in individuals, as well as an observable cellular activity (LaSalle *et al.,* 2015; Frolich *et al.,* 2019). Due to the predominance of human isoform 1 in cells, it was suggested that loss of isoform 1 alone may be responsible for AS characteristics (Avagliano Trezza *et al.,* 2019). A recent study using human cells shows that when only isoform 1 is knocked out, the cells do show some

characteristics of AS cells, although much less than was expected (Sirois *et al.,* 2020). Similarly, three AS individuals have been identified in a clinical setting who lack only the isoform 1 form of the protein. These individuals fit enough of the criteria to be diagnosed with Angelman Syndrome, but their symptoms appear much less severe than standard AS cases, with each demonstrating a much greater capacity for speech and communication than their typical AS counterparts (Sadhwani *et al.,* 2018). This is further underpinned by another mouse study using overexpression of the mouse isoform that correlates to human isoform 3 (labelled in the paper as *Ube3a 2* but in this report as *Ube3a 1*). Overexpression of this form of the protein replicates the phenotypes of Dup15q syndrome, a neurodevelopmental disorder caused by duplication of the 15q11.2-13.3 region containing *UBE3A* (Copping *et al.,* 2017). Together these observations suggest that although isoform 1 plays a large role in UBE3A signalling in both healthy and pathological states, isoforms 2 and 3, are far from irrelevant and more work is needed to understand the roles of all three forms of the protein.

1.3 Neurodevelopmental Disorders

UBE3A is active in all cells of the human body, yet many disease states caused by dysregulation of the protein are neurodevelopmental disorders. This is a direct result of the paternal imprinting of *UBE3A* in neurons, meaning that if the maternal allele is defective in some way, the paternal allele is unable to compensate in neurons. The most prominent disorder associated with UBE3A is Angelman syndrome (AS), a neurodevelopmental disorder associated with loss of UBE3A. UBE3A is also implicated in Dup15q syndrome, caused by duplication of the 15q11.2-12.4 chromosomal region containing *UBE3A*, as well as less defined autism spectrum disorders (ASD). The chromosomal region containing *UBE3A*, though not necessarily UBE3A itself, is associated with another neurodevelopmental disorder known as Prader-Willi syndrome (PWS), which shares some features with AS (LaSalle *et al.,* 2015).

1.3.1 Angelman Syndrome

Angelman syndrome was first characterised in 1965 by a British paediatrician, Harry Angelman (Angelman, 1965), and is characterised by a global developmental delay leading to severe intellectual disability, speech impairment, ataxia, and a unique behavioural profile including an exceptionally happy demeanour and a fascination with water (Tan and Bird, 2016). Other symptoms, such as sleep disruptions, clinical seizures, and microencephaly are frequently observed, but not in all patients (LaSalle *et al.,* 2015). Many people with AS will never talk, and those that do only use a few words, with an average 2 words used by children with the disorder, and up to 5 words for adults. However, their receptive language skills are significantly stronger than their expressive language skills, and they are often able to communicate through body language, gestures, and non-verbal vocalisations (Wheeler *et al.,* 2017). Most patients with a chromosomal deletion will never

walk, and though many with AS caused by other genetic causes can walk independently, they will display an abnormal gait compared to their neurotypical peers (Wheeler *et al.,* 2017). The disorder is caused by loss of functioning UBE3A in neurons, but the way this occurs can be through one of four mechanisms (Fig. 5), with the different genetic classes conferring different phenotypic presentations of AS.

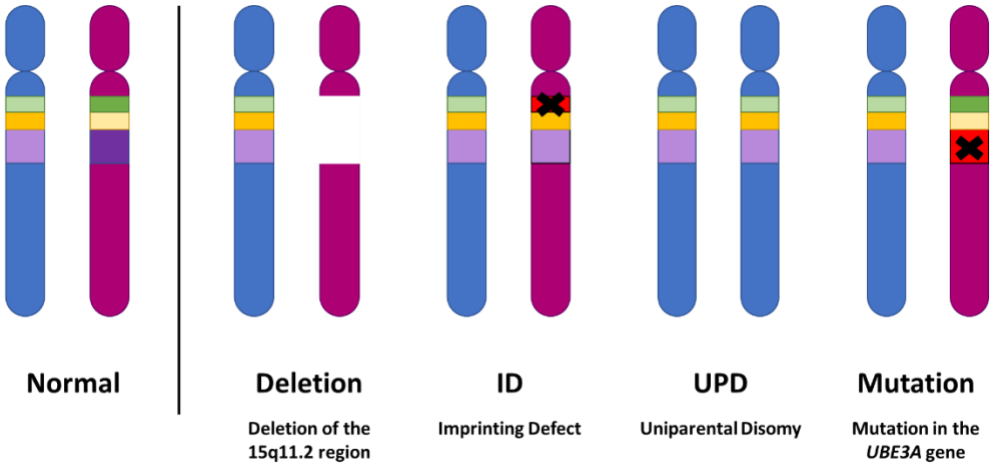


Figure 5: Genetic classes of Angelman syndrome. In each diagram the blue shape represents the paternal chromosome 15 and the purple shape represents the maternal chromosome. The green square represents the AS-IC, the yellow square the SNURF/SNRPN region upstream of UBE3A, and the purple box represents the UBE3A gene. Bright colours represent an active region while pastel shades represent an inactivated region. A red square and a black cross show an inactivating mutation.

The majority of AS cases are caused by a deletion of the entire maternal 15q11.2-13 chromosomal region (Fig. 3), along with a downstream region containing the *GABRB3, GABRA5, GABRG3* and *OCA2* genes, however, the size of the deletion can differ between patients. In the region of the chromosome surrounding the *UBE3A* gene there are six defined areas, known as break points (BP), where the DNA is more susceptible to breaking, causing the deletion of the chromosome region. 90% of deletions end at BP3 and start at either BP1 (class I deletions) or BP2 (class II deletions), although the deletion can include up to BP4, BP5, or BP6 in some cases (Fig. 6).

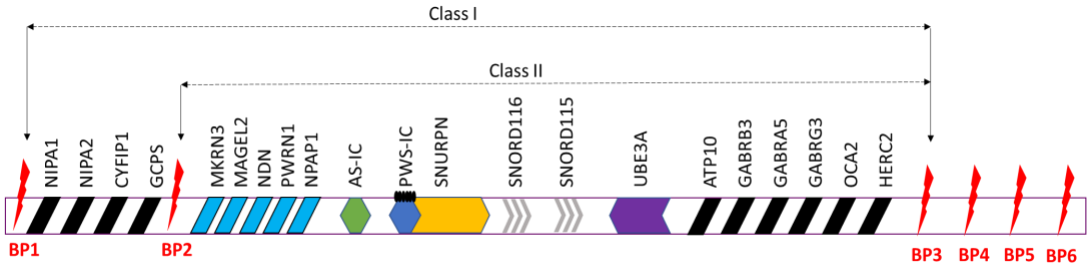


Figure 6: Six break points in the chromosome 15q11.2-13 region lead to different classes of deletions found in Angelman syndrome patients. DNA elements in the chromosome region are shown as coloured blocks, following the same scheme as figure 3, with the UBE3A gene shown in purple, the SNORD 116 and SNORD115 loci shown in grey, the maternally imprinted genes shown in light blue, and the AS-IC and PWS-IC are shown is green and dark blue respectively. The black blocks represent other genes in the chromosome region that are not imprinted as part of the bipartite imprinting centre, and the red lightning bolt shapes represent the six defined break points in the DNA.

Class II cases are more common than class I cases, with class I deletions accounting for 40% of all deletion cases and class II cases accounting for 50%. Deletion cases present a more severe phenotype than the other forms of AS, and class I deletions are more severe than class II, due to the loss of extra genes contributing to the phenotype (Frohlich *et al.,* 2019; Gentile *et al.,* 2010).

The phenotypic distinction between the other AS classes is less clear than that observed between deletion cases and all other forms of AS, although uniparental disomy (UPD) patients do appear to display a slightly reduced phenotype compared to deletion phenotypes (Gentile *et al.,* 2010). They experience better physical growth and reduced likelihood of microcephaly, along with fewer movement abnormalities, ataxia, and seizures (Dagli *et al.,* 2011). UPD occurs when an individual inherits two copies of a chromosome from one parent, instead of a single copy from each parent. In AS, the patient has two copies of the paternal chromosome 15, each of which display the epigenetic marks causing imprinting of the *UBE3A* gene in neurons (Fig. 5).

Individuals with AS caused by imprinting defects (ID) (Fig. 5) either have a deletion of the imprinting centres responsible for activation of the maternal *UBE3A* gene, or they have an epimutation displaying the epigenetic marks characteristic of the paternal 15q11.2 region on the maternal DNA (Dagli *et al.,* 2012). Possibly 10% of all ID AS cases are caused by small deletions in the imprinting control region, and most of these appear to be inherited mutations (Dagli *et al.,* 2011). The majority of ID cases are caused by epimutations, and the relatively high occurrence of mosaic ID cases, a less severe presentation of AS where only some of the cells expressing *UBE3A* contain the pathogenic epigenetic marks and a proportion of cells display the healthy epigenotype (Le Fevre *et al.,* 2017), suggests that the epimutations occur post-zygotically

(Buiting *et al.,* 2016; Dagli *et al.,* 2011). This means that many cases of AS caused by ID are *de novo* rather than inherited.

The final major class of AS is caused by genetic mutations within the *UBE3A* gene (Fig. 5). These mutations can take various forms with differing effects on the resultant UBE3A protein (Sadikovic *et al.,* 2014; Dagli *et al.,* 2011). Mutation cases can be the least severe presentation of AS, although the severity depends on the effect of the mutation on the protein (Sell and Margolis, 2015; Gentile *et al.,* 2010). One type of mutation is a splice variant, introducing irrelevant introns into the translated protein or removing exons containing integral domains of the protein. Another form is a nonsense mutation, wherein a single nucleotide change in the DNA creates a premature stop codon, resulting in a truncated form of the protein. Frameshift mutations occur when there is either a deletion or insertion of any number of nucleotides not divisible by three, which interferes with the way ribosomes read transcripts, leading to the sequence downstream of the mutation being transcribed in a different reading frame and no longer resembling the conventional UBE3A enzyme. The final form of *UBE3A* mutation that leads to AS is a missense mutation. This describes a class of mutation where a single nucleotide is substituted for another, causing a change in the resulting amino acid. Many missense mutations are non-pathogenic, or expected non-pathogenic, but several missense mutations do lead to the AS phenotype (Sadikovic *et al.,* 2011). Missense mutations identified in AS often help to identify the key catalytic residues of UBE3A, as they highlight the areas where a single amino acid change can disrupt the activity of the entire protein. Some AS cases display in-frame mutations, which result in insertion or deletion of amino acids in the protein without altering most of the protein sequence. These mutations are less common, and more resemble missense mutations than other deletion/insertion mutations (Sadikovic *et al.,* 2014). Many AS mutations occur within the HECT domain of UBE3A, or create a frameshift or truncation that precede the HECT domain, but there are also several missense mutations identified within other regions of UBE3A that do not obviously affect the catalytic activity of the HECT domain (Sadikovic *et al.,* 2014; Kuhnle *et al.,* 2018). It has been estimated that roughly 30% of AS causing mutations are inherited, while the remainder are *de novo*.

Of all the genotypic classes of AS, deletions are the most common, accounting for roughly 70% of all AS cases. An estimated 9% of cases are caused by UPD, 8% are caused by ID, and 11% are caused by mutations in the *UBE3A* gene (Tan and Bird, 2016). However, roughly 10% of people who display an AS phenotype do not fit into any of these classes (Wheeler *et al.,* 2017), they are often referred to as ‘AS-like’ cases rather than AS and could arise from loss of or mutations in downstream proteins in UBE3A signalling processes.

Although the genetic causes of the disorder are well understood, the mechanisms at a cellular level are not as well characterised. Several proteins identified as downstream effectors of UBE3A have been implicated in cellular mechanisms related to the AS phenotype (Tan and Bird, 2016) (Fig. 7).

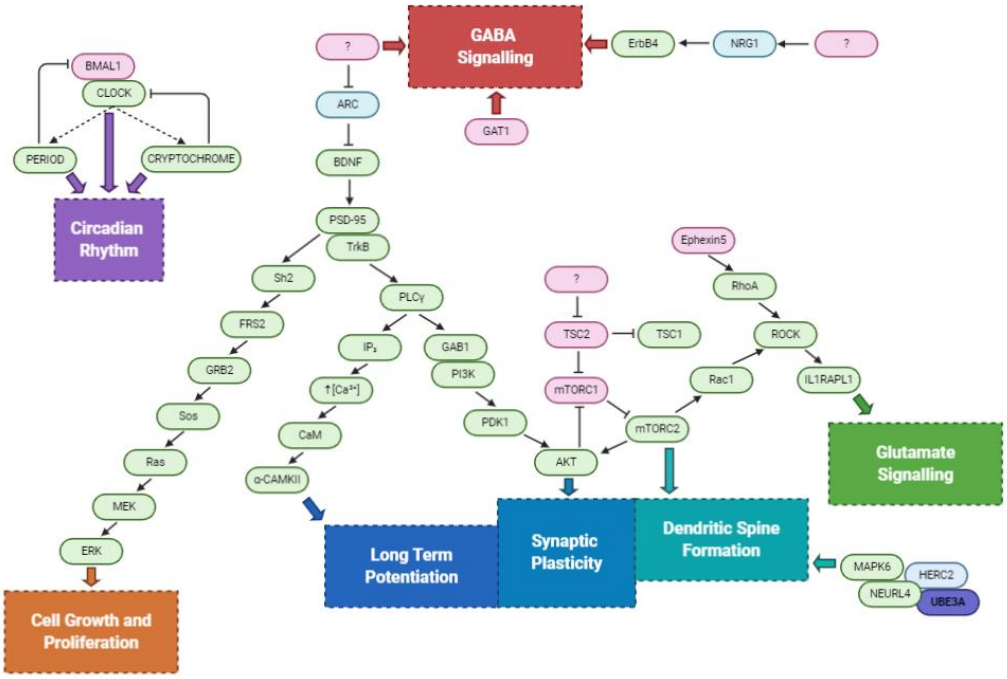


Figure 7: Some examples of downstream signalling processes regulated by UBE3A. Proteins in pink are confirmed substrates of UBE3A, proteins in blue are potential substrates of UBE3A, and proteins in green are downstream enzymes that are affected by UBE3A activity without interacting with the protein itself. Solid black arrows show an activating affect from one protein to another, dashed arrows show transcriptional activation, and solid black lines ending in a thick perpendicular line show an inhibitory effect from one protein to another. Large coloured arrows link enzymes at the end of a signalling pathway with the observable process it is involved in.

Arc is a synaptic protein responsible for internalisation of AMPA-type glutamate receptors (AMPARs) (Tan and Bird, 2016). Arc is regulated at the transcriptional level by UBE3A (Kuhnle *et al*., 2013), so impaired UBE3A function leads to increased Arc levels and subsequent internalisation of AMPARs, resulting in impaired glutamate signalling (Tan and Bird, 2016) (Fig. 7). Arc also downregulates BDNF (brain-derived neurotrophic factor) (Tan and Bird, 2016), a hormone involved in development of dopaminergic, GABAergic, cholinergic, and serotonergic neurons, as well as regulation of synaptic plasticity through TrkB-PSD-5 signalling (Cao *et al.,* 2013). BDNF facilitates the association of PSD-95 to the TrkB receptor, which acts through PLCγ and Gab1 to activate PLCγ-αCAMKII and PI3K-Akt signalling cascades, essential for long term potentiation (LTP), the molecular mechanism of learning and memory (Cao *et al.,* 2013) (Fig. 7).

Abnormal dendritic spine formation is a characteristic feature of AS neurons (Dindot *et al.,* 2007). Dendritic spines are short protrusions lining dendrites that interact with axons of another neuron to form synapses. Immature dendrites usually have thin, “filopodia-like” spines, which mature into short spines with large “mushroom-shaped” heads. Dendritic spines in AS neurons tend to show more variability, with less uniform spine lengths and head sizes, as well as appearing at a lower density (Tan and Bird, 2016) (Fig. 8).

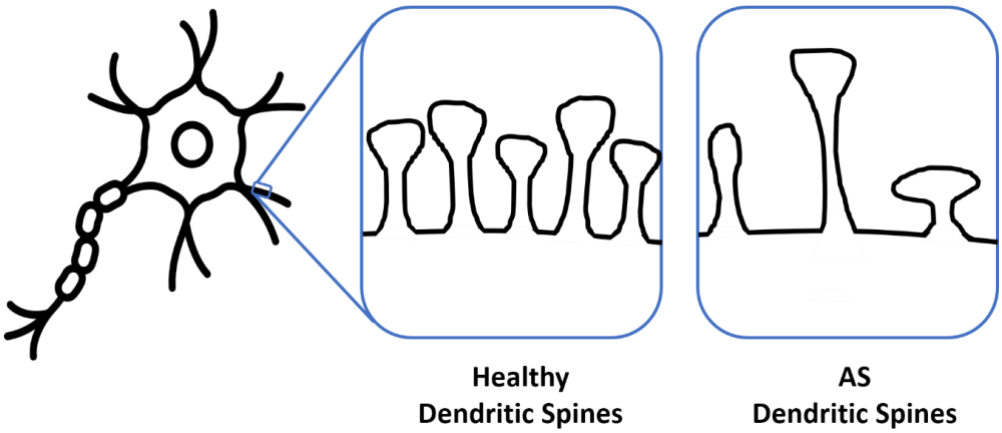


Figure 8: Dendritic spines in AS neurons compared to healthy neurons. Dendritic spines line the dendrites of neurons and form the synapses with connecting neurons. In AS, spines are more variable in length and head size than their healthy counterparts and show less density in general.

Ephexin5 is a Rho-GEF (guanine-nucleotide exchange factor) (Tan and Bird, 2016) that negatively regulates excitatory synapse development and is a substrate for UBE3A. When UBE3A is absent, Ephexin5 becomes elevated, resulting in dysregulation of excitatory synapse development (Tan and Bird, 2015) (Fig. 7). Dendritic spine malformation can be linked to LTP defects in AS neurons, as the consolidation and stabilisation of LTP relies on actin polymerisation in dendritic spines in response to LTP induction stimuli (Tan and Bird, 2016). The polymerisation of actin is regulated at least in part by mTOR signalling (Sun *et al.*, 2017), which is another downstream target of UBE3A. mTOR signalling involves two complexes, mTORC1 and mTORC2 (mammalian target of rapamycin complex 1 and 2), both of which contain mTOR as the main catalytic unit, acting as a serine/threonine kinase in mTORC1 and a tyrosine kinase in mTORC2 (Sun *et al.*, 2017). Loss of UBE3A leads to increased mTORC1 activation and decreased mTORC2 (Sun *et al.*, 2017; Tan and Bird, 2016) (Fig. 7). mTORC2 is involved in regulation of the actin cytoskeleton, while mTORC1 relates more to cell growth, proliferation, survival, and protein synthesis. mTORC1 is most likely the substrate of UBE3A, leading to upregulation in the absence of UBE3A, and repression of mTORC2 (Sun *et al.,* 2017) (Fig. 7).

UBE3A also affects the neurotransmitter receptors found on synapses. GAT1 is targeted for degradation upon ubiquitination by UBE3A and is responsible for

uptake of GABA (γ-Amino Butyric Acid) from the extra-synaptic space (Tan and Bird, 2016). GABA is the main inhibitory neurotransmitter in the central nervous system, essential for regulation of muscle tone and neuronal excitability. Ineffective reuptake reduces the tonic inhibition, which interferes with proper regulation of signalling (Tan and Bird, 2016). AS has also been linked to dysfunctional NRG1-ErbB4 signalling (Tan and Bird, 2016) (Fig. 7), which is responsible for downregulation or internalisation of AMPA and NMDA-type GABA receptors (Kwon *et al*., 2005; Gu *et al.,* 2005). NRG1 is elevated in AS mice, but other evidence suggests that it is not a substrate for UBE3A, so there may be an intermediate step upregulating levels of NRG1 as a result of UBE3A dysfunction (Tan and Bird, 2016).

Seizures and sleep disturbances occur in most but not all AS patients, and the mechanisms behind both are not well understood. Recent studies, however, suggest that seizures could be related to the effect of UBE3A loss on GABAergic neurons causing imbalance in local excitatory and inhibitory circuits (Tan and Bird, 2016). A potential link between UBE3A and sleep is BMAL1, a transcription factor critical in regulation of the circadian rhythm (Gossan *et al*., 2014). BMAL1 dimerises with CLOCK, another transcription factor, to activate transcription of PERIOD and CRYPTOCHROME proteins, which then suppress BMAL1 and CLOCK in a periodic manner (Fig. 7). BMAL1 is targeted for degradation by UBE3A to ensure tight regulation of the circadian system, so loss of this regulatory mechanism may be involved in the abnormal night-time behaviours of AS children (Gossan *et al.,* 2014; Shi *et al.,* 2015).

There are many more proteins that display reduced levels in AS models, but have not yet been identified explicitly as UBE3A substrates. These may represent potential substrates or downstream targets of other as yet unidentified substrates of UBE3A (Sell and Margolis, 2015). One such potential substrates is HERC2, which has been identified as a binding partner of UBE3A (Martínez-Noël *et al*., 2018), and implicated in a neurodevelopmental disorder very similar to AS (Harlalka *et al.,* 2012). However, it is likely that the interaction between UBE3A and HERC2 is more complicated than an enzyme-substrate interaction, since it has been suggested that HERC2 acts as an activator of UBE3A (Kühnle *et al.,* 2011).

Another set of proteins involved in AS are TSC1 and TSC2, a pair of enzymes that form a complex implicated in tuberous sclerosis (TS) and autism spectrum disorders (ASD)(Sell and Margolis *et al.,* 2015).TSC2 has been to shown to undergo UBE3A-dependent degradation in cell lines, but in AS TSC2 has been shown to be inhibited, preventing its inhibition of mTORC1 along with TSC1. Although TSC2 is a potential substrate of UBE3A, it is inhibited in the absence of functional UBE3A by increased methylation of an inhibitory site, which is likely caused by another unidentified substrate of UBE3A (Bi *et al.,* 2015) (Fig. 7).

Despite current understanding of the genetic mechanisms behind AS, treatments are focused on managing the symptoms rather than targeting the cause (Tan and Bird, 2016; Wheeler *et al.,* 2017). However, research into treatments is still ongoing, and methods of reinstating the missing UBE3A enzyme are being developed. This includes using gene therapy to reintroduce the protein, restoring expression from the paternal chromosome, and targeting of downstream substrates of UBE3A to restore the signalling mechanisms. Gene therapy involves introducing the *UBE3A* gene to cells where it is lacking, which allows the cells to translate it into functional protein. This has been approved for other genetic disorders, but would involve regular re-admission of the genetic material, with the effects of the treatment wearing off if the treatment is not repeated indefinitely (Stefflin *et al.,* 2019). Studies in AS mice suggest that gene therapy is not as effective as predicted, so activation of paternal *UBE3A* along with targeting downstream effectors is currently the most promising approach (Tsagkaris *et al.,* 2020). Another consideration for the treatment of AS is when the treatment should be administered. A study in mice that successfully reinstated expression of the paternal *UBE3A* transcript in neurons suggests that there are certain 'developmental windows' during which reinstatement of *UBE3A* had different effects. They show that while hippocampal synaptic plasticity could be restored by reinstatement at any age and motor deficits could be restored if treated by adolescence, other features including epilepsy, anxiety, and repetitive behaviours could only be rescued during early development (Silva-Santos *et al.,* 2015). Similarly, another study demonstrated that if UBE3A could be produced as normal during early development but then lost later, many of the phenotypic features of AS were avoided (Sonzogni *et al.,* 2019).

1.3.2 Dup15q Syndrome

While AS is caused by a loss of UBE3A, other neurodevelopmental disorders have been identified and characterised as associated with an increase in UBE3A levels (LaSalle *et al*., 2015). The most characterised of these is Dup15q syndrome, a disorder caused by duplication of the entire 15q11.2-q13.1 region, although the majority of the symptoms are attributed to loss of UBE3A-specific functions. Dup15q syndrome shares many characteristics with AS, including a developmental delay, intellectual disability, speech and language impairment, and epilepsy. Like with AS, almost all Dup15q patients experience hypotonia, a decrease in muscle mass, which can lead to difficulties with feeding, joint hyperextensibility, excessive drooling, and difficulty walking, although most patients are able to walk independently unlike many AS patients (LaSalle *et al.*, 2015). The key difference between AS and Dup15q syndrome is the probability of an autism spectrum disorder (ASD) diagnosis. While many AS patients fit the criteria for an ASD diagnosis, it is contentious as to whether this is a true reflection of the phenotype, or if the lack of verbal speech and degree of intellectual disability observed in AS patients renders current ASD diagnostic tests inaccurate (Trillingsgaard and

Østergaard, 2004). In contrast to this, Dup15q syndrome is strongly associated with an ASD phenotype, with most patients fulfilling the diagnostic criteria (LaSalle *et al.,* 2015) and Dup15q cases making up 1-3% of all identified ASD cases (DiStefano *et al.,* 2016). While the 15q11.2-q13.1 region encodes more than just the *UBE3A* gene, a case has been identified where duplication of the *UBE3A* gene alone was enough to cause developmental delay in the affected individual (Noor *et al.,* 2015). This along with the maternal inheritance of the disorder (Finucane *et al.,* 2016) implies that UBE3A is highly involved in the pathogenesis of the disease.

As with AS, there are multiple genetic causes of Dup15q syndrome, with the different genetic classes displaying different phenotype severity. An estimated 20% of Dup15q cases are causes by an interstitial duplication of the 15q11.2 region, where the region is duplicated within the maternal chromosome. The remaining 80% of Dup15q cases are caused by isodicentric duplications, where the patient has the standard, healthy, maternal and paternal chromosomes, but also an isodicentric chromosome, with each chromosome containing two centromeres, mirrored around the 15q11.2 region (Finucane *et al.,* 2016) (Fig. 9).

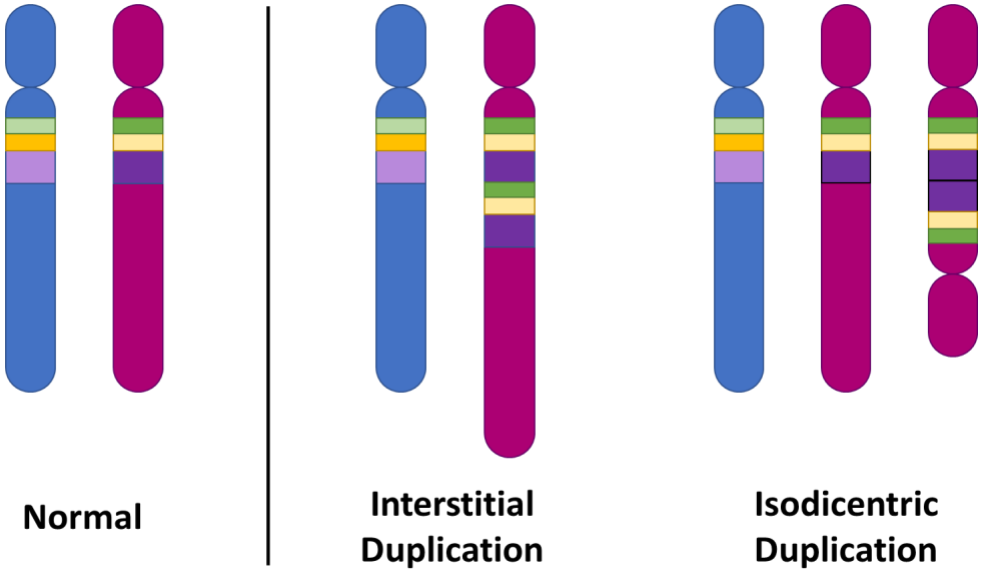


Figure 9: Genetic classes of Dup15q syndrome. In each diagram the blue shape represents the paternal chromosome 15 and the purple shape represents the maternal chromosome. The green square represents the AS-IC, the yellow square the SNURF/SNRPN region upstream of UBE3A, and the purple box represents the UBE3A gene. Interstitial duplication describes a duplication of the whole 15q11.2 region within the maternal chromosome, while isodicentric duplication describes the presence of an isodicentric copy of the maternal chromosome, duplicated just below the 15q11.2 region.

Isodicentric duplication of chromosome 15q in Dup15q syndrome is caused by the break points in the 15q11.2-13 region, as identified in AS deletion classes (Fig. 6). Most Isodicentric 15 chromosomes are caused by BP3:BP3 or BP4:BP5

recombination events (Wang *et al.,* 2008; Wang *et al.* 2004), which means that either two chromosomes will split at BP3 and recombine together, or one will split at BP4 and another at BP5 to recombine to form the Isodicentric chromosome. As with AS, the different classes of Dup15q are associated with different phenotype severities. Interstitial duplications show less severe symptoms compared to Isodicentric duplication cases (Finucane *et al.,* 2016; DiStefano *et al.,* 2016), presumably due to the extra copy of the core 15q11.2-13.1 region containing *UBE3A*.

1.3.3 Autism Spectrum Disorders

UBE3A has been identified as the most probable cause of ASD in Dup15q patients (Smith *et al.,* 2012), which fits with observations that Isodicentric duplication cases of Dup15q show a more pronounced ASD phenotype than inverted duplication cases, and Isodicentric duplication results in more copies of the *UBE3A* gene than inverted duplication (Smith *et al.,* 2012; DiStefano *et al.,* 2016). However, UBE3A has also been implicated in non-syndromic ASD, where the patient displays the ASD phenotype but without any of the other symptoms associated with other neurodevelopmental disorders. A whole genome sequencing study identified a missense mutation in the *UBE3A* gene that appeared to cause an ASD phenotype with a normal IQ, as opposed to the intellectual disability observed in AS and Dup15q patients (Iossifov *et al.,* 2014; Yi *et al.,* 2015). The mutation involved a threonine residue at position 485 (using human isoform 1 numbering) replaced by an alanine (T485A). The amino acid sequence upstream of this site was identified as a canonical protein kinase A (PKA) consensus motif, with T485 acting as the phospho-receptor, so mutation of the threonine to an alanine residue prevented phosphorylation of the site. UBE3A mutants containing T485A, which cannot be phosphorylated, or T485E, a phospho-mimetic of T485, were added to cells and their effects were studied. This led to the realisation that phosphorylation at T485 by PKA acts to inhibit UBE3A activity by targeting it for self-degradation (Yi *et al.,* 2015). The T485A mutant identified in an individual displaying ASD is therefore an over-active form of the UBE3A protein (Yi *et al.,* 2015; Vatsa and Jana, 2018).

1.3.4 Neuropsychiatric Disorders

As well as AS, Dup15q and ASD, UBE3A is also involved in schizophrenia (Bassett, 2011; Salminen *et al.,* 2019). Varied psychiatric phenotypes seem to segregate with increased *UBE3A* expression, with psychoses similar to aspects of schizophrenic spectrum disorders prevalent amongst a PWS population, but with a higher penetrance within those with a maternal UPD rather than deletion of the paternal 15q11.2-13 region (Bassett, 2011; Salminen *et al.,* 2019). The maternal inheritance of this predisposition to neuropsychiatric disorders suggests an involvement of *UBE3A* as it is the only paternally imprinted gene in the affected area, but a case of a family showing duplication of only the *UBE3A* gene and displaying a variety of neuropsychiatric disorders

along with learning disabilities further implicates *UBE3A* in this role (Noor *et al.,* 2015).

UBE3A is a multi-functional enzyme, involved in both transcriptional activation and ubiquitination processes with a wide range of cellular targets. Its role in the development of the central nervous system in early development and the activities of both GABAergic and glutaminergic neurons implicates UBE3A in a range of neurodevelopmental disorders. Dysregulation of UBE3A levels in the brain are involved in well-defined disorders such as AS and Dup15q, as well as more idiopathic conditions including ASD and schizophrenia. Interestingly, the effect of altered UBE3A activity may have different effects in different individuals, as a case of partial trisomy 15q11-q13, also referred to as Isodicentric duplication, typically a cause of Dup15q syndrome (Fig. 9), has been identified in both a mother and child who each display a different manifestation of the condition, and neither display the characteristic intellectual disability (Michelson *et al.,* 2011). The mother appeared asymptomatic until age 25, at which point she was diagnosed with schizophrenia, while the child presented with intractable seizures. This atypical presentation of a well characterised chromosomal abnormality suggests that there may be a more complicated degree of control regarding UBE3A signalling than is currently understood (Michelson *et al.,* 2011).

1.4 Cancer

Although *UBE3A* is implicated in various neurodevelopmental disorders due to its interesting epigenetic regulation in neurons, it is biallelically expressed in all tissue types (Sirois *et al.,* 2020) and has been identified in many key cellular processes that when disrupted can lead to cancer. UBE3A was first identified as a factor in the oncogenesis of cervical carcinomas resulting from human papillomavirus (HPV) infection (Beaudenon and Huibregtse, 2008; Scheffner et al., 1993). However, it has also been implicated in other forms of cancer, including oropharyngeal cancer associated with HPV (Berman and Schiller, 2017) and liver cancer associated with hepatitis C (Munakata *et al.,* 2005), as well as non-viral prostate cancer (Raghu *et al.,* 2017), breast cancer (Band *et al.,* 1991), B-cell lymphomas (Wolyniec *et al.,* 2012), and non-small-cell lung cancer (Bandilovska *et al.,* 2019; Gamell *et al.,* 2017) (Fig. 10).

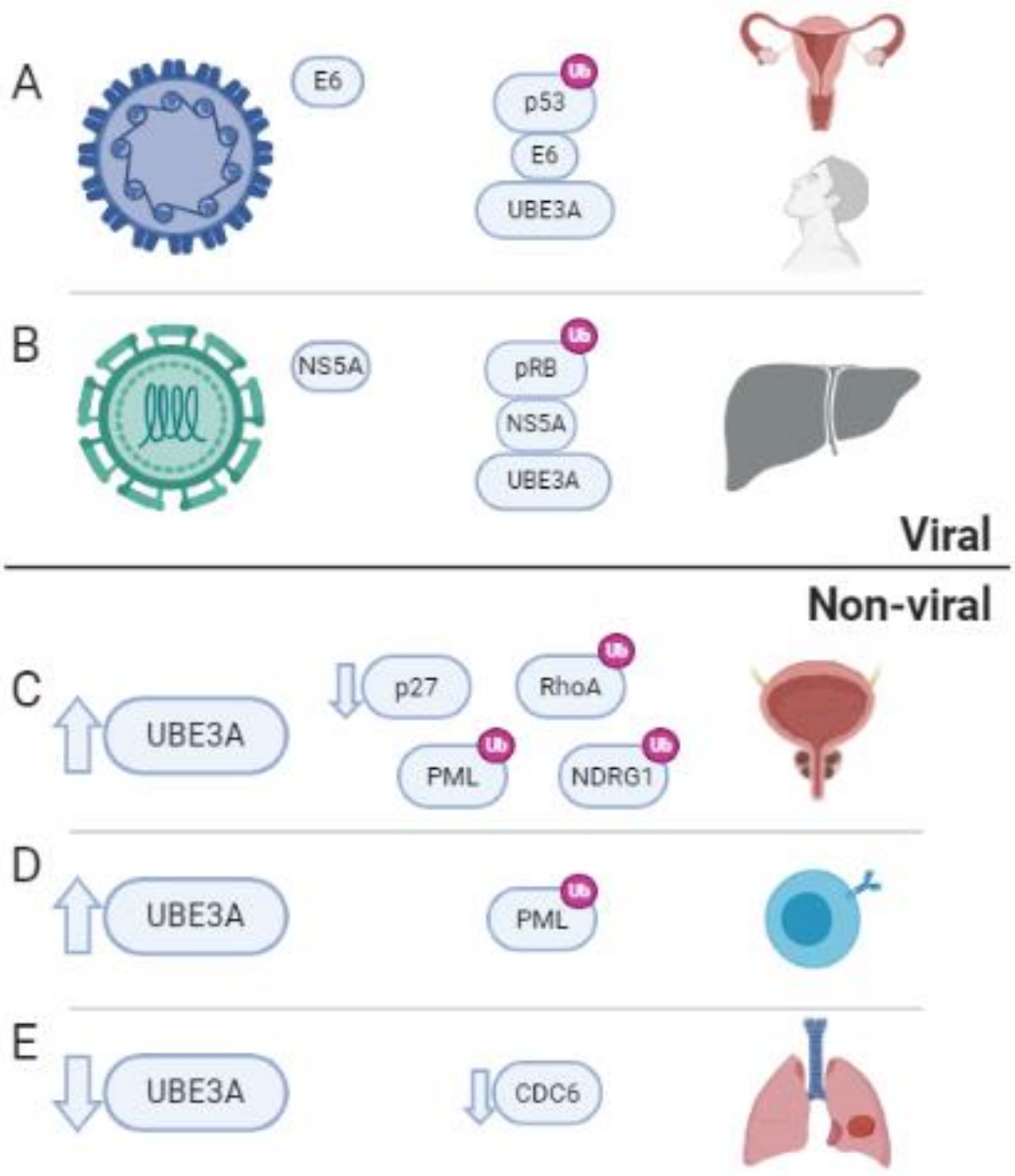


Figure 10: UBE3A is involved in several types of cancer, some are associated with a viral oncogene while others are triggered by changes in normal protein expression levels. A) HPV encodes the E6 protein, which targets p53 for degradation through ubiquitination by UBE3A, leading to cervical cancer or oropharyngeal cancer. B) The NS5A protein encoded by hepatitis C causes UBE3A to ubiquitinate pRB targeting it for degradation, leading to liver cancer. C) Increased UBE3A protein levels contribute to prostate cancer through targeting the RhoA, PML and NDRG1 proteins for degradation following ubiquitination, as well as downregulating p27 at the transcriptional level. D) Increased UBE3A leads to downregulation of PML in the oncogenesis of B-cell lymphomas, particularly Burkitt's lymphoma. E) Decreased levels of UBE3A are associated with a poor prognosis in non-small-cell lung cancer due to the downregulation of CDC6, a transcriptional coactivator that regulates several tumour suppressors.

There are hundreds of strains of HPV that can be differentiated into high-risk and low-risk strains based on their carcinogenicity (Bzhalava et al., 2015; Doorbar et al., 2015). Low-risk cutaneous strains only affect the skin, causing common warts, and are unlikely to lead to cancer, low-risk mucosal means that the virus is able to infect and affect surfaces inside the body, leading to genital warts, but are still unlikely to lead to cancer, while high-risk strains are able to infect surfaces inside the body and can lead to cancer. Of the hundreds of HPV strains only 12 are classed as high-risk (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, and 59) (Bouvard et al., 2009), and of these HPV16 and HPV18 account for most cervical cancer cases (Schiffman et al., 2011; Beaudenon and Huibregste, 2008).

1.4.1 Cervical Cancer

Cervical cancer caused by HPV makes up 99.7% of all cervical cancer cases (Bandilovska *et al.,* 2019; Yim and Park, 2005; Beaudenon and Huibregste, 2008), but only 8-10% of high-risk HPV infections lead to cervical cancer (Yim and Park, 2005). Co-evolution of papillomaviruses and their hosts means that HPV can persist in a host for long periods with no obvious signs of infection (Doorbar *et al.,* 2015). In most HPV cases the human immune system is able to clear the infection within a few years of initial infection (Schiffman *et al.,* 2011), but those that are not cleared survive by evading apoptosis and the immune response, increasing replication, and deregulating host cellular energetics, all of which are considered hallmarks of cancer (Hanahan and Weinberg, 2011; Mesri *et al.,* 2014).

Amongst the various proteins encoded by HPV to carry out its functions, the E6 and E7 proteins are responsible for HPV's oncogenicity. E7 binds and inhibits the endogenous pRB protein, a known tumour suppressor that inhibits E2F-family transcription factors, to increase transcription of genes that encourage cell-division and replication (Yim and Park, 2005). The E6 protein, however, binds to UBE3A to alter its substrate specificity, primarily targeting the p53 tumour suppressor for degradation (Yim and Park, 2005; Scheffner *et al.,* 1993; Huibregtse *et al.,* 1991).

E6 interacts with an LxxLL motif in UBE3A, a peptide-recognition motif involved in protein-protein interactions, primarily found in proteins that interact with nuclear hormone receptors, or transcription factors and coactivators (Plevin *et al.,* 2005). An isolated peptide comprised of 12 amino acids from UBE3A including the LxxLL motif (ELTLQELLGEER) is sufficient to stabilise the E6 protein and allow the formation of a ternary complex with p53 *in vitro* (Chen *et al.,* 1998; Martinez-Zapien *et al.,* 2016), but the full ligase activity required for degradation of p53 requires regions spanning from UBE3A N-terminal domain to the central LxxLL motif, as well as the C-terminal catalytic HECT domain (Drews *et al.,* 2020). Interestingly, the association of p53 to a pre-formed E6/UBE3A complex triggers ubiquitination of E6, which is required for targeting of p53 for degradation but also targets the E6 protein

for degradation. This could suggest that the conformational change induced in UBE3A upon E6 binding is sufficient to induce binding of p53, but a further change is required for p53 ubiquitination (Li *et al.,* 2019).

Although the primary mechanism of oncogenesis associated with the E6 protein is the degradation of p53, the E6 protein also interacts with a range of cellular proteins to enact the hallmarks of cancer (Tungteakkhun and Duerksen-Hughes, 2008). One of these other targets is UBE3A itself, as UBE3A has been shown to undergo intramolecular auto-ubiquitination upon E6 binding (Kao *et al.,* 2000) and substrates of UBE3A have been identified within cell growth and proliferation mechanisms (Fig. 7), as well as cell senescence and apoptosis in response to oxidative stress (Wolyniec *et al.,* 2012). siRNA mediated knock down of E6 expression had an almost identical effect to knock down of UBE3A in HPV-positive cells, so it is likely that almost all of E6-associated oncogenic effects are mediated through the E6/UBE3A complex (Kelley *et al.,* 2005).

Some of E6's p53-independent targets are the pro-apoptotic proteins Bak, TNFR1 (Tumour Necrosis Factor Receptor 1), FADD (Fas-Associated Death Domain), and procaspase 8 (Thomas and Banks, 1998; Filippova *et al.,* 2002; Filippova *et al.,* 2004; Filippova *et al.,* 2007), which help HPV to evade apoptosis. E6 has also been shown to target the transcriptional activator CBP/p300 and its interactor Gps2 to downregulate p53 at the transcriptional level (Zimmermann *et al.,* 1999; Degenhardt and Silverstein, 2001). E6/UBE3A has been shown to target the NFX1-91 protein for degradation, which would otherwise repress transcription of hTERT, the catalytic subunit of telomerase (Gewin *et al.,* 2004). The E6/UBE3A complex also interacts with MCM7 (Multicopy Maintenance protein 7), a subunit of the RLF (DNA Replication Licensing Factor) complex, to promote dysregulation of DNA replication and cell proliferation (Kühne and Banks, 1998; Beaudenon and Huibregtse, 2008).

As well as initiating cancer through dysregulation of cell cycle control and apoptosis, the E6/UBE3A complex is also involved in the ability of hrHPV to escape the innate immune response in the early stages of HPV infection. E6/UBE3A targets pro-IL-1β, the inactive precursor to interleukin-1β, disrupting IFN-κ signalling (Niebler et al., 2013).

HPV-Associated Cervical Cancer	
Viral Cofactor	**Effect on UBE3A**
E6	Changes the substrate profile
UBE3A Substrate	**Target Protein Function**
p53	Tumour suppressor, “Guardian of the genome”
UBE3A	Involved in regulating cell growth and proliferation signalling mechanisms
Bak	A tumour suppressor
TNFR1	A tumour suppressor
FADD	A tumour suppressor
Procaspase-8	A tumour suppressor
CBP/p300	A co-activator or p53 transcription
Gps2	An interactor of CBP/p300 in co-activating p53 transcription
NFX1-91	Represses transcription of hTERT, the catalytic subunit of telomerase
MCM7	A subunit of the RLF (DNA replication licensing factor) that ensures the cell genome is replicated only once per cell cycle
Pro-interleukin-1β	The precursor to IL-1β, a key cytokine in IFN-κ signalling in the innate immune response

Table 1: A summary of the key interaction partners and targets of UBE3A in HPV-associated cervical cancer.

1.4.2 Oropharyngeal Cancer

Although HPV is well-known in cervical cancer, hrHPV infection has also been associated with various head and neck squamous cell cancers (HNSCCs), particularly oropharyngeal cancer, affecting the back of the throat, base of the tongue, and tonsils. (Berman and Schiller, 2017). While almost all cervical cancer is caused by HPV infection, oropharyngeal cancers are associated with environmental factors, including smoking and alcohol consumption, as well as viral causes. In relatively recent years various screening programs and vaccinations against hrHPV strains have led to a decrease in cervical cancer cases worldwide (Vaccarella *et al.,* 2013), while HPV-associated oropharyngeal cancers have seen an increase (Chaturvedi *et al.,* 2011). HPV-associated oropharyngeal cancers are caused by the same mechanism of HPV-associated cervical cancer, the degradation of p53 and inhibition of pRB by E6 and E7. However, the HPV strains most associated with oropharyngeal differ to those most associated with cervical cancer, with over 90% of cases associated with HPV16 alone, and the next most prevalent strain being HPV35 at 4% of cases. The HPV strains associated with oropharyngeal cancers also show a greater propensity for mutations in the E6 protein, particularly on the surface of the protein, suggesting that there are subtle difference in the aetiology of the two conditions (LeConte *et al.,* 2018; Berman and Schiller, 2017).

HPV-Associated Oropharyngeal Cancer	
Viral Cofactor	**Effect on UBE3A**
E6	Changes the substrate profile
UBE3A Substrate	**Target Protein Function**
p53	Tumour suppressor, “Guardian of the genome”

Table 2: A summary of the key interaction partners and targets of UBE3A associated with HPV-associated oropharyngeal cancer

1.4.3 Hepatocellular Carcinoma

As well as HPV, UBE3A is involved in the viral carcinogenesis of hepatocellular carcinomas (HCC) following hepatitis C (HCV) infection. Hepatitis C is a blood-bourne virus that often causes cirrhosis of the liver, and after over a decade of sustained HCV infection, it can also lead to HCC. Although HCV is responsible for 85-90% of all HCC cases, only 2-6% of all HCV infections lead to cancer (El-Serag and Rudolph, 2007; Andrade *et al.,* 2009). UBE3A is involved in hepatocellular carcinogenesis in a similar way to its involvement in cervical cancer, it forms part of a complex comprised of UBE3A, a viral protein, and an endogenous tumour suppressor protein (Munakata *et al.,* 2007). Several HCV proteins have been suggested in the development of HCC, but it isNS5A that exerts its main effect, down-regulation of the tumour suppressor pRB, through UBE3A (Munakata *et al.,* 2007). Whereas the ternary complex in HPV cancers is formed through an interaction between UBE3A and the viral E6, which then allows UBE3A to interact with p53, In HCV-mediated HCC the viral NS5B protein binds to the tumour suppressor pRB first and sequesters it to the nucleus, where it then recruits UBE3A to target it for degradation by the proteasome (Munakata *et al.,* 2007). However, despite the integral role of UBE3A in hepatocarcinogenesis, UBE3A is also involved in the body’s natural defence against HCC. UBE3A has been shown to ubiquitinate the core protein of HCV to target it for degradation, limiting the ability of the HCV virus to reproduce within cells (Shirakura *et al.,* 2006). In response to the downregulation of HCV by UBE3A, the HCV core protein is able to downregulate UBE3A levels at the transcriptional level, by causing hypermethylation of the promoter region for the *UBE3A* gene in human hepatocytes (Kwak *et al.,* 2016). This suggests that the role of UBE3A in HCV infection and HCC progression is complicated, and regulation of UBE3A may contain a temporal element (Bandilovska *et al.,* 2019).

HCV-Associated Hepatocellular Carcinoma	
Viral Cofactor	**Effect on UBE3A**
NS5A	Changes the substrate profile
UBE3A Substrate	**Target Protein Function**
pRB	Tumour Suppressor
HCV core protein	A component of the HCV nucleocapsid

Table 3: A summary of the key interaction partners and targets of UBE3A in HCV-associated hepatocellular carcinomas.

1.4.4 Prostate Cancer

While UBE3A has been shown to be involved in viral cancers through alteration of its substrate specificity, it has also been implicated in non-viral cancers, particularly prostate cancer, where its oncogenic effects seem to be caused by an upregulation of its normal activity profile (Srinivasan and Nawaz, 2011). UBE3A appears to be crucial in the normal development of the prostate, where downregulation of expression led to an underdeveloped prostate gland in mice (Khan *et al.,* 2006), whereas overexpression led to preneoplastic lesions, a precursor to prostate cancer (Srinivasan and Nawaz, 2011). This regulation of prostate development is enacted through the PI3K-Akt pathway, where UBE3A acts as both a ubiquitin ligase to target RhoA, a negative regulator of Akt activity, for degradation, and as a coactivator of the androgen receptor, possibly causing upregulation of both PI3K and Akt levels at the transcriptional level (Srinivasan and Nawaz, 2011; Khan *et al.,* 2006).

As well as the increase in cell proliferation and growth through upregulated PI3K-Akt signalling, UBE3A targets the tumour suppressors p27 and PML (promyelocytic leukaemia protein) for degradation in the aetiology of prostate cancer. (Raghu *et al.,* 2017; Louria-Hayon *et al.,* 2009). While PML has been identified as an endogenous substrate of UBE3A in a non-cancerous environment (Louria-Hayon *et al.,* 2009), increased levels of UBE3A and subsequent decreased levels of PML in the cells of prostate cancer patients is an indicator of poor prognosis and is associated with a higher likelihood of cancer recurrence (Birch *et al.,* 2014). p27 was identified as a target of UBE3A from a subset of prostate cancer patients that showed overexpression of UBE3A without the associated decrease in PML levels (Raghu *et al.,* 2017). However, rather than downregulate p27 through its E3 ligase activity, p27 appears to be regulated at the transcriptional level, and this regulation is dependent on E2F1, a known transcriptional activator of p27. UBE3A appears to interact with E2F1 to prevent it binding to the p27 promoter, but it does not target it for degradation (Raghu *et al.,* 2017). UBE3A has also been implicated in the progression to a metastatic phenotype through the ubiquitin-dependent proteasomal degradation of NDRG1 (N-myc downstream regulated gene 1), a known metastasis suppressor implicated in several human cancers (Gamell *et al.,* 2019). Several studies have shown that decreasing UBE3A levels in cells leads to increased levels of PML, p27, and NDRG1 and an associated restoration of the innate tumour suppression strategies, which highlights the potential of targeting UBE3A specifically as a treatment for prostate cancer (Paul *et al.,* 2016; Raghu *et al.,* 2017; Gamell *et al.,* 2019).

Prostate Cancer	
Viral Cofactor	**Effect on UBE3A**
-	Upregulation of its normal activity profile
UBE3A Substrate	**Target Protein Function**
RhoA	A negative regulator of Akt activity, it decreases cell proliferation and growth regulated by the PI3K-Akt signalling cascade.
E2F1	A transcriptional activator of p27
PML	A tumour suppressor
NDRG1	A metastasis suppressor

Table 4: A summary of the interaction partners and targets of UBE3A in the development and progression of prostate cancer.

1.4.5 B-cell Lymphoma

The first non-viral cancer that was shown to directly involve UBE3A was B-cell lymphoma (Bandilovska *et al.,* 2019; Wolyniec *et al.,* 2012). In a mouse model for Burkitt's lymphoma, a form of B-cell lymphoma where 60% of cases are associated with an increase in UBE3A levels, UBE3A was shown to prevent cell senescence in the presence of excess c-Myc, an oncogenic stress signal. It does this through ubiquitin-mediated degradation of the PML tumour suppressor (Wolyniec *et al.,* 2012).

B-Cell Lymphoma	
Viral Cofactor	**Effect on UBE3A**
-	Increased protein levels in the presence of excess c-Myc
UBE3A Substrate	**Target Protein Function**
PML	A tumour suppressor

Table 5: A summary of the key interaction partners and targets of UBE3A associated with B-cell lymphoma

1.4.6 Non-Small Cell Lung Cancer

Interestingly, another study has shown that loss of UBE3A, rather than overexpression, is associated with pathogenesis of non-small cell lung cancer (NSCLC) (Gamell *et al.,* 2017). UBE3A interacts with E2F1 to prevent the transcriptional activation of its target genes, similar to how UBE3A downregulates p27 in prostate cancer, but in NSCLC it downregulates CDC6 (cell division control protein 6), a transcriptional suppressor of the *INK4/ARF* locus. The *INK4/ARF* locus encodes several key tumour suppressors, including p15/INK4b, p16/INK4a, and p14/ARF, which regulate pRB and p53 signalling to prevent cancer development. This tumour suppressor role for UBE3A appeared to be subtype specific, as UBE3A levels show a higher correlation to survival rates in NSCLC adenocarcinomas than in squamous cell carcinoma patients (Gamell *et al.,* 2017).

Non-Small Cell Lung Cancer (NSCLC)	
Viral Cofactor	**Effect on UBE3A**
-	Loss of UBE3A is associated with NSCLC pathogenesis
UBE3A Substrate	**Target Protein Function**
E2F1	A transcriptional activator of CDC6 (cell division control protein 6), which suppresses transcription of the INK4/ARF locus of tumour suppressors.

Table 6: A summary of the key interaction partners and targets of UBE3A associated with NSCLC progression.

1.5 Structures and Key Residues

The three dimensional structure of a protein is often crucial to determining how proteins carry out their cellular functions. Identifying the residues at key interfaces between proteins and their substrates allows the chemical reactions and processes to be mapped out in atomic detail. Not only does this add to understanding of physiological processes at a molecular level, it can also prove invaluable in the design and development of novel therapeutics. Although UBE3A has been identified as a key part of several signalling processes and is critical in a variety of clinical contexts, no full-length structure has yet been produced for the enzyme. The first piece of structural information for UBE3A was a crystal structure of the catalytic HECT domain in 1999 (Huang *et al.,* 1999; 1C4Z), but in the intervening years the N-terminal region of the protein has remained difficult to crystallise. In 2011 Lemak *et al.* provided an NMR structure of a zinc finger domain at the N-terminus of UBE3A (2KR1), termed the AZUL domain for Amino-terminal Zinc-finger of UBE3A Ligase. As recently as 2020 Buel *et al.,* showed this domain in complex with a cognate domain within PSMD4 (6U19), which solidified its suggested role as a protein interaction interface. The only other structural data for the entire region of UBE3A between these two domains was provided by Martinez-Zapien *et al.* in 2016, where they show a 12-residue α-helix of UBE3A in a ternary complex with the HPV16 E6 protein and the core region of human p53 solved through x-ray crystallography (4XR8).

Although there is no published structure for large portions of the enzyme, several studies have been able to identify other key regions involved in protein interactions and regulation of UBE3A's ubiquitin ligase activity through the use of binding kinetics, *in silico* modelling, and truncation mutants (Ronchi *et al.,* 2017; Kühnle *et al.,* 2011; Drews *et al.,* 2020). While the identification of these regions allows a better understanding of how some identified clinical mutations contribute to the pathogenic state, they introduce more questions as to how all of these regions come together in the 3D structure of the enzyme. Many of the regions identified in co-ordinating various interactions sit distal to each other in the protein sequence, further

supporting the need for a full-length structure of UBE3A in order to fully understand the mechanism of activity (Fig. 11).

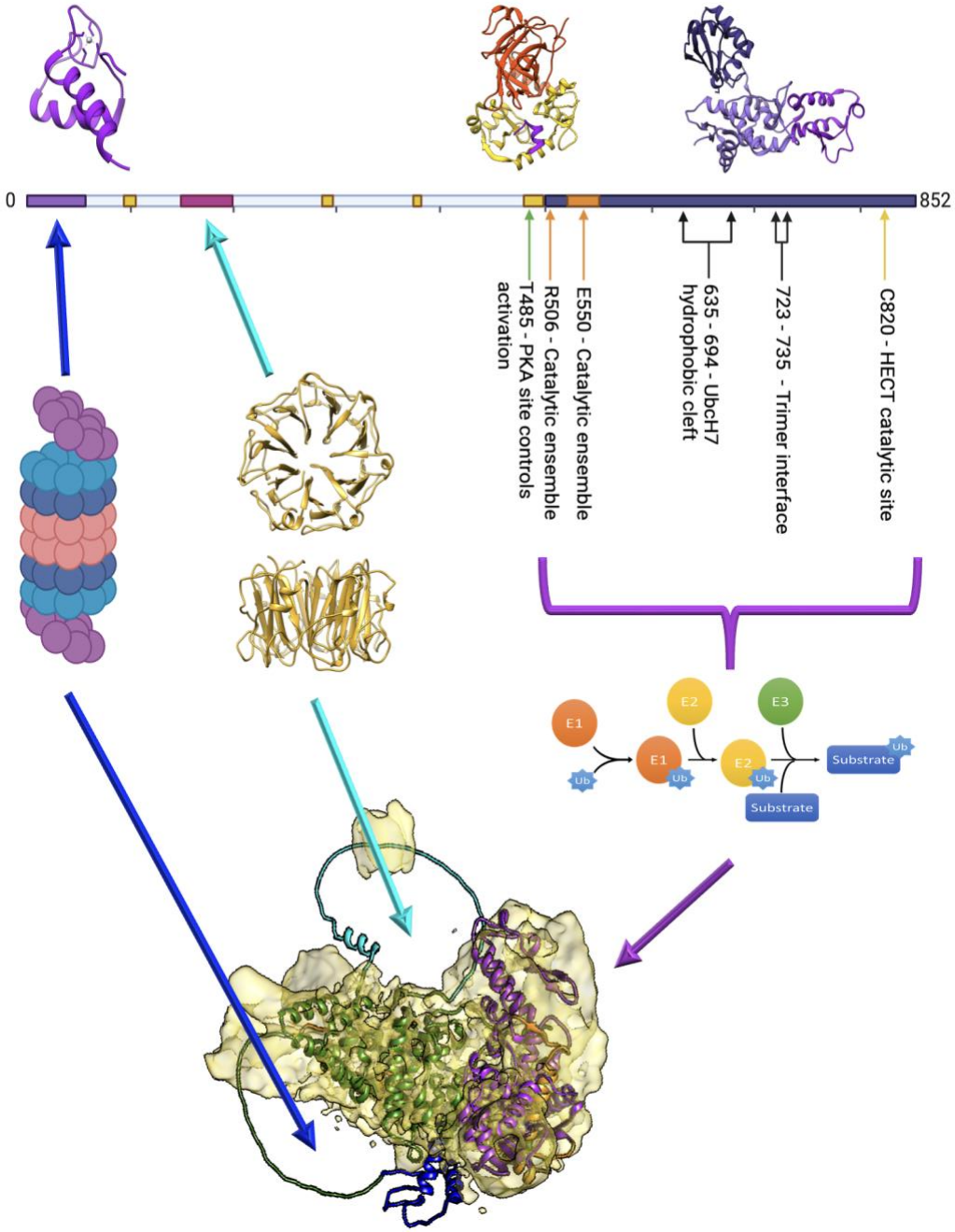


Figure 11: An overview of the different identified regions within UBE3A, their locations within the amino acid sequence, cellular roles, and locations within the 3D protein. Previously solved structures for regions within UBE3A can be seen at the top of the diagram, with the NMR structure of the AZUL domain (2KR1) on the left, the crystal structure of the LxxLL motif in complex with E6 and p53 in the middle, and the crystal structure of the HECT domain (1C4Z) on the right. Below the structures is a schematic showing the locations of these domains in the protein sequence, with the AZUL domain in light purple, the HERC2-binding region shown in pink, the E6 binding regions shown in yellow and orange (yellow for hrHPV E6 binding areas and orange for lrHPV E6 binding), and the HERC domain in dark purple. Arrows and labels show the key residues and areas involved in the catalytic activity of UIBE3A and the possible trimer interface. Below the protein sequence schematic are demonstrations of the various functions of each domain. The AZUL domain is known to coordinate the interaction between UBE3A and the proteasome via PSMD4, The HERC2-binding region coordinates the interaction with the RLD2 domain of HERC2 with uncharacterised cellular functions, and the HECT domain is responsible for catalysing the E3 ubiquitin ligase activity of UBE3A, where a ubiquitin moiety or chain is added to a cellular target in the final stage of the E1-E2-E3 enzyme cascade (see section 1.1). The locations of these regions within the 3D structure of UBE3A are shown in the bottom portion of the diagram, using the model generated in this project and discussed in chapter 7.1.3.

The first crystal structure for UBE3A showed the HECT domain of the enzyme in complex with its cognate E2, UbcH7 (Huang *et al*., 1999; PDB 1C4Z). The HECT domain spans residues 495 to 852 in UBE3A isoform 1, with the cysteine at position 820 (C820) acting as the catalytic site. E2 enzymes are a more homogenous group than E3s, and all are based around a 150-residue catalytic core unit. UbcH7 is a 154-residue protein that differs very little from the conserved form. UbcH7 forms an elongated α/β structure, with four α-helices and four β-strands. The UBE3A HECT domain is split into a smaller C-lobe and a larger N-lobe, which is then further split into large and small N-terminal domains. The UBE3A-UbcH7 complex determined crystallographically displays a U-shaped structure, with the UBE3A HECT domain forming an L-shape and the UbcH7 completing the other side (Fig. 12) (Huang *et al*., 1999).

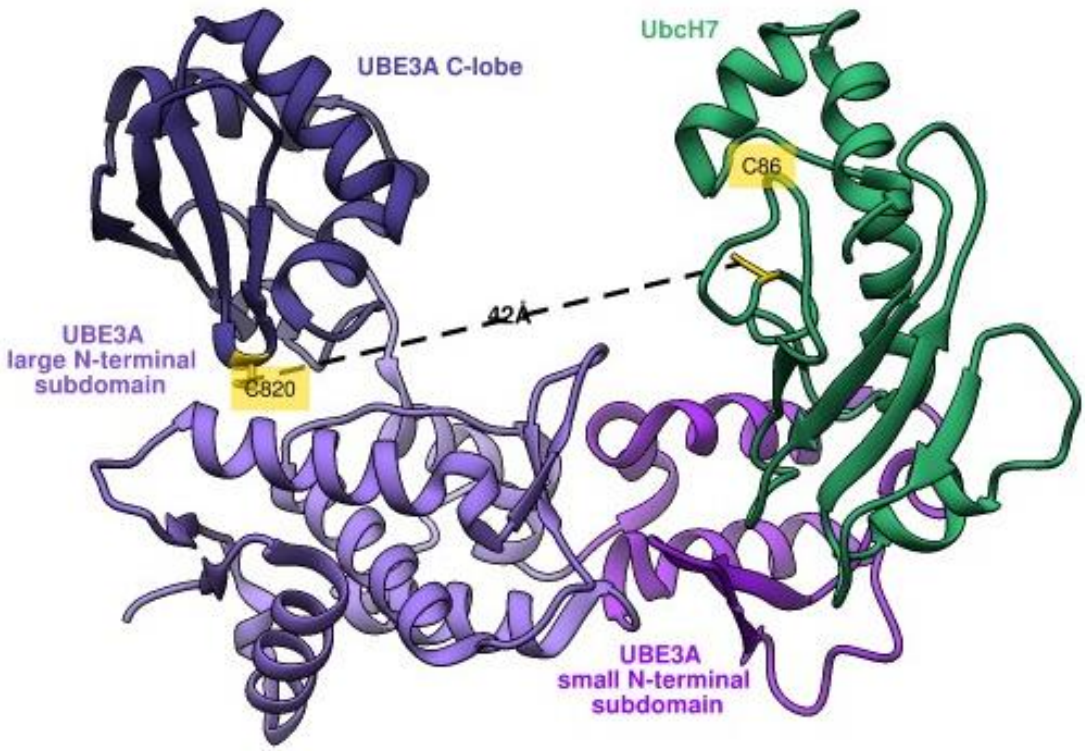


Figure 12: The crystal structure of the UBE3A HECT domain in complex with the UbcH7 E2 enzyme (1C4Z). UbcH7 is shown in green, UBE3A is shown in shades of purple where dark purple shows the C-lobe, the lavender section is the large N-terminal domain, and the bright purple section is the small N-terminal domain. The catalytic cysteine residue of each enzyme is highlighted in yellow and labelled according to the isoform 1 residue numbering. The distance between the catalytic residues is indicated with a dashed line and labelled in angstroms to show the separation of the active sites in the crystal form.

Within the small N-terminal subdomain of UBE3A is a hydrophobic groove, with two β-sheets on one side and two α-helices on another forming a V-shape. The specific residues within this region are not tightly conserved, although the hydrophobic characteristic is maintained in all HECT ligases (Huang *et al.,* 1999). This hydrophobic cleft appears to be key in maintaining the specificity of the E2-E3 interaction, as a conserved phenylalanine at position 63 (F63) is found in all E2 enzymes that function specifically with HECT ligases, but is absent in E2s specific for other E3 subtypes. The F63 residue of UbcH7 in the crystal structure sits in the deepest portion of the hydrophobic groove on UBE3A, forming van der Waals contacts with at least six hydrophobic and aromatic side chains in the hydrophobic interface (Fig. 13). Deletion of this residue has been shown to prevent complex formation, demonstrating the importance of interactions within the hydrophobic groove for the association of the E2 to UBE3A (Huang et al., 1999).

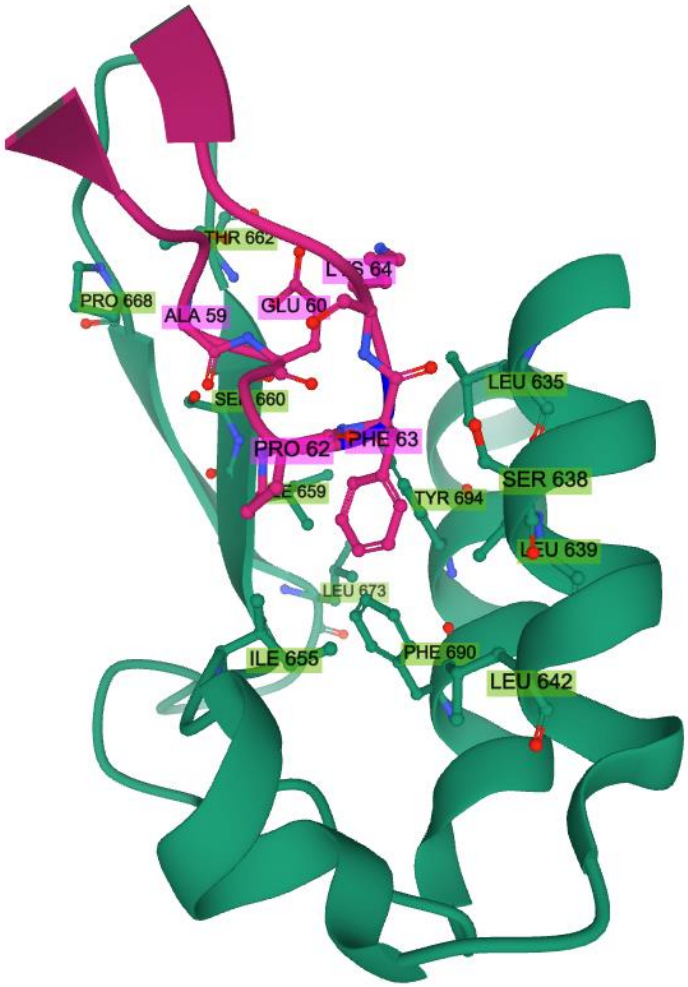


Figure 13: The hydrophobic groove of the UBE3A small N-terminal subdomain, with the F63 residue of UbcH7 at the deepest point. UBE3A residues are shown in a green colour, while UbcH7 residues are shown in pink, key residues are displayed in a ball-and-stick format and labelled.

A key feature of this structure is the distance between the catalytic cysteine residues of each enzyme of over 41 Å (Fig. 12). Transfer of ubiquitin from the E2 to the E3 occurs through nucleophilic attack on the E2~Ub thioester bond by the active site cysteine residue of UBE3A, but for this to occur the separation of the active cysteine residues of the two enzymes must be significantly reduced. The structures of other HECT domain proteins suggest that there is a flexible hinge region between the N- and C-lobes that could bring the two sites closer, such as in HUWE1 (3G1N), NEDD4-2 (3JVZ), and SMURF2 (1ZVD), but this still may not be close enough for a spontaneous reaction to occur (Ronchi *et al.,* 2013). Kinetic studies and subsequent *in silico* modelling identify a potential second site on the UBE3A HECT domain that can interact with the Ub-loaded UbcH7 molecule, which might solve the problem of the distance between the sites (Ronchi *et al.,* 2013; Ronchi *et al.,* 2017). This second binding site sits on the surface of the large N-terminal subdomain

of UBE3A, allowing UbcH7 to make contacts with key residues in both the large N-terminal subdomain and the C-lobe, and bringing the catalytic sites close enough for spontaneous nucleophilic attack (Ronchi *et al.,* 2017) (Fig. 14).

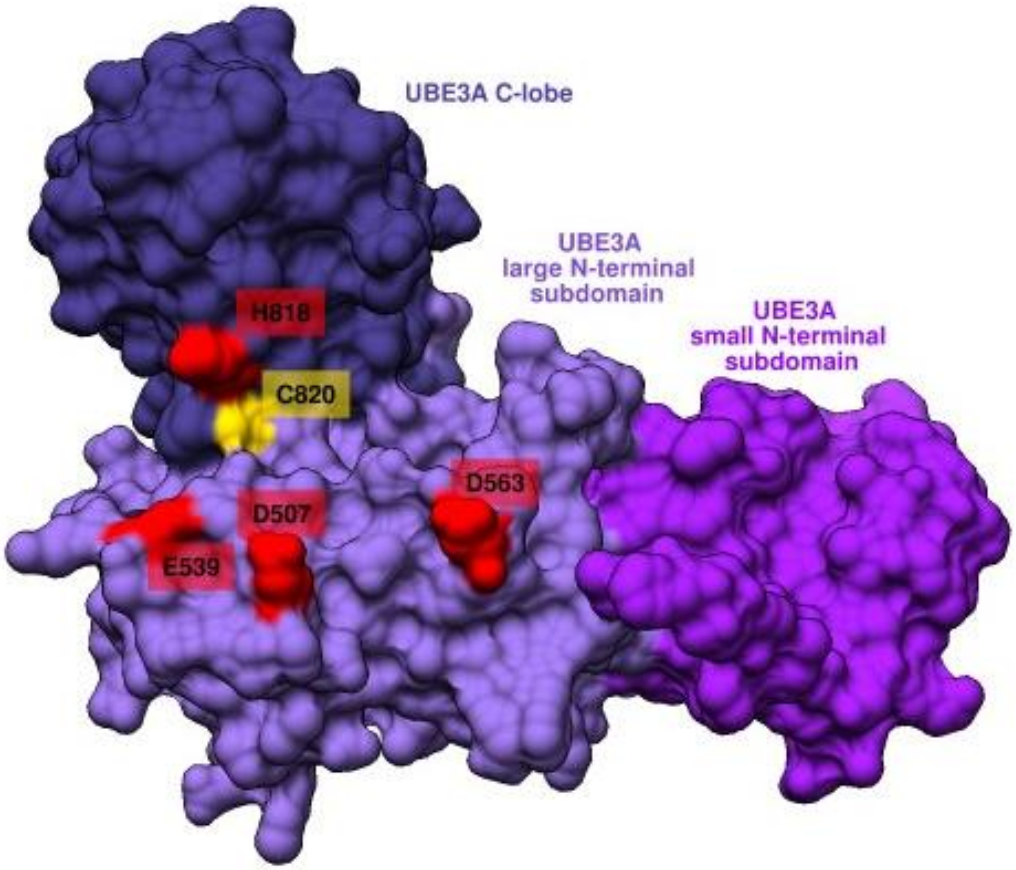


Figure 14: A surface representation of the UBE3A HECT domain with the key residues involved in the putative second E2~Ub binding site highlighted. The residues involved in the interaction surround the catalytic residue of the UBE3A HECT domain and would bring the catalytic residue of the E2 enzyme within atomic distance. The UBE3A residues involved in the interaction are coloured red, and the catalytic C820 residue is coloured yellow.

It has also been suggested that UBE3A acts as a trimer (Ronchi *et al.,* 2014). This is supported by some biochemical and biophysical evidence, including activity assays using different fractions from a size exclusion chromatography experiment (Ronchi *et al.,* 2014), but the most convincing argument comes from the observation of a trimer in the original 1999 crystal structure. This was originally thought to be an artefact caused by crystal packing forces, but a more recent analysis of the subunit interfaces suggests that it may represent an energetically favourable stable form of the enzyme (Ronchi *et al.,* 2014). A proposed mechanism for UBE3A that combines these observations is a two-step proximal indexation model (Fig. 15), which involves UBE3A units in an

oligomer working *in trans* to build a ubiquitin chain by joining new ubiquitin units to the G76 C-terminal residue of the previous ubiquitin in the chain as it is still attached to the C820 catalytic site. In this model, the ubiquitin chain is formed entirely through this proximal addition mechanism on the UBE3A active site, and is then moved *en bloc* onto the substrate protein once the chain is complete (Ronchi *et al.,* 2017).

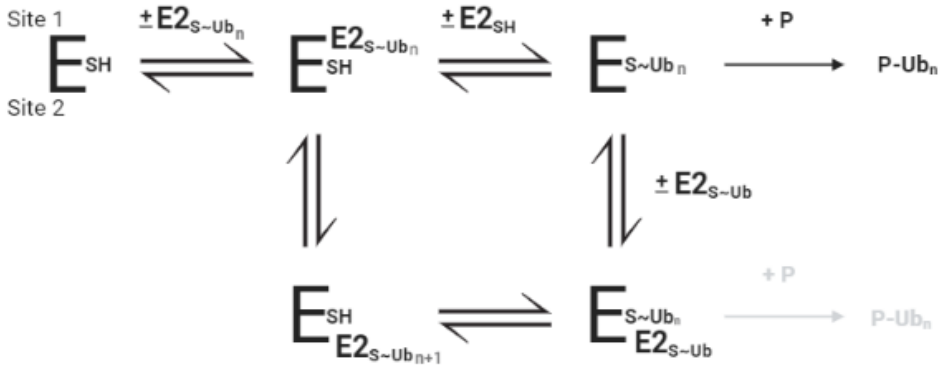


Figure 15: The proximal indexation model for polyubiquitin chain assembly catalysed by UBE3A. A ubiquitin-bound UbcH7 enzyme binds to site 1 of the UBE3A HECT domain, where the ubiquitin is transferred to the active site and the E2 enzyme leaves. A second ubiquitin-bound UbcH7 enzyme then binds to site 2 of another HECT domain, where the ubiquitin bound to the active site is transferred to the ubiquitin bound to the new E2 enzyme. The E2 enzyme then shifts to site 1 of UBE3A, where the chain is transferred to the active site and the E2 enzyme leaves. The ubiquitin chain can be transferred onto the substrate at any point where there is ubiquitin bound to the active site, but this mostly occurs when there is no E2 bound also (figure adapted from Ronchi et al., 2017).

In this model site 1 refers to the *in silico* predicted site close to the active site of UBE3A, while site 2 refers to the original UbcH7 binding site from the 1999 crystal structure (Ronchi *et al.,* 2017). For this model to work, with the 41Å gap between site 1 and site 2, UBE3A must function *in trans* as a multimer, so the E2 and ubiquitin are passed from the site 1 of one HECT domain onto the site 2 of a second HECT domain (Fig. 16).

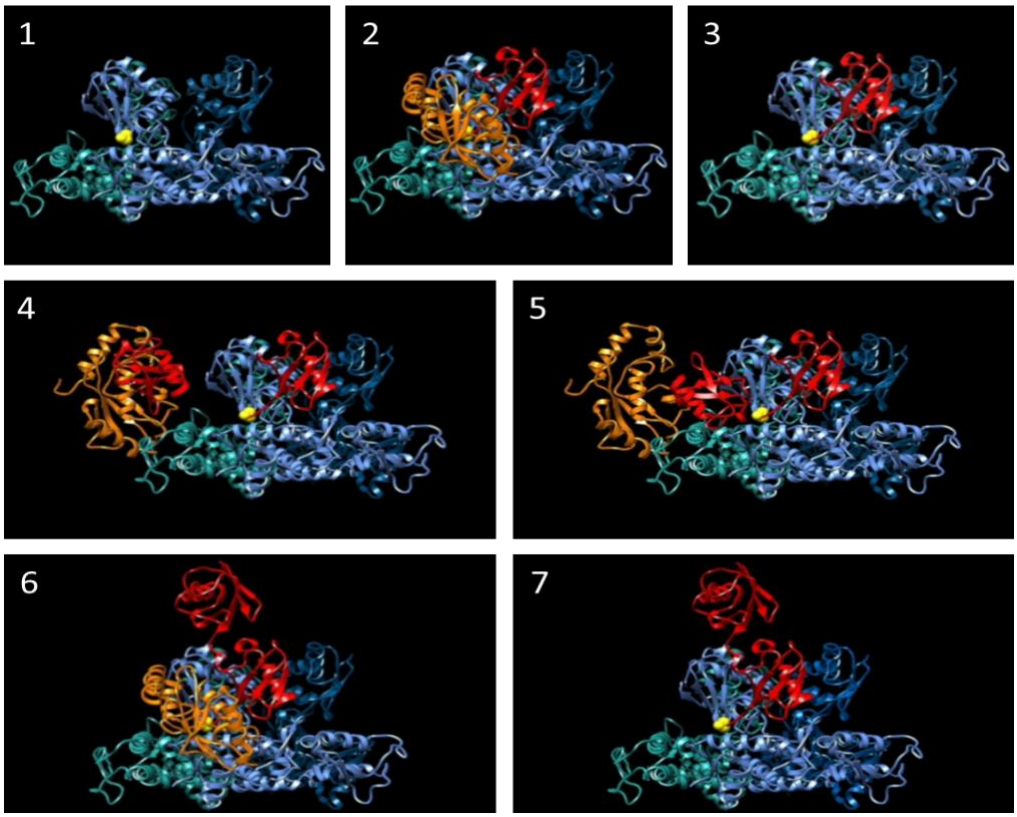


Figure 16: A simulated demonstration of the proximal indexation model in action. The numbers 1-5 match the numbering of each step in figure 14, each UBE3A unit in the oligomer represented by either blue or turquoise with the catalytic cysteine highlighted in yellow, UbcH7 is shown in orange, and ubiquitin is shown in red. (1) The HECT domains in a UBE3A multimer sit so that the E2 binding site 2 of one domain is in proximity of the catalytic cysteine reside of the other. (2) When a ubiquitin-bound E2 enzyme binds to site 1 of the forward facing UBE3A HECT domain, the G76 residue of ubiquitin is within atomic distance of the catalytic site. (3) Transthioesterification occurs, transferring ubiquitin from the E2 enzyme onto the HECT domain C820 residue. The E2 enzyme is then free to dissociate from the complex. (4) Steric hindrance from the ubiquitin residue bound to the UBE3A active site causes a second ubiquitin-bound E2 enzyme to bind to site 2 of the adjacent HECT domain. (5) The E2-bound ubiquitin unit rotates to place its K48 residue in proximity to the C820 active site and the G76 residue of the bound ubiquitin unit. This results in transfer of the G76 residue of the HECT-bound ubiquitin onto the K48 residue of the E2-bound ubiquitin. (6) The transthioesterification process releases the ubiquitin from the active site, which relieves the steric hindrance, allowing the E2 enzyme to undergo a translocation event from site 2 to site 1 of the adjacent HECT domain. (7) The positioning of the E2-bound ubiquitin residue over the active site allows transfer of the ubiquitin chain from the E2 enzyme onto the HECT domain itself, leading to dissociation of the unloaded E2 enzyme, replicating step 3 of the mechanism. The cycle then continues until the ubiquitin chain is of sufficient length, at which point the chain is transferred to a substrate protein. (Figure adapted from Ronchi et al., 2017)

Although the majority of structural studies involving UBE3A focus on the HECT domain at the C-terminal end of the protein, it also has a large region N-terminal to this which has been less thoroughly studied. Although E2-E3 specificity is determined by residues within the catalytic domain (Huang *et al*., 1999), substrate specificity may be determined by residues within the uncharacterised N-terminal region. One section of the N-terminal region has been identified as a zinc-finger domain, or more specifically an AZUL domain, standing for "Amino-terminal Zinc-binding domain of ubiquitin ligase E3A" (Lemak et al., 2011) (PDB 2KR1) (Fig. 17a). This domain appears to be specific for HECT-ligases, and the documented role of zinc-finger domains in protein-protein interactions suggests a role for this region in substrate specificity of the E3 ubiquitin ligase (Lemak *et al.,* 2011). This is supported by a recent study that shows that AS point mutations within the AZUL region contribute to the AS phenotype by preventing UBE3A from binding to the proteasomal subunit PSMD4, a substrate of UBE3A (Kühnle *et al.,* 2018). Even more recently, NMR experiments show that the AZUL domain of UBE3A is able to induce the formation of a similar structured domain, termed the RAZUL domain, in the otherwise disordered C-terminus of PSMD4 (Fig. 17b) (Buel *et al.,* 2020) (PDB 6U19).

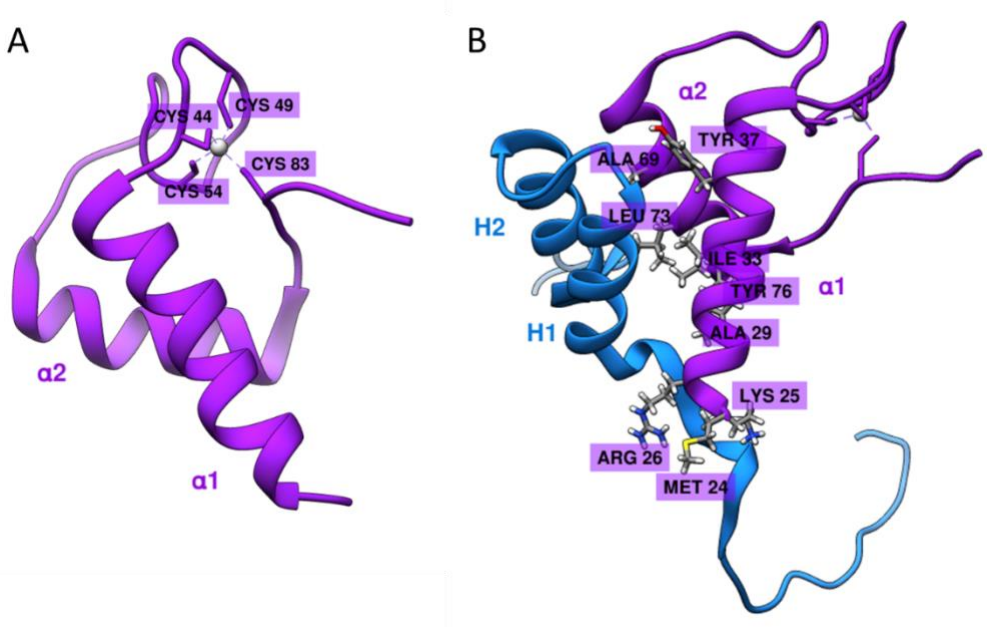


Figure 17: The structure of the UBE3A AZUL domain. A) The AZUL domain of UBE3A only, showing the four cysteine residues involved in coordinating the zinc atom (shown in grey) (2KR1). B) The UBE3A AZUL domain in complex with the PSMD4 RAZUL domain (6U19). UBE3A is shown in purple and PSMD4 in blue. The key residues of UBE3A are shown in a ball and stick format and labelled.

The key residues involved in the interaction are found on the surface of the H1 helix of PSMD4, and the distal surface of the α1 and α2 helices of UBE3A (Buel *et al.,* 2020). Although PSMD4 has been shown to be a substrate of UBE3A, it has also been suggested that the interaction between UBE3A and

PSMD4 may acts to recruit UBE3A to the proteasome, where it can add several short ubiquitin chains to proteins that have already been targetted for degradation in order to improve the efficiency of proteasomal degradation (Buel *et al.,* 2020). This makes it unclear as to whether the interaction between UBE3A and PSMD4 is typical of substrate binding to UBE3A or not.

The only other region of UBE3A that has been solved structurally is a small alpha helix region involved in the formation of the UBE3A-E6-p53 ternary complex (Fig. 18) (Martnez-Zapien *et al.,* 2016) (PDB 4XR8).

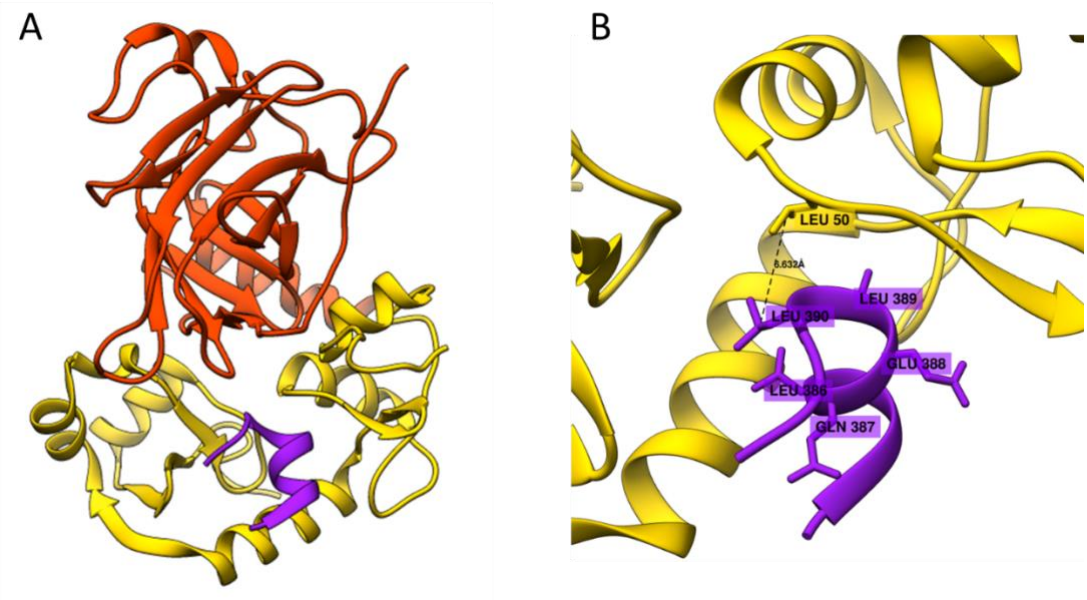


Figure 18: The ternary complex of UBE3A-E6-p53 solved by x-ray crystallography, showing the small LXXLL motif helix of UBE3A. A) The deposited model for the structure (4XR8), p53 is shown in dark orange, E6 in yellow, and the UBE3A peptide in purple. B) A close-up view of the UBE3A-E6 interaction, with key interaction residues shown in ball-and-stick format and labelled. The UBE3A peptide sits between the two lobes of E6, which brings the E6 protein into position to interact with the core region of p53, without the UBE3A peptide interacting with any residues of p53. B) The key interaction between the UBE3A peptide and E6 is between L413 of the LxxLL motif, and L50 in E6.

The LxxLL motif is a small consensus sequence identified in several proteins that interact with the hrHPV E6 proteins (Zanier *et al.,* 2013). In UBE3A, this LxxLL motif is found in a small alpha-helix N-terminal to the catalytic HECT domain with the sequence 'ELTLQELLGEER' at positions 383-394 (according to isoform 1 numbering). An isolated peptide with this sequence was sufficient to stabilise the interaction between the HPV E6 protein and the core domain of p53 (Martinez-Zapien *et al.,* 2016). However, more recent analysis of the the regions involved in the ternary complex suggests that although the LxxLL motif region is sufficient to interact with p53, other regions within UBE3A, other than the HECT domain, are required for p53 ubiquitination (Drews *et al.,* 2020). Experiments involving truncated protein and single amino acid

substitutions identify a region N-terminal of the LxxLL motif, a predicted alpha helix spanning residues 287-297 (isoform 1 numbering) that allows UBE3A to bind to E6 and subsequently ubiquitinate p53 even when the LxxLL motif is unable to interact with E6. They also identify another two separate regions, one spanning residues 98-105 of the N-terminal region and another spanning residues 474-498 immediately upstream of the HECT domain, that enable UBE3A-E6 interaction when the LxxLL-L50 interaction is prohibited (Drews *et al.,* 2020). Although the LxxLL region is a key part of the interaction between UBE3A and hrHPV E6 proteins such as 16E6, low risk E6 proteins don't appear to bind to the isolated LxxLL peptide at all. Instead, residues 538-572 within the HECT domain appear to coordinate the interaction, alongwith the 287-297 helix that is still required for degradation of cellular targets (Drews *et al.,* 2020).

Although only the AZUL domain, the LxxLL motif, and the HECT domain have been solved structurally, the identification of other regions involved in protein-protein interactions through the use of truncation mutants and single amino-acid substitutions also provides valubale insights into the mechanism of UBE3A activity. As well as the E6-binding regions (Drews *et al.,* 2020), the region involved in the interaction with HERC2 has also been identified using this technique (Kühnle *et al.,* 2011). HERC2 is another E3 ubiquitin ligase, it is part of a the HERC protein family where HERC stands for HECT and RLD domain containing proteins. RLD stands for RCC1-like domain, due to the structural similarity between the defined region within the HERC proteins and the RCC1 enzyme, a regulator of chromosome condensation during DNA replication. RCC1 is also a guanine exchange factor (GEF) for the nuclear import protein Ran and the RLD1 domain of another HERC protein, HERC1, potentially acts as a GDP releasing factor (GRF) for cellular proteins, although no GEF or GRF activities have been identified for HERC2 (García-Cano *et al.,* 2019). The HERC family is further subdivided into large HERC and small HERC protein families, where small HERC proteins contain only a single RLD domain, while large HERC proteins can contain several (García-Cano *et al.,* 2019). HERC2 contains 3 RLD domains, and it is the most central RLD domain, RLD2, that interacts with UBE3A (Kühnle *et al.,* 2011). The use of truncation mutants has led to the identification of a region N-terminal to the HECT domain in UBE3A, spanning residues 150-200 in isoform 1, that is responsible for the interaction with RLD2 (Kühnle *et al.,* 2011). There are currently no published structures for this region, and no known protein domains have been predicted from its sequence data, but the observation that the isolated RLD2 domain is able to stimulate UBE3A activity despite its binding site being distal from the catalytic HECT domain (Kühnle *et al.,* 2011) (Fig. 19) suggests an important role for the N-terminal region of UBE3A in the regulation of its ubiquitin ligase activity.

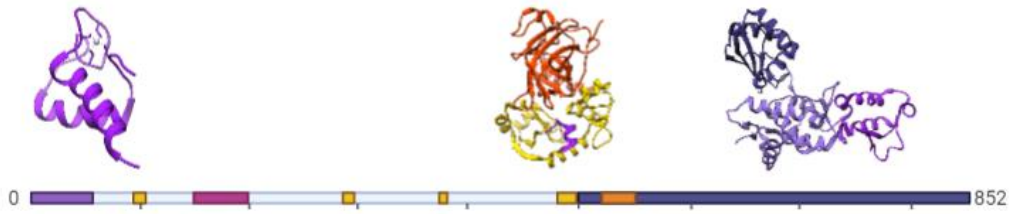


Figure 19: An overview of the current structural information available for UBE3A isoform 1. The light purple box represents the residues of UBE3A isoform 1 that form the AZUL domain, the yellow boxes represent the E6- binding regions within the UBE3A N-terminal region, the pink box represents the region involved in the interaction with HERC2, the dark purple box spanning the C-terminal region of UBE3A represents the HECT domain, and the orange box represents the region within the HECT domain involved in binding lrHPV E6 proteins. Structures for the AZUL domain (2KR1), the LxxLL motif E6-binding region (4XR8), and the HECT domain (1C4Z) are shown in ribbon format above their representative box in the diagram.

1.6 Project Aims

The ultimate goal of this project was to determine the structure of full-length UBE3A at atomic or near-atomic resolution. Previous attempts to crystallise UBE3A have failed because of the predicted flexibility of the disordered N-terminal region, so the most promising technique to generate a high resolution structure is cryo-EM. One line of investigation involves purification of UBE3A alone for cryo-EM imaging, but this is also challenging due to the relatively small size of UBE3A at 100 kDa. Alternative targets that could facilitate cryo-EM reconstruction and also illuminate different roles for UBE3A include PSMD4, E6 and p53, and HERC2. Each of these proteins appear to interact with UBE3A in different places within both the HECT domain and N-terminal regions, so information on each interaction would allow us to develop a better understanding of the mechanism of UBE3A activity.

Although the main goal is to solve structures of UBE3A and its complexes by cryo-EM, it was also the plan to use a variety of biophysical techniques to identify key features of each complex. Analytical ultracentrifugation (AUC), size exclusion chromatography (SEC), circular dichroism (CD), and isothermal titration calorimetry (ITC) can all provide important characteristics of a sample. These include potential stoichiometry of complex formation, effect of complex formation on the structure of UBE3A, and basic shape profile information that will complement cryo-EM analysis, as well as allowing insights into the complexes even if cryo-EM doesn't work as intended.

Alongside the structural and biophysical properties of UBE3A, I have designed an *in vitro* assay to demonstrate the ability of the purified UBE3A to ubiquitinate either itself or other substrate proteins, and I can use this to observe the effects of partner protein interactions on the catalytic activity of

UBE3A. The combination of structural, biophysical, and biochemical data on UBE3A both alone and in several complexes could help elucidate its role at a molecular level, providing insights into the aetiology of both AS and cancer, which could eventually lead to improved treatment options for both disorders.

2 Materials and Methods

2.1 Cells

2.1.1 DNA Propogation

Commercial Top10 *E.coli* cells *(F- mcrA Δ(mrr-hsdRMS-mcrBC) Φ80lacZΔM15 Δ lacX74 recA1 araD139 Δ(araleu)7697 galU galK rpsL (StrR) endA1 nupG)* from Invitrogen were used for plasmid propagation throughout the project.

2.1.2 Protein Expression

BL21 (DE3) pLysS cells were used as a default protein expression vector for all proteins, unless otherwise specified. Originally, commercial stocks were used from various sources, but stocks of competent BL21 pLysS cells were also generated and maintained within the lab. Stocks of BL21 pLysS cells with resistance to both T1 and IME253 phage strains were also provided as a gift from the Membrane Protein Laboratory (MPL) group at Diamond Light Source.

Rosetta (DE3) pLysS cells were used for expression of the UBE3A isoform 1 construct. A mixture of various commercial stocks and $CaCl_2$-generated commercial cell stocks were used throughout, and stocks of phage-resistant Rosetta (DE3) pLysS cells were also provided by the MPL group.

ArcticExpress (DE3) cells (Agilent) were used for the co-expression of the UBE3A-E6-p53 complex.

BL21(DE3) cells with no accessory plasmids were also used for various experiments; these were acquired from various sources.

2.1.3 Mammalian Cells

All mammalian cell expressions performed as part of the work presented here were carried out by members of the Protein Production UK (PPUK) using their published protocols.

2.2 Plasmids

2.2.1 UBE3A

The UBE3A plasmid was a gift from Prof Martin Scheffner (University of Konstantz).

UBE3A was expressed as a gene construct in an unspecified proprietory vector containing a T7 promoter, an N-terminal His tag, a TEV cleavage site, the UBE3A isoform 1 gene, and a kanamycin resistance marker (Fig. 20). A plasmid map was not provided for the proprietory vector, but a sequence was provided for the flanking region of the UBE3A gene.

BglII...T7 promoter...XbaI...*His-tag*...E6AP...**BamHI**...*XhoI*

Figure 20: The fragment of the UBE3A plasmid for which a sequence was provided. The full sequence can be seen in Appendix 2.

2.2.2 UbcH7

The UbcH7 isoform 1 gene was obtained from EuroFins Genomics in the form of a synthesised, *E. coli* codon optimised, gene within a basic vector. The *UbcH7* gene was then cloned into a protein expression vector, as described in section 3.2.2 for use in this project.

2.2.3 E6 and p53

The *E6* gene from HPV16 was synthesised by EuroFins Genomic using the canonical sequence from UniProt (UniProt ID: P03126-1) codon optimised for expression in *E. coli* cells. The *E6* gene was then cloned into expression plasmids, as described in section 3.5, for use in this project.

The *p53* gene was synthesised using the canonical sequence from UniProt (UniProt ID: P04637-1) codon optimised for expression in *E. coli* cells. It was provided in a pET-28a vector from Twist Bioscience. The *p53* gene was then cloned into various vectors for co-expression with other proteins, as described in section 3.5.

2.2.4 PSMD4

The *PSMD4* gene was purchased from the MRC PPU unit at the University of Dundee in the form of a His-3C-*PSMD4* construct in a pET15b plasmid (DU20525; Fig. 21).

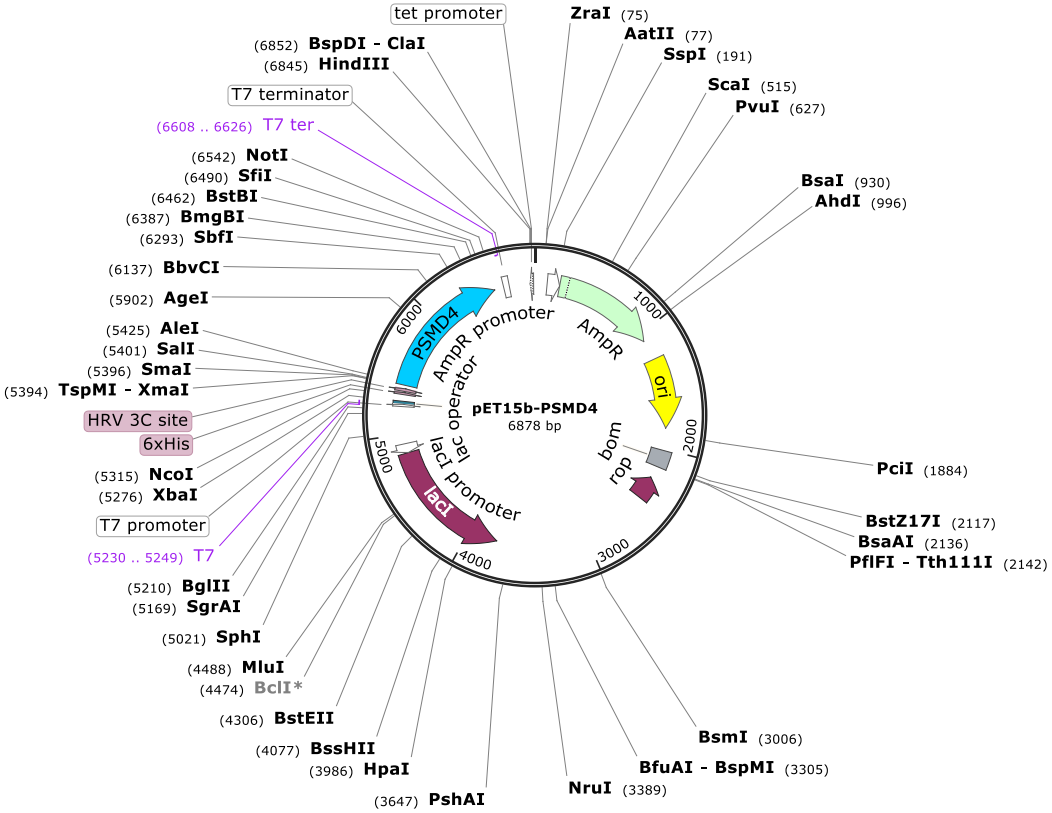


Figure 21: The pET15b-PSMD4 plasmid that was used to express a His-3C-PSMD4 construct in BL21 cells.

2.2.5 HERC2

The full-length *HERC2* gene sequence was purchased from Addgene in a pcDNA5 FRT/TO plasmid. The whole gene was cloned into the proprietary pOpinENeo-GFP-StrepII-His-3C plasmid provided by PPUK using the Gibson Assembly cloning method. The pOpinENeo-*HERC2* construct allowed expression of a GFP-StrepII-His-3C-HERC2 construct in HEK93 mammalian cells.

2.2.6 RLD2

Expression of the RLD2 protein construct throughout this project was done using the pETM11-RLD2 plasmid, described in section 3.2.4.

2.2.7 Ufrag

Expression of the Ufrag region of UBE3A was done using the pETM41-Ufrag plasmid, described in section 3.2.5.

2.3 Materials

2.3.1 Primers

All primers used during this project were synthesised by EuroFins.

Primer Name	Primer Sequence	Primer Purpose	TM (°C)	TARGET SEQUENCE	AMPLICON SIZE (NT)

				NUCLEOTIDES	
E6 Forward	GCAGCACCATGGCCATGCATCAGAAACGTACGGCG	Restriction Digest Cloning of E6 into pACYC	62	1-22	574
E6 Reverse	GCAGCAGGATCCTTATTAGAGCTGGGTTTCGCGACG	Restriction Digest Cloning of E6 into pACYC	63	550-568	574
UbcH7 Forward	GCGAGAATCTTTATTTTCAGGGCGC	Sequencing UbcH7 in pETM40 / pETM41	59	4182-4206	/
UbcH7 Reverse	GTTAGCAGCCGGATCTCA	Sequencing UbcH7 in pETM40 / pETM41	57	18	/
Forward Mid-Sequence Primer	CCATTAGAAACAGAACTGGG	Sequencing full-length pUBE3A	54	1377-1396	/
Reverse Mid-Sequence Primer	GGTTCGTTAATGAATTCCTC	Sequencing full-length pUBE3A	52	1437-1456	/
Reverse Primer	TGGTGGTGCTCGAGGATC	Sequencing full-length pUBE3A	59	2749-2766	/
T485A Forward	GAACGCCGTATTGCGGTGCTGTACTCATTG	QuickChange PCR of the UBE3A T485A mutation	68	1608-1637	/
T485A Reverse	CAATGAGTACAGCACCGCAATACGGCGTTC	QuickChange PCR of the UBE3A T485A mutation	68	1608-1637	/
T485E Forward	CAGCGAACGCCGTATTGAGGTGCTGTACTCATTGG	QuickChange PCR of the UBE3A T485E mutation	72	1604-1638	/
T485E Reverse	CCAATGAGTACAGCACCTCAATACGGCGTTCGCTG	QuickChange PCR of the UBE3A T485E mutation	72	1604-1638	/
OE PCR Forward	GCTGCACCATGGCCAAACGTGCAGCGGCCAAGCAT	Overlap extension PCR of T485 mutants	70	171-191	1,473
OE PCR Reverse	GCAGCTGAATTCTTACAGCATGCCGAAGCCTTTGG	Overlap extension PCR of T485 mutants	65	2704-2726	1,026
T485 BamHI Reverse	GTACCTGGATCCTTACAGCATGCCGAAGCCTTTGG	Overlap extension PCR with a BamHI site	65	2704-2726	1,026
T485 XbaI Forward	GCTGCATCTAGAATGAAACGTGCA	Overlap extension PCR with an XbaI site	69	168-191	1,473

	GCGGCCAAGCAT				
UBE3A Sequencing Forward	TCCAAGTGCTTGAAGATGGT	Sequencing over the T485 region of pUBE3A	58	1212-1231	/
UBE3A Sequencing Reverse	CCATTGCGATCATTTCTAAGCG	Sequencing over the T485 region of pUBE3A	59	1713-1734	/
HERC2 InFusion Forward	AGGAGATATACCATGCCCTCTGAATCTTTCTGTTTGGCTGC	PCR of HERC2 with overlaps for InFusion	65	1-29	14,532
HERC2 InFusion Reverse	GTGATGGTGATGTTTGTGTCCTGTTAAATAATCTTGTGTAGAGTCCGAAGC	PCR of HERC2 with overlaps for InFusion	67	14467-14502	14,532
HERC2 Restriction Forward	GCACAGCCATGGCCCCCTCTGAATCTTTCTGTTTGGCTGC	Add an NcoI site for restriction digest cloning of HERC2	68	4-29	14,524
HERC2 Restriction Reverse	GCACAGGTTTAAACGTGTCCTGTTAAATAATCTTGTGTAGAGTCCGAAGC	Add a PmeI site for restriction digest cloning of HERC2	67	14467-14502	14,524
HERC2 Forward	TTATTTACAATCAAAGGAGATATACATGCCCTCTGAATCTTTCTGTTTGG	PCR of HERC2 with overlaps for Gibson / HiFi assembly	62	1-25	14,552
HERC2 Reverse	GGCCCTGAAACAGAACTTCCAGTTTGTGTCCTGTTAAATAATCTTGTGTAGAGTCC	PCR of HERC2 with overlaps for Gibson/HiFi assembly	62	14472-14502	14,552
cDNA Primer	CGCAAATGGGCGGTAGGCGTG	HERC2 Sequencing in the cDNA plasmid	67	769-789	/
Primer 1	AGACTGACAATGAGCGTTCC	HERC2 Sequencing	58	968-987	/
Primer 2	GCTTCAAGACTTGGATGTGG	HERC2 Sequencing	57	1959-1978	/
Primer 3	CATAGAAGCAGGACTCCACTG	HERC2 Sequencing	58	2944-2964	/
Primer 4	AGACACAGAGAGGAATCTGGG	HERC2 Sequencing	59	3909-3929	/
Primer 5	TTTATCCGCAGTCTCCACTCC	HERC2 Sequencing	60	4886-4906	/
Primer 6	CACCCCACTGCAATGATG	HERC2 Sequencing	57	5872-2889	/

Primer 7	TCAGTTGGTGAACCTCGCT	HERC2 Sequencing	60	6852-6870	/
Primer 8	AGAAAGGGGGCACCTACT	HERC2 Sequencing	58	7838-7855	/
Primer 9	GCGGAAGCCTCATTAGAAAGA	HERC2 Sequencing	59	8819-8839	/
Primer 10	CAGGTGTATGCTTGGGGTGA	HERC2 Sequencing	60	9829-9848	/
Primer 11	TTGGAGGATGTGGCCACAGA	HERC2 Sequencing	62	10783-10802	/
Primer 12	GAGAGCAAGACGAACAACTTG	HERC2 Sequencing	58	11780-11800	/
Primer 13	GAAGAAAGTCATCGCCATCGC	HERC2 Sequencing	60	12744-12764	/
Primer 14	CTGGACTCATGTACATCCGAGAC	HERC2 Sequencing	61	13742-13764	/
Primer 15	ATCCTTATCAACCTCACTGAGGT	HERC2 Sequencing	59	13712-13764	/
Primer 16	CGCAATGATGGGGATCTCCT	HERC2 Sequencing	60	13016-13035	/
Primer 17	CCAATGCCTAGTCTGCCAC	HERC2 Sequencing	59	12041-12059	/
Primer 18	ATAGACGGTGAAGCGCCA	HERC2 Sequencing	59	11104-11121	/
Primer 19	GGCCAGATTTCCATCCAATG	HERC2 Sequencing	57	10157-10176	/
Primer 20	ATCGTCACCTTCGCCCCA	HERC2 Sequencing	62	9211-9228	/
Primer 21	ATACCGCAGACTGACCTGG	HERC2 Sequencing	59	8257-8275	/
Primer 22	TTGCGCCTCTTCACTCTG	HERC2 Sequencing	58	7338-7355	/
Primer 23	AGCGTGGACTCTCTGAGTA	HERC2 Sequencing	58	6350-6368	/
Primer 24	AGCATGCCGGAATTGAGC	HERC2 Sequencing	59	5415-5432	/
Primer 25	GAGCGAACATTTTGCTTGGTA	HERC2 Sequencing	58	4453-4473	/
Primer 26	GGTGCCAGATGGTTGAACC	HERC2 Sequencing	59	3509-3527	/
Primer 27	CTGCCCAGACCTAAACCAAG	HERC2 Sequencing	59	2557-2576	/
Primer 28	CGGAGATCACCTTAGGCTCC	HERC2 Sequencing	60	1608-1627	/
Primer 29	CAACAGCTCACTGCAGAG	HERC2 Sequencing	56	658-675	/
2958-3362 Forward	GCACACCCATGGCCAGAACCAAGGTGTTTGTGTGG	Isolation of RLD2 construct 1	63	8872-8892	1.239
RLD Construct 1 Reverse	GGTAGAGAATTCTTAAGCAGAAG	Isolation of RLD2 construct 1	62	10063-10086	1,239

	AATCAGCATCTG AAGG				
2959-3327 Forward	GCACACCCATGG CCACCAAGGTGT TTGTGTGG	Isolation of RLD2 construct 2	75	8875-8892	1,131
RLD Construct 2 Reverse	GGTAGAGAATTC TTACACAGTTGT CCACGCC	Isolation of RLD2 construct 2	67	9966-9981	1,131
3010-3323 Forward	GCACACCCATGG AAGGGAAGGTG TATGCC	Isolation of RLD2 construct 3	71	9026-9045	966
RLD Construct 3 Reverse	GGTAGAGAATTC TTACGCCACACT GTGGG	Isolation of RLD2 construct 3	67	9956-9970	966
UBE3A Fragment Forward	GCACACCCATGG CCCACACTAAGG AAGAGCTTAAAT CACTG	Isolation of the 'Ufrag' fragment	60	615-641	176
UBE3A Fragment Reverse	GGTAGAGAATTC TTAGTTGTTATC GCCCTGGGAG	Isolation of the 'Ufrag' fragment	60	749-767	176

Table 7: A list of the primers used during this project.

2.3.2 DNA Sequencing

All DNA sequencing was performed by Source Bioscience. Most reactions were performed by providing 5 µl of purified plasmid at 100 ng/µl and 5 µl of the chosen primer at 3.2 pmol/µl. The DNA sequences provided were viewed using the SnapGene Viewer software (from Insightful Science; available at snapgene.com), and sequences were translated into protein sequences using the ExPasy translation server (Gasteiger *et al.,* 2003) for confirmation.

2.3.3 Gene Synthesis

The genes for the human UbcH7 and HPV16 E6 proteins were synthesised by EuroFins Genomic and provided in a pET-28a vector as a lyophilised powder containing 2.4 and 2 µg of DNA.

The human p53 gene was synthesised by Twist Bioscience and provided in a proprietary pTwist vector as a lyophilised powder containing 2 µg of DNA.

2.3.4 Kits and Consumables

Qiagen Miniprep Kit

Plasmids were purified from bacterial cultures using Qiagen's QiaPrep Spin Miniprep Kit. The kit contains several premixed buffers and filter columns. Buffer P1 is comprised of 50 mM Tris, 10 mM EDTA, pH 8, and 100 µg/ml RNase A. Buffer P2 is comprised of 200 mM NaOH and 1 % SDS (w/v). Buffer N3 contains 4.2 M guanidine chloride and 3M potassium acetate, pH 4.8. Buffer PE is proprietory to Qiagen, but it contains 80 % ethanol. The elution buffer, buffer EB, consists of 10 mM Tris at pH 8.5.

Bacterial cultures were prepared by inoculating a single colony in 5 ml LB media supplemented with the corresponding antibiotic, and incubating at 37°C 200 rpm for 16 h. The cells were pelleted by centrifuging at 4,000 xg for eight min. The cells were resuspended in 250 μl P1 and transferred to an Eppendorf tube, then 250 μl P2 was added to encourage cell lysis and the tube was inverted several times. 350 μl N3 was added to neutralise the mixture and the tube was inverted again to mix the reagents. The tubes were centrifuged at 12,000 xg for 10 min to pellet the cell components, and the supernatant was transferred to a filter cartridge inside a collection tube. This was centrifuged at 12,000 xg for 1 min, and the flow through was discarded. 750 μl buffer PE was added to wash the filter, it was centrifuged at 12,000 xg for 1 min and the flow through was discarded. The filter was centrifuged for a further 1 min at 12,000 xg to ensure all of the wash buffer was removed, and then 50 μl buffer EB was added to the centre of the filter. The elution buffer was incubated on the filter for 1 min on the bench, and then it was centrifuged for 1 min at 12,000 xg and the elution was collected in a fresh tube. The purity and concentration of the extracted DNA was measured on a ThermoFisher Nanodrop instrument.

Qiagen Plasmid Plus Midi-Prep Kit

A starter culture was prepared by inoculating a single colony from a transformation plate into 50 ml LB with 50 μg/ml carbenicillin, and incubating at 37°C 200 rpm for 16 h. The cells were separated from the media by centrifugation at 4500 xg for 20 min, the media was discarded, and the pellet was resuspended in 2250 μl buffer P1 from the Qiagen kit containing RNase A. 2250 μl of buffer P2, comprised of primarily NaOH, was then added and left at RT for 3 min to carry out the alkaline lysis of the cells. 2250 μl of buffer S3 was then added to neutralise the reaction and precipitate out salts, large chromosomal DNA strands, proteins, and cell debris. The solutions were mixed by inverting, left at RT for 10 min, and centrifuged at 4500 rpm for 17 min. The supernatant was transferred to a fresh Falcon tube and the pellet was discarded. 2 ml buffer BB was added to adjust the buffer conditions to encourage the plasmid DNA to bind to the filter, and the mixture was added to the column with an extender attached. This was then subjected to the vacuum pump until the solution had flowed through, leaving the DNA bound to the column filter. The filter was washed by adding 700 μl of buffer ETR and pulling it over the filter with the vacuum pump. This step removes any endotoxins that were present. A further wash was carried out using buffer PE, and a residual buffer was removed from the filter by centrifugation at 10,000 xg in a benchtop centrifuge for 1 min. The purified DNA was eluted by incubating 200 μl buffer EB on the filter for 1 min, before centrifuging at 17,000 xg for 1 min and collecting the elute in a fresh tube. The elute contains the plasmid DNA solubilised in buffer EB, which is simply 10 mM Tris, pH 8.5. 1.5 μl of the elute was subjected to the nanodrop instrument for concentration determination prior to any further use of the DNA sample.

NEB PCR Cleanup/Gel Extraction Kit

The NEB Monarch PCR and DNA cleanup kit contains a series of proprietary vectors. However, the binding buffer is guanidine and isopropanol based to ensure that the DNA is free of proteins of other contaminants, the wash buffer is an ethanol based buffer, and the elution buffer consists of 10 mM Tris, 0.1 mM EDTA, pH 8.5. The gel dissolving buffer is a guanidine thiocyanate and sodium iodide based buffer to dissolve the agarose gel.

The DNA sample to be purified was mixed with binding buffer in a 1:2 ratio for constructs above 2 kb, or 1:5 for constructs below 2 kb. The solution was loaded into a filter cartridge inside a collection tube and then centrifuged for 1 min at 12,000 xg. The flow through was discarded, and 200 μl wash buffer was added. The filter was centrifuged at 12,000 xg for another minute, the flow through was discarded, and another 200 μl wash buffer was added. The filter was transferred to a new Eppendorf tube and 15 μl elution buffer was added to the centre of the filter. The buffer was left to incubate with the filter for 1 min, and then the filter was centrifuged at 12,000 xg for 1 min to collect the eluted DNA.

For the gel extraction protocol, the sample was subjected to a 1% agarose gel with TAE. Large wells were used in the gel so that the sample formed a shallow band with minimal spread as the gel runs. The band was excised from the gel using a scalpel on a transilluminator plate, including as little excess gel as possible. Gel dissolving buffer was added to the gel fragment, adding 400 μl buffer per 100 mg of gel. The gel was incubated in the buffer at 50^{o}C and mixed periodically using a vortex mixer until the gel slice was completely dissolved. The dissolved gel solution was added to the filter cartridge and purified following the standard DNA purification protocol as described above.

Takara MightyMix Ligation Kit

The Takara mighty mix solution is a 2X reaction mixture that includes the ligase enzyme and buffer components in a single solution. Digested vector and gene insert DNA fragments are mixed in a ratio of 1:3, with the vector at 25 fmol and a total reaction volume between 5 – 10 μl. An equal volume of ligation mix is added, and the mixture is incubated at RT for 15 – 30 min. The ligation mix can then be used immediately for transformation into chemically competent cells.

Vazyme ClonExpress II Kit

The ClonExpress II kit from Vazyme includes the Exnase II enzyme and 5X CE II buffer solution. Primers were designed following the instructions from the kit to generate a gene insert product with overlapping sequences with the linearised vector. The linearised pOpinENeo-GFP-StrepII-His-3C used for cloning of *HERC2* was provided by the Membrane Protein Laboratory (MPL) group at Diamond Light Source. The cloning was carried out as described in section 2.4.4.

Invitrogen Gibson Assembly Kit

The Invitrogen Gibson Assembly kit includes a single reaction mix that includes the enzyme and buffer components required for the reaction. 0.08 pmol of linearised vector and 0.08 pmol of gene fragment were mixed in a volume of less than 10 µl, 10 µl Gibson assembly mix was added, and then Milli-Q purified water was added to a final volume of 20 µl. The reaction was incubated at 15°C for 15 – 60 min, and the products were identified using agarose gel electrophoresis.

NEB HiFi Assembly Kit

The NEB HiFi Assembly kit included a single reaction mix containing the enzyme and buffer components required for the reaction. 0.2 pmol gene insert was mixed with 0.1 pmol linearised vector in a volue of less than 10 µl. 10 µl HiFi assembly mix was added, and deionised water was added to bring the reaction mixture up to 20 µl. The reaction mix was incubated at 50 °C for 15 – 60 min, and the products were identified using agarose gel electrophoresis.

Acrylamide Gels

Acrylamide gels for SDS-PAGE analysis of protein samples were purchased from BioRad for use in the BioRad mini Protean tetra system. Unless otherwise stated, 4-20% gradient TGX acrylamide gels were used. The gels were run using a tris-glycine buffer system, comprised of 25 mM Tris, 192 mM Glycine, and 0.1% SDS for a denaturing system or no SDS for a native system.

Gel Stains

Unless otherwise stated, SDS-PAGE gels were stained with the Coomassie-based InstantBlue gel stain (SigmaAldrich). Gels were removed from the cassette after an electrophoresis run and placed directly into a container with ~20 ml gel stain. The gel was left in the solution on a rocker for 15-30 min, and then the gel stain solution was replaced with water. The gel could be imaged immediately without the need for a destaining step, although gels were typically left for at least 1 h in water to remove any background stain.

Media

Luria broth (LB) and Terrific broth (TB) (Formedium) were used for *E. coli* cultures throughout the project. They were both provided as a pre-mixed powder containing all of the nutrients required for *E. coli* growth, although the TB mix required supplementing with 4 ml glycerol per litre before autoclaving.

2.4 Cloning

2.4.1 Transformations

Purified DNA was transformed into chemically competent cells using the heat shock method. The cells commonly used are described in section 2.1. Plasmids were used at varying concentrations between 10 – 200 ng/µl and 2 – 5 µl DNA was added to a 50 µl aliquot of competent cells. The cell and DNA mixture was

left on ice for 30 min. The sample was subjected to a heat block at 42°C for 30 s, followed by 2 min on ice. 250 µl LB media was then added to the transformation mixture, and the culture was incubated at 37°C 200 rpm for 1 h. The cell suspension was then spread on an agar plate containing LB agar and the required antibiotic for plasmid selection, and the plate was incubated at 37°C overnight. The antibiotic used was dependant on the plasmid being transformed (section 2.2).

2.4.2 PCR Mutagenesis

QuickChange Mutagenesis

For a QuickChange PCR reaction primers were designed to include an overlapping region including the desired mutation. Each primer was ~25-30 residues long, with an overlapping region of ~10-15 residues on either side of the mutated residue. The primers were then used to amplify the whole plasmid, resulting in a circular product (Fig. 22).

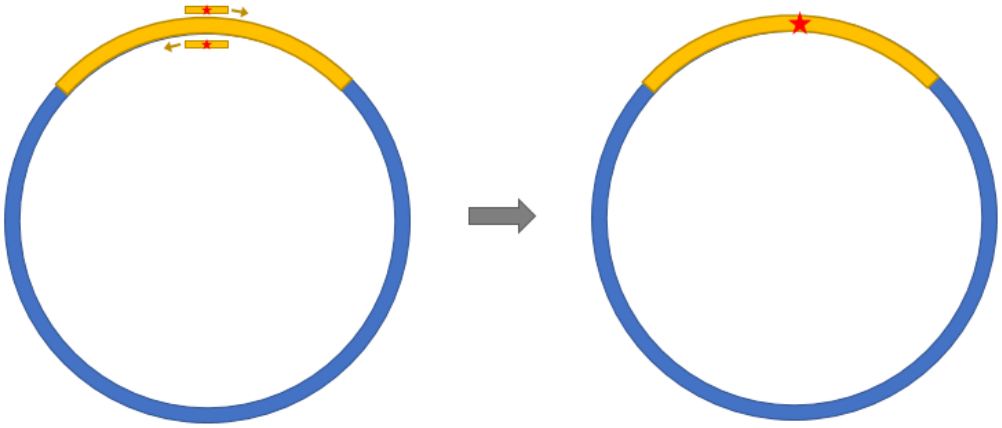


Figure 22: A diagram representing the mechanism of the QuickChange mutagenesis protocol. A pair of overlapping primers are generated that contain the desired mutation in the middle. A single PCR reaction is run to amplify the entire plasmid, incorporating the mutation into the sequence.

QuickChange PCR reactions were carried out using the Phusion HF polymerase from Thermo Scientific, and the PCR conditions were varied according to the requirements of the template. Typically, an initial denaturation step was carried out at 98°C for 30 seconds. The next stage of the PCR involved 30 cycles, each of which involving a 10 s denaturing step at 98°C, an annealing step at 45-72°C for 15-30 s, and an extension step at 72°C for a time determined by the length of the plasmid. The annealing temperature was set at 3 degrees below the lowest Tm of the primers, and the extension time was calculated as 20 s per kb of template. After the 30 cycles, a final annealing step occurred for 7 min at 72 °C, and then the reaction was held at 4°C or placed on ice until use.

Following PCR, at least some of the sample was subjected to an agarose gel to confirm that the size of the product is as expected before the amplified plasmid was purified and transformed into Top10 cells. The cells were

subjected to a miniprep purification and the DNA samples were sent for sequencing before further use.

Overlap-Extension PCR

Overlap-Extension PCR protocols involve several separate PCR reactions to amplify only the gene of interest, which then needs to be re-inserted into a plasmid of choice. The first step of an overlap extension protocol is to create two (or more if required) fragments of the gene with an overlapping region in the middle (Fig. 23). This overlapping region will contain the desired mutation. Once the two fragments have been created through PCR, a second PCR reaction is performed without primers, using the overlapping regions of the gene fragments as their own primers, to extend each fragment to cover the full gene over 15 cycles. After this reaction, the terminal primers containing the restriction sites are added and another 20 cycles are completed in order to amplify the complete gene product.

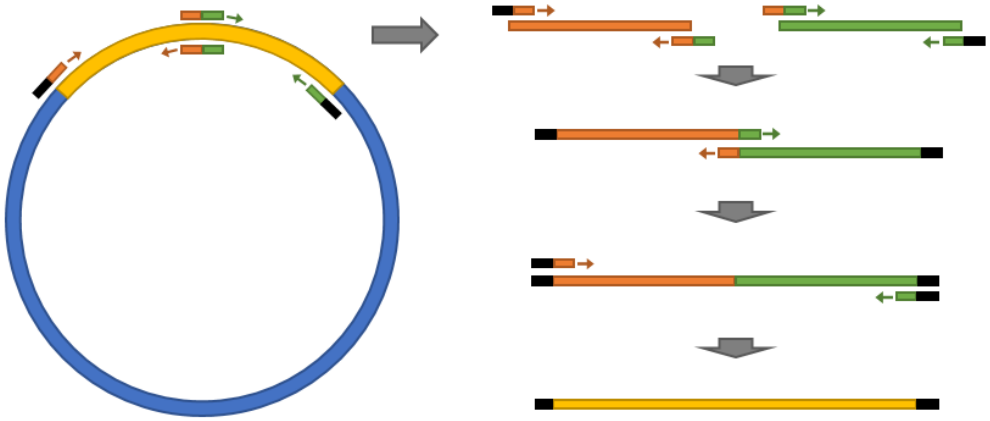


Figure 23: A diagram demonstrating the overlap extension PCR process. The coloured blocks represent portions of the gene, the blue area represents the vector backbone, and the black blocks represent restriction enzymes introduced to allow re-insertion of the gene into a vector of choice.

The final PCR product is purified using the NEB gel purification kit, and then it can be used in a restriction digest reaction to create the final plasmid construct.

2.4.3 Restriction Enzyme Digests

Both the plasmid containing the gene of interest and the desired vector plasmid were digested by mixing 1 µg DNA with 0.5 µl each restriction enzyme at 20,000 units/ml, along with NEB CutSmart buffer (New England Biolobs, MA, USA) diluted to 1X in a final volume of 20 µl. This was incubated at 37°C for 30 min. For the vector plasmid, if heat inactivatable enzymes were used the reaction mix was heated to 80°C for 20 min to stop the cleavage reaction, and 1 µl quick CIP enzyme from NEB at 5,000 units/ml was added to dephosphorylate the end of the DNA fragment. The dephosphorylation reaction was incubated at 37°C for 5 min, and then the enzyme was

deactivated at 80°C for 20 min. If heat in-activatable restriction enzymes were not used, the DNA digest reaction mixture was purified using the NEB PCR cleanup kit as described in section 2.3.4 and the purified DNA was used to prepare a fresh reaction mixture in 1X CutSmart buffer to a final volume of 20 µl. Following dephosphorylation, the cleaved and dephosphylated DNA was purified from the mixture using an NEB PCR cleanup kit, while the cleaved gene of interest was separated from the corresponding plasmid backbone by running the sample down an agarose gel at 1% in TAE buffer at 70V for 1 h. The band corresponding to the expected size for the gene was excised from the gel and purified using the NEB gel extraction kit, while the larger band corresponding to the empty plasmid remained in the gel and was discarded.

The sample was measured on a Nanodrop instrument from ThermoFisher to ascertain both the purity of the DNA and the concentration of the sample.

The purified insert and dephosphorylated vector were ligated using Takara's Mighty Mix ligation mastermix (section 2.3.4).

2.4.4 InFusion Cloning

InFusion cloning involves designing primers to produce a gene insert containing regions on either end that will overlap with the plasmid into which it will be inserted. For the ClonExpress II kit from Vazyme used in this work, the optimal overlap length was 15 - 20 bp, so primers were designed to be ~30 bp in length, with the first half of the primer overlapping with the vector sequence and the second half overlapping with the insert sequence. These primers were used to amplify the gene from its original vector using PCR. The vector plasmid was linearised by cleaving with *NcoI* and *PmeI* restriction enzymes in a double digest reaction for 30 min at 37°C. In theory, the PCR product and the linearised vector can be combined with the Exnase enzyme from the ClonExpress II kit without the need for a PCR clean-up step, incubated for 30 min at 37°C to recombine the plasmid, and then used immediately for bacterial transformations. However, as *HERC2* is a very large gene at over 14kb, extra steps were taken to ensure that the reaction could be as efficient as possible.

The *HERC2* gene was amplified with PCR to create a product that contained 15 bp regions on either end that would overlap with the linearised vector plasmid. The linearised vector was provided by the Membrane Protein Laboratory (Diamond Light Source, Oxfordshire, UK), and the both the amplified gene product and the linearised vector were visualised using a 1% TAE agarose gel. Both products produced clear bands at the expected molecular weights, so the bands were excised and the DNA was extracted from each sample using the NEB Monarch gel extraction kit (see section 2.3.4). The purified samples were quantified using the nanodrop instrument, and were then mixed in a 2:1, 1:1, and 1:2 ratio, ensuring that the concentration of either component did not exceed 200 ng and the total

volume did not exceed 14 µl. 2 µl of the Exnase II enzyme was added at the provided concentration, along with 4 µl of 5x CE buffer, and the reaction was incubated at 37^{o}C for 30 min. After 30 min it was immediately moved to ice, 5 µl was transformed into competent Top10 cells, and the remainder of the reaction mix was analysed using an agarose gel to observe the products.

The ClonExpress II kit used in this work is advertised to be effective for inserts between 50 bp and 10 kb, but *HERC2* alone was 14 kb and the vector plasmid added another ~8 kb, so regardless of how much optimisation was attempted, this experimental route was ultimately unsuccessful.

2.4.5 Gibson/HiFi Assembly

Gibson and HiFi cloning are similar to InFusion cloning in that they rely on an overlapping region between the insert and the vector to insert the gene. Whereas the InFusion method is only intended for inserting a single gene insert into the vector, the Gibson and HiFi protocols are designed to allow assembling of multiple gene fragments in a single reaction. Although the assembling of several fragments was not necessary for cloning of the *HERC2* gene, the increased capacity of this kit for dealing with larger volumes of insert DNA suggested that it may work more effectively than the ClonExpress II kit had done.

Gibson assembly was attempted using the GeneArt Gibson Assembly HiFi Master Mix kit from Invitrogen. This kit advertises the ability to insert gene fragments up to 32 kb, so the 14 kb *HERC2* should be possible. FFollowing the recommendations from the kit, new primers were designed to allow a 35 bp overlap following PCR of the gene. In theory, the Gibson assembly kit does not require purification of the PCR product or the linearised vector prior to including them in the reaction mix, but due to the difficulties with *HERC2* cloning previously both were gel-purified. The insert gene and vector plasmid were mixed in an equimolar ratio, using 0.08 pmols of each species, in a volume of no more 10 µl. To this, 10 µl 2X GeneArt Gibson Assembly HiFi Master Mix was added, and if necessary dH_2O was added to bring the reaction volume up to 20 µl. The reaction was incubated at 50^{o}C for 15 min to 1 h, and then the reaction mix was placed on ice and transformed into competent Top10 cells.

The NEBuilder HiFi Assembly kit was similar to the Gibson kit but was designed more for assembling a diverse range of fragment or oligo species than it was long single gene inserts. The recommended overlap length for a reaction involving 2-3 fragments was 15-20 bp, while the recommended overlap length for 3+ fragments was 20-30 bp. As *HERC2* was only a single insert that was the length of multiple typically sized inserts, it was conducted using both the primers from the original InFusion cloning attempt with 15 bp overlap, and the primers designed for the Gibson assembly reaction with a 35 bp overlap. Once the PCR products had been created and the vector was

linearised and the samples were both gel-purified. Insert and vector samples were mixed in 2:1, 1:1, and 1:2 molar ratios due to the size of the *HERC2* gene compared to the pOpin vector, aiming for a total DNA amount of 0.2 pmols, in a volume no greater than 10 µl. To this DNA mixture, 10 µl 2X Master Mix solution was added, and dH_2O was added if required to bring the reaction volume up to 20 µl. The reaction mix was incubated at 50°C for 15 min to 1 h, and then it was transferred to ice and transformed into chemically competent Top10 cells.

For each transformation, a colony PCR (see section 2.4.6) was used to identify the presence of the insert, and then positive colonies were cultured and the DNA was extracted using the miniprep kit (section 2.3.4). The purified DNA was then sent for sequencing using the vectors designed for the pOpinE vector. Ultimately, the GeneArt Gibson Assembly HiFi kit was able to produce the pOpinENeo-GFP-StrepII-His-HERC2 plasmid that was required.

2.4.6 Colony PCR Screening

A colony PCR screen uses a colony from a transformation plate as the template for a PCR reaction. As the Phusion polymerase was the preferred polymerase used through the project, the cells were lysed during the initial denaturation step at 95 °C. For a colony PCR reaction, a standard PCR reaction mix was assembled, with 2.5 µl 10x Phusion HF buffer, 1 µl 10 mM dNTP solution, 1.5 µl each primer at 10 µM, and 0.5 µl Phusion polymerase at 2 units/µl. The reaction was made up to 25 µl with the addition of dH_2O, and a single colony from the plate was picked using a pipette tip and swirled in the PCR mix. The end of the tip was squeezed with a finger to ensure that none of the reaction mix remained in the pipette tip, and then the reaction mix was subjected to PCR. If a band appears on the gel then the insert region must be present in the cells from that colony, but it must still be sequenced to confirm that no errors have been introduced in the process.

2.5 Protein Expression

2.5.1 Overexpression in *E. Coli*

Proteins were expressed using either Luria Broth (LB) or Terrific Broth (TB) media (section 2.3.4). For either media type, starter cultures were created by inoculating a single colony from an agar plate into LB media containing the appropriate antibiotics, ensuring 10 ml media per 1 L culture to be innoculated,. The starter cultures were incubated at 37 °C and 200 rpm for 16 h, and then 10 ml of each starter culture was added to a 1L culture containing either LB or TB supplemented with the appropriate antibiotic. These large cultures were incubated at 37 °C 200 rpm until the the absorbance at 600 nm (OD_{600}) reach 0.3 for LB media or 0.6 for TB media. At this point the temperature was dropped to the chosen post-induction temperature. Cultures were induced with IPTG at OD_{600} 0.6 for LB and 1.2 for TB media. After induction, cultures grown at 37 °C were incubated at 200 rpm for a further 3 h, wheras cultures incubated at lower post-induction temperatures were left incubating at 200 rpm overnight. After incubation, the cells were centrifuged in a Beckman JA

8.1000 rotor at 5000 xg for 10 min to pellet the cells, the supernatant was removed, and the cell pellet was resuspended in 15ml lysis buffer before being stored at -80°C.

UBE3A

For UBE3A cell cultures starter cultures were set up using a UBE3A Rosetta PlysS transformation plate using 50 µg/ml kanamycin and 37 µg/ml chloramphenicol. Cultures were grown using both LB and TB on different occasions, both supplemented with 50 µg/ml kanamycin and 37 µg/ml chloramphenicol, and using a post-induction temperature of 25 °C.

In LB cultures UBE3A expression was induced with a final concentration of 200 µM IPTG, whereas in TB media a final concentration of 1 mM IPTG was used.

The lysis buffer used for UBE3A samples was referred to as Buffer A, and consisted of 100 mM Tris, 600 mM NaCl, 40 mM Imidazole, pH 8.

UbcH7

For expression of UbcH7, starter cultures were set up using a UbcH7 BL21 PlysS transformation plate, and were supplemented with 50 µg/ml kanamycin. UbcH7 was expressed solely in LB media supplemented with 50 µg/ml kanamycin, and with a constant temperature throughout of 37 °C. The culture was induced at OD_{600} 0.6 with 1 mM IPTG (final concentration), and cells were harvested after 3 h.

The lysis buffer for UbcH7 samples was Buffer C, which consisted of 50 mM Tris, 150 mM NaCl, pH 8.

PSMD4

For expression of PSMD4, starter cultures were set up using a PSMD4 BL21 PlysS transformation plate, and were supplemented with 50 µg/ml kanamycin. UbcH7 was expressed solely in LB media supplemented with 50 µg/ml kanamycin, and with a constant temperature throughout of 37 °C. The culture was induced at OD_{600} 0.6 with 1 mM IPTG (final concentration), and cells were harvested after 3 h.

The lysis buffer for PSMD4 samples was Buffer C, which consisted of 50 mM Tris, 150 mM NaCl, pH 8.

RLD2

For RLD2 cell cultures starter cultures were set up using an RLD2 BL21 PlysS transformation plate using 50 µg/ml kanamycin and 37 µg/ml chloramphenicol. Cultures were grown using both LB and TB on different occasions, both supplemented with 50 µg/ml kanamycin and 37 µg/ml chloramphenicol, and maintaining a constant temperature of 37 °C. In both LB and TB cultures the expression was induced with a final concentration of 1 mM IPTG.

The lysis buffer for RLD2 samples was Buffer C, which consisted of 50 mM Tris, 150 mM NaCl, pH 8.

UBE3A-E6-p53

Co-expression of the pUBE3A and pACYC-E6-p53 plasmids involved successive transformation of the plasmids into BL21 cells, and a single colony from this was used to set up a starter culture of 10 ml LB, 50 µg/ml kanamycin, and 37 µg/ml chloramphenicol for incubation at 37°C 200 rpm for 16 h. This starter culture was then inoculated into 1L LB supplemented with 50 µg/ml kanamycin and 37 µg/ml chloramphenicol. The culture was incubated with a pre-induction temperature of 37°C and a post-induction temperature of 16°C. Cells were induced with 1 mM IPTG, and then harvested 24 h after induction.

Co-expression of the pUBE3A and pACYC-E6-p53 plasmids in ArcticExpress cells was carried out in the same way as in BL21 cells, other than the addition of 50 µg/ml gentamicin in the cultures.

Co-expression of E6 and p53 from the single pACYC-E6-p53 plasmid involved transforming the plasmid into BL21 cells and using a single colony to set up a 10 ml starter culture consisting of LB with 37 µg/ml chloramphenicol. This culture was incubated at 37°C 200 rpm for 16 h, and then inoculated into 1L LB with 37 µg/ml chloramphenicol. The large-scale culture was incubated at 37°C pre-induction and 16°C post induction with 1 mM IPTG. The cells were harvested 24 h after induction.

Co-expression of E6 and p53 from the single pCDF-E6-p53 plasmid was carried out in almost exactly the same way, other than that the plasmid was transformed into Rosetta pLysS cells. For these cells the 10 ml LB starter culture was supplemented with 37 µg/ml chloramphenicol and 100 µg/ml streptomycin, and the large scale culture was supplemented with 100 µg/ml streptomycin only. The temperatures and timings were kept the same as the pACYC-E6-p53 expression protocol.

Co-expression of UBE3A, E6, and p53 using the pUBE3A, pACYC-E6, and pCDF-p53 plasmids involved sequentially transforming all three constructs into BL21 cells. A single colony from this was used to inoculate a starter culture containing 10 ml LB, 30 µg/ml kanamycin, 20 µg/ml streptomycin, 30 µg/ml chloramphenicol, and 20 µg/ml gentamycin. This was incubated at 37°C, 200rpm for 16 h, and then 50 µl of the culture was spread on an LB agar plate with the same antibiotics. This plate was incubated at 37°C for 8 h, and then all colonies were scraped into 40 ml LB with no antibiotics. This 40 ml suspension was then used to inoculate 1L TB with 30 µg/ml kanamycin, 20 µg/ml streptomycin, and 30 µg/ml chloramphenicol, which was incubated at 16°C, 200 rpm. Protein expression was induced with 400 µM IPTG at OD_{600} = , and cells were harvested 24 h later.

Co-expression of UBE3A, E6, and p53 using the pUBE3A, pACYC-E6, and pCDF-p53 plasmids was also conducted in ArcticExpress cells. This used the same protocol as above, but with the addition of 50 µg/ml gentamicin alongside the antibiotics mentioned.

Co-expression using the plasmids provided by Dr Masuda was carried out using the same protocol as with the pUBE3A, pACYC-E6, and pCDF-p53 strategies outlined above. The only difference was the plasmids used in the intitial transformations.

Cells were harvested by centrifugation and any pellet was resuspended in lysis buffer for storage at -80 °C.

The lysis buffer used to resuspend the UBE3A-E6-p53 cell pellets consisted of 50 mM HEPES, 1M NaCl, 10 mM β-ME, 0.1 mM EDTA, pH 7.5.

2.5.2 Overexpression in Mammalian Cells

The pOpinENeo-*HERC2* construct was passed on to members of the PPUK group for expression in HEK293 cells. All cell culture work was carried out by PPUK, and cells were lysed and clarified before returning to us for small-scale purification trials.

2.6 Protein Purification

Cells were typically grown in large batches and the resulting cells were stored at -80 °C. The pellets could then be removed from the freezer for purification when a fresh protein sample was required without the need to express the protein anew each time. For each 1L cell culture the cells were harvested and stored as a 15 ml suspension of cell pellet in a specified buffer.

2.6.1 Cell Lysis

Sonication

A 15 ml cell pellet was defrosted, and 200 μl of a Roche cOmplete EDTA-free protease inhibitor tablet dissolved in 1 ml dH_2O was added. The falcon tube was placed into a beaker of ice, and the cells were sonicated at full power in 30 s cycles for 2 min. The cells were then centrifuged at 50000 xg in a JA 25.50 rotor for 30 min at 4°C. The supernatant was saved and the pellet was discarded.

All protein samples other than UBE3A were supplemented with DNase along with the protease inhibitor prior to sonication. For UBE3A samples, the DNase was omitted and the clarified lysate was subjected to syringe filtration through a 0.45 μm membrane to remove the DNA component prior to subsequent purification steps.

Cell Press

For larger scale cell preparations the cells were lysed in a cell disruptor (CF2 model from Constant Systems Ltd.) rather than a sonicator. For this, the cells were diluted further in lysis buffer to approximately 50ml per 10g of cell pellet, and protease inhibitor and DNAse (where applicable) was added. The cells were further diluted with lysis buffer during the run if the viscosity appeared too much for the machine. The cell disruptor was pre-chilled to 4°C, and washed through with 100 ml dH_2O at 15 kpsi. The diluted *E. coli* solution was then added to the chamber, pumped through at 25-28 kpsi, and collected

in a glass bottle on ice. The samples were then passed back through the cell disruptor, again at 25-28 kpsi, in order to ensure full cell lysis. The cell disruptor was washed with 100 ml dH_2O, then 150 ml 20% ethanol, and then 200 ml RBS and a final 100 ml dH_2O all at 40 kpsi after use.

2.6.2 HisTrap Purification

For UBE3A purifications, a 5 ml HisTrap column was equilibrated in 10 CV binding buffer (buffer A), and then the clarified lysate was loaded at 2 ml/min on a peristaltic pump. The column was then washed with 10 CV buffer A'1 (100 mM Tris, 600 mM NaCl, 60 mM imidazole, pH 8), and then 50 CV buffer A at 5 ml/min. The protein was eluted in 3 CV buffer B (100 mM Tris, 600 mM NaCl, 600 mM imidazole, pH 8). The first 5 ml of the flow through was collected in one tube, and the remainder in another. The first 5 ml of the buffer A'1 wash was then collected, and then the next 10 ml, and the remaining 35 ml was collected separately. The 50 CV buffer A wash was collected, and then the 15 ml elution was collected separately.

For all other His-tagged proteins, the column was equilibrated and the lysate was loaded as with above, but the column was then washed with 10 CV buffer A and then 10 CV buffer A'1 , before elution with 3 CV buffer B.

2.6.3 MBPTrap Purification

A 5 ml MBPTrap column was first equilibrated in 10 CV binding buffer (buffer C), and then clarified lysate was loaded at 5 ml/min. The column was washed with 10 CV buffer C, and then the protein was eluted in 15 ml maltose elution buffer (50 mM Tris, 150 mM NaCl, 20 mM Maltose, pH 8). The first 5 ml of the flow through was collected, and then the remainder of the flow through was collected separately. The first 5 ml, the next 10 ml, and then the remaining 35 ml of wash were collected individually as well. The elution was collected separately.

2.6.4 TEV/3C Digest

The 15 ml elution from either a HisTrap or MBPTrap purification was added to a ~20 cm piece of dialysis tubing that had been soaked in dialysis buffer for ~5-10 min, along with 1 mM DTT, 1 mM EDTA at pH 8, and 100 µl of either TEV or 3C protease at 4 mg/ml. The tubing was closed at both ends with clips and added to 2L buffer C and 5mM β-mercaptoethanol. The dialysis was left overnight, stirring slowly, at 4°C.

2.6.5 Reverse HisTrap Purification

A 5 ml HisTrap column is first equilibrated in 10 CV buffer C, and then the sample from the TEV digest is loaded onto the column at 5 ml/min on the peristaltic pump. The column is washed with 10 CV buffer C, and then 3 CV buffer B. The first 5 ml of the flow through is collected, and then the remaining 10 ml is collected separately. Then the first 5 ml, the next 10 ml, and the remaining 35 ml of the wash are all collected individually. The final 15 ml elution fraction is collected separately.

2.6.6 Anion Exchange Chromatography

A 1 ml HiTrap Capto Q anion exchange column was equilibrated in 10 CV buffer C. The sample was loaded onto the column at 1 ml/min on a peristaltic pump, and the column was transferred to the AKTA purifier system. The column was washed with 10 CV buffer C, and then a gradient elution was run over 20 CV to a final salt concentration of 1 M NaCl. The column was washed through with 5 CV 1 M NaCl, and then re-equilibrated with 10 CV buffer C. 0.5 ml fractions were collected for the entire 20 CV gradient and the first 2CV of the 1 M NaCl wash segment.

2.6.7 Size Exclusion Chromatography

A Superdex 200 10/300 increase column, or a Superdex 75 10/300 increase column, was loaded onto the AKTA purifier system and equilibrated in 30 ml buffer C at 0.5 ml/min, with a pressure alarm set for 4 mPa maximum pressure. The 500 µl injection loop was washed with a 5 ml syringe full of buffer C, and then a ~300 µl concentrated sample was added to the injection loop ensuring no air bubbles were present. The syringe was left in place for the run to prevent air entering the system. Sample was eluted over 1.2 CV with a 0.5 ml/min flow rate and collected in either 500 µl or 200 µl fractions. The column was then washed in 30 ml dH2O and 20% ethanol for storage.

2.6.8 Small-Scale Gravity Affinity Purifications

The mammalian cell expression cultures were purified using a small-scale high-throughput method that involved gravity purification in 96-well block. The mammalian cells were grown as 3 ml cultures in a 24 deep well block by members of the Protein Production UK (PPUK) group at Diamond and were provided to us as cell pellets. Each pellet was resuspended in 1 ml lysis buffer with protease inhibitor. The cells were lysed by sonication using the 24-well probe at 60% amplitude for 2 min in 5s cycles on ice. 950 µl of each lysate were transferred to individual wells of a 96 deep well block, and the block was centrifuged at 3500 xg for 30 min at 4^{o}C.

The Talon and Streptactin XT resins were equilibrated before by adding the required amount of resin suspension to a falcon tube and centrifuging at 300 xg for 2 min at 4^{o}C. The ethanol storage solution was removed, the resin was resuspended in 15 ml dH_2O, and the mixture was centrifuged at 300 xg for 2 min to separate the resin. This was repeated twice, followed by three more times using lysis buffer, to ensure equilibration of the resin in the appropriate buffer. The resin was centrifuged a final time and resuspended in an equal volume of lysis buffer.

For the GFP-nanobody purification, an aliquot of streptactin resin was first equilibrated as described above, and then an aliquot of purified GFP-nanobody, provided by the MPL group, was bound by passing the sample over the resin several times in succession.

100 µl of each resin was transferred to each well of a 96 well block alongside the clarified lysates. A foil seal was added and the lysate and resin was left to incubate at 4°C for 1 hour on a platform shaker at 450 rpm. The incubated resin and lysate mixtures were added to a 96 well filter plate placed above another 96 deep well block, and the flow through for sample was collected. 100 µl wash buffer was added to each well, and this time the filter plate and deep well block setup was centrifuged at 300 xg for 1 min. This was repeated a total of three times, and then it was centrifuged a final time at 500 xg for 3 min to remove non-specifically bound species and any remaining wash buffer. The filter block was then placed over a 96 well microtitre plate, 50 µl elution buffer was added to each well, and the plate was sealed and left to incubate at 4°C for 20 min on a platform shaker at 450 rpm. The seal was removed and the filter block and microtitre plate setup was centrifuged at 500 xg for 3 min at 4°C to elute any remaining protein. Samples of the initial flow through, the wash steps, and the elution were all analysed using SDS-PAGE and InstantBlue dye.

The lysis buffer for the small-scale purifications was 20 mM HEPES, 500 mM NaCl, 10 mM Imidazole, pH 8, 10% glycerol. Protease inhibitor was added as 1 cOmplete tablet (EDTA-free) per 50 ml buffer.

The wash buffer was 20 mM HEPES, 500 mM NaCl, 10% glycerol, pH 8 with 25 mM Imidazole for the Talon buffer or no imidazole for the streptactin or GFP-nanobody resins.

The elution buffer used was 20 mM HEPES, 500 mM NaCl, 10 % glycerol, pH8, with 500 mM imidazole for the talon purification and 50 mM biotin or the streptactin and nanobody purifications.

2.6.9 Strep-tag Gravity Purification

1 ml resin was equilibrated by centrifuging 2ml of resin slurry at 300 xg for 2 min at 4°C. The ethanol solution was removed with a pipette and the resin was resuspended in 15 ml wash buffer. This was centrifuged again at 300 xg for 2 min at 4 °C, the wash buffer was removed, and the resin was again suspended in 15 ml wash buffer. This was repeated a third time, and then the resin was resuspended in 1ml wash buffer and transferred to a gravity flow column. The sample to be purified was added to the resin in the gravity flow column, and the flow through was collected. The column was washed with 5 ml wash buffer, followed by 10 ml wash buffer, and then 3 ml elution buffer was added to the column. The elution fraction was collected, and all samples were subjected to SDS-PAGE and an InstantBlue stain to visualise the results.

The wash buffer was comprised of 20 mM HEPES, 500 mM NaCl, 10% glycerol, pH 8.

The elution buffer was comprised of 20 mM HEPES, 500 mM NaCl, 10% glycerol, and 50 mM Biotin, pH 8.

2.7 Protein Gel Electrophoresis

2.7.1 SDS-PAGE

Tris-Glycine buffer system

SDS-PAGE samples were prepared by mixing 15 µl of a protein sample with 5 µl of 4x LDS sample dye, and heating the mixture for ~2 min at 95°C. Running buffer was comprised of 25 mM tris, 200 mM glycine, and 0.1% SDS w/v. Preprepared 4-20% gradient TGX (tris-glycine) gels were used (BioRad, UK), containing 10 wells per gel, each capable of holding up to 30 µl of sample. 15 µl of sample was typically loaded, along with 5 µl of a coloured, broad protein range ladder. Multiple ladders were used throughout the project due to availability, all covering a protein range between 10 – 250 kDa. Gels were run at 300 V for 15 min, and then stained with an InstantBlue dye for 15 min.

Tris-Tricine/Tris-Acetate buffer system

Tris-Acetate gels were hand-poured using different acrylamide percentages as necessary. The gels used during this project were 9%, 7%, 5%, or 3% acrylamide. Two different sizes of gels were also used for this gel type, in an attempt to improve the resolution at the highest molecular weight region. For either size, the gel solution was made using 30% acrylamide solution to the percentage required, Tris-acetate pH 8 to a final concentration of 200 mM, and Milli-Q purified water to make up the volume. TEMED and APS were added immediately before pouring the gel, to 0.12% v/v and 0.042% final concentrations.

The running buffer for the Tris-Acetate gels was a Tris-Tricine buffer, rather than the standard Tris-Glycine system. Tris-tricine buffers are usually associated with gel used to resolve small peptides, but along with the tris-acetate gels it should enable resolution across a wider range of proteins at both ends of the molecular weight scale (Cubillos-Rojas *et al.,* 2010). The tris-tricine running buffer consisted of 50 mM Tris, 50 mM Tricine, and 0.1 % SDS. The low pH of the tricine meant that the solution was naturally at pH 8.2 with no further need for alterations of the pH with either HCl or acetic acid.

The sample buffer for Tris-Acetate gels was prepared so that when in it's 1x form mixed with the sample, the final concentrations were 250 mM Tris-HCl pH 8.5, 2% w/v SDS, 100 mM DTT, 0.4 mM EDTA, 10 % v/v glycerol, and Brilliant Blue R. Samples in this buffer were heated to 70°C for 10 min prior to loading on the gels.

Small gels were prepared using the BioRad Mini Protean Tetra system, which results in gels 7 cm x 8 cm in size, with optional thicknesses of 1 mm or 1.5 mm. The gels used in this project were typically 1 mm, unless otherwise specified. These gels were made using a single acrylamide percentage throughout, and no stacking gel was used as the protein that I hoped to observe was a higher molecular weight than the stacking gel would be useful for. The gels were loaded with sample in a 5X SDS-containing dye solution

and a prestained PrecisionPlus protein ladder and run at 130 V for ~75 min, until the top band of the protein ladder was almost halfway down the gel. The gels were imaged using a UV filter to detect any GFP that had managed to remain folded and fluorescing after the run, and then the gels were stained in the Coomassie-based dye, InstantBlue, for 30 min. They were soaked in dH_2O until imaging to remove any residual background staining, although destaining is not typically required for this gel stain. The stained gels were also imaged without the UV filter to produce the images seen in this report.

Large gels were prepared using the BioRad Protean II xi system, resulting in gels 16 cm x 20 cm in size. The thickness of these gels could also be set to 1 mm or 1.5 mm. As with the smaller gels, the acrylamide percentage remained fixed across each gel, and no stacking gel was used. The minimum acrylamide percentage used for the large gels was 5%, as the 3% gels were too difficult to handle even in the smaller size, and the larger size relative to the thickness of the larger gels made even the higher percentage gels more difficult to handle. The gels were run at 100 V in the cold room for ~2 hr, until the top bad of the prestained marker had migrated into the gel sufficiently. The gels were imaged using a UV filter before overnight staining in a typical Coomassie stain, comprised of 10% methanol, 10% acetic acid, and Coomassie R-250 dye. The stained gels were then destained in a destain solution, made of 10% methanol and 10% acetic acid with no dye added, for 2 hr – overnight, before images were taken.

2.7.2 Native PAGE

Native PAGE gels are similar to SDS-PAGE gels in that they separate proteins in a sample across a polyacrylamide gel through electrophoresis, but whereas SDS-PAGE involves denaturing the protein sample before application through the use of SDS in the sample buffer and heating the samples prior to loading, native PAGE allows the protein to retain their folded forms and instead separates proteins based on their isoelectric points. Theoretically, native PAGE gels can allow resolution of larger molecular weight species than SDS-PAGE gels (Roelofs *et al.,* 2018), and they could also allow for the retention of GFP-fluorescence.

Commercial 4-20% TGX gels from Bio-Rad were run using a native tris-glycine running buffer. The running buffer for native tris-glycine gels was comprised of 25 mM tris, and 200 mM glycine, and the 2x sample buffer was comprised of 500 mM Tris-HCl pH 8.5, 200 mM DTT, 0.8 mM EDTA, 20% glycerol, and a small amount of Brilliant Blue R dye. Native PAGE gels were run in the cold room and at a lower voltage so that the heat of the electrophoresis process did not denature the proteins in the samples. For the commercial 4-20% gels native PAGE gels were run at 150 V for 1 h.

The Tris-Acetate/Tris-Tricine native gels involved preparing gels as described in 2.7.2 but ommitting SDS in all of the buffers. The gels were also run in the

cold room and at a lower voltage, so the small size gels were run at 100 V while the larger gels were run at a constant current of 35 mA for 4-5 h.

All native gels were imaged first using a UV filter, and then stained with either InstantBlue or a standard Coomassie solution depending on the size of the gel before imaging.

2.7.3 BN-PAGE

Blue-Native PAGE is a technique that uses Coomassie Blue in the sample dye and running buffer of a native PAGE gel to increase the resolution relative to a standard native PAGE gel, while retaining the folded structure of each sample. The Coomassie Blue particles bind to the proteins, as they do when Coomassie Blue is used as a post-electrophoresis gel stain, but in BN-PAGE gels the slight negative charge of the Coomassie particles bound to the protein samples allows more effective separation of different species based on their size (Schägger, 2001).

In this project, the BN-PAGE method was attempted using a commercial Invitrogen nativePAGE 4-16% gel, run using the Invitrogen XCell SureLock Mini equipment. The sample dye used was comprised of 50 mM BisTris, 6N HCl, 50 mM NaCl, 10% glycerol, and 0.001% Ponceau S. The pH of the solution was 7.2, with no further pH adjustment necessary. Different running buffers were used for the anode (outer chamber) and cathode (inner chamber) of the gel electrophoresis arrangement. For the anode buffer, the standard tris-glycine native buffer was used, 25 mM Tris, 200 mM glycine. For the cathode buffer, the tris-glycine buffer was supplemented with 20x nativePAGE cathode additive, which is simply a solution of 4% w/v Coomassie Blue G-250 in water. The wells of the gel were rinsed with the cathode buffer, samples were loaded into the wells, and then the inner chamber was filled with cathode buffer. The outer chamber was then filled with anode buffer, and the gel was run in the cold room at 150 V for 1 h followed by a further 90-120 min at 250 V. The gel was imaged using a UV source, and then destained using the 10% methanol and 10% acetic acid solution overnight.

2.7.4 Horizontal Agarose Gel Electrophoresis

Horizontal agarose gels for protein electrophoresis were created by making a solution of 1% agarose in tris-glycine SDS running buffer solution, and then setting the gel in a horizontal gel casting system as used typically for agarose gels. Once set, the gel was placed in the centre of the horizontal gel electrophoresis chamber, and the chamber was filled with 1x running buffer. The samples were mixed with the standard SDS-containing sample dye and heated at 95 ^{o}C prior to the run to denature the proteins. The gels were run at 100 V for ~2 h, until the prestained ladder had sufficiently migrated through the gel

Horizontal agarose gels were tried using a tris-glycine buffer under denaturing conditions, as described above, as well as with tris-glycine under native

conditions, and also with a tris-borate buffer under both native and denaturing conditions. The tris-borate running buffer consisted of 90 mM Tris and 90 mM boric acid, with 0.1% SDS added for denaturing conditions but left out for native gels. For native gels using either the tris-glyicne or tris-borate buffers, samples were mixed with the native sample dye described in 2.7.2 and were not heated prior to loading.

Following the electrophoresis run, each horizontal agarose gel was imaged initially using a UV source, then stained in InstantBlue for 3 h to overnight and re-imaged using a standard light source.

2.7.5 Vertical Agarose Gel Electrophoresis

Vertical agarose gels were used to separate protein samples by size, following the protocol from (Greaser and Warren, 2012). These gels were made and run using the BioRad Mini Protean Tetra setup, resulting in gels 7cm x 8 cm in size. The first step of pouring these gels, other than assembling the cassette, was to pour a small section of acrylamide gel, 1 cm high. This was composed of a final concentration of 10 % acrylamide, 75 mM Tris-HCl pH 9.3, 5% glycerol, 0.2824 % v/v 10% APS, and 0.153% TEMED. ~1ml dH_2O was added to the top of this 1 cm gel solution to create the barrier from the air to allow it to set. The next step of the gel was the agarose portion. This was made of 1% agarose, 30% glycerol, 50 mM Tris, 384 mM glycine, and 0.1 % SDS. The mixture was made up in a beaker, saran wrap was added to cover the top, and holes were poked in the wrap. The solution was heated in a microwave in 10 – 30 s pulses to dissolve the agarose into the solution. The layer of dH_2O was removed from the top of the acrylamide plug with a KimWipe tissue, and the agarose solution was poured into the gel cassette up to ~0.5 cm from the top. The comb was soaked in paraffin oil prior to allow for easier removal, and then placed into the top of the gel cassette, ensuring that no air bubbles formed between the gel solution and the comb. Once set, the top of the gel was rested over a heat block at 60^{o}C for 5 min, and the comb was carefully removed. The wells were rinsed out thoroughly with dH_2O before use. The acrylamide block was essential to keep the gel in place during the run, or the agarose gel would float out of the glass frame. However, the comb must be removed very carefully after setting the gel, or the agarose gel would stick to the comb and separate from the acrylamide portion. This introduced small air bubbles between the two gel layers that could not be removed by simply pushing the gel back in place and led to extremely high resistance during the run. These gels were trialled using both denaturing conditions and native conditions. The denaturing runs used the buffers listed above, while the native version used the same buffers but without SDS

The vertical agarose gels were run using the standard tris-glycine running buffers, comprised of 25 mM Tris and 200 mM glycine, with 0.1% SDS included in the denaturing runs but not the native runs. For denaturing gels, 10 mM βME was also added to the upper chamber buffer but not the lower

chamber buffer to improve the resolution of the run. Both denaturing and native gels were run in the cold room at 200V for several hours, until the top of the ladder could be seen to have migrated into the gel sufficiently. Samples were dissolved in the same sample buffers used for standard tris-glycine SDS-PAGE and native PAGE experiments.

Following each run, the gels were imaged using a UV filter before being stained in InstantBlue stain for 15 – 30 min and imaged again without UV.

2.8 Ubiquitination Assay

2.8.1 *In vitro* Assay

A 100 µl reaction was prepared, containing final concentrations of 50 mM Tris, 5 mM ATP, 50 mM MgCl2, 10 µg Ub, 8 µg His6-E1, 4 µg UbcH7, and 16 µg UBE3A. Control reactions of 50 µl were also prepared, each lacking a single enzyme from the E1-E2-E3 cascade. In each control the enzyme was replaced with dH2O to maintain the reaction volume and buffer concentrations. The reaction and the control reactions were all left at 37oC for 90 min, and time points were taken from the 100 µl reaction after 5 min, 15 min, 30 min, 60 min, and 90 min. At each time point 20 µl was removed and mixed with SDS sample dye to quench the reaction. After 90 min, all assay samples and the controls were subjected to an SDS-PAGE gel.

2.8.2 Western Blot

A Western blot was carried out using a Vu-1 anti-ubiquitin antibody against the time points of several *in vitro* ubiquitination assays. For each Western blot, the assay samples were mixed in a 1:1 ratio with a 2x SDS sample dye and heated to 95^{o}C for 2 min. The heated, dyed samples were then subjected to SDS-PAGE using a 4-20% tris-glycine acrylamide gel run at 300V for 15 min.

Following SDS-PAGE, the protein bands were transferred from the acrylamide gel onto a nitrocellulose membrane through the use of the iBlot 2 system from ThermoFisher. The iBlot transfer stack was separated into its bottom stack and top stack constituents, and the acrylamide gel was placed onto the nitrocellulose membrane of the bottom stack. The entire bottom stack, including the plastic tray, was then placed into the iBlot machine, ensuring that the electrical contacts match up. A piece of filter paper was pre-soaked in dH_2O for several minutes before being applied on top of the acrylamide gel, the roller was used to remove any air bubbles, and the top stack of the transfer stack was placed on top. A plastic roller was used to remove any air bubbles from the stack, the iBlot absorbant pad was placed on top with the electrical contact lined up with the electrical contact in the iBlot machine, the roller was used to remove any air bubbles for a final time, and the lid was closed. The transfer was run at 10V for 7 min, and then the transfer stack was disassembled, the nitrocellulose membrane was removed, and the remainder of the stack was discarded.

For the remainder of the Western Blot procedure, the iBind system from ThermoFisher was used. A stock of 1X iBind solution was made by mixing 6 ml 5X iBind buffer, 300 µl 100X additive, and 23.7 ml dH_2O. The nitrocellulose membrane was left to soak in 5 ml iBind solution for 5 min immediately following the protein transfer. The iBind card was placed into the iBind cassette with the stack region at the front, furthest from the wells. 5 ml 1X iBind solution was pipetted evenly across the flow region of the iBind card, that's everywhere other than the stack, and a further 1ml was pooled in the membrane region immediately above the stack. After the 5 min soak, the nitrocellulose membrane was placed face down onto the iBind card with the lowest molecular weight region closest to the stack, and the roller was used to remove any air bubbles. The lid was closed and the western blot solutions were added to the wells at the top. 2 ml primary antibody solution, containing the VU1 anti-Ub antibody diluted to 1/1000 in semi-skimmed milk powder prepared with TBST, was added to well 1. 2 ml 1X iBind solution was added to well 2. 2 ml secondary antibody solution, which contained an anti-mouse antibody diluted to 1/5000 in TBST-milk, was added to well, 3, and 6 ml 1X iBind solution was added to well 4. The well cover was put in place and the system was left overnight with no further interference.

To visualise the results of the Western blot, the membrane was removed from the iBind system, 500 µl each of reagent 1 and reagent 2 from the Pierce ECL Western blotting substrate kit (Thermo Fisher) were mixed, and the solution was pipetted evenly over the membrane in a shallow dish. The dish was covered with foil to protect the membrane from light and the membrane was imaged after ~1 minute using a chemidoc imager.

2.8.3 Densitometry Analysis

Densitometry of the UBE3A assay gels was conducted using the FIJI image processing package (Schindelin *et al.,* 2012). Each gel image was opened in FIJI, and a rectangle was drawn around the first lane. 'Ctrl + 1' was pressed to set it as the first lane, and the rectangle was dragged to cover the second lane. 'Ctrl + 2' was then pressed to set this lane. This process was repeated until all of the lanes had been defined. After the final lane, 'ctrl + 3' was pressed to open a new window containing a plot for each lane, with any bands represented as peaks in each plot. The line tool was used to define each peak, and then the 'magic wand' tool was used to select each defined area and generate a list of values for the total area of each defined region. These area values were plotted using the OriginPro software (OriginPro, 2021) and fitted to an exponential model to generate the graphs shown in chapter 5.

2.9 Biophysical Techniques

2.9.1 SV-AUC

Sedimentation Velocity-Analytical Ultracentrifugation (SV-AUC) uses high speed centrifugation and optical tracking to measure the time it takes for a macromolecule sample to sediment. It can be used on proteins, lipids, or

peptides, but for this project it was used solely with proteins and protein complex samples. They key principle behind SV-AUC is that molecules with different masses and different shapes will sediment at different speeds, so by tracking the speed of sedimentation you can get an idea of the size of any species within a sample (Svedberg and Fåhraeus, 1926). The rate at which a species sediments divided by the rotor speed is referred to as its sedimentation coefficient (S), and the relationship between the mass of the sample and the sedimentation coefficient can be described by:

$$S = \frac{v}{\omega^2 r} = \frac{M(1 - \bar{v}\,\rho_0)}{N_A f}$$

Where S is the sedimentation coefficient, v is the boundary terminal velocity, ω is the angular velocity in radians per second, r is the radius from the centre of rotation, M is the molecular mass, $\bar{v}$ is the partial specific volume, ρ_0 is the solvent density, N_A is avogadro's number, and f is the frictional coefficient. The partial specific volume of a sample is defined as the change in volume of a solution when a set amount of solute is added, and can be predicted from the protein sequence. The solvent density can be measured using a DMA 5000 density meter from Anton-Paar. The frictional coefficient relates to the shape of the molecule, a less spherical shape will have a higher frictional coefficient. This can be estimated during the data processing, but an average value for proteins is ~1.2.

As the rate of sedimentation will be affected by the properties of the buffer the sample is in as well as its hydrodynamic properties, the sedimentation coefficient is corrected to the expected observation under standard conditions, which is defined as water at 20°C, in order to compare different species across different buffer systems. The standardised sedimentation coefficient can be calculated as:

$$S_{20,w} = \frac{(1 - \bar{v}\rho)_{20,w}}{(1 - \bar{v}\rho)_{T,B}} \times \frac{\eta_{T,B}}{\eta_{20,w}} \times S_{T,B}$$

Where the subscript (20,w) represents the value for water at 20°C, and the subscript (T,B) represents the value for the actual buffer at the actual temperature of the experiment. $\bar{v}$ represents the partial specific volume, ρ represents the density of the solvent, and η represents the viscosity measurement.

The core theory behind AUC data analysis is contained in the Lamm equation, which describes the shape of the boundaries formed through sedimentation:

$$\frac{dc}{dt} = D\left[\left(\frac{d^2c}{dr^2}\right) + \frac{1}{r}\left(\frac{dc}{dr}\right)\right] - s\omega^2\left[r\left(\frac{dc}{dr}\right) + 2c\right]$$

Where D is the diffusion coefficient, c is the concentration, r is the radial position, t is time, and ω is the angular velocity (Lamm, 1929). When SV-AUC data can be processed in the SEDFIT software to measure determine the

continuous distribution (c(S)) fit of the data. Based on the Lamm equation, the c(s) analysis attempts to account for diffusion of samples within the AUC cells and so results is sharper peaks than other processing approaches and the ability to estimate the average molar mass of species. The diffusion coefficient of a sample is related to its frictional ratio, which is the ratio between the observed frictional coefficient and the expected frictional coefficient of a perfect sphere with the same mass (Brown and Schuck*,* 2006).

In this project, SV-AUC was used for individual protein purification samples to investigate the distribution of multimeric states present after purification, and also to identify a stoichiometric ratio of protein complexes.

Sedimentation velocity experiments were performed in standard 2 sector AUC cells. 396 µl of protein sample was loaded into the sample side, and 400 µL of buffer in the other channel so that a meniscus could be observed in the collected data. The small holes in the cell that the solutions are loaded through were covered by a small piece of plastic, the hole seal, and then a small screw. The rotor was then loaded into a Beckman Coulter Optima AUC instrument (Beckman Coulter, USA), the centrifuge was set to a speed of 40,000 rpm and a set temperature of 4°C, with both absorbance and interference scans taken. AUC data sets were analysed and partial specific volumes were calculated using the SEDFIT software (Schuck, 2000; www.analyticalultracentrifuge.com), and sedimentation coefficient distributions were determined using the c(s) method (Chaturvedi *et al.,* 2018).

2.9.2 ITC

Isothermal titration calorimetry (ITC) uses the thermal change of a binding reaction to determine the thermodynamic properties of an interaction between two species. It involves two cells, one to act as a reference and another in which the components are mixed. All reactions are either exothermic or endothermic, so heat is either produced or absorbed by the reaction. Both cells start off at the same temperature, and then one component is slowly added to the other in the sample cell, causing a change in the temperature of that cell. The energy of the reaction is then measured by the feedback current required to keep the sample cell at the same temperature as the reference cell. From the isotherm formed by a series of injections I can derive thermodynamic parameters (Bundle and Sigurskjold, 1994; Lewis and Murphy, 2004; Velazquez-Campoy *et al*., 2015). The first of these, enthalpy (ΔH^0) is derived from the total heat of the reaction divided by the number of moles of protein. The second, the Gibbs free energy (ΔG^0) is derived from fitting the integrated heat for each injection to a binding model from which the association constant, K_a, is derived, along with n, the stoichiometry of ligand binding (Christensen *et al.,* 1966). Using equation 1:

$$\Delta G^0 = -RT\ln(K_a) \qquad (1)$$

Where *R* is the universal gas constant and *T* is the thermodynamic temperature.

The final parameter, the entropy of the reaction (ΔS^0) is derived from the second law of thermodynamics:

$$\Delta G^0 = \Delta H^0 - T\,\Delta S^0 \qquad (2)$$

The superscript ^{0} designates that the system is at equilibrium.

ITC experiments were performed using a Malvern MicroCal ITC 200 instrument. 60 µl of one protein sample at a high concentration was loaded into the syringe, making sure that the chamber filled completely with no air gaps. The cell was washed by filling and emptying with dH_2O five times, and then a final 300 µl of the second protein sample at a lower concentration was added to the cell with a Hampton syringe. For each ITC experiment, 17 injections were carried out at 25°C with a reference power of 6 and a stirring speed of 750. The first injection of each run contained 0.4 µl of sample one and was injected for a duration of 0.8 s, while each subsequent injection comprised of 2.4 µl with a duration of 4.8 s. All injections were performed with a spacing of 180 s between each for the entirety of the ITC run.

For each experiment, one run was carried out with protein sample in both the cell and syringe, and another was run with sample in the syringe but buffer in the cell. The trace for the buffer run was used as a blank to adjust the baseline of the sample run when processing the data to create the final thermogram. The data was processed initially using the NitPic software (Keller *et al.,* 2012), and then the data package was exported and further processed using either the Origin ITC 200 software or the SedPhat programme (Zhao *et al.,* 2015; www.analyticalultracentrifugation.com/sedphat).Typically, the concentration of the sample in the syringe was set to 400 µM while the sample in the cell was set at 20 µM, unless otherwise stated. Most experiments were carried out with both proteins suspended in Buffer C (50 mM Tris, 150 mM NaCl, pH 8), but samples of RLD2 were subjected to ITC in Buffer A instead (100 mM Tris, 600 mM NaCl, 40 mM Imidazole, pH 8) as the sample was unstable in the low salt buffer at the sample concentrations required.

2.9.3 Circular Dichroism

Polarisation of light occurs when light is filtered to limit the geometrical orientations of the oscillations of the electromagnetic component. The electric field is a vector field, which means it is composed of x, y, and z components. In an electromagnetic wave the z component of the electric field defines the direction of wave propagation, so orientation of the polarised wave is formed by summation of the electric field's x (horizontal) and y (vertical) components. If the x and y components are in equal magnitude and equal phase then the resulting wave is a linearly polarised wave, but if the linear and horizontal polarisation states are out of phase by 90° then the resulting electromagnetic wave becomes circularly polarised (Fig. 24).

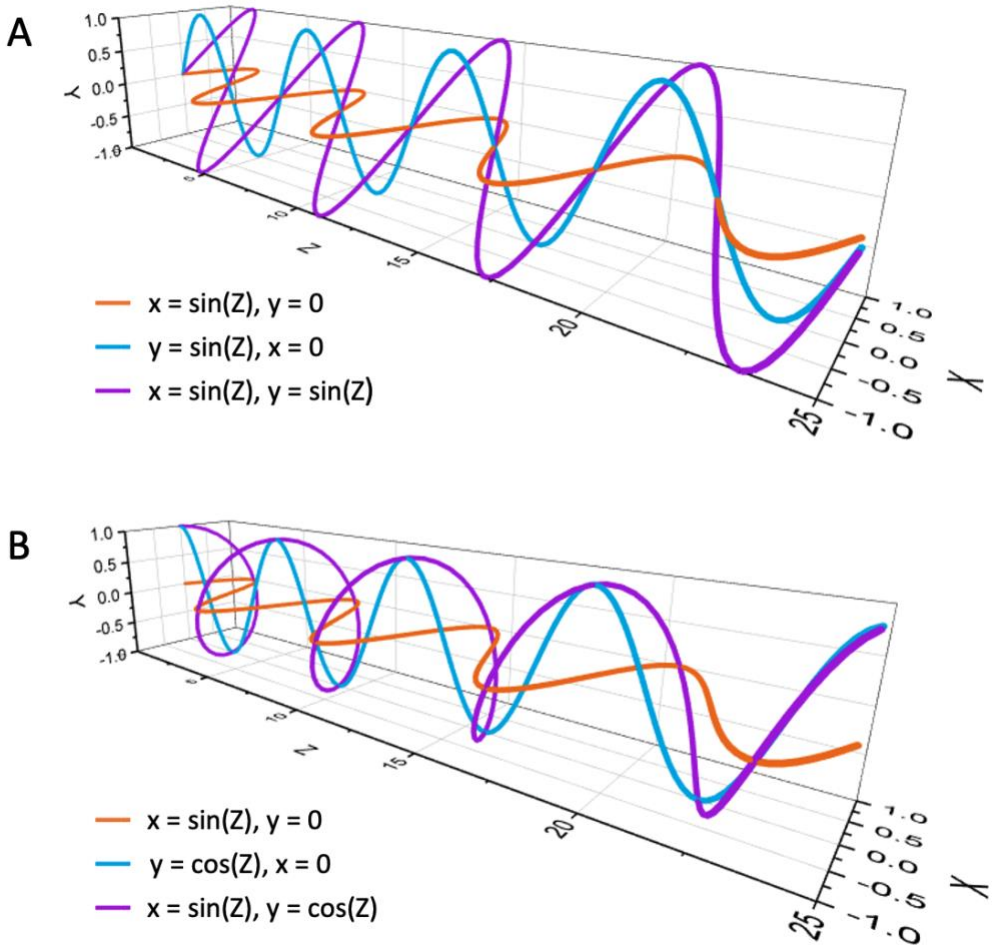


Figure 24: Circularly polarised light is formed by a 90° phase shift between the x and y components of the electric field. A) If the horizontal and vertical components are in phase then the resulting wave is linearly polarised. B) A 90° phase shift of the horizontal and vertical phases results in a circularly polarised wave.

Circularly polarised light can be left or right handed, depending on whether the electric field rotates in a left-hand sense (anti-clockwise) or a right-hand sense (clockwise) with respect to the direction of wave propagation. When a circularly polarised beam of light interacts with an asymmetric object it will absorb the left-handed and right-handed circularly polarised light differently, so circular dichroism (CD) measures the unequal absorbance of the different forms to gather information on the structure of the sample (Greenfield, 2006).

Practically, CD is used to identify changes in the secondary structure of proteins, as alpha helices and beta sheets produce distinct and identifiable spectra when subjected to circularly polarised light (Greenfield and Fasman, 1969), so any changes in the alpha-helical or beta-sheet composition of the macromolecule in question can be easily identified by any changes to these distinctive spectra (Greenfield, 2015).

Due to COVID19 restrictions, the CD experiments during this project were carried out by the beamline staff a beamline B23 at Diamond Light Source. Samples were prepared for UBE3A, RLD2, and a UBE3A+RLD2 complex as described in Chapter 3, before each sample was buffer exchanged through dialysis and the use of a spin concentrator into a 50 mM sodium phosphate buffer at pH 8. A full-wavelength scan was performed using the Nanodrop instrument to ensure that the buffer did not introduce any signal in the far or near UV range that could interfere with the experimental results. The beamline staff were responsible for running the experiments and for formulating the data, which was then used by myself to plot the graphs shows in this report.

2.9.4 Thermal Melt Measurements

Thermal melt measurements of proteins can be used to probe the stability of the sample (Senisterra *et al.,* 2012). A protein sample is heated in regularly increasing increments and optical measurements are taken continuously. When the temperature is sufficient to cause the protein to unfold, the aromatic residues within the structure are exposed, causing the absorbance measurements to increase. The greater the stability of a protein, the higher the temperature required to unfold it. Thus thermal melt curves can be used to determine the relative stability of a sample (Gao *et al.,* 2020; Senisterra *et al.,* 2012). Thermal melt measurements were used in this project to determine the stability of protein complexes, the effects of different buffers for each sample, and the effect of crosslinking on a protein complex.

Thermal melt measurements were taken using the Nanotemper Prometheus instrument (Nanotemper, USA). 5 µl of a sample is loaded into a capillary tube through capillary action, ensuring that the sample fills the tube across the scanned area without bubbles. An absorbance reading was taken without adjusting the temperature of the sample to allow optimisation of the intensity of optical measurements, and then the full thermal melt measurement scan was begun. The temperature was set to ramp up by 1^{o}C every 2 s, and measurements were taken using both absorbance and interference data. The data was plotted to show the thermal melt curve, and the curve was integrated to show the half-point of the inflection point. The temperature value at half the peak optics value is referred to at the T_m and is taken as the melting temperature of that sample.

2.9.5 Glutaraldehyde Crosslinking

Glutaraldehyde crosslinking of UBE3A was first optimised with a range of small scale trial reactions. UBE3A was purified as previously described, and then dialysed and buffer exchanged into a buffer comprised of 50 mM HEPES and 150 mM NaCl. Half of the sample was added to this buffer at pH 6, and half was at pH 8. 16 reactions were prepared with a final volume of 25 µl per reaction. Each reaction contained a different combination of pH, concentration of UBE3A, glutaraldehyde concentration, and reaction time

(table 8 section 4.5.4). Each reaction was quenched with 2.5 µl 1M Tris-HCl pH 8.5, and samples were subjected to SDS-PAGE and an InstantBlue stain to observe the extent of the crosslinking. The most promising samples were then subjected to a thermal melt analysis to confirm the stabilisation of each sample.

The UBE3A crosslinking reaction was scaled up to produce samples for further experiments. For the 'UBE3A crosslink 1' sample UBE3A was dialysed, buffer exchanged, and then diluted into HEPES buffer at pH 8 to a final concentration of 2.5 mg/ml in 14.5 ml. Glutaraldehyde was added to a final concentration of 0.025 % v/v, and the reaction was left for 10 min before quenching with 1.5 ml 1M Tris-HCl pH 8.5.

For the 'UBE3A crosslink 2' sample UBE3A was dialysed and buffer exchanged into 14.5 ml HEPES buffer at pH 8 with a final concentration of 0.25 mg/ml. Glutaraldehyde was added to 0.05 % v/v and the reaction was left for 15 min before quenching with 1.5 ml 1 M Tris pH 8.5.

For each of the crosslinked complexes, UBE3A+PSMD4 and UBE3A+RLD2, the crosslinking reactions were carried out using the same reaction conditions as the 'UBE3A crosslink 2' sample, but scaled up to reaction volumes of 34.5 ml and 28.8 ml.

Following crosslinking, each of the large-scale reactions was passed through a PD-10 desalting column to remove the deactivated glutaraldehyde and then concentrated for SEC and SDS-PAGE to determine the effects of crosslinking on each sample (see sections 4.5.4, 4.5.5, and 4.5.6).

2.10 Electron Microscopy

2.10.1 Negative Stain Sample Preparation

300 mesh Cu + carbon film grids from Agar Scientific were used for negative stain samples. Grids were glow discharged using the Quorun GloCube for 2 min, carbon side up, before use. 3 µl of sample was pipetted onto the carbon side of a grid, left for 1 min, and then removed by blotting perpendicular to the grid. 3 µl dH_2O was then pipetted onto the grid and removed with blotting paper immediately. 3 µl of a 1% uranyl acetate solution was then pipetted onto the same grid and left for 1 min before any excess was also removed with blotting paper. Prepared grids were left suspended in reverse-action tweezers for several minutes to dry, and then stored in a grid box until use. Grids were imaged on the JEOL 2100 TEM microscope at the RCaH.

2.10.2 Preparing Cryo-EM Grids

QuantiFoil R2/2 grids were used initially, QuantiFoil R1.2/1.3 grids were used for data collection once ice conditions had been optimised, and UltrAuFoil R2/2 and R1.2/1.3 grids, and Lacey-Carbon ultrathin carbon film grids were used as part of the optimisation steps.

Glow Discharged Grids

QuantiFoil grids were placed carbon side up on a parafilm wrapped glass slide and glow discharged for 45 – 60 s, immediately prior to use. UltrAuFoil grids were glow discharged, carbon side up, on a folded piece of filter paper for 1 – 2 min before use.

Graphene Oxide-DDM Coated Grids

QuantiFoil grids, either R2/2 or R1.2/1.3, were not glow discharged before addition of graphene oxide and detergent. A stock solution of 0.3 mM DDM was prepared, along with a graphene oxide solution comprised of 0.15 mM DDM and 0.01 mg/ml graphene oxide in a volume of 1 ml. A grid, suspended in fine point tweezers, was suspended in the DDM solution for 1-3 s, shaking gently, ensuring it did not touch the sides of the tube. The grid was then blotted with blotting paper applied to the back face (non-carbon side) of the grid. 5 µl of the graphene oxide-DDM solution was then pipetted onto the carbon surface of the grid, and then removed with blotting paper from the non-carbon face. The coated grids were placed onto a piece of blotting paper to dry and used within an hour.

Blotting and Plunge Freezing

Cryo-EM samples were prepared using a Vitrobot (ThermoFisher) set to 100% humidity and 4^{o}C. The sample volume was altered as part of the ice optimisation, but the optimised volume used for data collection was 2.5-3 µl per grid, pipetted directly onto the carbon side of the grid. For graphene oxide coated grids, the sample was applied using a swirling motion to apply the sample across the entire face of the grid. The grid was then blotted, and while the blot force remained constant per machine (a setting of 3 for the RCaH vitrobot, 3 for the eBIC academic user vitrobot, and 8 for the eBIC industry user vitrobot) the blot time was optimised. 3-5 s blot times were used for graphene oxide grids, and 5-7 s blots were used for glow discharged grids for optimal ice thickness. Glow discharged grids were plunged into liquid ethane at liquid nitrogen temperature immediately after blotting, while there was a 10 s wait time for the graphene oxide grids to allow dissipation of the sample across the entire grid. Once the grid was submerged in liquid ethane, the vitrobot tweezers were retained in the liquid ethane while the nitrogen and ethane container was removed from the machine, and the grid was transferred directly to a grid box held under liquid nitrogen for storage.

2.10.3 Clipping and Loading Grids

The FEI Cryo-EM microscopes, such as the Titan Krios and the Glacios, use a cartridge grid loading system, which allows loading of up to 12 grids on the microscope at once. To allow for automated handling of each grid within the microscope each grid must be clipped. This involves securing a copper support around the rim of each grid, so that grippers within the microscope can handle them without piercing the sample areas.

A clipping chamber was cooled with liquid nitrogen and clip rings were placed into a well in the copper area of the central metal platform. C-clips were

picked up at the back of the C shape using the forceps, and pushed into a clipping tool at an angle. The clipping tool was then placed with the opening on a clean flat surface and the tool was engaged, pushing the c-clip to the edge of the opening. The end of the loaded clipping tool was then submerged into the liquid nitrogen in clipping station to cool. Grid boxes containing grids to be clipped were placed into the first few wells of the central metal platform, and empty autogrid boxes were placed into the last few wells, with the lids screwed in between each well to keep them in place.

Once everything had cooled, a clip ring was placed into the notch under the cut-out section of the copper platform with the flat bottom edge down and the sides facing upwards. A grid was then transferred from the grid box into the clip ring. For TEM the orientation of the carbon side of the grid is not important. Cooled autogrid tweezers were slotted into the centre of the copper platform and used to turn the copper platform one notch, passing over the clip ring and grid and positioning a hole in the copper only slighter larger than the grid over the arrangement. A loaded and chilled clipping tool is placed into this hole, the top is pushed to eject the c-clip into the clip ring, and the empty clipping tool is warmed and dried in order to be used again. The copper platform is turned back to its original position, with the now-clipped grid sitting in the cut out section of the platform. The clipped grid is picked up using the autogrid tweezers, and it is placed into a groove in the metal between the grid-box wells to be flipped over and handled in order to confirm the integrity of the clipping. It is then transferred into an autogrid box. This is repeated for all grids, and the filled autogrid boxes are closed and returned to storage in liquid nitrogen.

Once clipped, autogrids can be loaded into the loading cassette, which is then loaded into a nanocab, which can then be placed into the microscope for docking of the grids. The clipped grids will sit in the cassette inside the microscope throughout the session, while the nanocab is removed once the grids have been docked.

2.10.4 Screening Grids on the FEI Glacios

Grids were screened on a 200 kV FEI Glacios microscope to assess the sample quality before taking grids forward for data collection. Up to 11 grids are loaded at once, and once the vacuum and temperature levels have returned to a satisfactory state, an inventory of the grids is taken to ensure that they have all been loaded correctly. Each grid is examined at a range of magnification levels, from an Atlas image of the whole grid to a high magnification Data Acquisition image within a foil hole for data collection (more detailed description is available in appendix 6). This allows thorough determination of the ice conditions, sample concentration, and particle contrast across each grid. The image acquisition settings can be altered, e.g. exposure time and defocus values, to determine optimal data collection conditions, but if the grids will be collected on a 300 kV Titan Krios microscope

then the image acquisition settings cannot be directly carried across. Following screening of all grids, suitable grids with well distributed homogeneous particles were recovered from the microscope at the end of the session to await a full data collection session.

2.10.5 Data Collection on the Titan Krios

Pre-screened grids are loaded onto a 300 kV Titan Krios microscope, and a short amount of time can be spent altering the data collection settings to optimise them. The ideal GridSquare conditions should have been identified during screening, so the GridSquares for collection can be selected from the Atlas image, and the optimal holes within each square can be selected. Ideally a histogram can be configured to select holes based on the ice gradient across each square, but this can also be performed manually. A eucentric height value is determined for each grid square, and a data acquisition template is configured to determine the location of data collection areas in the foil holes, the auto-focus area, and the periodicity of each event. The amount of images taken over a data collection session varies depending on the collection settings, but modern detectors and the AFIS software can allow acquisition of over 10,000 micrographs in a 24 hour session.

2.10.6 Data Processing

A more detailed description of the data processing stages can be found in Appendix 6.

2.10.6.1 RELION

RELION is a computer program that uses a Bayesian approach to determine a high resolution 3D structure from raw micrographs with fairly minimal user input required. The Bayesian approach allows the program to iteratively determine many of the parameters for concurrent steps in the process directly from the data used and generated by previous steps, which reduces the chance of errors introduced by inexperienced users (Scheres, 2012). It was the first cryo-EM data processing software to incorporate the gold-standard Fourier shell correlation (FSC) value as a measure of the resolution of the calculated map. The gold-standard value is calculated by splitting the data into two random halves, finding the Fourier shell correlation (FSC) between the two maps, and the point at which the FSC value = 0.143 is taken as the overall resolution. This prevents overfitting of noisy data, and allows for validation of the map generated (Rosenthal and Henderson, 2003; Scheres, 2012). RELION was initially developed solely for the reconstruction of a 3D model from 2D particle images, but now it incorporates all stages of the cryo-EM data processing pipeline up until map building (Kimanius *et al.,* 2021).

2.10.6.2 CryoSPARC

CryoSPARC is an alternative program to RELION that includes all of the processes required for full data analysis of raw cryo-EM or negative stain data. One of the key features of cryoSPARC is the increased efficiency of many of its

processes relative to RELION. The cryoSPARC software was designed for effective use on commercially available computing hardware, without the need for large computing clusters. The focus was on streamlining the algorithms involved in computational process to allow processes to run using CPU hardware rather than relying on GPU-acceleration (Punjani *et al.,* 2017).

CryoSPARC also introduced the stochastic gradient descent (SGD) method for *ab initio* model generation from curated particles, allowing a first approximation of the shape of the structure without the need for a pre-existing template model (Punjani *et al.,* 2017). The SGD method has since been incorporated into the RELION software as well (Zivanov *et al.,* 2018).

Another new implementation in cryoSPARC is the branch-and-bound method of structural refinements (Punjani *et al.,* 2017). It allows more efficient calculations of values to minimise errors in 3D reconstructions. Rather than attempting to calculate the lowest value from all possible values, it iteratively determines the lower bound of the values to reduce the range of possible results. This is repeated until the remaining subset of possible values is small enough that it become computationally effective to more accurately determine the final value with the lowest error rate.

2.10.6.3 Motion Correction

As vitrified cryo-EM grids are exposed to the electron beam over several seconds during data collection, the irradiation causes the ice to begin to melt and allow drift of species contained within it. In order to observe the high resolution features of target proteins this drift must be corrected during data processing through a process known as motion correction. Throughout this project the MotionCor2 implementation within RELION was used to do this. Modern data collection strategies include splitting each micrograph acquisition into a series of frames, resulting in a short movie for each micrograph rather than a single still image. This allows tracking of movement of objects within the ice across each frame to determine the overall motion of each micrograph, and subsequently correct for it.

2.10.6.4 Contrast Transfer Function (CTF) Estimation

The contrast transfer function (CTF) describes the Fourier transform of the point spread function of the microscope (Thon, 1966). When data is collected on a precisely aligned modern microscope, it will have a near perfect signal. However, due to the weak scattering nature of biological specimens, a near-perfect signal will result in a very poor contrast in resulting images, rendering particles indistinguishable from noise. In order to overcome this, images are taken with an imperfect signal, and this results in a point spread function where each point becomes convoluted into a larger, fuzzier object. The source of the imperfect signal is the inherent spherical aberration of the microscope, and the defocus value of the objective lens. Cryo-EM images are taken at a series of defocus values to improve the phase contrast of the images, but also

to ensure that the modulation of the CTF is different for each image so that the zero-crossings occur in different places, and data can be obtained for all points across a whole dataset. However, the point-spread function convolutes the individual particle images, so the effects must be de-convoluted in order to reconstruct the higher resolution information. This is done using a CTF estimation job. In this project, CTF estimation was carried out using either RELION's implementation of the CTFFIND-4 program (Rohou and Grigorieff, 2015), or the 'patch CTF estimation' process in CryoSPARC.

2.10.6.5 Particle Picking

RELION Reference-Based Autopicking

RELION features a reference-based autopicking function that allows identification of particles within micrographs based on a resemblance to a series of provided reference images. The reference-based picking program allows identification of particles from micrographs with a lower signal-to-noise ratio (SNR) than is required for non-template based approaches, but it also introduces template bias. If the actual particles do not sufficiently resemble the template they will not be picked, even if they are clear by eye.

RELION LoG Autopicking

An alternative to the reference-based picking within Relion is the Laplacian-of-Gaussian based picking job (Zivanov *et al.,* 2018). This uses an edge-detection program to identify particles by the step change in contrast around each object. The LoG function first applies a Gaussian filter to smoth the noise profile across the whole image, and then a Laplace operator is applied. The Laplace operator is a second order derivative function, which means that rather than measuring the change in image values (i.e. the gradient), it detects the rate of change of the change (i.e. an increase/decrease in the gradient) (Jain *et al.,* 1995; See Fig. 97 in section 7.1.2). The LoG autopicking program removes the template bias, but it requires much less noise than is acceptable in reference-based approach.

Deep Learning Methods

As well as these more basic methods of particle picking, deep learning methods can also be used to pick particles. The programs used in this project, crYOLO and TOPAZ, both implement deep learning particle picking alogorithms to allow rigid selection of optimal particles. Both were used with either a predetermined model designed to work with most protein particle presentations, a model trained directly on the real data (Wagner et al., 2019; Bepler et al., 2019). Training a model involves manually picking several micrographs to provide the program with a series of references for both true and false particle images, and then observing its attempt at picking similar particles. Settings can be altered to improve the picking results until the model is sufficient. Once a model has been chosen, the program will search through each micrograph image in turn to identify objects resembling the true particles. Deep-learning methods allow both the freedom from template bias,

and applicability to noisier images as the model is trained to distinguish between particle images and noise within the dataset before picking is attempted.

2.10.6.6 Particle Extraction

Once the particles have been picked the individual particle images were extracted from the whole micrographs. Particles were downscaled at this stage to reduce the computing cost of future processing steps enabling them to run quicker.Typically 4x downsampling can bebeneficial, but as UBE3A is a particularly small protein by usual cryo-EM standards, most of the processing was carried out with particles rescaled to only half of the original size. Once most of the processing has been carried out, the particles involved in the final model are re-extracted without any downsampling to retain as much high-resolution information as possible.

2.10.6.7 2D Classification

2D Classification of similar particles orientations was performed in RELION using a regularised maximum-likelihood based approach (Scheres, *et al.,* 2005). This involves comparing particle images to group similar particles, and averaging over each identified class. Averaging particle images together increases the SNR of each image as the noise will be more randomly distributed across particle images than the true signal areas will be.

2.10.6.8 3D Classification

3D Classification was implemented in RELION using the same maximum-likelihood method as the 2D classification job to simultaneously perform alignment and classification assignments. (Scheres *et al.,* 2007). The number of classes specified for each calculation depended on the heterogeneity of the sample, which is typically not known prior to successful completion of this job, so this parameter required a trial-and-error approach of simply running the job with different numbers of classes and seeing which produced the best results.

2.10.6.9 Ab initio Model Generation

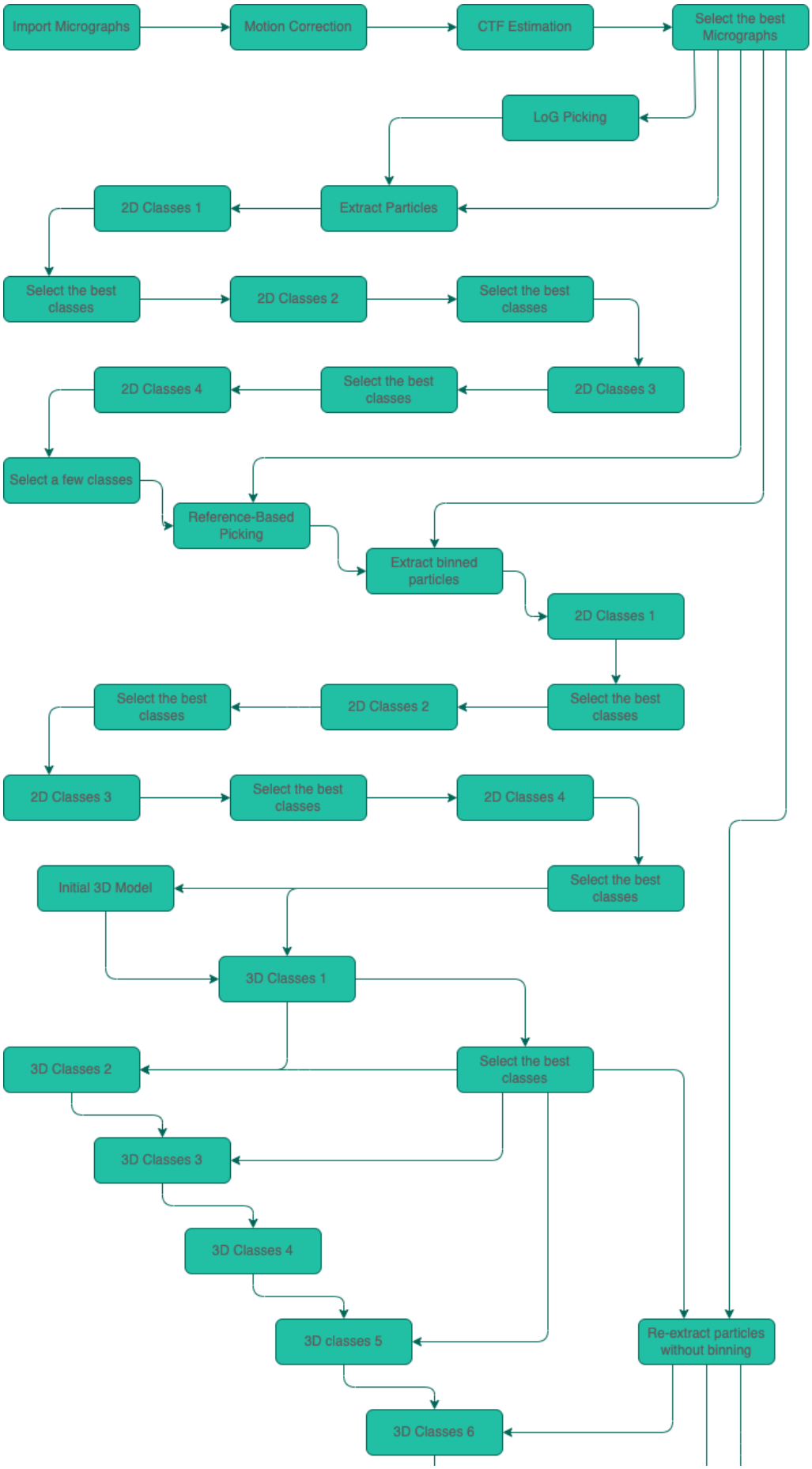
Although the 2D classification job can be run without an initial reference, the 3D classes require a more precise reference model in order to generate models with reasonable accuracy. Many proteins have homologs with already solved structures available in the PDB, and these can be used as an initial model, however, for many proteins, such as those contained in this thesis, there is no prior structural information available and *ab initio* 3D models must be generated from the data at hand (Scheres, 2016). This was carried out in either RELION or CryoSPARC for each dataset, but both use the same SGD approach to model generation.

2.10.7 UBE3A-only Pipeline

UBE3A was purified and applied to a grid and vitrified immediately following gel filtration. The grids used to collect the data presented here were

QuantiFoil R1.2/1.3 grids, glow discharged using a Harrick plasma cleaner. The sample was applied as a 2.5 µl aliquot at 0.3 mg/ml. Data was collected on the Titan Krios microscope housed at the University of Leicester, using the Falcon III camera in counting mode. A VPP was used for contrast, so the defocus was kept constant at -0.5 µm. Micrographs were collected with a dose rate of 0.7 e^-/pix/s over 35 s and 40 fractions, and a magnification level of 96kx. Data was processed solely using Relion.

The particles were picked initially using the LoG picking job in relion, then subjected to several rounds of 2D classes. The best classes for a range of orientations from the final round were then used as templates for relion's reference-based picking job. The newly picked particles were then subjected to more rounds of 2D classes before the best classes were used to generate two *ab initio* initial 3D models, one of which was used as a reference for 3D classification. Several rounds of 3D classes were run with eight classes each time, using the best class as a template for the next round. The particles were re-extracted without any binning and put into a final round of 3D classes, before a series of refinement steps to attempt to push the resolution of the model (Fig. 25).


Import Micrographs
Motion Correction
CTF Estimation
Select the best Micrographs
LoG Picking
Extract Particles
2D Classes 1
Select the best classes
2D Classes 2
Select the best classes
2D Classes 3
Select the best classes
2D Classes 4
Select a few classes
Reference-Based Picking
Extract binned particles
2D Classes 1
Select the best classes
2D Classes 2
Select the best classes
2D Classes 3
Select the best classes
2D Classes 4
Initial 3D Model
Select the best classes
3D Classes 1
3D Classes 2
Select the best classes
3D Classes 3
3D Classes 4
3D classes 5
Re-extract particles without binning
3D Classes 6

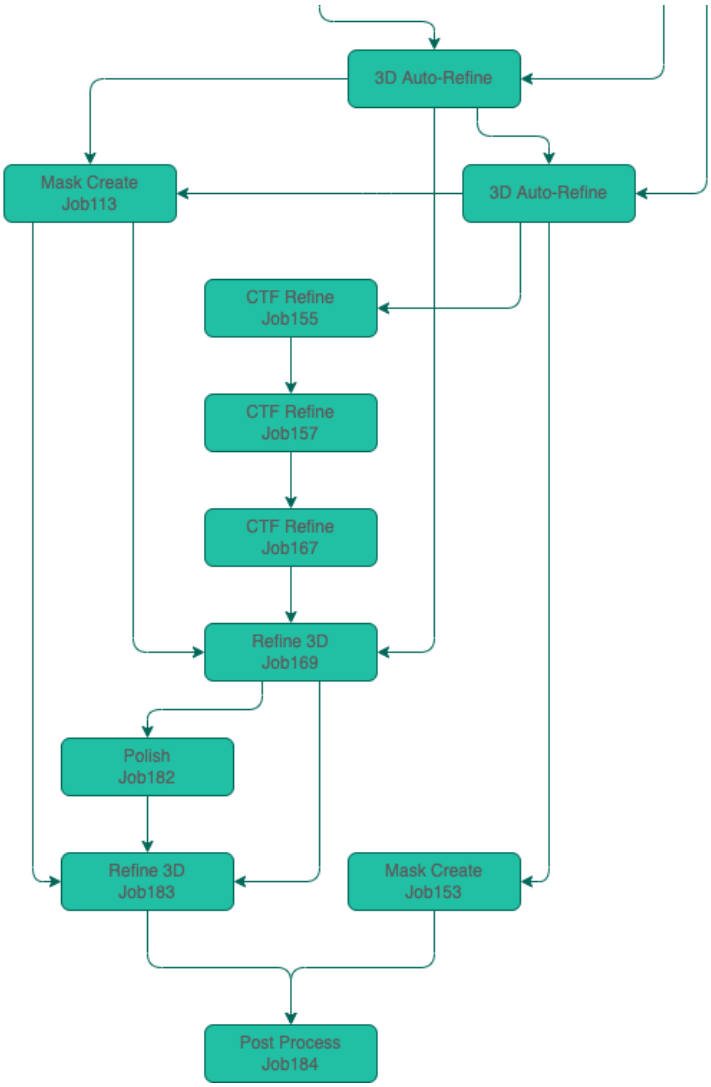


Figure 25: A flowchart for the data processing steps involved in obtaining the UBE3A structure shown in chapter 7. The process involved several iterative rounds of 2D classifications and particle picking, followed by several stages of 3D classifications in an attempt to refine the particle sets included in the model. The model was then subjected to several refinement steps, resulting in the model described in chapter 7.

The FSC plot for the final map was generated in RELION (Fig. 26).

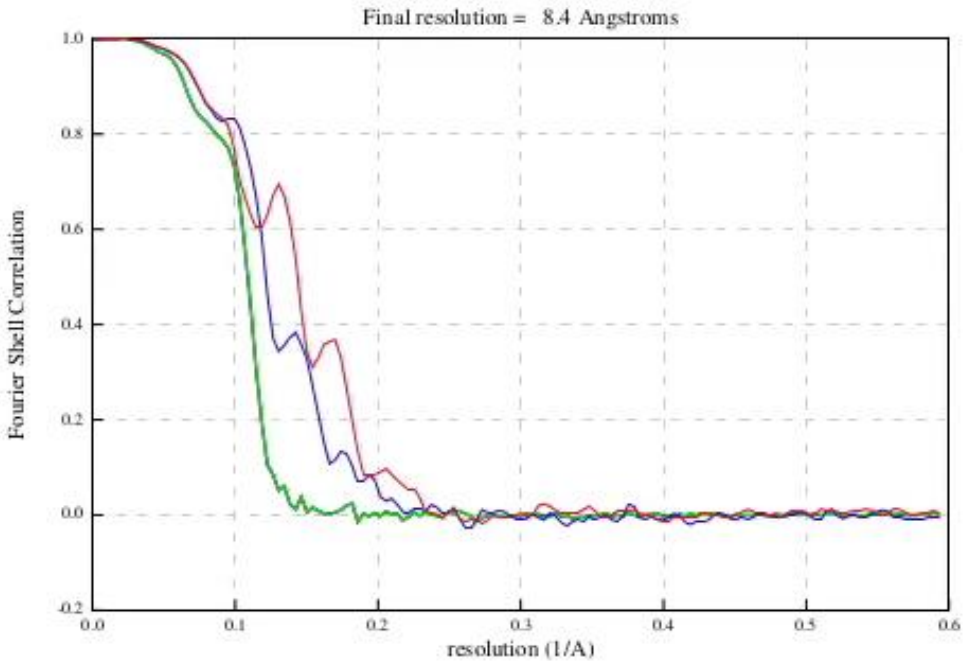


Figure 26: The FSC plot for the final model of UBE3A, generated in RELION. The red line shows the Fourier shell correlation (FSC) for the masked map at each given resolution, the blue line shows the FSCs for unmasked maps, and the green line shows the corrected FSCs for phase randomised half maps. The final resolution is taken as the point at which the FSC of the randomised half maps drops below 0.143.

2.10.8 UBE3A+PSMD4 Pipeline

The UBE3A+PSMD4 complex was purified as described in chapter 3, and then applied to QuantiFoil LaceyFoil grids with an ultrathin carbon film. The grids were glow discharged before applying sample using a Harrick plasma cleaner, and the vitrified grids were prepared using an FEI vitrobot. 2.5 µl purified sample was applied at 0.25 mg/ml. Data was collected on Krios 2 at eBIC, using the phase plate and the K3 camera. crYOLO was used for particle picking, but all other processing was carried out using relion.

The micrographs from the two datasets were motion corrected in separate Relion directories, before the motion corrected micrographs from dataset 1 were imported into the Relion directory for dataset 2. CTF estimation was then carried out separately for the two sets of micrographs, and particle picking was carried out in crYOLO using a model trained on dataset 1. The two sets of particles were extracted separately, and then the particle files were joined. Several rounds of 2D classes were carried out, and the best classes were used to generate several initial models and subsequent 3D classes in order to define a consensus model from 3D classes. The consensus model was used as a template to generate more 3D classes using all particles, resulting in a second consensus model. This was done because 2D classes can restrict the number of rare views included in the dataset when the particles are particularly small or lacking in contrast. The particles from this consensus

model were used to run a round of 2D classes without alignments, to attempt to identify and filter out any noise that has been mistaken as a particle, and then a reference-free 3D model was generated. This model was refined using Relion's auto-refine job, and then used as a reference to generate another 3D model incorporating more particles. The particles from this final model were re-extracted without binning, and the newly extracted particles were combined with the model itself to run a final round of 2D classes with a single class (Fig. 27).

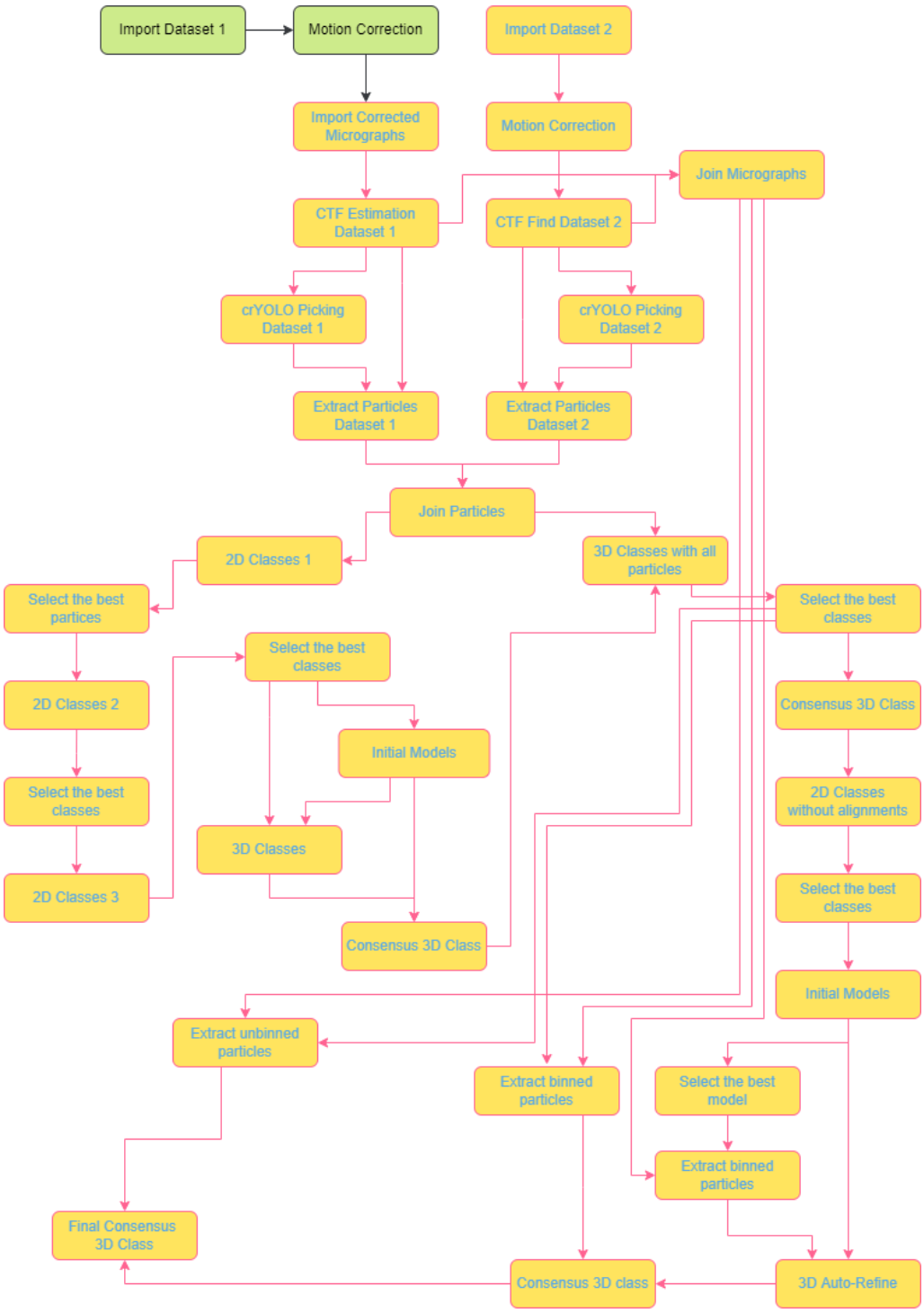


Figure 27: A flowchart for the data processing process involved in generating the final UBE3A+PSMD4 structure shown in chapter 7. The UBE3A+PSMD4 data was collected as two separate datasets and then merged after particle picking. Several rounds of 2D and 3D classes were run with varied settings to optimise the sampling, and limited refinements were carried out to confirm the limitations of the data.

The UBE3A+PSMD4 model was not processed any further than the consensus 3D class shown in section 7.2.3, so there is no corresponding FSC plot. The

model was judged to be too low resolution, too noisy, and too dubious on whether it was a true UBE3A+PSMD4 sample, so processing was stopped after the various attempts to better resolve various 3D classes were unsuccessful.

2.10.9 UBE3A+RLD2 Pipeline

A UBE3A+RLD2 complex sample was purified and crosslinked as described in sections 4.3.2 and 4.5.6. The sample that was applied to the grid was obtained after gel filtration of the crosslinked complex. QuantiFoil R1.2/1.3 grids were used, with a sample volume of 2.5 µl and a concentration of 0.3 mg/ml. The grids were glow discharged using a Quorum GloCube, and the vitrified grids were prepared on an FEI vitrobot. The data was collected on the Titan Krios at the University of Leicester, using the K3 camera in counting mode without the phase plate. The micrographs were taken at 105kx magnification, with a total dose of 50 e^-/Å over 2 s and 50 fractions. The GIF energy filter was used with a slit width of 20 eV. three images were taken per hole, and a defocus range of -1.5 µm to -3.0 µm in 0.3 µm increments. AFIS was also used during the data acquisition, to allow an acquisition speed of ~330 micrographs per hour. The data was processed initially using both Relion and CryoSparc simultaneously, but CryoSparc provided better particle picks using its 'blob picker' job and higher resolution 2D classes so the Relion processing pipeline was dropped. Micrographs were curated after motion correction and CTF estimation jobs before particle picking, and then the particles were extracted with 2x binning. Six rounds of 2D classes were carried out to generate a homogeneous set of classes, and then six rounds of *ab initio* models were run with several classes generated for each job. Once the particles were curated to form a consensus model, homogeneous refinement was performed. The particles involved in the refined model were re-extracted without binning to improve the resolution, and the homogeneous refinement was repeated. A final refinement round was carried out to see if the resolution could be improved, but the resolution was low enough that further refinements and particle polishing steps would not have improved it any further, so the processing was stopped (Fig. 28).

Figure 28: A flowchart for the data processing stages involved in generating the final UBE3A+RLD2 model shown in section 7.3.3. The data was processed in CryoSparc, usingthe CryoSprac 'Blob Picker' program to pick particles. Successive rounds of 2D classes and ab initio models were run to optimise the subset of particles included, followed by minimal refinement steps to confirm the limitations of the model.

The FSC plot for the final model was generated in CryoSPARC (Fig. 29).

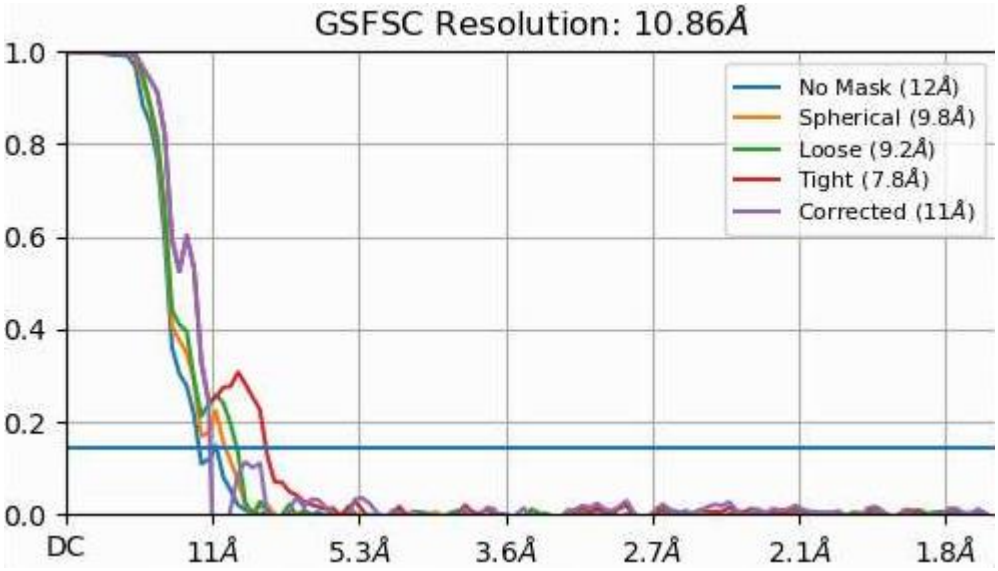


Figure 29: The FSC plot for the final model of the UBE3A+RLD2 complex, generated in CryoSPARC. The FSC values are shown on the Y axis, with reolution on the X axis. The various traces shown demonstrate the FSC traces for various masks, as listed in the legend. The final resolution (GSFSC resolution in the figure) was calculated by determining the resolution at which the FSC value drops below 0.143.

2.11 X-Ray Crystallography

2.11.1 Hampton PCT

A pre-crystallisation trial was conducted using the Hampton PCT kit, comprised of four crystallisation reagents. Reagent A1 contains 100 mM Tris, 2 M ammonium sulfate, pH 8.5; reagent B1 contains 100 mM Tris, 1 M ammonium sulfate, pH 8.5; reagent A2 is comprised of 100 mM Tris pH 8.5, 200 mM magnesium chloride hexahydrate, and 30% w/v PEG 4000; and reagent B2 is comprised of 100 mM Tris pH 8.5, 200 mM magnesium chloride hexahydrate, and 15 % w/v PEG 4000. The sample that will be subjected to the PCT is prepared in a range of concentrations, typically between 5 mg/ml and 10 mg/ml to start with. In a 24 well VDX plate with sealant, 500 µl of reagent A1 is applied to one well and 500 µl of reagent A2 is applied to another for each protein concentration sample. 0.5 µl of protein sample is applied to the centre of a square glass cover slide, and 0.5 µl reagent A1 from the well is added to the drop. The glass slide is then placed over the well, with the drop on the inside of the well and pushed to seal in place. This is repeated for reagent A2. After 30 minutes or longer the drops are viewed under a light microscope. If heavy amorphous precipitate is seen in both drops then the protein concentration is too high, if both drops are clear then the protein concentration is too low. If one drop contains a heavy amorphous precipitate but the other drop is clear then the process is repeated using the B1 and B2 reagents. If a light granular precipitate is observed in either drop, then the sample is at a suitable concentration for crystallisation and crystal screens can be prepared.

2.11.2 Setting Crystal Screens

While the PCT suggests a concentration of the sample that may be amenable to crystallisation, it does not confirm that a sample will form crystals, nor does it give an idea of the buffer conditions required for crystallisations. For this, more in-depth crystal screens are prepared. Commercial screens are sets of reagents that are mixed in various ways to form 96 different crystallisation buffers to cover a span of as many potential crystallisation conditions as possible. At the RCaH, stocks of several commercial screens are maintained by the MX group at Diamond Light Source for communal use, so crystal screens can be easily prepared by simply transferring the solutions from a 96 deep-well block to a crystal screen tray using the Hydra 96 liquid handler (Robbins Scientific). For all crystal screens conducted as part of this project crystal screens were prepared in CrystalQuickX plates, using the SG1, Morpheus, Wizard 1+2, Wizard 3+4, and JCSG+ screens from Molecular Dimensions, and the Index screen from Hampton Research. Once the screens were transferred into the large well of each position in the CrystalQuickX plates, the crystal drops were prepared using the Mosquito instrument. The concentrated protein sample that was used for the PCT was added to a 96-well conical bottom plate so that 2.5 µl sat in each well of a single column of the plate.

This 96 well plate was then placed into the mosquito chamber along with the CrystalQuickX plate containing the screen, and the program was set to transfer 200 nl each of buffer from the reservoir and protein sample into a single crystal drop position of the crystal plate, for each of the 96 conditions. The crystal screen plate was sealed with a clear vacuum seal, using the plastic tool to remove any air between the bars of the CrystalQuickX plate and the film to secure it in place. This was repeated for each crystal screen, the plates were loaded into the RockMaker hotel, and automated image acquisition of each drop was scheduled. All screens in this project were prepared and further incubated at 20^{o}C. Any drops that showed signs of containing crystals in the RockMaker images were then imaged using the UV filter, and any potential crystals with a UV signal were looped and frozen for data collection. In this project, crystal screens were prepared for a full-length UBE3A sample at 5 mg/ml using all six screens listed. A full-length UBE3A+RLD2 complex sample was subjected to all six crystal screens at 5 mg/ml. Samples of a Ufrag+RLD2 complex were subjected to all six crystal screens at 6 mg/ml. All six crystal screens were also prepared for both His-tagged and untagged RLD2 samples at 5 mg/ml and 6 mg/ml respectively.

Further optimised crystal screens were generated for His-tagged Ufrag+RLD2, cleaved Ufrag+RLD2, and full-length UBE3A+RLD2 samples. The tagged Ufrag+RLD2 optimised crystal screen was based on well A6 of the Wizard 3+4 plate, which had a buffer composition of 20% PEG 3350 and 200 mM $BaCl_2$. The screen was generated using the Scorpion instrument, resulting in conditions varying from 14-26% PEG 3350 across the numerical axis of the 96-

well block, and 0 – 300 mM $BaCl_2$ down the alphabetical axis. The optimisation screen for the cleaved Ufrag+RLD2 sample was based off well B12 of the Wizard 1+2 plate, which was comprised of 35% MPD (2-Methyl-2,4-pentanediol), 100 mM Tris pH7, and 200 mM NaCl. The optimisation screen was again created using the Scorpion instrument to create a gradient of MPD from 29 – 41 % across the numerical axis, and a pH gradient from pH 7 to pH 8 down the alphabetical axis. The concentrations of Tris and NaCl were kept constant across the grid at 100 mM and 200 mM respectively. The final optimisation screen, produced for the full-length UBE3A+RLD2 sample, was based on the B10 position of the SG1 plate. This condition was comprised of 200 mM $CaCl_2$, 100 mM HEPES pH 7.5, and 28% PEG 400. The optimisation screen was generated on the Scorpion to feature a gradient of PEG 400 from 22 – 34 % across the numerical axis and a pH gradient from pH 7 to pH 8 down the alphabetical axis. The concentrations of $CaCl_2$ and HEPES were kept constant across the screen at 200 mM and 100 mM respectively. As with the original commercial screens, the crystal plates were prepared using the Mosquito at 20^{o}C and they were stored in the 20^{o}C RockMaker hotel for automated image acquisition in increasing intervals up to 3 months after the creation of the plate.

2.11.3 Data Collection

Once suitable crystals had been observed in the crystallisation screen plates, the crystals were looped, frozen, and taken to beam I24 at Diamond Light Source for data collection. In order to loop and freeze crystals, cryo-protectant solutions were first prepared for each relevant buffer condition. In most cases this was as simple as making a small stock of the buffer condition containing 10% glycerol, although for conditions with high PEG contents the glycerol could be omitted. The looping tool, comprised of a magnetic wand and removable loops, was used to fish crystals. The process was conducted over a light microscope, where the vacuum sealed film was cut away from the drop containing crystals using a scalpel. The loop was placed into the crystal drop and the crystal was transferred into the centre of the loop through capillary action. The crystal was then moved from the original crystal screen drop to a fresh drop of cryo-protectant solution, where the crystal was left for a few seconds to take on as much of the cryo-protectant as possible. The crystal was then again picked up with the loop and transferred directly to liquid nitrogen, into a specialised puck for holding looped crystals. A fresh loop was added to the wand, and the process was repeated until all potentially collectable crystals were stored in the puck under liquid nitrogen. The puck was transferred to a charged dry shipper and taken to the I24 beamline for storage until the session could be arranged.

During a beamline session, the looped crystals are loaded into the equipment by the beamline staff, so that individual crystals can be placed into the beam using an automated system remotely. Once a loop was loaded into the beam

area, the crystal had to be centred on the beam. This involved setting the central point across several orientations of the loop so that once the beam was exposed and the crystal was rotated in place, the beam remained centred on the beam throughout. The next step is screening, where three images were taken with an exposure time of 0.04 s each and a 45^{o} shift between each image. This results in images that show the extent of possible diffraction spots. This is used to set the resolution limit when collected a full dataset. The predicted resolution determines how far away from the detector the crystal is positioned in the path of the beam, as a high-resolution dataset may include diffraction spots that extend to the edge of the detector, while low-resolution datasets will not extend out as far. Positioning of low-resolution crystals closer to the detector allows for better resolution of the spots closest to the source, while positioning high-resolution crystals further away allows for capture of the highest resolution data points. Screening a crystal also allows the software to suggest data collection parameters that may best suit that crystal. However, each dataset collected in this project, all of which resulted from crystals of cleaved RLD2, were collected using the same collection parameters. There was an oscillation of 0.1, 0 delta, an exposure time of 0.01 s and a total of 3600 images taken per crystal. Whereas during the screening run the delta was set to 45^{o} to allow sampling of different areas of the crystal, during a data collection run the crystal is not moved between images, denoted by a delta of 0, but instead the oscillation of 0.1 during each image exposure results in sampling of the entire crystal across the 3600 images. Once a dataset has been generated, various different software packages installed on the beamline provide estimates on the image statistics and resolution limit of the data, including the DIALS software that was used for RLD2 (Winter *et al.,* 2018). This process also determines the crystal lattice form, which is essential for accurate deconvolution of the diffraction spots through Phaser.

2.11.4 Data Processing

Following data collection on I24, crystal datasets were processed using the CCP4 software suite. Initially, the raw data was passed through Phaser to perform a molecular replacement process to deconvolute the diffraction spots in Fourier space into an electron density map in real space. Molecular replacement involves using a pre-solved homolog structure to determine the phase of the Fourier transform. If a homologous structure is not available for a given protein then other methods must be used, such as anomalous dispersion or isomorphous replacement techniques (Taylor, 2010). These involve the use of heavy metals to introduce a phase shift in the data that can be used to determine the phase of the non-anomalous data points. However, molecular replacement uses the phase information of the pre-solved homolog structure to estimate the phase of the new dataset. RLD2 is one of three RLD domains within the HERC2 protein, and structures had already been solved for RLD1 and RLD3, so molecular replacement was an ideal technique to solve the

RLD2 structure. The RLD3 structure (PDB accession code: 3kci) was used as a homologous structure and inputted into Phaser (McCoy *et al.,* 2007) as a pdb file. The sequence similarity of the two domains was 56.56 %.

Following generation of the electron density map by Phaser, an automated model was built in using the Arp/Warp software (Langer *et al.,* 2008). A manual approach was also initiated simultaneously using Coot, but the Arp/Warp method was much faster than the manual building method so the Arp/Warp model was used going forward. The model generated by Arp/Warp was visualised using the Coot software (Emsley *et al.,* 2010), and any discrepancies or ambiguous areas were adjusted manually. The model was passed through the RefMac software (Murshadov *et al.,* 2011) after each round of adjustments to determine the effect on the resolution and the quality of the fit to the data. The PDB-REDO (Joosten *et al.,* 2014) and MolProbity (Williams *et al.,* 2018) programs were also used to assess the quality of the model and identify any outliers, but the overall model statistics were generated through RefMac. Once the data was processed to completion, the resolution was validated using pairwise refinement in RefMac. This involved putting the model through RefMac a final time, but this time it was put through twice, once with refinements and once without. This produced the statistics tables as usual, but if the model statistics for the refined model were significantly different to those of the refined then it would indicate that further refinement was possible. When the values are similar for both processes, the model can be considered validated.

3 Protein Expression and Purification

3.1 Expression of Proteins in a Bacterial vs Mammalian Cell System

Most *in vitro* studies of protein, whether for structural biology or biophysical methods, tend to require milligram quantities of protein purified to homogeneity. The easiest way to obtain such large amounts of protein is to introduce the gene of interest into a vector organism, most commonly *E. coli*, yeast cells, insect cells, or immortalised mammalian cells. These are then propagated to generate large cultures of cells each overexpressing the protein of interest, which are then harvested and lysed to separate the insoluble cell membranes from the soluble proteins contained within. The overexpressed protein of interest is then separated from any endogenous cellular proteins through a series of chromatography steps, resulting hopefully in the required milligram yields of typically greater than 95% purity.

Of the different organisms chosen for protein overexpression, *E. coli* is usually the easiest and cheapest method of expressing relatively large quantities of protein due to the quick doubling rate and minimal requirements for growth. Although almost all of the proteins used in this project are derived from humans, they are reasonably small and do not require post-translational modifications to be active so were amenable to expression in *E. coli*. These proteins include UBE3A human isoform 1 (98 kDa); the human proteasomal shuttle protein PSMD4 (isoform 1) (40.7 kDa); the isolated domain of HERC2 that interacts with UBE3A, RLD2 (40.5 kDa); and the small fragment of UBE3A (residues 150-200) that has previously been shown to interact with RLD2, referred throughout this work as Ufrag (5.5 kDa). I also attempted to co-express the human tumour suppressor protein p53 (44 kDa), the viral HPV16 E6 protein (19 kDa), and human UBE3A isoforms 1 (98 kDa) and 2 (101 kDa) in an *E. coli* system in an attempt to allow formation of the complex in a cellular environment.

Although *E. coli* expression systems are the most convenient method when applicable, many human proteins require post translational modifications such as glycosylation to perform optimally, and *E. coli* lack the cellular machinery to carry out these functions. To overcome this, proteins requiring such modifications are usually expressed in more complex organisms, such as yeast cells, insect cells or mammalian cells. The human HERC2 protein explored in this project is very large (527 kDa) containing 4834 amino acids in a single chain, and so it necessitated a mammalian cell expression system. Due to the increased operational intensity of mammalian cell propagation compared to bacterial cells, coupled with the decreased capacity for collaborative working due to the current pandemic, HERC2 was only ever expressed transiently in an existing stock of HEK293 cells rather than through a stable HERC2-expressing cell line. There was also a much-reduced opportunity for optimisation of the mammalian expression system due to the

reliance on others carrying out the cell culture steps of this work on my behalf alongside their own experiments due to COVID19 restrictions.

3.2 Purification by Affinity Chromatography

3.2.1 UBE3A

The expressed UBE3A protein contained an N-terminal 6xHis tag, with a TEV protease cleavage site between the tag and UBE3A. This enabled initial separation of UBE3A from the whole cell lysate using high throughput nickel affinity chromatography. The cell lysate from a 1L UBE3A expression was passed over a HisTrap column (GE healthcare; section 2.6.2) to separate out the tagged UBE3A protein. The sample was then cleaved with TEV protease, to remove the tag, followed by a reverse HisTrap purification step as described in sections 2.6.4 and 2.6.5. This allows separation of UBE3A from any endogenous products that bound non-specifically during the initial HisTrap purification step. It also removes both the cleaved His-tag and any un-cleaved product from the sample, which would otherwise be difficult to identify and separate at a later stage. This produced a sample that appeared adequately pure upon SDS-PAGE analysis (Fig. 30).

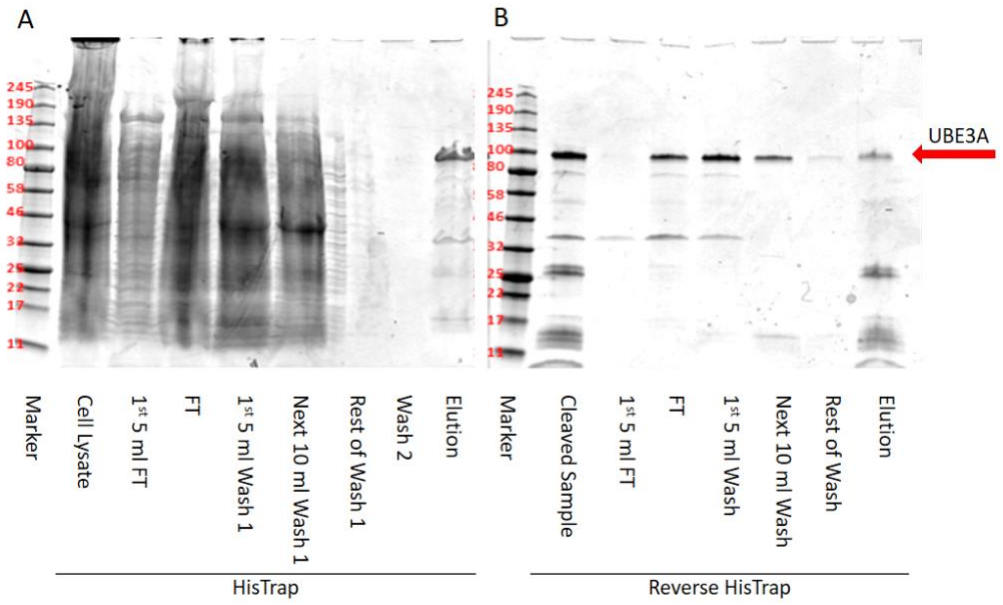


Figure 30: Affinity chromatography of UBE3A using a 3-step HisTrap system. A) The overexpressed UBE3A construct is first separated from the clarified cell lysate by utilising the affinity of the N-terminal His-tag for the nickel resin. The purified product from this step is then subjected to simultaneous dialysis and digestion with TEV protease. B) The cleaved sample is subjected to a second round of affinity chromatography. This time the untagged UBE3A passes through the column in the flow though (FT) and subsequent wash fractions, while any untagged product or endogenous protein with affinity for the nickel resin remains bound and is cleared in the elution fraction. This results in a more homogeneous sample that can be further purified using different chromatographic techniques. The samples were run on a 4-20% tris-glycine gel with a tris-glycine-SDS buffer system.

Although the two HisTrap purifications separate UBE3A from the majority of cellular proteins, a contaminant bands appears to co-elute with UBE3A at the reverse HisTrap stage. This was not an issue as the sample was taken forwards for further purifications using either ion exchange chromatography (section 3.3.1) or size exclusion chromatography (section 3.4.1), either of which resulted in the removal of this contaminant. The fractions from the reverse HisTrap purification that were taken forward for further purifications were the flow-through, the first 5 ml wash, and the next 10 ml wash samples. Occasionally the 35 ml fraction containing the rest of the wash sample was taken forward as well, although this depended on the yield of the purification and the amount of protein required for further experiments.

When overexpressing proteins in *E. coli* it is easy to assume that any band of the expected size of the target protein represents the target protein, but this is not always the case. At 98 kDa UBE3A is within the range of endogenous proteins that *E. coli* typically express, so it cannot be assumed that it is the correct product solely based on its size. However, as the sample progresses

through various purification stages, the behaviour of the sample allows a reasonable level of confidence that the protein being purified is the correct protein. It is not unusual for endogenous *E. coli* proteins to bind to a HisTrap column and elute alongside the target protein, as histidine is a naturally occurring amino acid and many proteins may naturally contain histidine-rich regions. However, the chances of an endogenous protein eluting from the HisTrap column, but then becoming cleaved with TEV protease and failing to interact with the reverse HisTrap column, become much lower. For UBE3A, I was also able to perform *in vitro* ubiquitination assays to confirm its catalytic activity, which also confirms my assumption that it is the correct enzyme.

Although the primary band across the HisTrap purifications is the expected size of UBE3A, *E. coli* also express endogenous proteins at around that molecular weight, so the size alone is not enough to confirm that the product is the target product. For UBE3A I was able to conduct *in vitro* ubiquitination assays (see section 5) to confirm that the product that I had purified was most likely UBE3A, although the behaviour of the sample through the TEV cleavage and subsequent reverse-His purification also supported this.

3.2.2 UbcH7

UbcH7 was obtained as a synthesised product in a basic vector, so several cloning steps were required to generate the final product that allowed us to obtain the purified UbcH7 enzyme. The gene product was synthesised using the protein sequence for UbcH7 isoform 1 as found on UniProt (UniProt ID: P68036-1), and the corresponding DNA sequence was codon optimised for optimal expression in *E. coli*. The gene was initially cloned into a pETM40 vector, resulting in expression of an TEV-cleavable N-terminally MBP-tagged protein, but the high affinity of maltose, the eluting agent for an MBPTrap purification, for the MBPTrap column is significantly higher than that of imidazole for the HisTrap column, so the elution sample would have had to have been subjected to a much more thorough buffer exchange process in order to remove sufficient maltose from the sample buffer to allow for a meaningful reverse-affinity purification. Rather than alter the digest and dialysis protocol that was used for every other purified protein sample (section 2.6.4), the UbcH7 gene was cloned from pETM40 into a pETM41 vector, that contained an N-terminal dual His-MBP tag with a TEV cleavage site (Fig. 31).

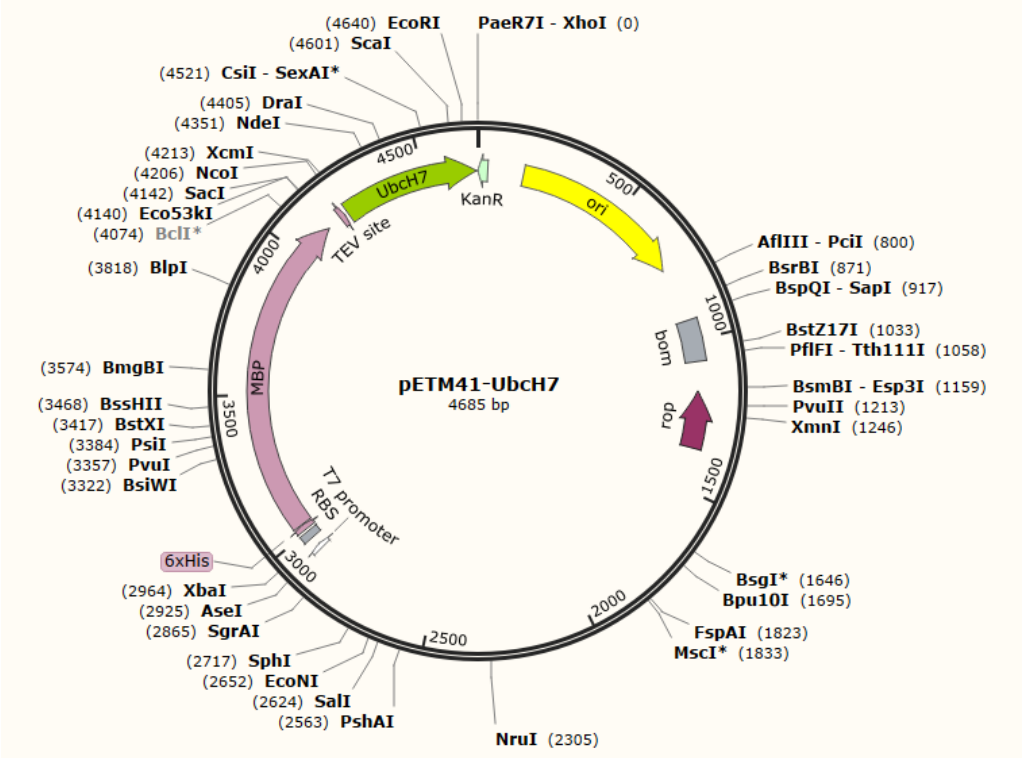


Figure 31: The pETM41-UbcH7 plasmid used for expression of the UbcH7 protein with an N-terminal His tag, MBP tag, and TEV cleavage site.

UbcH7 was therefore purified initially through a high-throughput MBP-Trap column (2.6.3), before cleavage of both tags with TEV protease (2.6.4), and a final reverse HisTrap purification (2.6.5) to separate the isolated UbcH7 protein from the 6xHis-MBP construct. Representative samples throughout the process were visualised through SDS-PAGE and a Coomassie-based dye (Fig. 32)

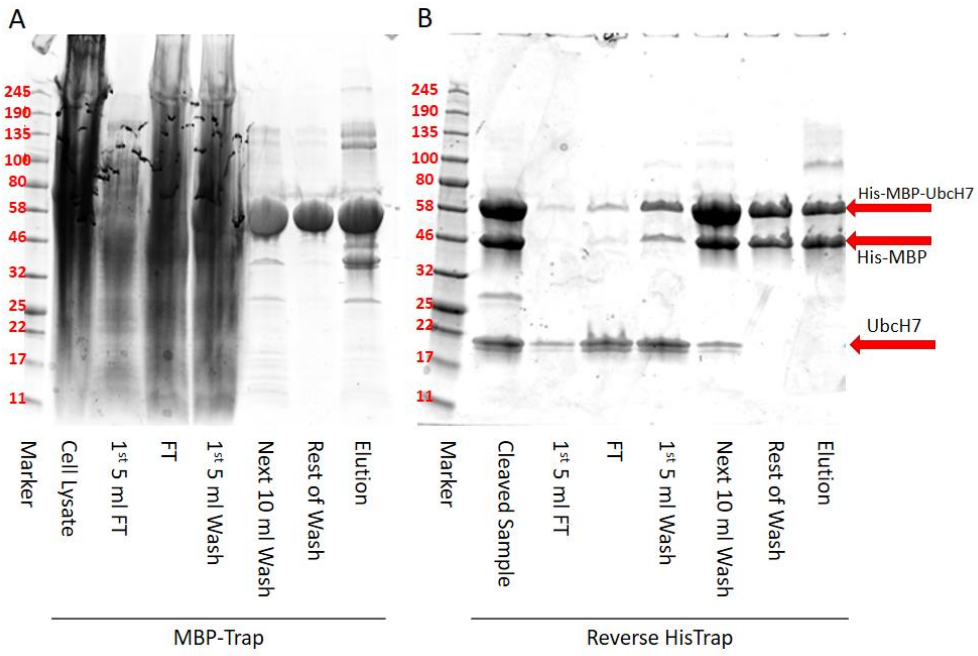


Figure 32: Affinity chromatography of UbcH7 using a 3-step MBPTrap and HisTrap system. A) The overexpressed UbcH7 construct is first separated from the clarified cell lysate by utilising the affinity of the N-terminal MBP-tag for the amylose resin in the packed MBPTrap column. The purified product from this step is then subjected to simultaneous dialysis and digestion with a TEV protease. B) The cleaved sample is subjected to a second round of affinity chromatography using a HisTrap column. This time the untagged UbcH7 passes through the column in the flow through and subsequent wash fractions, while any remaining tagged product and the cleaved tag itself remains bound and is cleared in the elution fraction. The samples were run on a 4-20% tris-glycine gel with a tris-glycine-SDS buffer system.

As with the UBE3A sample (section 3.2.1) the fact that a band appears at the expected molecular weight of the gel for the tagged UbcH7 sample is not enough to confirm that it is my expected product. However, the UbcH7 construct was expressed in such large quantities relative to the endogenous *E. coli* proteins that it is much more likely that this product is my purposefully overexpressed protein rather than an endogenous contaminant. The behaviour of the sample after TEV cleavage also supports this, as the predominantly single species sample seen after the MBP-Trap purification (Fig. 32a) becomes three distinct species, with only the band at the expected weight of the cleaved UbcH7 sample eluting from the reverse HisTrap column in the FT and wash samples, as would be expected on an untagged sample. The probability of the *E. coli* cells expressing a protein of the expected molecular weight of the tagged protein in such large quantities, that is then cleavable with TEV to result in samples of the expected molecular weights of the expected components, is so low that it is reasonable to assume that the sample shown is indeed UbcH7.

3.2.3 PSMD4

The PSMD4 expression plasmid was purchased from MRC PPU Dundee (Fig. 21). PSMD4 was expressed as a construct with an N-terminal His tag and a 3C protease cleavage site. It was purified as in sections 2.6.2, 2.6.4, and 2.6.5 by an initial high throughput HisTrap purification, followed by overnight 3C protease cleavage and dialysis, before a reverse HisTrap purification. This allowed efficient separation of the protein of interest from the whole cell lysate, non-specific histidine-containing proteins, and the tag. The various fractions collected throughout the purification were subjected to SDS-PAGE and visualised with a Coomassie-based stain (Fig. 33)

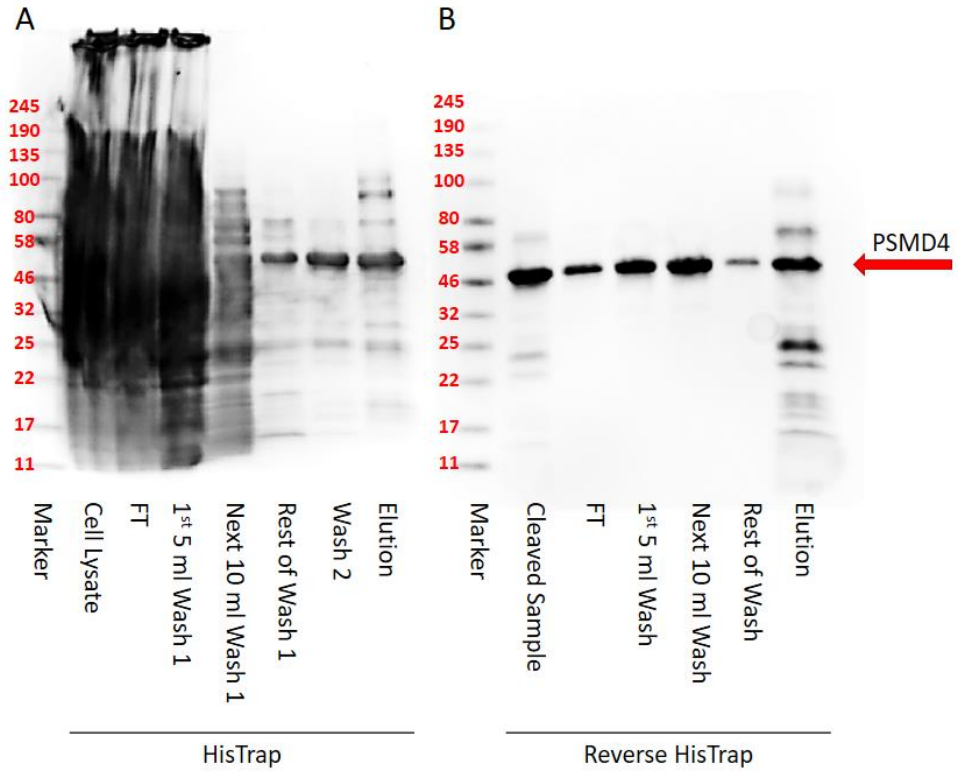


Figure 33: Affinity chromatography of PSMD4 through a 3-step HisTrap system. A) His-tagged PSMD4 is first separated from the clarified cell lysate with a HisTrap purification utilising two wash steps, the first with 40 mM imidazole and the second with 60 mM. Although a final high concentration elution step was carried out, the second wash step (60 mM imidazole) was consistently cleaner and a larger volume, and so was carried forward for 3C cleavage and dialysis. B) The cleaved sample was purified further through a reverse HisTrap method, resulting in large quantities of very pure sample. The samples were run on a 4-20% tris-glycine gel with a tris-glycine-SDS buffer system.

As with previous proteins (sections 3.2.1, 3.2.2), I first assumed that the product of the HisTrap purification was PSMD4 due to the size of the band on SDS-PAGE, but this was also confirmed by TEV cleavage and subsequent insensitivity of the sample to the reverse HisTrap column. The elution of the PSMD4 sample at the expected location in the ion exchange purification also

suggested that the pI of the product was at least very similar to that expected for PSMD4 (section 3.3.2). This gave me enough confidence to continue with this purified sample for further experiments,

3.2.4 RLD2

The RLD2 domain (residues 2959 to 3327) of HERC2 was isolated from a full-length HERC2 construct by designing PCR primers to amplify this gene segment along with a 5' NcoI site and a 3' stop codon and EcoRI site. The new construct was cloned via restriction digest into the pETM11 vector, creating a construct of the RLD2 domain with an N-terminal His tag and TEV cleavage site (Fig. 34).

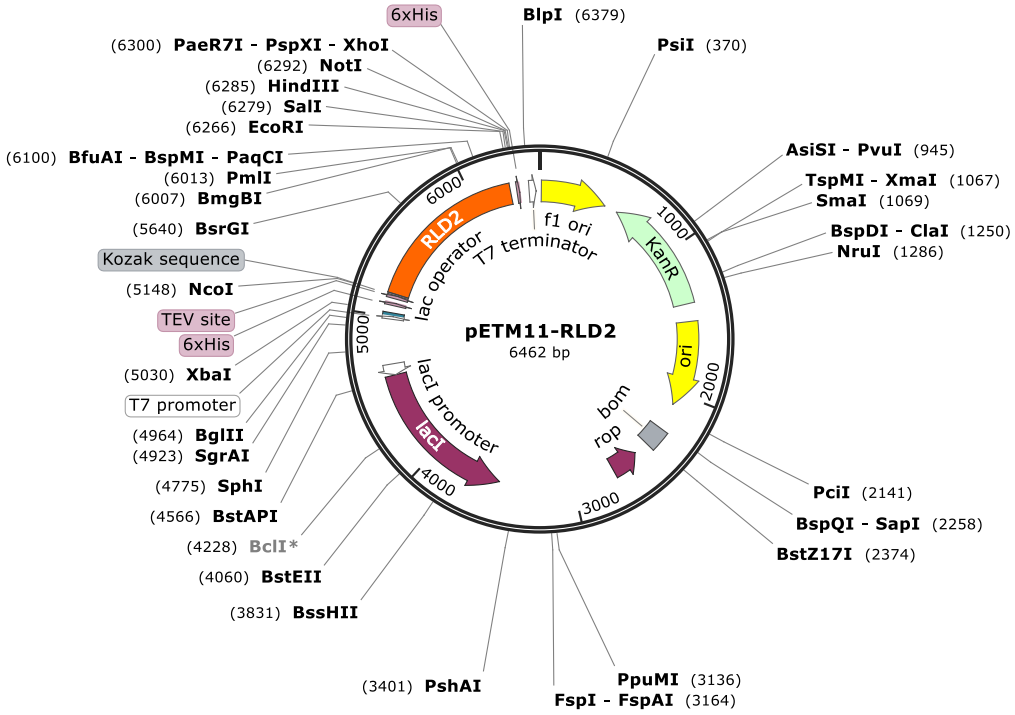


Figure 34: The pETM11-RLD2 plasmid for expression of a His-TEV-RLD2 construct in E. coli cells.

The fusion protein was then expressed in a bacterial system. Initially it was constructed as a His-MBP-tagged product and was purified as described in sections 2.6.3, 2.6.4, and 2.6.5, but it showed a resistance to cleavage by TEV protease (Fig. 35).

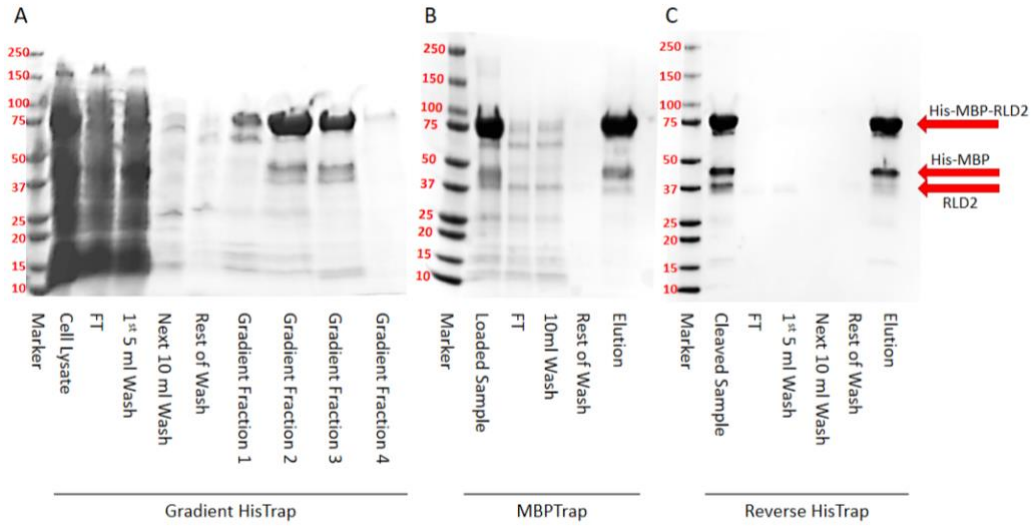


Figure 35: Affinity chromatography purification of His-MBP-RLD2 through a 4-step HisTrap plus MBPTrap system. A) The tagged RLD2 construct was first purified using a gradient HisTrap step. B) This was followed by an MBPTrap purification step as described in 2.6.3. C) The product was then cleaved with TEV protease and subjected to a reverse HisTrap purification to separate any cleaved product from uncleaved sample and the isolated tag. The samples were run on a 4-20% tris-glycine gel with a tris-glycine-SDS buffer system.

The lack of cleavage in the MBP construct was attributed to MBP contributing to a steric hindrance effect, preventing the TEV protease from accessing the cleavage site. In an attempt to overcome this the RLD2 construct was cloned again to produce a protein with only an N-terminal His-tag instead. This construct was then subjected to an initial HisTrap purification, TEV cleavage, and subsequent reverse HisTrap step. This process was visualised using SDS-PAGE and a Coomassie-based dye (Fig. 36).

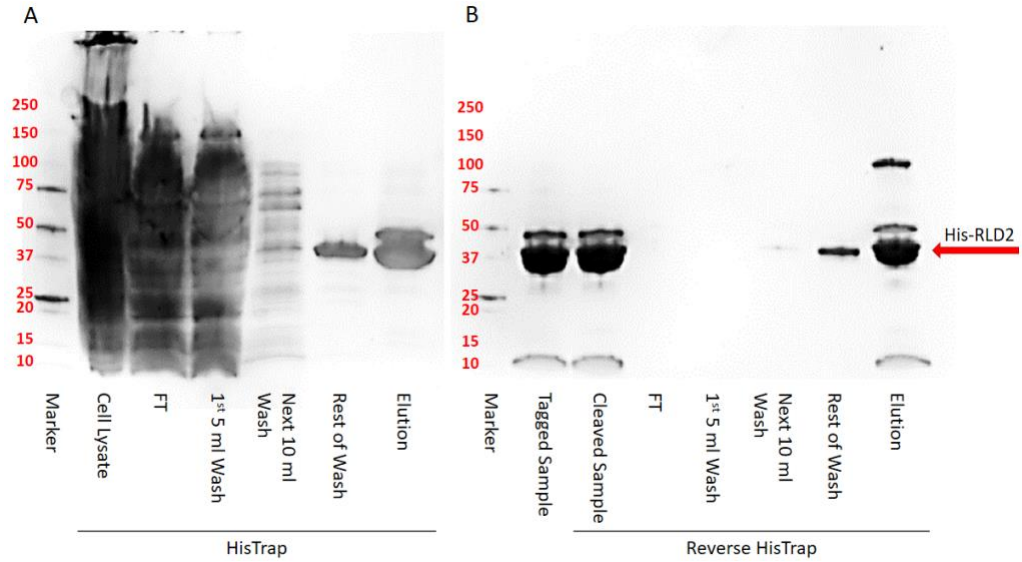


Figure 36: Affinity chromatography purification of His-RLD2. A) The sample was separated from the clarified cell lysate with a HisTrap column, before overnight

cleavage with TEV protease. B) The sample was subjected to a reverse HisTrap purification in an attempt to separate the RLD2 protein from the histidine tag. However, no or little sample was present in the flow through or wash fractions, indicating the absence of any cleaved products in the sample. The samples were run on a 4-20% tris-glycine gel with a tris-glycine-SDS buffer system.

Although the His-tagged construct removed the theoretical obstruction, it was no more amenable to cleavage than the initial MBP-tagged construct. This is indicated by the intensity of the band for the reverse HisTrap elution fraction, and the absence of bands in the flow through and wash fractions (Fig. 36b). A His-tag is a small and fairly unobtrusive tag as it consists of typically only 6-9 histidine residues that do not form a large secondary structure. However, they can still have an effect on protein function in some cases. Histidine residues are positively charged amino acids, so the tag itself is not completely inert and could contribute to electrostatic interactions between proteins. Particularly if the N-terminal or C-terminal ends of the protein are key to its cellular interactions, the presence of a His-tag may preclude its activity. However, as RLD2 is predicted to form a conserved structural motif, and it is only a domain within a larger protein rather than a functional enzyme on its own, I could predict that the presence of a 9His-tag should not affect the structure of RLD2 in any way, and the N-terminal residues were unlikely to be crucial to its interactions. Due to this, the His-tagged form of RLD2 was retained for use in various experiments.

A small amount of cleaved product could be produced when only a small amount of sample was subjected to a large excess of TEV protease, but the yield was still less than ideal (Fig. 37) so it was not used as often as the His-tagged construct.

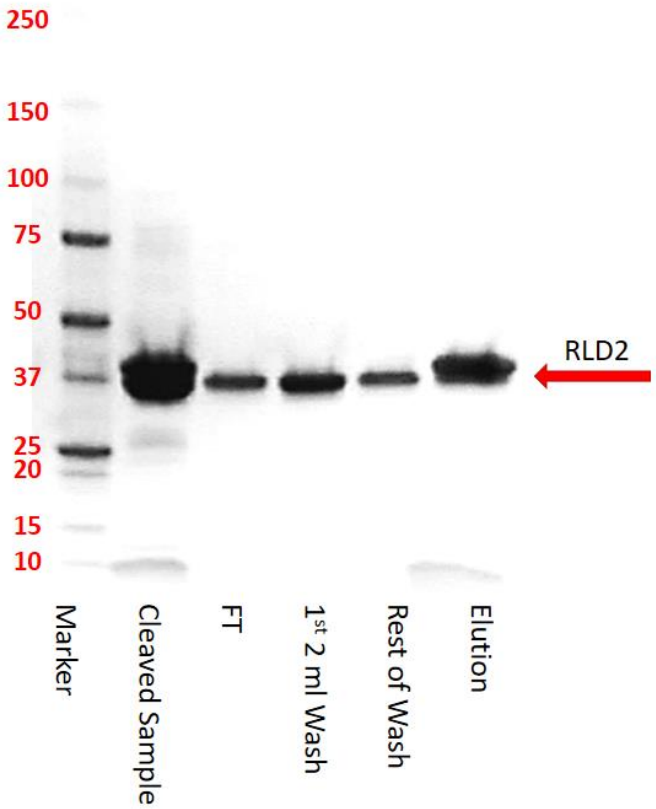


Figure 37: Reverse HisTrap purification of RLD2 following small-scale TEV cleavage. Cleaved sample eluted in the FT and wash fractions, while the portion of the sample that retains the tag elutes in the high imidazole elution fraction. The samples were run on a 4-20% tris-glycine gel with a tris-glycine-SDS buffer system.

The cleaved sample was obtained by subjecting only a small amount of the MBPTrap elution sample to a digest and dialysis protocol with a large excess of TEV. The volume of RLD2 sample used in the small-scale cleavage reaction was 500 μl, and the volume of TEV used was 200 μl. The concentration of the initial HisTrap elution sample was not determined prior to this step, and based on the much increased intensity of the RLD2 band relative to the TEV band at around 25 kDa it is possible that the actual ratio of the two proteins in solution was not sufficient to cleave the tag previously. However, the TEV protease was effective at cleaving large concentrations of previous protein samples at the ratios described in section 2.6.4, so it is still likely that the TEV cleavage site was precluded in the structure in some way. In either case, the quantity of TEV that would have been required to ensure sufficient cleavage of the larger quantities of RLD2 required for the various experimental techniques described in this report would have been excssive, and the retention of the His-tag did not seem to abrogate the interactions characterised in this report.

As with the previous protein samples, the first indication that the sample contained the RLD2 construct was the band at the predicted molecular weight

of the protein. However, unlike with the previous samples, the tagged RLD2 construct was not very amenable to TEV cleavage, which increases the possibility that it could be a conveniently sized contaminant. One supporting feature of the RLD2 purification was the increased level of expression relative to other endogenous proteins in the lysate and FT sample. Once the sample was assumed to be RLD2 and further experiments were carried out, the high resolution crystal structure of the RLD2 sample sufficiently confirmed that it was the correct protein.

3.2.5 Ufrag

The DNA sequence corresponding to the region of UBE3A involved in the interaction with HERC2 RLD2 (amino acids 150 – 200 of isoform 1, identified by Kühnle *et al.*, (2011)) was isolated from the full-length *UBE3A* gene using PCR primers designed to amplify this region and add a 3' NcoI site with a 5' stop codon and EcoRI site. This construct was cloned into the pETM40 vector using a restriction enzyme digest, resulting in a *Ufrag* construct with an N-terminal MBP tag and a TEV cleavage site (Fig. 38).

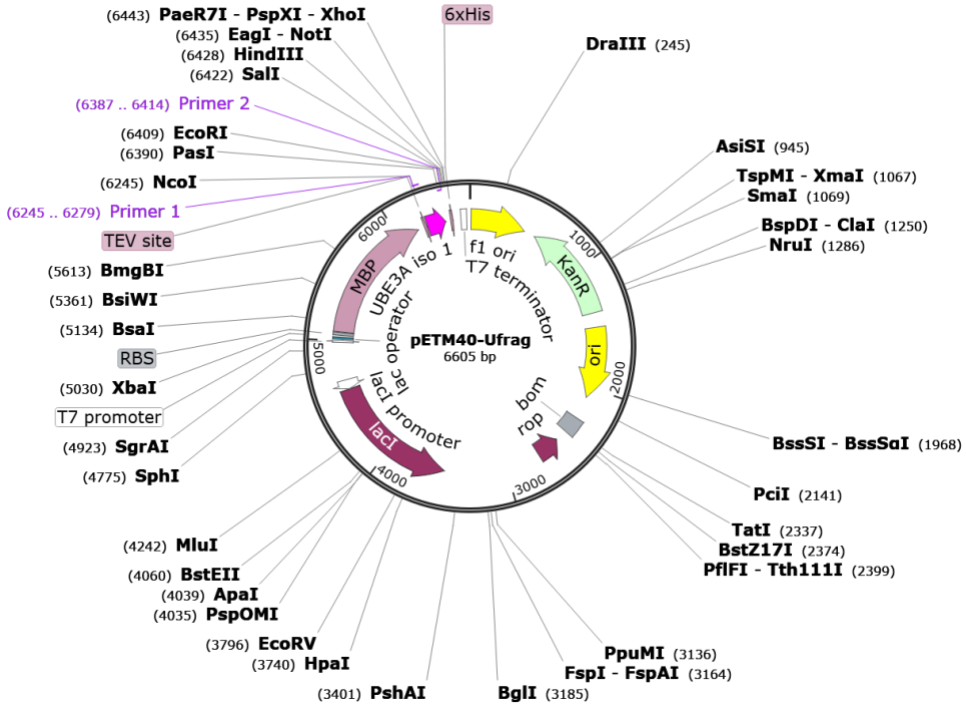


Figure 38: The pETM40-Ufrag plasmid for expression of an MBP-TEV-Ufrag construct in E. coli cells.

The MBP-Ufrag construct was purified in a single step with an MBPTrap column, and the resulting fractions were analysed with SDS-PAGE and a Coomassie-based stain (Fig. 39).

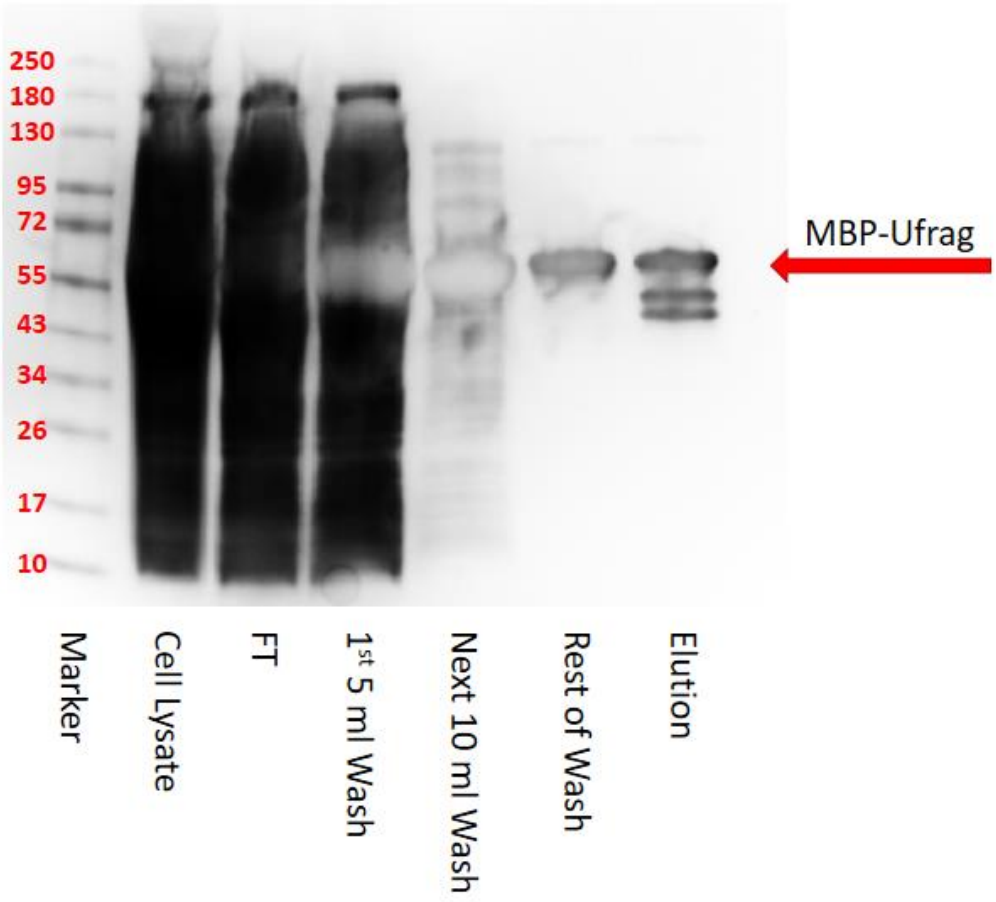


Figure 39: Affinity chromatography purification of MBP-Ufrag. The MBP-Ufrag construct was separated from the clarified cell lysate through an MBPTrap column as described in 2.6.3, resulting in a relatively large quantity of clean product in both the elution and final wash fraction. The samples were run on a 4-20% tris-glycine gel with a tris-glycine-SDS buffer system.

The product of the MBP-Ufrag purification appears to run at a higher molecular weight on the gel than you would expect from a Ufrag-MBP conjugate, as the predicted molecular weight is ~48 kDa. However, the Ufrag fragment is quite acidic with a predicted pI of 4.29, so it is possible that the apparent increase in molecular weight is due to the acidity of the protein causing a slower migration through the gel (Guan et al., 2015).

For most of the proteins purified during this project I was able to confirm the identity of the sample through removal of the purification tag, but this was not possible for the MBP-Ufrag sample as the cleaved Ufrag sample was too small to confidently identify on the SDS-PAGE gels used. For this construct, the large level of overexpression coupled with the observed molecular weight of the product were the sole determinants of its identity in the protein stage.

3.3 Purification by Ion Exchange Chromatography

3.3.1 UBE3A

UBE3A has a predicted pI of 5.12, so following the reverse HisTrap purification it was subjected to anion exchange on a Capto Q column, as described in

section 2.6.6. The fractions relating to the peak on the AKTA trace gave clean bands on SDS-PAGE (Fig. 40).

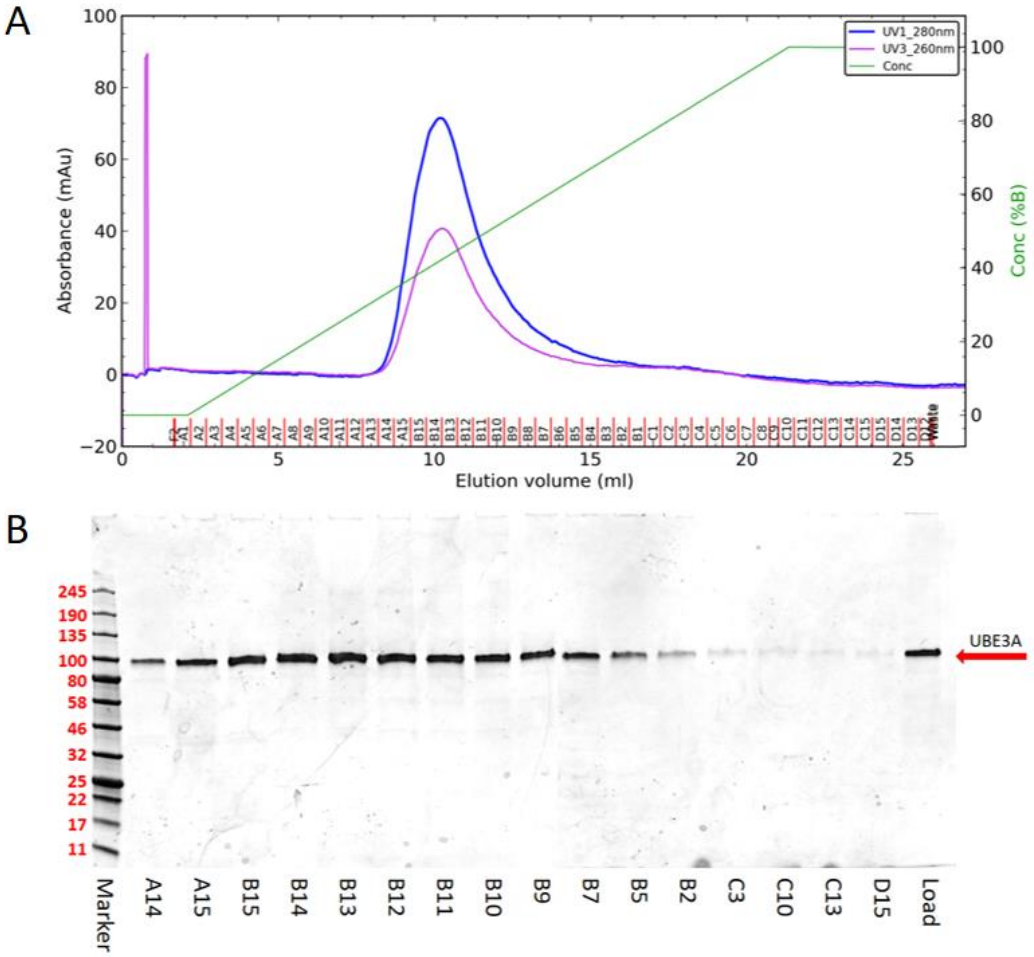


Figure 40: The anionic exchange purification stage of UBE3A purification. A) The sample of UBE3A following affinity purification was subjected to anion exchange chromatography on the AKTA explorer system, producing the absorbance trace shown in A. The blue line shows absorbance measured at 280 nm and the pink line shows absorbance measured at 260 nm with a scale bar to the left. The green line on the graph represents the proportion of high salt buffer washed through the column at any time, as described in 2.6.6, with the scale bar on the right. B) The fractions relating to potential UBE3A elution fractions from the anion exchange run were visualised using a 4-20% tris-glycine SDS-PAGE gel stained with a Coomassie dye, showing the final homogeneity of the sample.

3.3.2 PSMD4

The initial affinity chromatography purification of PSMD4 resulted in a nice clean sample with a fairly high yield (Fig. 33), but when the sample was concentrated for SEC or for other downstream uses it appeared to form higher order multimeric states that could not be separated by SEC (Fig. 44). Ion exchange was trialled to see if the potential multimers could be separated based on the increased interaction between higher ordered species for the resin. As PSMD4 has a pI of 4.68 a HiTrap Capto Q column was used, as

described in section 2.6.6. The fractions collected from the AKTA run were subjected to SDS-PAGE and stained with a Coomasie-based dye (Fig. 41).

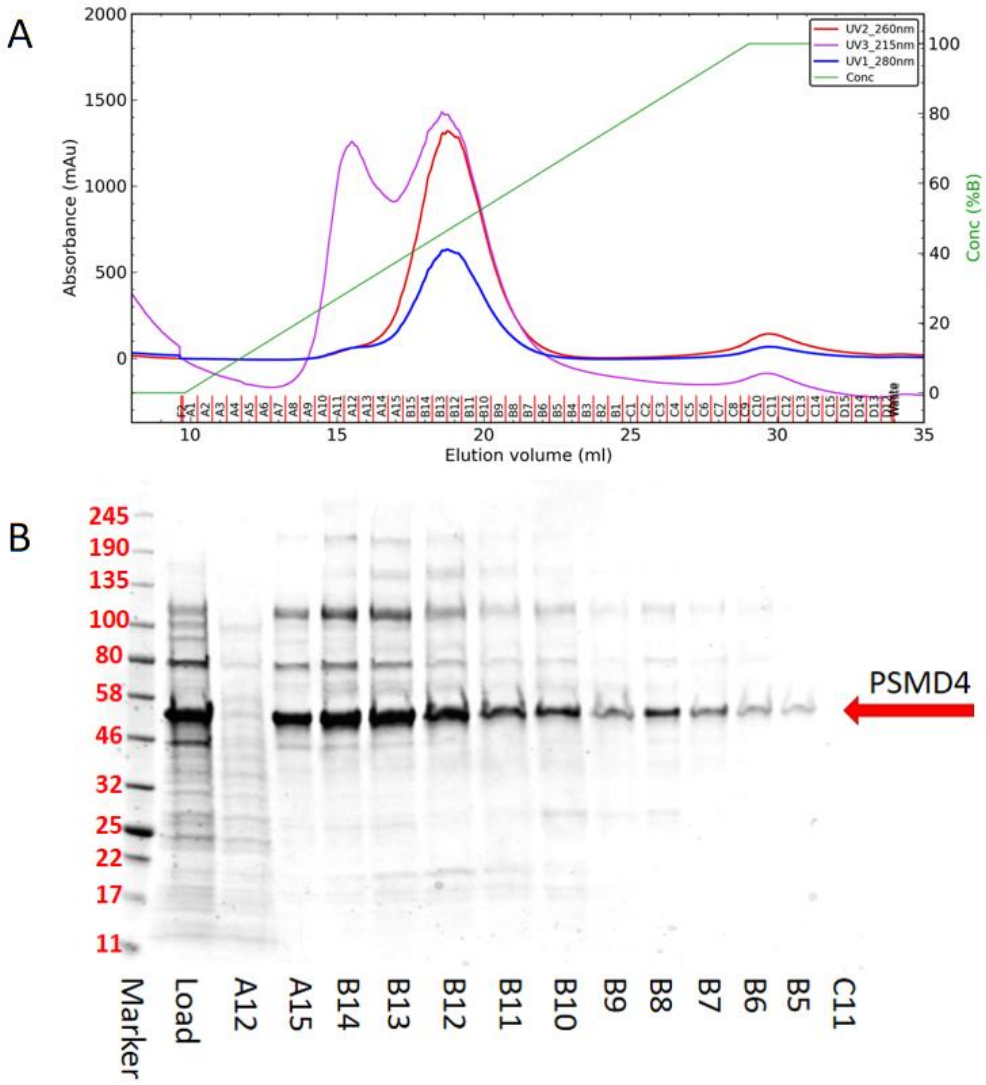


Figure 41: Anionic Exchange of PSMD4 following affinity chromatography purification and SEC. A) The absorbance of the sample was measured at 280 nm, 260 nm, and 215 nm as it eluted off the column. The absorbance at 280nm is shown in blue, the absorbance at 260 nm is shown in red, and the absorbance at 215 nm is shown in pink. The scale bar for the 215 nm trace is not shown, while the scale bar for both the 260 and 280 nm traces is shown on the left. The ratio of 260 to 280 readings is abnormal for this sample as PSMD4 contains no tryptophan residues, and so has a very low absorv=bance profile at 280 nm. B) The fractions relating to peaks in the absorbance trace were subjected to SDS-PAGE analysis on a 4-20% tris-glycine gel to determine the protein composition at each point.

Although monomeric PSMD4 appears to be the main component of each sample, higher molecular weight species of the predicted molecular weight of

PSMD4 multimers are also present, particularly at lower salt concentrations. The higher order contaminants are less present in the later, higher salt concentration samples, but the concentration of monomeric PSMD4 is also reduced in this region so it is not obvious whether these samples are truly cleaner or not.

3.4 Purification by Size Exclusion Chromatography

3.4.1 UBE3A

Size exclusion chromatography (SEC) separates macromolecular structures by their size. It involves a column tightly packed with porous beads, so larger proteins pass through the relatively large spaces between the beads and come off the column earlier, while smaller proteins get caught in the smaller pores within the beads and take longer to be flushed through. This form of protein purification is not very specific, but the sample will usually have been passed through other purification steps beforehand, such as affinity or ion exchange chromatography, so the size exclusion chromatography acts as a polishing step to remove any species that have managed to remain throughout the previous steps. It also provides some initial biophysical information about the sample, including its size, shape, and the presence and proportion of distinct multimeric states. UBE3A was subjected to SEC as described in section 2.6.7 after the affinity purification steps, and the relevant fractions were subjected to an SDS-PAGE gel (Fig. 42).

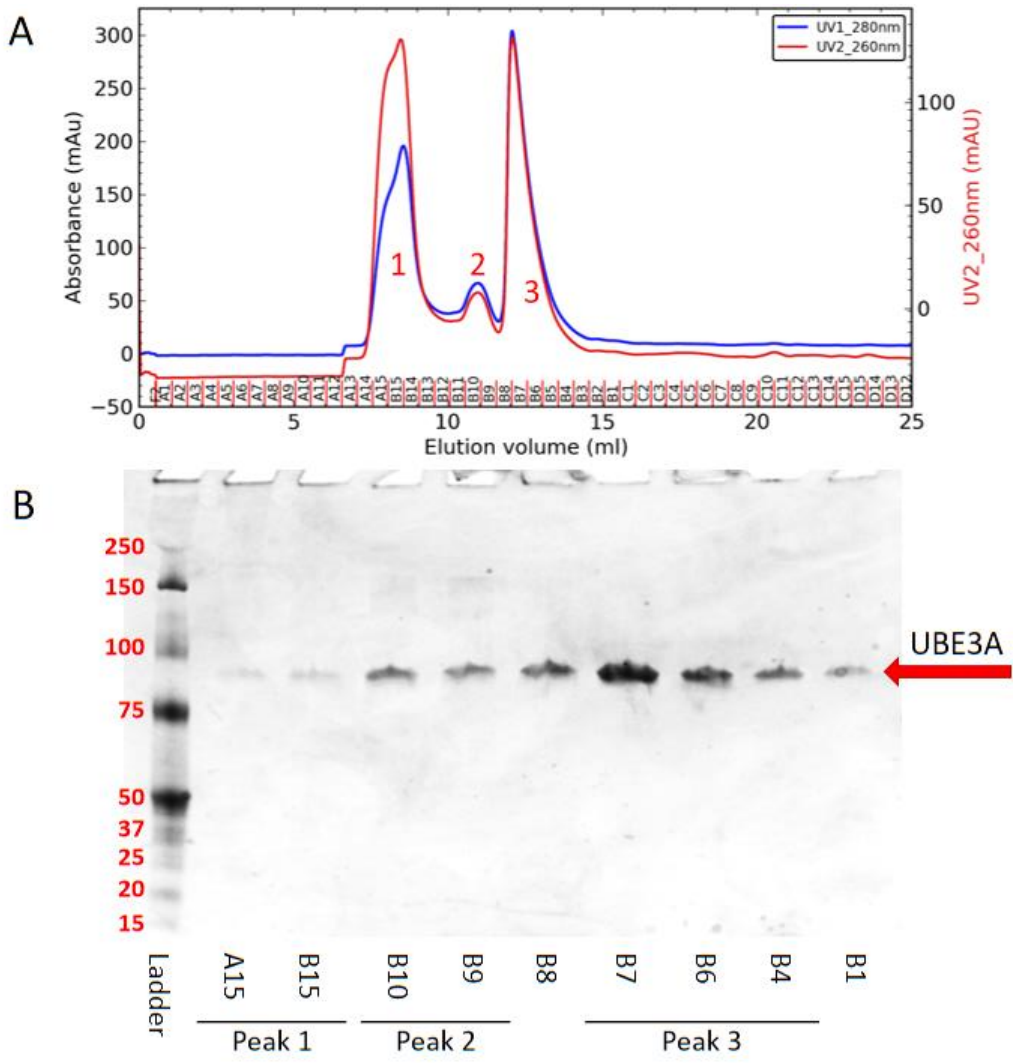


Figure 42: The SEC profile for UBE3A. A) The absorbance profile of UBE3A as it passes through the S200 column. The blue line represents the absorbance at 280 nm, while the red line represents the absorbance at 260 nm, with the scale and units for both the 280 nm reading on the left and the scale bar and units for the 260 nm reading on the right. The red dashes along the bottom show the fractions collected throughout the run, and the observed peaks are numbered. B) The fractions relating to points of interest in the absorbance traces were visualised using a 4-20% tris glycine SDS-PAGE gel and a Coomassie-based dye.

Bands of the expected molecular weight for UBE3A are observed primarily in peak 3, and also to a lesser extent in peak 2, correlating with the intensity of the absorbance measurements for the respective peaks. The fractions representing peak 1 show only a lower molecular weight component.

While peaks 2 and 3 of figure 42 show a species of the same molecular weight on SDS-PAGE, the separation of the sample into two distinct peaks in the absorbance profile represent different oligomeric states of the protein. While previous studies have suggested the presence of UBE3A as a trimer (Ronchi *et al.,* 2014), the separation of species in that size range by the S200 column

used here is insufficient to determine whether the oligomeric state represented by peak 2 is a trimer or dimer of UBE3A. Further characterisation of the nature of these multimeric states can be found in section 4.1.1. Peak 1, however, elutes at a position suggestive of a large molecular weight, while its SDS-PAGE visualisation suggests only the presence of a smaller molecular weight species. This is indicative of an aggregate comprised of a contaminate species but not UBE3A itself.

3.4.2 UbcH7

The cleanest samples following affinity purification of UbcH7 were subjected to SEC, as described in section 2.6.7. The fractions related to peaks in the absorbance profile were then subjected to SDS-PAGE and a Coomassie-based dye (Fig. 43).

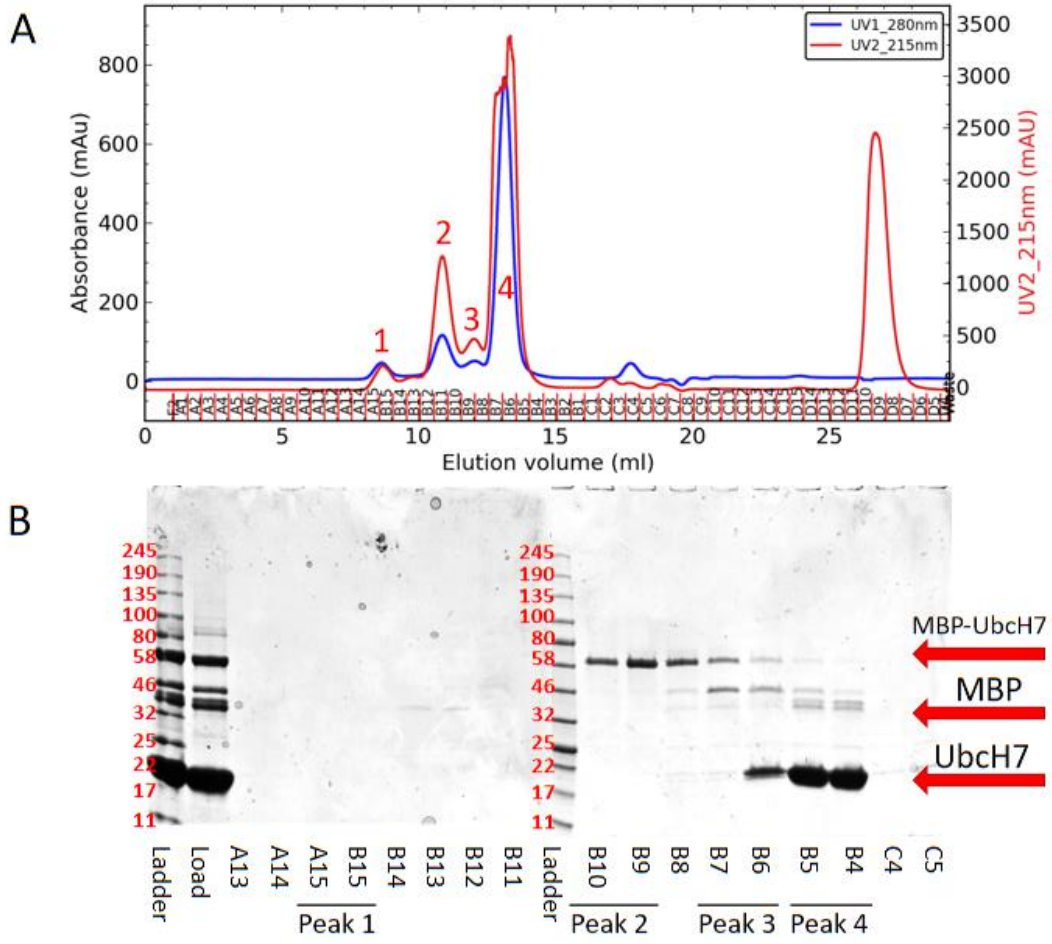


Figure 43: A) The SEC profile for UbcH7. The blue line shows the absorbance at 280 nm with the scale bar on the left, while the red solid line shows the absorbance at 215 nm with the scale bar on the right. The dashed red lines along the bottom show the fractions collected throughout the run. B) The fractions relating to peaks were visualised using a 4-20% tris-glycine SDS-PAGE gel and a Coomassie-based dye.

Bands for the expected molecular weight of UbcH7 are observed primarily in samples from peak 4, while a species at the expected molecular weight for the

His-MBP-UbcH7 construct appears primarily in peak 2, but to a lesser extent across peak 3 and also peak 4. Bands representing the expected molecular weight of isolated His-MBP are observed in peak 1, but also fairly prominently in peak 4.

Although the sample was reasonably heterogeneous before SEC, the majority of the sample appears to have eluted in peak 4 (Fig. 43). SDS-PAGE analysis of this peak does not appear to be very clean initially, but when you consider the overwhelming majority that the UbcH7 band represents in the sample the contaminating species are relatively minor. Crucially, the tagged form of UbcH7 appears to have been almost completely separated from the cleaved form, leaving the majority of the contaminants in the peak 4 samples to be attributed to the isolated His-MBP species. Also, as this protein was purified for use in a series of biochemical analyses rather than for any biophysical or structural characterisations, the inertness of the tag construct renders its already minor presence in the resulting UbcH7 sample negligible.

3.4.3 PSMD4

Although the PSMD4 sample looked very clean after affinity chromatography purification (Fig. 33), it was subjected to SEC as described in section 2.6.7 on an S75 column to confirm the absence of any multimeric states or aggregate species. The absorbance of the elute was measured throughout the process and the interesting fractions were subjected to SDS-PAGE followed by a Coomassie-based stain (Fig. 44).

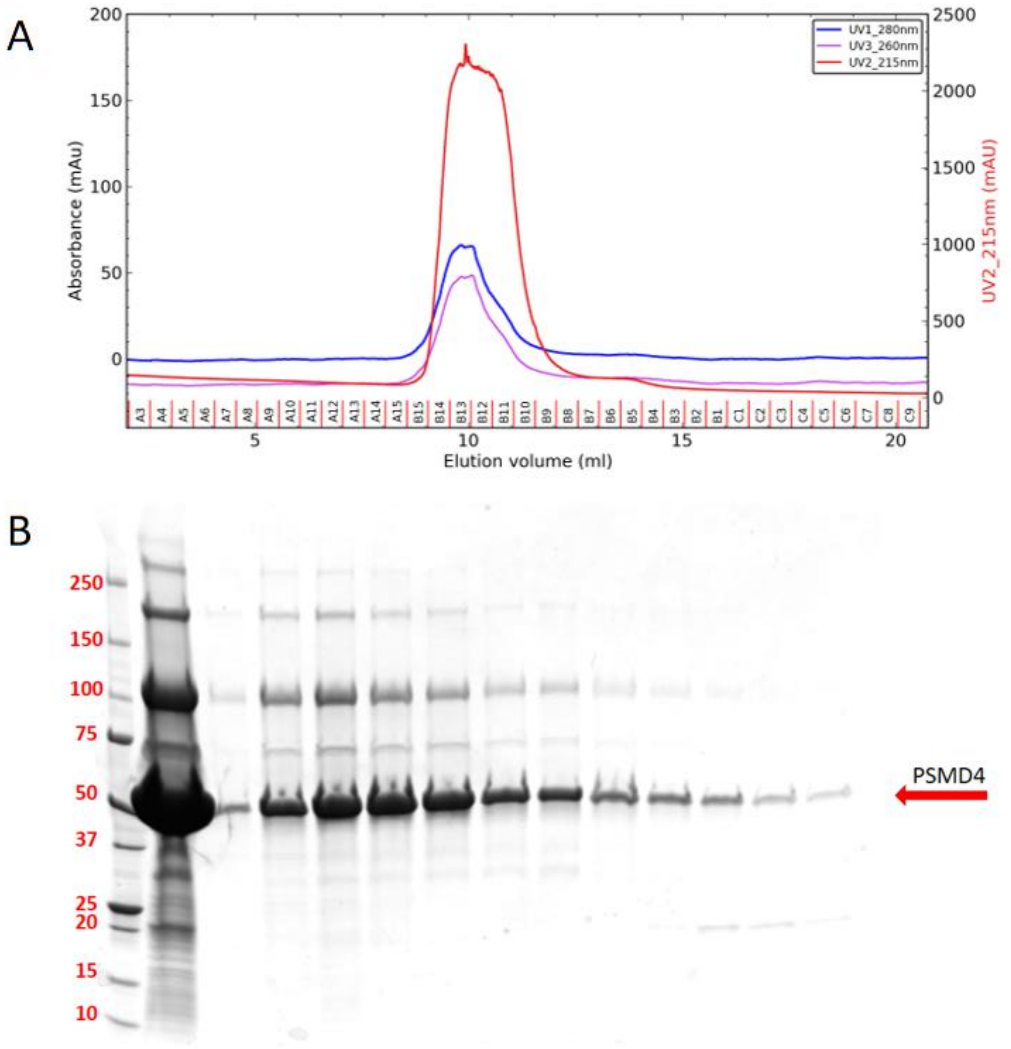


Figure 44: The SEC profile for PSMD4 following affinity chromatography. A) The absorbance readings at 280 nm, 260 nm, and 215 nm for the sample as it elutes from the column were plotted. The blue line shows the 280 nm trace and the pink line shows the 280 nm trace with the scale bar for both on the left, while the red line shows the 215 nm trace with the scale bar on the right. The red dashes along the x axis show the fractions that were collected throughout the run. B) The fractions relating to the peak on the trace were visualised through SDS-PAGE on a 4-20% tris-glycine gel with a Coomassie-based dye.

Although the AKTA trace shows only a single peak for the PSMD4 sample, indicating that only a single species in present, the SDS-PAGE lanes for the fractions suggest that the PSMD4 sample has formed multimers which co-elute with the monomeric sample.

One notable feature of the PSMD4 chromatogram is the similarity between the 260 nm and 280 nm traces. Typically, for purified proteins, the ideal 260:280 ratio would be 0.5, but the trace shown suggests a 1:1 ratio. However, the absorbance at 280 nm is coordinated by aromatic residues,

particularly tryptophan, and PSMD4 contains no tryptophan residues within its sequence.

3.4.4 RLD2

RLD2 was subjected to SEC using an S75 column as described in section 2.6.7 to assess the distribution of oligomeric states within the sample that may have been precluded during the affinity chromatography purification process due to the use of SDS in the SDS-PAGE sample preparation. The use of constant absorbance monitoring resulted in a trace demonstrating where the majority of the proteins elute during the run (Fig. 45a). This trace was used to identify key fractions to be analysed with SDS-PAGE and a Coomassie-based stain (Fig. 45b).

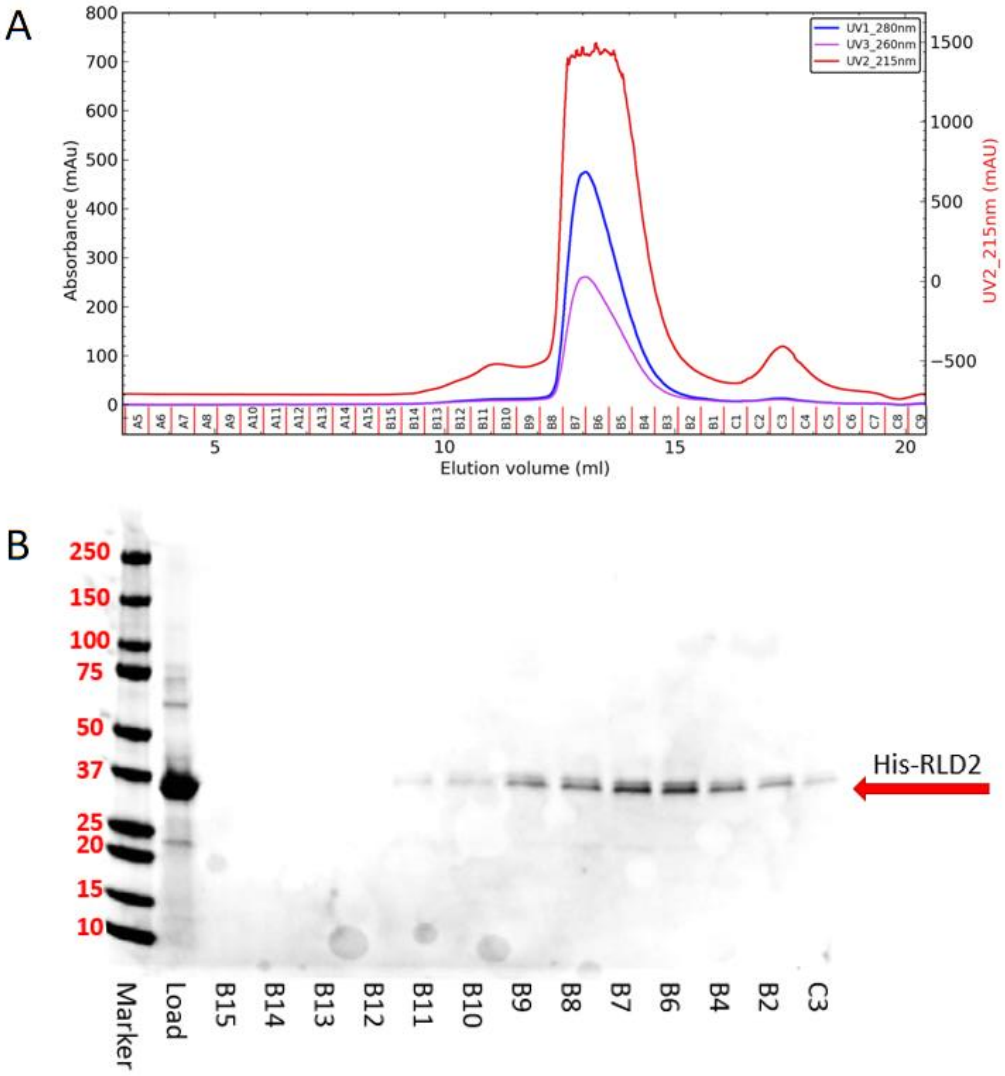


Figure 45: The SEC profile for His-RLD2 following affinity chromatography purification. A) The absorbance trace at 280 nm, 260 nm, and 215 nm during the run. The trace for the absorbance at 280 nm is shown in blue and the trace for absorbance at 260 nm is shown in pink with the scale bar for both shown on the left, while the trace for absorbance at 215 nm is shown in red with the scale bar on the right. B) The SDS-PAGE image for the fractions identified within the absorbance peak. Although the trace showed a slight rightwards skew, the SDS-PAGE lanes all show only a single, clean species within each fraction. The sample were run on a 4-20% acrylamide gel with a tris-glycine-SDS buffer system, and then stained with a Coomassie-based dye.

3.5 UBE3A+E6+p53

UBE3A isoform 1, p53, and the HPV16 E6 protein were cloned with the aim of co-expressing the three proteins in *E. coli* in an attempt to reproduce the work of Masuda *et al*. (2019). The complex formed by UBE3A, HPV16 E6, and p53 was approached in several different ways. Initially, the E6 and p53 genes were cloned into the pACYC-Duet-1 plasmid (Fig. 46), which would have resulted in a His-tagged UBE3A protein, a strep-tagged p53 protein, and un-tagged E6.

The *E6* gene was cloned into MCS1 of a pACYC-Duet1 vector using the *NcoI* and *EcoRI* restriction sites, and then the *p53* gene was cloned into MCS2 of the pACYCDuet1-*E6* plasmid using *NdeI* and *KpnI* restriction enzymes (Fig. 46).

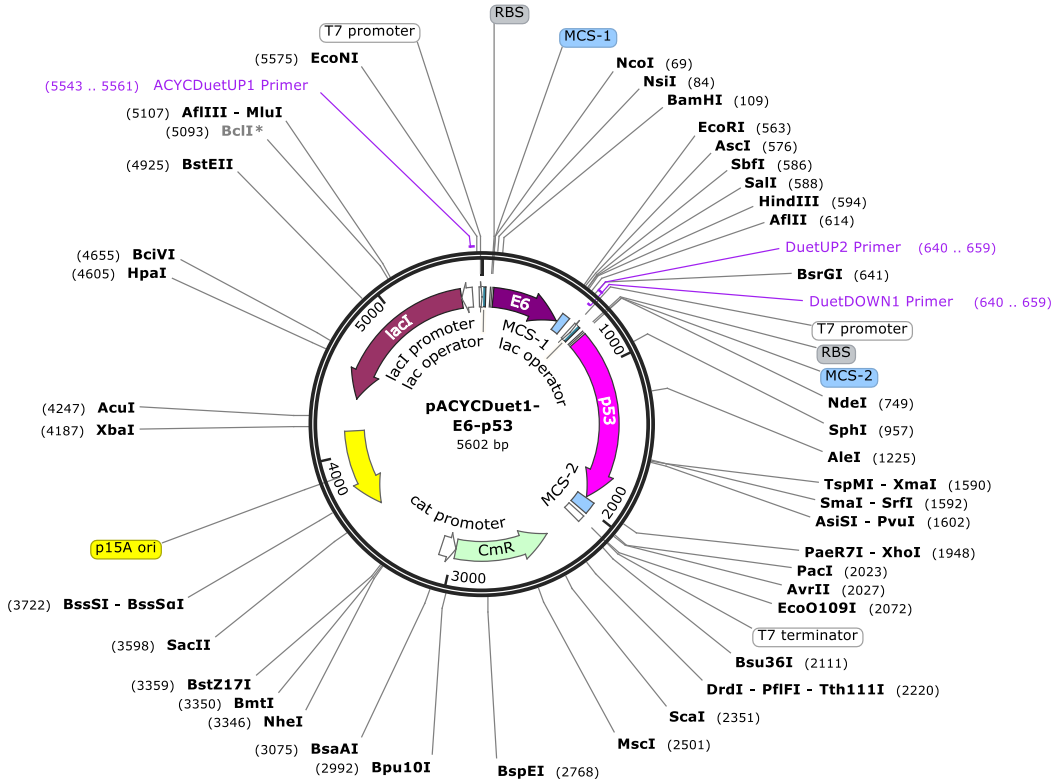


Figure 46: The pACYCDuet1-E6-p53 plasmid used for co-expression of untagged E6 and p53 in BL21 or ArcticExpress cells.

I hypothesised that the three proteins would assemble in the cells to form the ternary complex, which could then be purified using a HisTrap and a strep resin purification. To achieve this, the pACYC-E6-p53 plasmid (Fig. *46*) and the pUBE3A plasmid (section 2.2.1) were co-transformed into BL21 cells. Large scale cultures were grown using 1L LB media supplemented with kanamycin and chloramphenicol for plasmid selection. The cells were grown at 37°C initially, and then the temperature was dropped to 16°C prior to induction with 1mM IPTG. After 24 h post-induction the cells were harvested, resuspended (section 2.5.1), sonicated (section 2.6.1), and subjected to successive rounds of HisTrap and Strep-resin purifications (sections 2.6.2, 2.6.4, 2.6.5, and 2.6.9). However, the HisTrap purification from this showed the presence of the His-tagged UBE3A in the sample but the p53 and E6 proteins were not observed (Fig. 47). In addition to the obvious UBE3A band, a band at around 15 kDa could be seen in the initial HisTrap elution sample that could have been E6 (Fig. 47a), but, in the reverse HisTrap gel the potential E6 species remained bound to the column and did not co-elute with the cleaved UBE3A sample (Fig. 47a). The initial HisTrap purification showed a band close to 50 kDa in the 10 ml wash fraction which I supposed could be the

strep-tagged p53 protein that had been expressed but had not formed the complex. In order to test this, it was subjected to a strep resin purification (see section 2.6.9), which suggested an absence of anything strep-tagged in the sample, therefore precluding it from being the p53 construct (Fig. 47b).

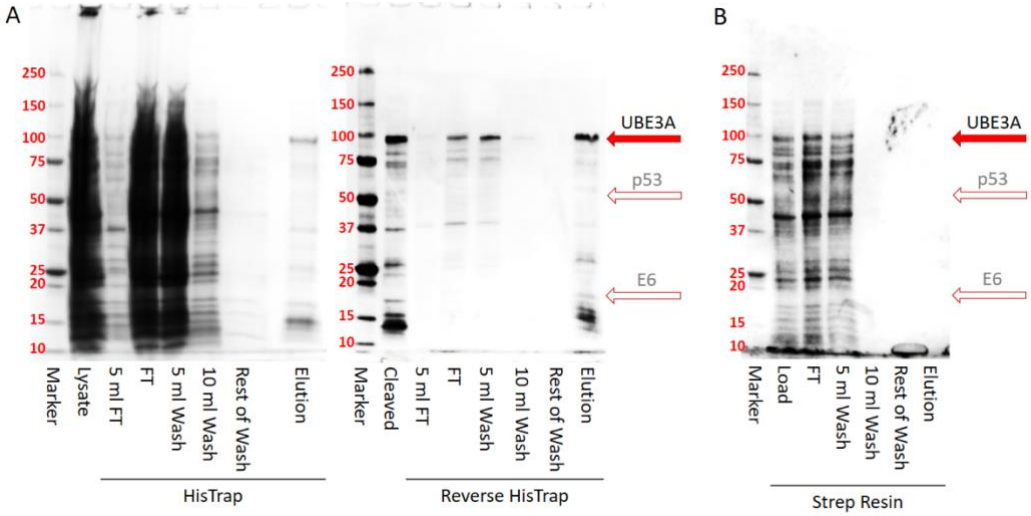


Figure 47: Initial purification of a co-expressed UBE3A-E6-p53 complex. The constructs were generated to express a His-tagged UBE3A protein and a Strep-tagged p53 protein, and the two proteins should not interact without the presence of the untagged E6 protein. A) The forward and reverse HisTrap purification steps. B) The 10 ml wash fraction from the HisTrap purification was subjected to a Strep resin gravity column purification. The band at the predicted size for UBE3A is indicated by a solid red arrow, and the predicted positions for p53 and E6 if they were present are shown by the hollow red arrows.

One theory at this stage was that the p53 and E6 proteins were not being efficiently expressed in the standard BL21 cells at the reduced temperature suggested by Masuda *et al.,* (2019), so the same plasmids (Fig. 46, section 2.2.1) were transformed into ArcticExpress cells from Agilent. These cells express chaperone proteins that help to stabilise proteins during expression at low temperatures. Unfortunately, all attempts at using these cells were ultimately unsuccessful, as the cell cultures did not seem to survive the extended incubation times necessary for the reduced temperatures.

As not much is known about the interaction between full-length UBE3A and p53 it was possible that the presence of the strep-tag on the p53 construct could be blocking the complex formation, so I attempted to express only the pACYC-E6-p53 construct in the BL21 cells. To do this, the pACYC-E6-p53 plasmid (Fig. 46) was transformed into BL21 cells, and a large scale culture comprised of 1L LB supplemented with chloramphenicol only was set up. This culture was incubated at 37°C until roughly 1h pre-induction, at which point the temperature was dropped to 16°C. The cells were induced with 1mM IPTG, and then harvested 24h post-induction. The cells were then resuspended and sonicated (section 2.5.1, section 2.6.1) before being

subjected to a strep-resin purification (section 2.6.9) to see if the strep-p53 protein could be extracted from the cell lysate. The fractions collected during the purification attempt were subjected to SDS-PAGE and stained with a Coomassie-based dye (Fig. 48), which suggested that there was no strep-p53 present in the sample.

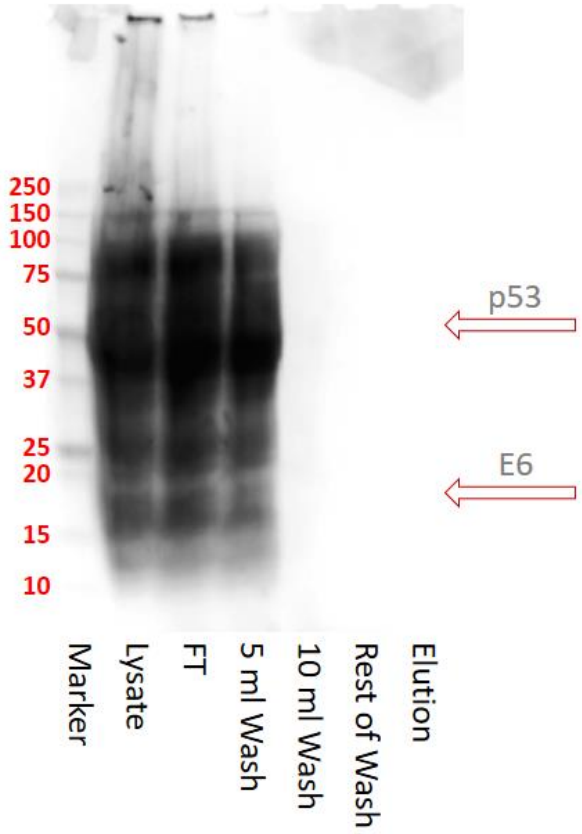


Figure 48: Strep resin purification of p53 co-expressed with E6. The predicted positions for E6 and p53 are shown by the hollow red arrows. The samples were run on a 4-20% acrylamide gel with a tris-glycine-SDS buffer system, and then stained with a Coomassie-based dye.

Bands at similar molecular weights to those of p53 and E6 could be seen in the cell lysate and FT fractions but not in the elution fraction, or even the washes. This suggests that neither E6 nor p53 have been successfully expressed in these cells.

The p53 and E6 proteins can be toxic to human cells when overexpressed so I attempted to express them in Rosetta pLysS cells instead of the BL21 cells used previously. Rosetta cells contain the pRARE plasmid to encourage expression of proteins whose sequences contain codons that are rare in *E. coli*, and the pLysS component inhibits expression of the cloned genes during the initial growth stage of bacterial cell protein expression, just in case they were able to exert a similar effect on the bacterial cell growth systems. The pACYC-Duet-1 plasmid was not suitable for expression in pLysS cells however due to both the pACYC plasmid and the pLysS plasmid using chloramphenicol

resistance as a retention strategy. To overcome this, the E6 and p53 genes were cloned into the pCDF-Duet-1 plasmid, which used a streptomycin resistance gene instead (Fig. 49).

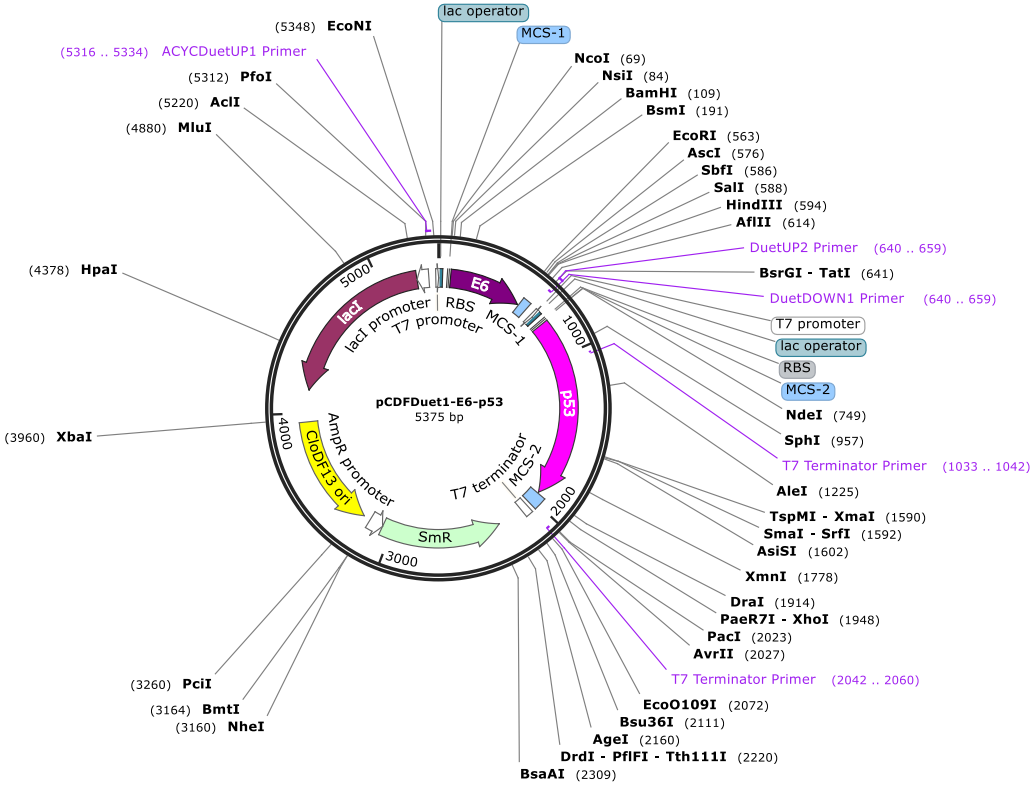


Figure 49: The pCDFDuet1-E6-p53 plasmid for co-expression of untagged E6 and p53 in BL21 pLysS or Rosetta cells.

In order to limit the number of antibiotics used in the cell expression the pCDF-E6-p53 and pUBE3A constructs were expressed separately in the Rosetta pLysS cells, and the clarified lysates were mixed prior to purification on a HisTrap. The pCDF-E6-p53 plasmid (Fig. 49) was transformed into the Rosetta pLysS cells and grown in a large culture consisting of 1L LB and streptomycin. Chloramphenicol was used as a selection marker for the pRare and pLysS plasmids during the starter cultures, but chloramphenicol was not added to the large scale cultures as it can interfere with the activity of bacterial ribosomes, and the pRare and pLysS plasmids are much smaller than typical protein expression plasmids so are not so susceptible to selection pressures. The culture was incubated at 37°C, and then the temperature was dropped to 16°C prior to induction. Protein expression was induced with 1 mM IPTG and the cells were harvested 24 h after induction. The UBE3A cultures were grown as described for UBE3A samples not used in complexes (section 2.5.1), in 1L LB with kanamycin, at 37°C pre-induction and 25°C post-induction, and cells were harvested roughly 20 h post-induction. Both sets of cells were resuspended (section 2.5.1), sonicated, and clarified (section 2.6.1) separately, and then the clarified lysates were mixed prior to a HisTrap

purification (section 2.6.2). Unfortunately, this attempt did not appear to increase the expression of either protein either. There are bands in the elution lane of similar molecular weights to p53 and E6, but there are also bands of these rough molecular weights in the diluted pUBE3A lysate lane prior to addition of the pCDF-E6-p53 lysate, which suggests that they could be contaminant proteins from the Rosetta cells themselves that have bound non-specifically to the purification column (Fig. 50).

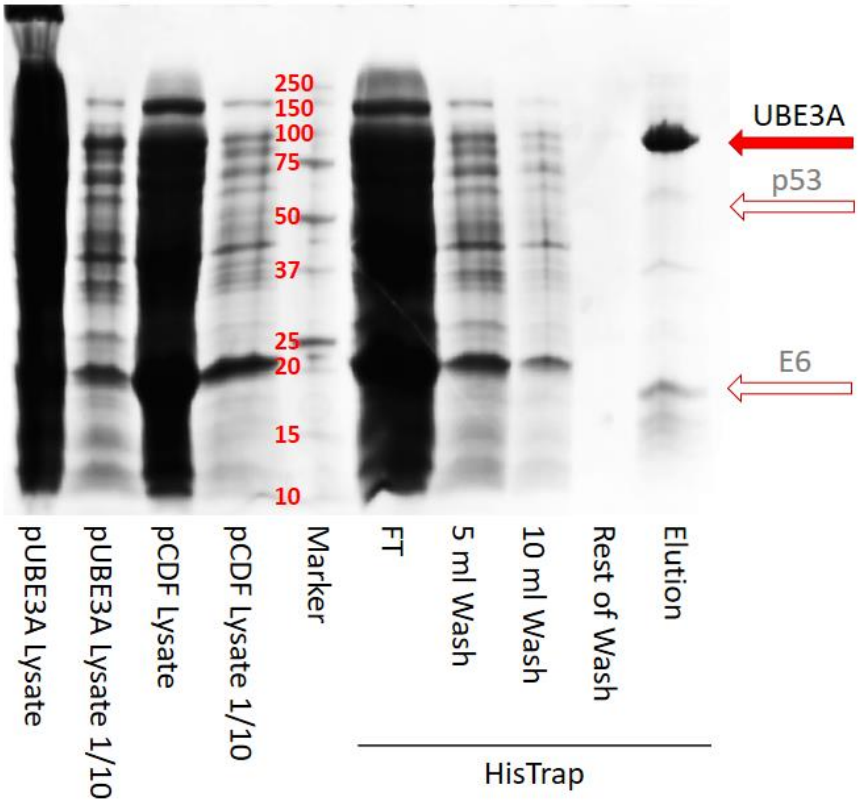


Figure 50: Attempted purification of the UBE3A-E6-p53 complex from separate expressions of His-UBE3A and the pCDF plasmid expressing E6 and p53. The left of the marker shows the clarified lysates from each expression at both full concentration and diluted to 1 in 10, and the HisTrap process is shown to the right of the marker. The expected locations of p53 and E6 are shown by the hollow red arrows, while the obvious UBE3A band is indicated by the solid red arrow.

Following these set-backs I contacted the authors of the previously published paper that had shown the co-expression and subsequent purification of the complex in question from an *E. coli* expression system (Masuda *et al.,* 2019). Their suggestion was to express each protein in an individual plasmid rather than attempting to utilise the Duet function of the pACYC and pCDF plasmids used so far. This was due to an observation by them that the dual promoter system in the Duet-1 plasmids caused a decrease in the efficiency of protein expression. In light of this, I next attempted to co-express the pUBE3A construct (section 2.2.1), the pACYC-E6 plasmid (Fig. 51), and the pCDF-p53

plasmid (Fig. 52) in cells.

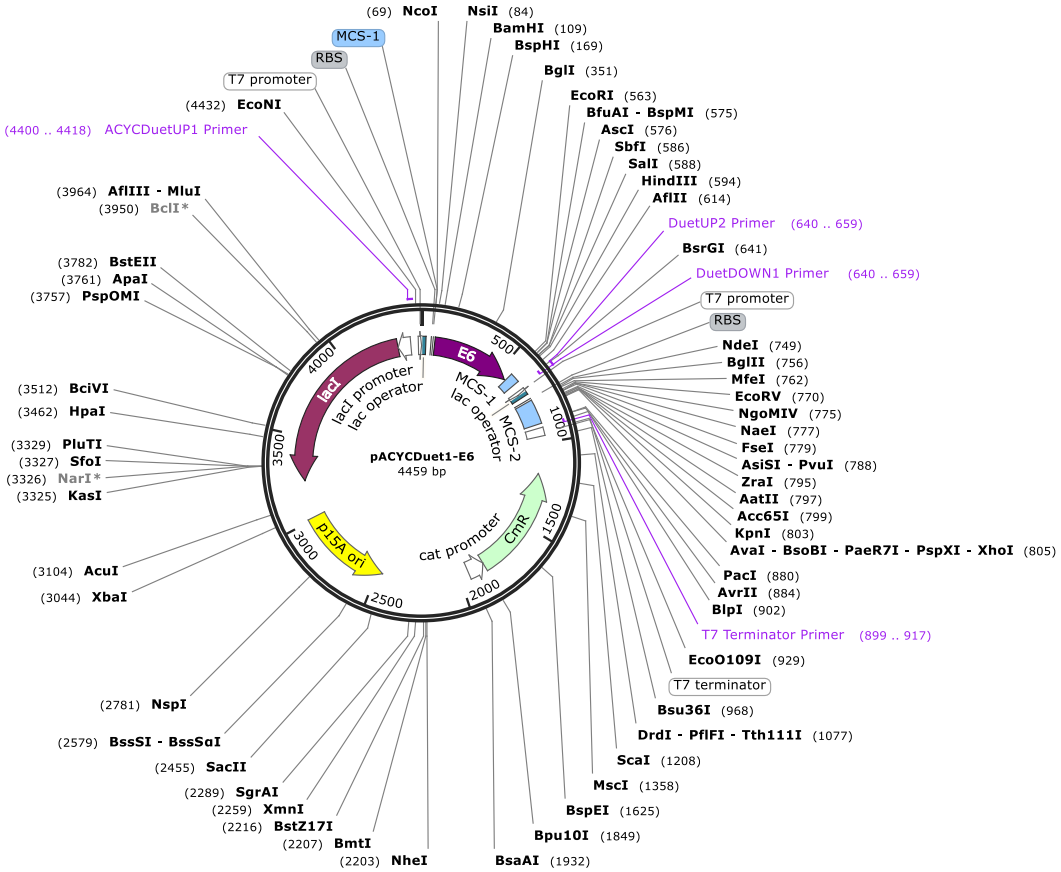


Figure 51: The pACYCDuet1-E6 plasmid for expressing the untagged HPV16 E6 protein in BL21 cells.

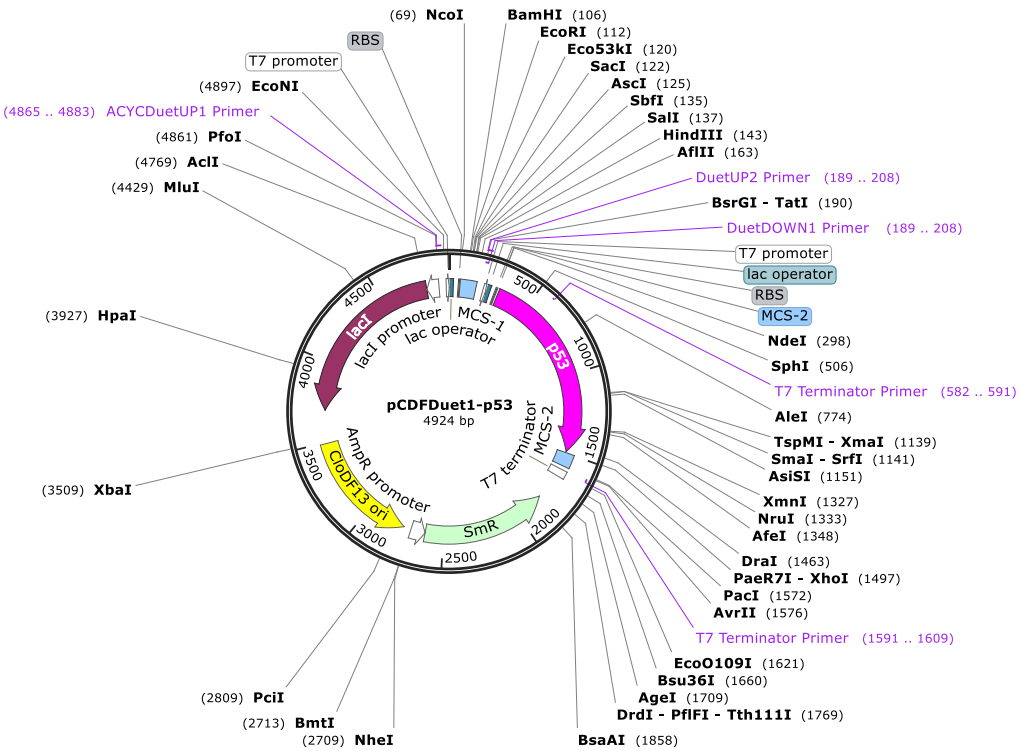


Figure 52: The pCDFDuet1-p53 plasmid to enable co-expression of p53, E6, and UBE3A across three different plasmids.

In this system, UBE3A retained the His-tag, E6 remained untagged, and p53 was cloned such that both a strep-tagged and untagged version could be used to identify whether the presence of the strep-tag was preventing the complex formation. This combination of plasmids was transformed into both BL21 and ArcticExpress cells, using the same protocol for either cell type with the only difference being the addition of gentamycin with the ArcticExpress cells to retain the chaperone protein expression. For each, cells were grown in 1L TB at 16^{o}C pre- and post-induction, and protein expression was induced with 400 µM IPTG. Cells were harvested 24 h post-induction, and the presence of the complex was tested for using a gradient elution HisTrap purification (section 2.6.2), followed by SDS-PAGE, attempting to repeat the work of Masuda *et al.* (Masuda *et al*., 2019; Fig. 53).

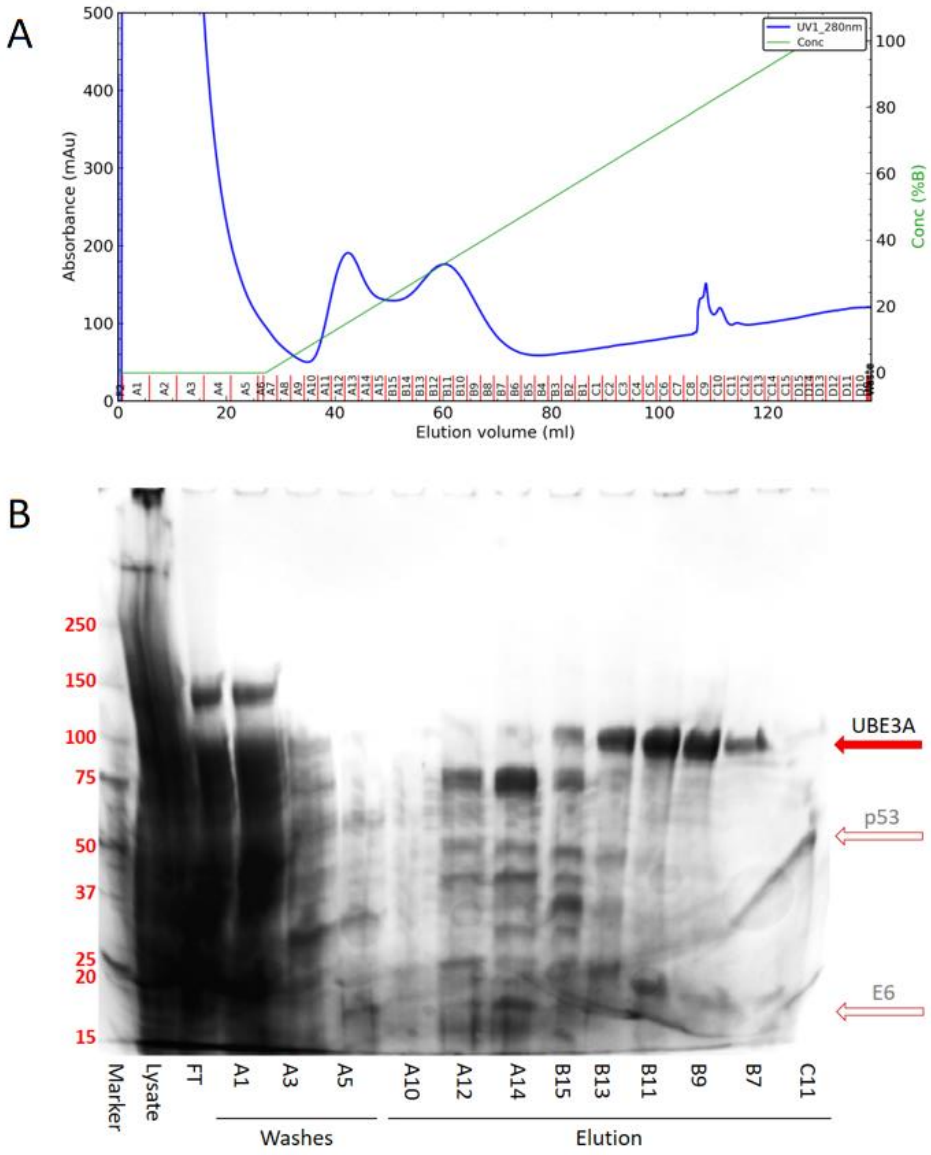


Figure 53: Co-expression of UBE3A isoform 1, p53, and E6 from three different plasmids, separated with a gradient HisTrap purification. A) The AKTA explorer trace for the run showing the absorbance at 280 nm (blue line), the concentration of high imidazole elution buffer (green line), and the collected fractions. The scale bar for the absorbance at 280 is shown on the left, whole the scale bar for the elution gradient is shown on the right. Fractions are indicated by the red dashes along the x-axis. B) SDS-PAGE analysis of key fractions from the run. The band for UBE3A is indicated by the red solid arrow, and the expected locations of p53 and E6 are indicated by the hollow red arrows. The samples were run on a 4-20% tris-glycine SDS-PAGE gel and stained with a Coomassie-based dye.

Unfortunately, neither cell-type showed a clear sign of complex formation. I hypothesised that the lack of expression at this point was due to the previously identified issues with a tandem promoter system. Although I had adjusted my constructs so that each gene had its own plasmid, I was still using the Duet-1 plasmids, so it was possible that the inefficiency observed by Dr

Masuda remained. Dr Masuda kindly provided us with the exact plasmids used in the Masuda *et al.,* 2019 paper, which featured a deactivated second promoter region in each plasmid but retained the distinct origins of replication. Another key difference between these plasmids and my own plasmids was that the UBE3A isoform used by them was the slightly longer isoform 2 gene rather than the previously used isoform 1 construct. These three plasmids were transformed into both BL21 cells and ArcticExpress cells, but unfortunately neither were successful. As with the co-expression of my own individual UBE3A, E6, and p53 plasmids, the large-scale cultures were incubated at 16^{o}C from the offset, rather than decreasing the temperature after induction. However, when following this protocol the cells grew very slowly. After three days the induction point still had not been reached but the antibiotics were no longer stable so the culture was discarded. I attributed the lack of growth in the ArcticExpress cells to an overloading of the cells with plasmids, as the addition of the ArcticExpress chaperone-expressing plasmid brings the total number of plasmids present in the cell to four. This meant that four separate antibiotics were present during the cell culture, which puts much more strain on the cells despite the presence of the resistance genes. The failure of the BL21 cells, however, was due to mutations among my BL21 cell stocks resulting in both chloramphenicol and streptomycin/spectinomycin resistance, removing the selection pressures for the pCDF and pACYC plasmids. The BL21 cells used for the co-expression of UBE3A, E6, and p53 were not commercial cells, they were stocks that had been subjected to a process to render them resistance to two different strains of phage that had been identified in the RCaH facility. The phage resistance process was carried out by a member of the MPL team, and then I generated a stock of these cells using the calcium chloride method. However, the process to make the cells resistant to phage infection seemed to cause this unexpected antibiotic resistance within the cells. While the streptomycin/spectinomycin resistance could be attributed to a mutation within the cells, the development of chloramphenicol resistance is typically much more complicated. This observed chloramphenicol resistance is more likely the result of the inclusion of a separate plasmid encoding the chloramphenicol resistance gene at some point during the process. Unfortunately, the extra antibiotic resistance was not identified until much of the work in this area of the project had been carried out, and an earlier phage infection in my plain BL21 cell stocks left us with no viable alternative. The pressures of both Brexit and the Covid-19 pandemic resulted in an inability to acquire fresh BL21 cell stocks within the time constraints of this project, and so I was unable to take this area of work any further.

3.6 HERC2

The giant E3 ligase HERC2 (4834 amino acids in length) has been identified as a key binding partner of UBE3A and the interaction between the two proteins

has been studied previously (Kühnle *et al*., 2011), but there is still not much information about the residues involved or the physiological importance of the interaction. As HERC2 is a 4834 amino acid protein the full-length product cannot be expressed in a bacterial system. As such, the *HERC2* gene was cloned into a variant of the pOpinENeo plasmid for transfection into HEK293 cells as part of a collaboration with the Protein Production UK (PPUK) group based at the Research Complex at Harwell (RCaH). However, due to the large size of the construct, at 14.5 kb for the HERC2 gene and a further ~8 kb for the vector, the process, as detailed below, was not straightforward.

3.6.1 Cloning

The pOpin vectors are proprietary vectors developed by PPUK (formerly Oxford Protein Production factory - OPPF) that are designed for cloning using the InFusion system, where two linear products are joined through recombination of an overlapping region at each end of the constructs (see section 2.4.4). PPUK routinely use the ClonExpress II kit for quick and easy insertion of a range of genes into a variety of pOpin vectors, making it easy to generate constructs with different tags and cleavage combinations for use across bacterial, insect, and mammalian cell expression systems. Unfortunately, at 14.5 kb, the HERC2 gene exceeded the manufacturer's recommendation of the maximum DNA fragment size of the enzymes in the kit, and necessitated a more thorough troubleshooting regime than typical. I first used the InFusion kit and attempted to optimise the reaction as suggested by the kit, but the desired product was over double the maximum length recommended by the kit and this ultimately led to no product being produced.

I next reverted to a classic restriction enzyme digest technique, as I already had HERC2 in a vector with compatible restriction sites to the pOpinENeo vector. Unfortunately, the first attempt at this method and subsequent sequencing of the HERC2 gene revealed the presence of seven internal *NcoI* sites within the HERC2 gene, while *NcoI* was the only appropriate 3' cloning site within the pOpinENeo vector that would allow expression of the target gene with the required tags in the correct reading frame. A partial digest was attempted in an attempt to overcome the issue of the internal *NcoI* sites, but this was ultimately unsuccessful.

The next attempt at cloning HERC2 into pOpinENeo involved the use of two new kits with similar principles to the InFusion kit, but with different overlap requirements. The two kits were Thermo Scientific's HiFi assembly kit and NEB's Gibson assembly kit (see sections 2.4.5 and 2.3.4). Both were designed for the assembly of multiple smaller fragments to generate a single, circular plasmid, but they both advertised efficiency to a larger size than the InFusion kit so both were trialled. As neither kit was designed specifically for the task of dealing with a fragment as large as the HERC2 gene both reactions required a significant amount of trial and error and troubleshooting at various steps, but

eventually the Invitrogen Gibson Assembly kit was successful in generating the 22.5 kb pOpinENEo-HERC2 construct.

Following positive identification of the required plasmid construct through antibiotic plate selection and subsequent colony PCR screening, as described in section 2.4.6, conclusive confirmation that the HERC2 gene was present, in-frame, and intact through Sanger sequencing was a further challenge. The large size of the HERC2 gene necessitated a significant number of sequencing reactions to cover the whole gene even in one direction, but additional complications resulted from the presence of very GC-rich regions and areas with a high degree of predicted secondary structure. The final accepted level of coverage involved the use of 31 individual sequencing primers and multiple sequencing reactions run for several sections due to different reaction conditions for high GC areas, and resulted in mapping of all but 132 of the 14,502 base pairs (Fig. 54). Ideally I would have worked for full coverage, but given the challenges associated with the HERC2 gene I decided that 99.1% error-free coverage was sufficient to progress with the construct as it was.

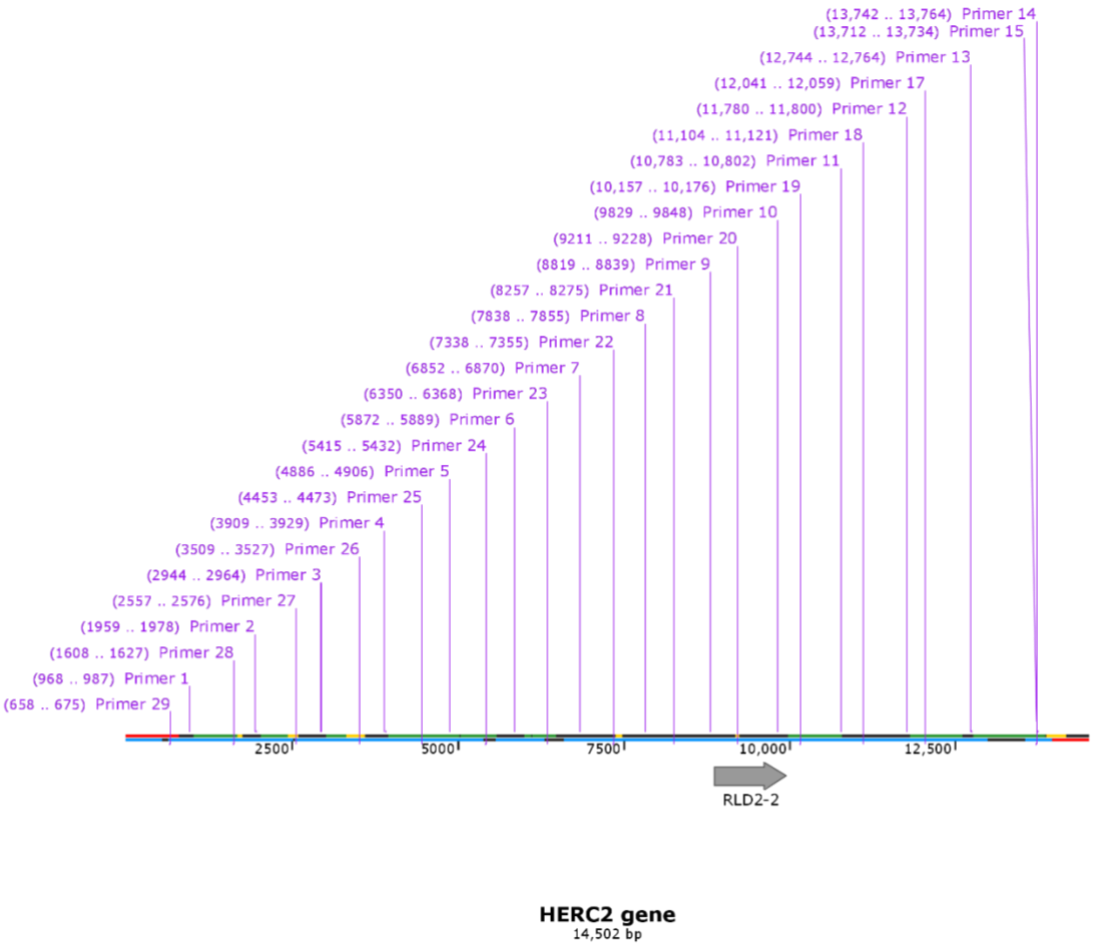


Figure 54: The coverage of the sequencing data for the HERC2 gene along with the placements of the designed primers for the gene. The red sections show the sequencing data from the standard forward and reverse sequencing primers for the pOpinE-Neo plasmid, the yellow, green and blue sections are sections covered by the HERC2-specific primers from different iterations of sequencing. Primers 1-14 cover the forward direction while primers 15-29 cover the reverse. 99.1% of the gene is covered in at least one direction by the provided sequencing data, but 132 bases, 6481-6612, are not sequenced. Figure created using Snapgene. The construct contains a C-terminal EGFP-His-Strep tag encoded by the plasmid that was not sequenced, but the C-terminal plasmid sequencing primer anneals to the N-terminus of the GFP sequence, confirming its presence.

Once the pOpinENeo-HERC2 construct was created, a Qiagen Plasmid Plus Midi-prep kit (section 2.3.4) was used to obtain sufficient HERC2 DNA for transient mammalian cell expressions. The DNA was passed on to PPUK for an initial transient expression in 3 ml HEK293 cells to produce a HERC2-TEV-His-Strep-GFP construct, with the tags at the C-terminus of HERC2 in order to allow observation of full transcript expression within cells by following the GFP fluorescence.

3.6.2 Affinity Chromatography Purification and Optimisation of Gel Electrophoresis Detection

The presence of the His-Strep-GFP tag on the HERC2 construct allowed three different affinity chromatography purification techniques to be tested. HERC2 was initially grown as a 3 ml transient expression, (with a 50 ml expression trialled later) and the cells were harvested and lysed by the PPUK group. The clarified lysates were then split into equal sections and subjected to gravity purifications using Talon resin, strepTactinXT resin, and a GFP-nanobody immobilised on strepTactinXT resin (see section 2.6.8). This allowed us to ensure that if any one tag was precluded in some way the HERC2 product could still be isolated. However, the real issue with the purification of expressed HERC2 was the visualisation of the process. Typically, fractions from an affinity purification are collected and subjected to a 4-20% gradient tris-glycine acrylamide gel, but this is only optimised for proteins between 10-250 kDa, and the HERC2 construct was predicted to be almost 600 kDa. In order to follow the process of the affinity purifications many different gel electrophoresis methods were attempted with varying levels of success (Appendix 8).

Across the range of protein gels used in an attempt to visualise the potential HERC2 product, several produced promising and yet ambiguous results. In the different gels some showed protein species of high molecular weight, either stuck in the wells or in the interface between the stacking gel and resolving gel (Fig. 55).

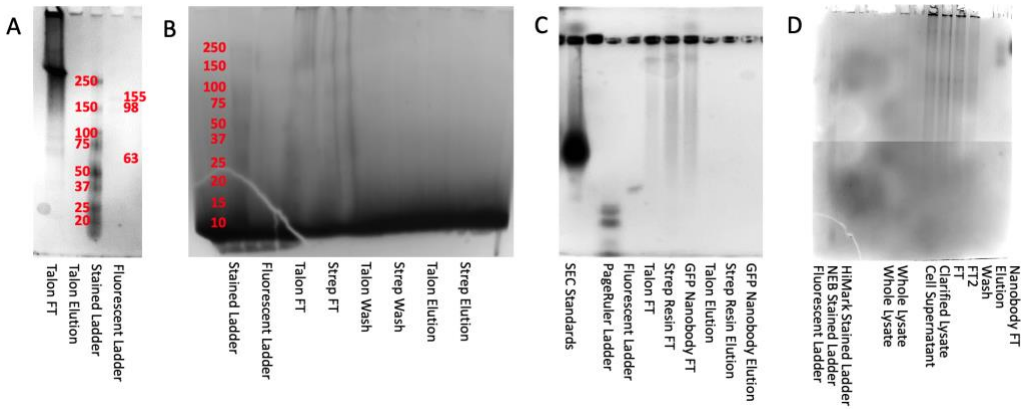


Figure 55: Electrophoresis gels of different types demonstrating high molecular weight species within a HERC2 purification sample. A) A commercial 4-20% PAGE gel run with native buffers and samples from a small-scale HERC2 Talon purification. B) A commercial blue-native PAGE gel with samples from a small-scale purification trial of HERC2. C) A 1% horizontal agarose gel with a native tris-borate buffer system, run with samples from a small-scale purification of HERC2. D) A large 9% PAGE gel with a tris-acetate gel composition and a native tris-tricine buffer.

Others showed bands at unexpected molecular weights, but with fluorescent properties that suggested the presence of the GFP tag. Since the GFP tag was

C-terminal in the HERC2 construct it is unlikely that the cells had expressed the GFP moiety without the attached HERC2 protein, and the sequencing results appeared to confirm that the gene was in frame with the tags, so the presence of fluorescence suggested that the construct was present (Fig. 56).

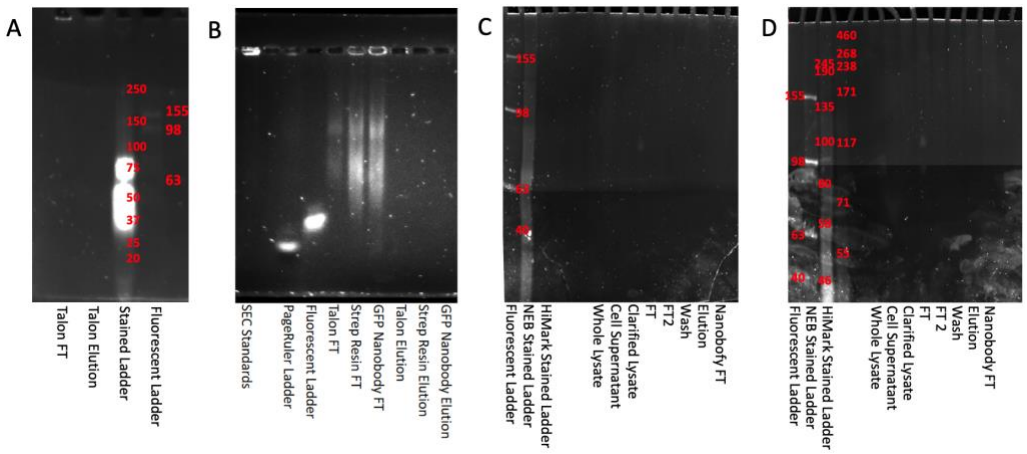


Figure 56: Various electrophoresis gels showing high molecular weight fluorescent products within HERC2 purification samples. A) A commercial 4-20% PAGE gel ran under native conditions, with a fluorescent sample stuck in the well of one lane. B) A 1% horizontal agarose gel with a native tris-borate buffer system. Fluorescent bands can be seen in the gel above the range of the markers, and also some fluorescing sample can be seen remaining in the wells of multiple lanes. C) A large 9% PAGE gel with a tris-acetate gel composition and a denaturing tris-tricine buffer.

However, none of the gels showed any unambiguous evidence of the GFP-tagged HERC2 construct, so the true level of expression is still unclear. Further work to improve the yields from the mammalian cell expression protocol may result in a less ambiguous result, but due to the high level of complexity of every stage of HERC2 development thus far, and limitations on room capacities due to Covid19, further optimisation was not possible within the time constraints of this project.

3.6.3 Identification through GFP Fluorescence

As the HERC2 construct featured a C-terminal GFP tag, theoretically, the presence of full-length HERC2 should co-localise with any GFP fluorescence in the cells. In one attempt to utilise this to identify protein expression without the need to visualise the result on an SDS-PAGE gel the clarified cell lysate from a 3 ml transient expression was subjected to a fluorescent-SEC (FSEC) run. This utilises the same principles as standard SEC methods, but rather than following the full A280 absorbance reading, the fluorescence measurements are recorded as the proteins travel through the column. This results in a trace similar to a SEC trace, but it only shows the presence of any fluorescent molecules within the cell lysate, which in this case should be the GFP-tagged HERC2 protein only. Unfortunately, the FSEC trace for HERC2 was almost completely flat, and what little signal there was, at a much later elution point

than would be expected for HERC2, didn't produce any bands on either SDS or native PAGE.

Although a suitable FSEC trace result would have confirmed the presence of HERC2 through the identification of the size of the fluorescent eluate, a more robust method of identifying the presence of GFP in the sample is to simply measure the fluorescence of the cells themselves before cell harvesting and lysis. Another 3 ml transient expression of HERC2 was imaged with a fluorescent microscope and the results were somewhat encouraging although fairly ambiguous (Fig. 57).

Figure 57: Fluorescence scans of a small scale HERC2 expression test. There were no observable puncta of high concentration GFP-tagged construct, but there is a clear green tint to the solution.

Due to the large size of the HERC2 protein compared to the size of the GFP tag, the relative fluorescence of any expressed product would be expected to be low. However, the level of fluorescence observed in the 3 ml culture above is still lower than hoped, which suggests that more work is required to optimise the cell culture process to increase the yield of HERC2. However, due to the social distancing requirements in place due to the Covid-19 pandemic, the opportunity to be trained in cell culture techniques was not available. The collaboration with PPUK allowed for continued expression using the basic protocol alongside their own cell maintenance, but their increased workload due to their involvement in covid-19-related research prevented any optimisation of cell expression techniques from being performed.

4 Biophysical Characterisation

4.1 Sedimentation Velocity Analytical Ultracentrifugation

Analytical ultracentrifugation (AUC) is a biophysical, solution-based technique that exploits the effect of molecular mass and buoyancy of macromolecules on the speed of sedimentation under a high centrifugal force (Schuck, 2013). As a sample is centrifuged at high speed, sensors within the analytical ultracentrifuge track the sedimentation of the protein through both absorbance and interference measurements, and mathematical analysis based on rigorous thermodynamics determines a sedimentation coefficient for any large particle within the sample (Lamm, 1926). From this, molecular weights of components can be determined (Svedberg and Fåhraeus, 1926). If a protein is present in several oligomeric states within a solution, each state will sediment at a different speed and can therefore be identified and analysed during data analysis (Schuck, 2013).

4.1.1 UBE3A

Previous publications have suggested that UBE3A may function as a trimer (Ronchi et al., 2014) based on the elution of isoform 2 from a size exclusion gel filtration column, but this observation appears not to have been noted by any other group. The SEC trace for my isoform 1 UBE3A sample did suggest the presence of two distinct oligomeric states (Fig 46), but the proportion of the sample in the potential higher order species is small compared to the main peak, and the resolution of the column used is not adequate at that size range to conclusively determine the stoichiometry of the multimer. UBE3A was therefore subjected to sedimentation velocity analytical ultracentrifugation (SV-AUC), as described in section 2.9.1, in order to gain more of an insight into the molecular weights and any interacting dynamics of the species within the sample (Fig. 58).

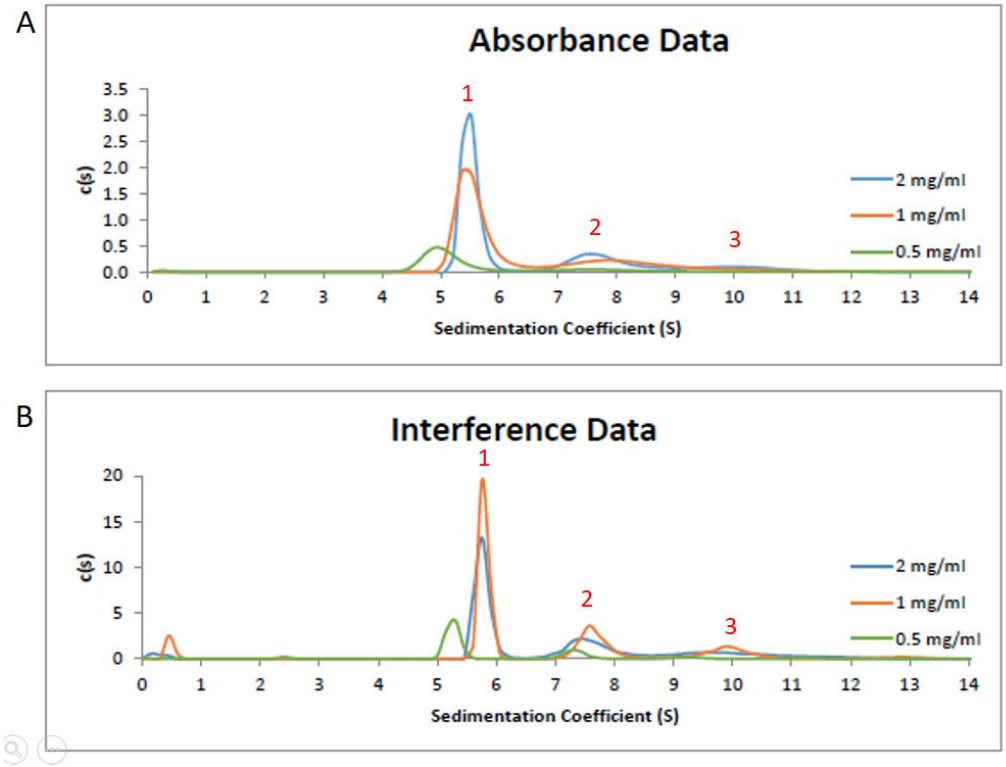


Figure 58: The SV-AUC distribution of species within a UBE3A sample. Both the absorbance and interference data for the run demonstrate the presence of at least two species with a possible third, indicated by the peak labels in red. The distribution between the species does not change across a concentration range, which suggests a stable ratio which is not affected by an increase in concentration.

The SV-AUC results appear to corroborate the findings from SEC that two different ordered species are present in the pure UBE3A sample. Although the previous publication suggests that the predominant multimeric state of UBE3A should be a trimer (Ronchi *et al.,* 2014) the calculation carried out by SedFit suggest a rough monomer/dimer ratio of the two species in peaks 1 and 2. However, this predicted molecular weight is not very reliable for a heterogeneous sample such as this as the frictional ratios of the two species will be different. The main peak is at 5.8 S. For the determined molecular weight of 98 kDa, this corresponds to a frictional coefficient of 1.3, which is a slightly extended conformation based on results from ULTRASCAN3 (Demeler, 2005).

4.1.2 UBE3A + PSMD4

As well as providing an insight into the oligomeric states within a sample, SV-AUC can provide information on the nature of protein-protein interactions involved in multiprotein complexes. UBE3A was mixed with PSMD4 in a range of stoichiometric ratios in order to see if the interaction between the two proteins was concentration dependent (Fig. 59).

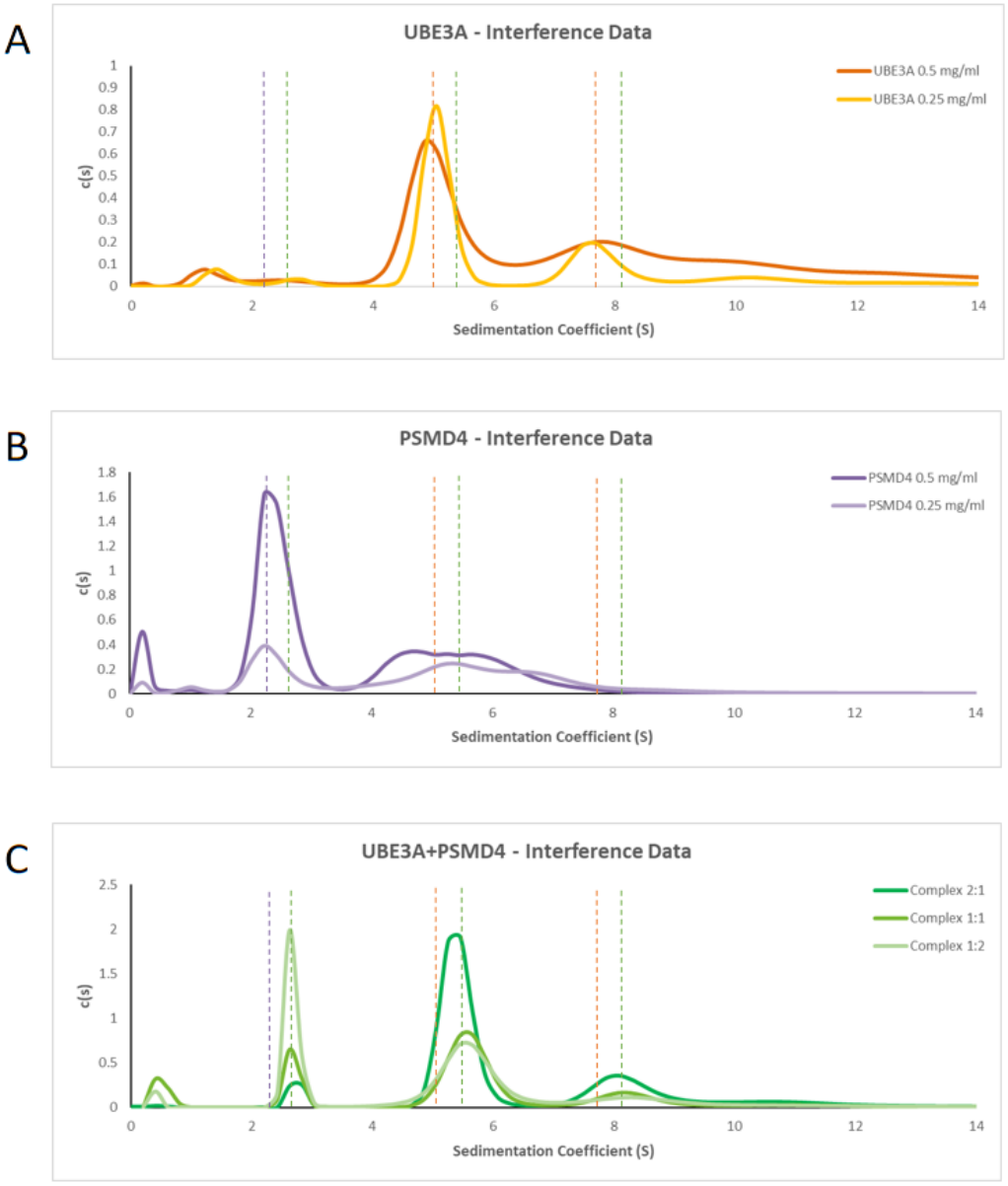


Figure 59: The SV-AUC distribution of a complex of UBE3A and PSMD4. A) The trace for UBE3A alone at two different concentrations. B) The trace for PSMD4 in two different concentrations. C) The trace for samples containing UBE3A and PSMD4 in different molar ratios. The ratios shown in the legend show the ratio of UBE3A:PSMD4 in each sample, with PSMD4 kept at 0.5 mg/ml throughout. The S values for each defined peak are shown by dashed lines in the colour of each dataset, showing the change in S values in the different samples.

The distribution of peaks in the UBE3A+PSMD4 samples (Fig. 59c) are similar to the peaks for UBE3A and PSMD4 individually (Fig. 59a and Fig. 59b), but they are offset by approximately one quarter of an S value. All seven samples shown in figure 59 were prepared concurrently and subjected to SV-AUC in the same run so any experimentally-derived offsets should be minimal, but the consistency of the shift between the complex samples and the individual samples would suggest that this is an artefact rather than a feature of the

complex. In addition, the distribution of the first two peaks of the complex traces shift with the ratio of UBE3A and PSMD4 in the sample, which suggests that they represent separate species of UBE3A and PSMD4 and no stable complex is present in the samples.

4.1.3 UBE3A + RLD2

The UBE3A + RLD2 sample was also subjected to SV-AUC as described in section 2.9.1. UBE3A and RLD2 were mixed in a range of stoichiometric ratios, but the concentrations of each sample were altered to show the stability of the complex across a concentration range (Fig. 60).

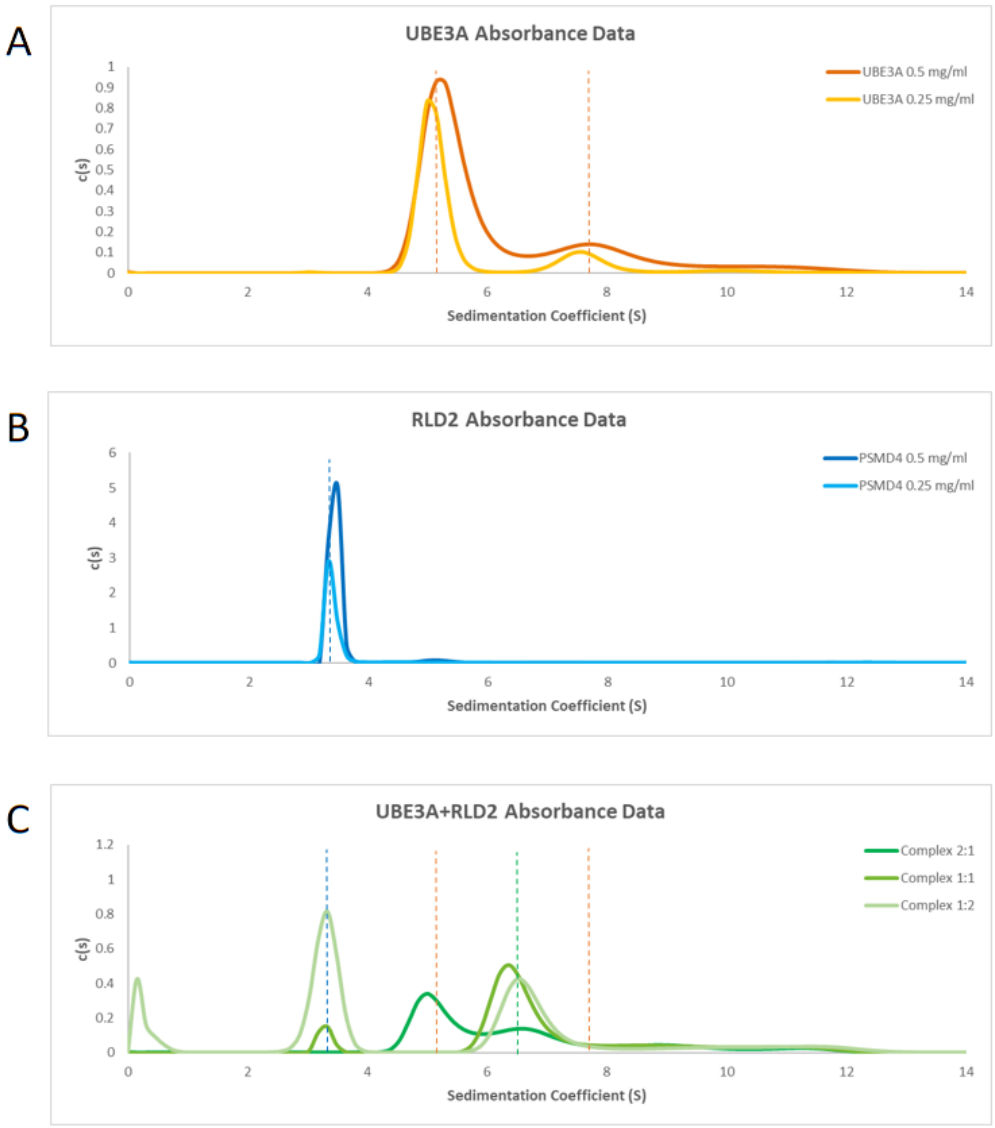


Figure 60: The SV-AUC distribution of a complex of UBE3A and RLD2. A) The trace for UBE3A alone at two different concentrations. B) The trace for RLD2 in two different concentrations. C) The trace for samples containing UBE3A and RLD2 in different molar ratios. The S values for each defined peak are shown by dashed lines in the colour of each dataset, showing the change in S values in the different samples.

The SV-AUC profiles for UBE3A, RLD2, and a UBE3A+RLD2 complex shows a clear 1:1 complex formation. UBE3A alone (Fig. 60a) produces two peaks, the monomer and probable dimer identified earlier (see section 3.4.1), while RLD2 alone (Fig. 60b) produces only a single peak at both concentrations. This suggests that it does not form any multimeric or aggregate species at the concentration range tested. When the two proteins are mixed in various ratios (Fig. 60c), a new species is formed that was not observed in either of the controls, at ~6-7S. This must be a stable UBE3A+RLD2 complex. When the proteins were mixed in a 2:1 excess of UBE3A (the dark green line in Fig. 60c), a clear peak of free-UBE3A is observed at 5 S, and when the proteins were mixed in an excess of RLD2 (the pale green line, Fig. 60c), a large monomeric RLD2 peak is observed. However, when the proteins were mixed in an equimolar ratio, there is no large excess peak for either protein, instead the majority of the sample forms the new UBE3A+RLD2 complex peak at around 7S. There is a small peak for monomeric RLD2 in the 1:1 complex sample, but this is likely to be due to either an inaccurate determination of the equimolar ratio when forming the sample, or it could also simply demonstrate the increased absorbance ability of RLD2 compared to UBE3A.

4.2 Isothermal Titration Calorimetry

Isothermal titration calorimetry (ITC) is a biophysical technique that allows an insight into the thermodynamic properties of an interaction between two species. It uses the thermal properties of the reaction, either exothermic or endothermic, as a measure of the reaction process. The instrumentation is comprised of two cells, one that contains the reaction process and another that acts as a reference cell. As the reaction progresses and the resulting thermodynamic effects occur, current is applied to maintain the temperature of the sample cell at that of the reference cell. The current that is required to maintain the temperature of the cell as the reaction progresses is used as a measure of the energy involved in the interaction. (Velazquez-Campoy *et al*., 2015)

4.2.1 UBE3A + PSMD4

UBE3A and PSMD4 were subjected to ITC as described in section 2.9.2 in a low salt buffer, with PSMD4 in the cell and a high concentration sample of UBE3A in the syringe. The data was analysed and plotted as described in section 2.9.2 (Fig. 61).

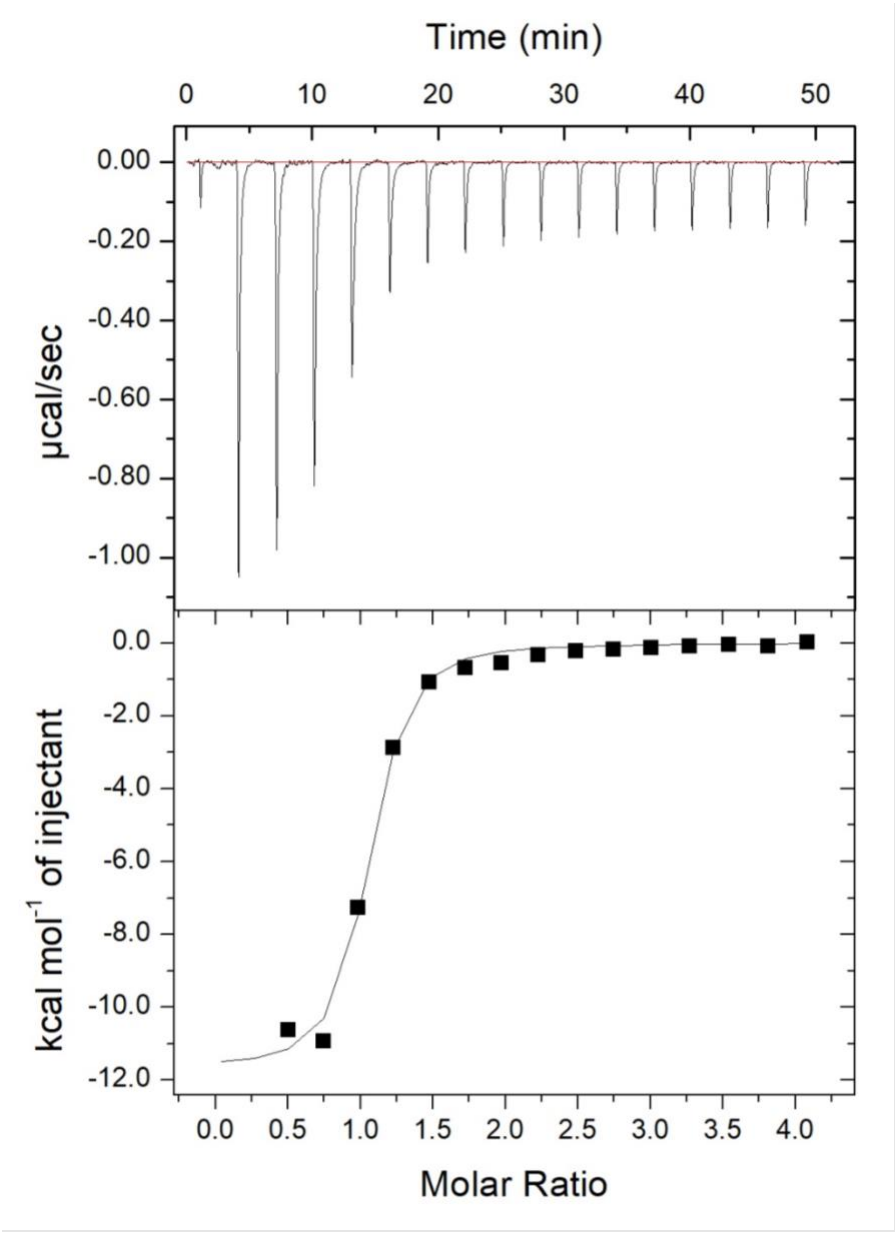


Figure 61: The ITC isotherm (upper) and integrated heat-per-injection plot for the interaction between UBE3A and PSMD4. UBE3A was injected into a PSMD4 sample resulting in the above trace, showing a tight interaction between the two species. The data was processed using the Origin ITC200 software to generate both the above figure and the thermodynamic features of the reaction.

For the interaction between UBE3A and PSMD4 the stoichiometry was calculated as 0.966 ±0.0156, the K_d was determined to be 340 nM, the ΔH was determined as -1.169×10^4 ±332.1 cal/mol, and the resulting ΔS was -9.62 cal/mol/deg. The stoichiometry refers to the number of PSMD4 moiety interacting with a single UBE3A moiety, in this case a single PSMD4 appears to bind to a single UBE3A, forming a heterodimer with a 1:1 ratio. The ΔH value describes the enthalpy of the interaction, defined as the total heat energy of a system. The negative sign of the ΔH value shows that the reaction in exothermic, meaning that energy is used up in the reaction. This suggests that

more new bonds are being created than pre-existing bonds are being broken. These parameters can be combined to derive the ΔS (section 2.9.2), the entropy of the reaction, a measure of the disorder in the system. The negative ΔS value derived for the UBE3A+PSMD4 reactions suggests that the system is becoming less disordered, which is indicative of complex formation.

The K_d value is one of the key parameters that can be derived from an ITC experiment. The K_d of a reaction is the dissociation constant, defined as yhe concentration of a ligand at which half of the ligand binding sites on the protein are occupied at equilibrium. A lower K_d value denotes a higher affinity for an interaction, meaning a tighter binding. For the UBE3A+RLD2 interaction the K_d was calculated to be in the nanomolar range, which suggests a fairly tight binding interaction and suggests that the complex could be a good target for cryo-EM.

4.2.2 UBE3A + RLD2

The AUC results for the UBE3A + RLD2 interaction suggest a 1:1 ratio, so ITC was used to both confirm this observation and to determine the thermodynamics of the reaction. The reaction was prepared as described in section 2.9.2, with UBE3A in the syringe and RLD2 in the cell (Fig. 62).

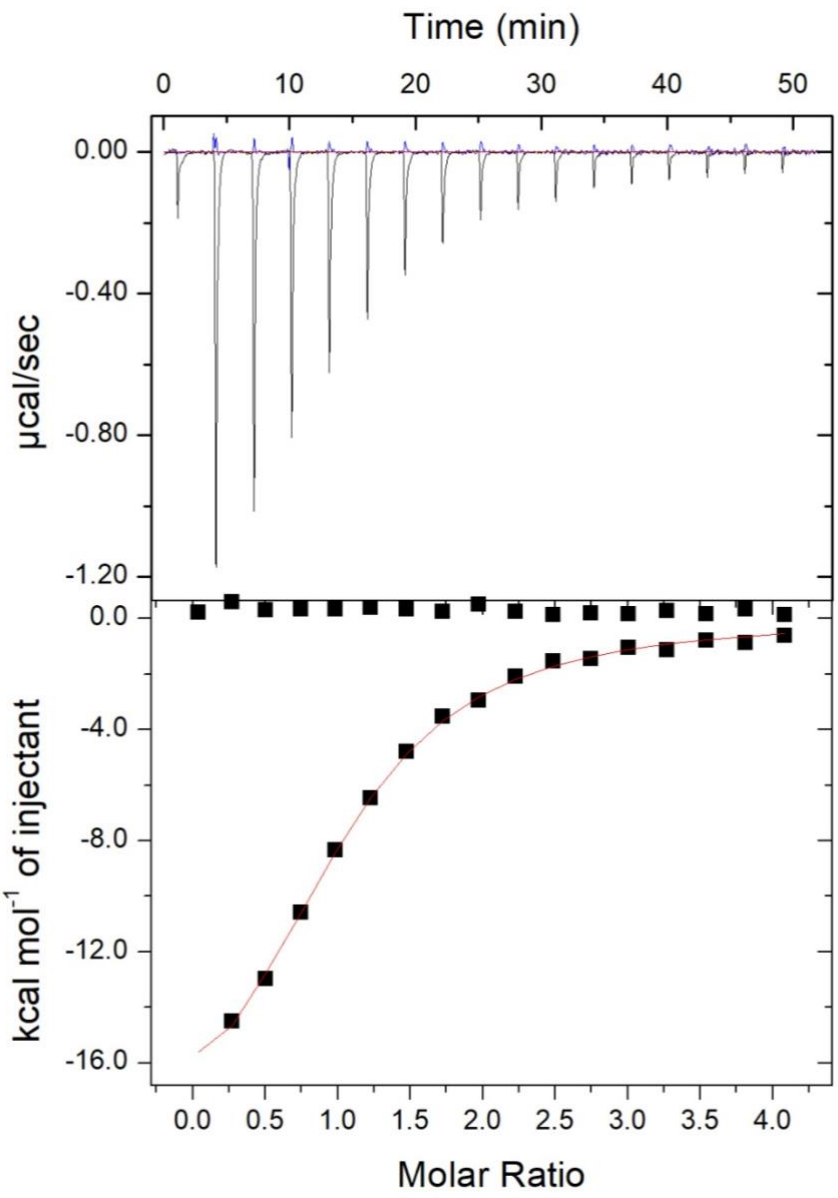


Figure 62: The ITC isotherm (upper) and integrated heat-per-injection plot for the interaction between UBE3A and RLD2. UBE3A was injected into an RLD2 sample resulting in the above trace, showing the interaction between the two species. The data was processed using the Origin ITC200 software to generate both the above figure and the thermodynamic features of the reaction.

The stoichiometry of the reaction was determined as 0.991 ±0.0206, the K_a was calculated as $1.56x10^5$ (K_d = 6.4 µM) ±$9.57x10^3$ M^{-1}, the ΔH was -$2.082x10^4$ ±573.7 cal/mol, and the resulting entropy (ΔS) of the reaction was determined as -46.1 cal/mol/deg. The stoichiometry suggests a 1:1 ratio, similar to the UBE3A+PSMD4 reaction (section 4.2.1). The enthalpy change of the reaction (ΔH) is negative, suggesting more bonds being formed than broken, and the entropy change (ΔS) is strongly negative, suggesting that the products are ben=coming more ordered as they interact. This is indicative of a complex formation.

Both the ΔS and ΔH values for the UBE3A+RLD2 interaction have a greater magnitude than the UBE3A+PSMD4 interaction, particularly the ΔH, which would suggest that more bonds are being formed and the components are forming a more stable complex, but the K_d of the UBE3A+RLD2 interaction is in the micromolar range, which suggests a relatively weak binding. It is in the low micromolar range, so it is still a reasonable interaction and is still a promising target for cryo-EM, although the ITC data would suggest that PSMD4 interacts more strongly with UBE3A than RLD2.

4.2.3 Ufrag + RLD2

The interaction between UBE3A and HERC2 has been narrowed down to an interaction between the RLD2 domain of HERC2 and 50 amino acids of UBE3A spanning residues 150-200 (Kühnle *et al.,* 2011). This region is distal to the catalytic HECT domain of UBE3A, so in order to have an effect on the ubiquitination activity of UBE3A it must either cause a large rearrangement of the internal UBE3A structure, or other areas of UBE3A must also be involved in the interaction. The 50 amino acid interaction region, termed Ufrag, was expressed and purified as an MBP-fusion protein. The MBP-tag was retained during subsequent biophysical analysis since the isolated Ufrag construct alone was difficult to identify (see section 3.2.5). I performed ITC experiments using the MBP-Ufrag species with His-RLD2 to identify whether the thermodynamics of the interaction were the same as those of the full-length UBE3A + RLD2 reaction (Fig. 63).

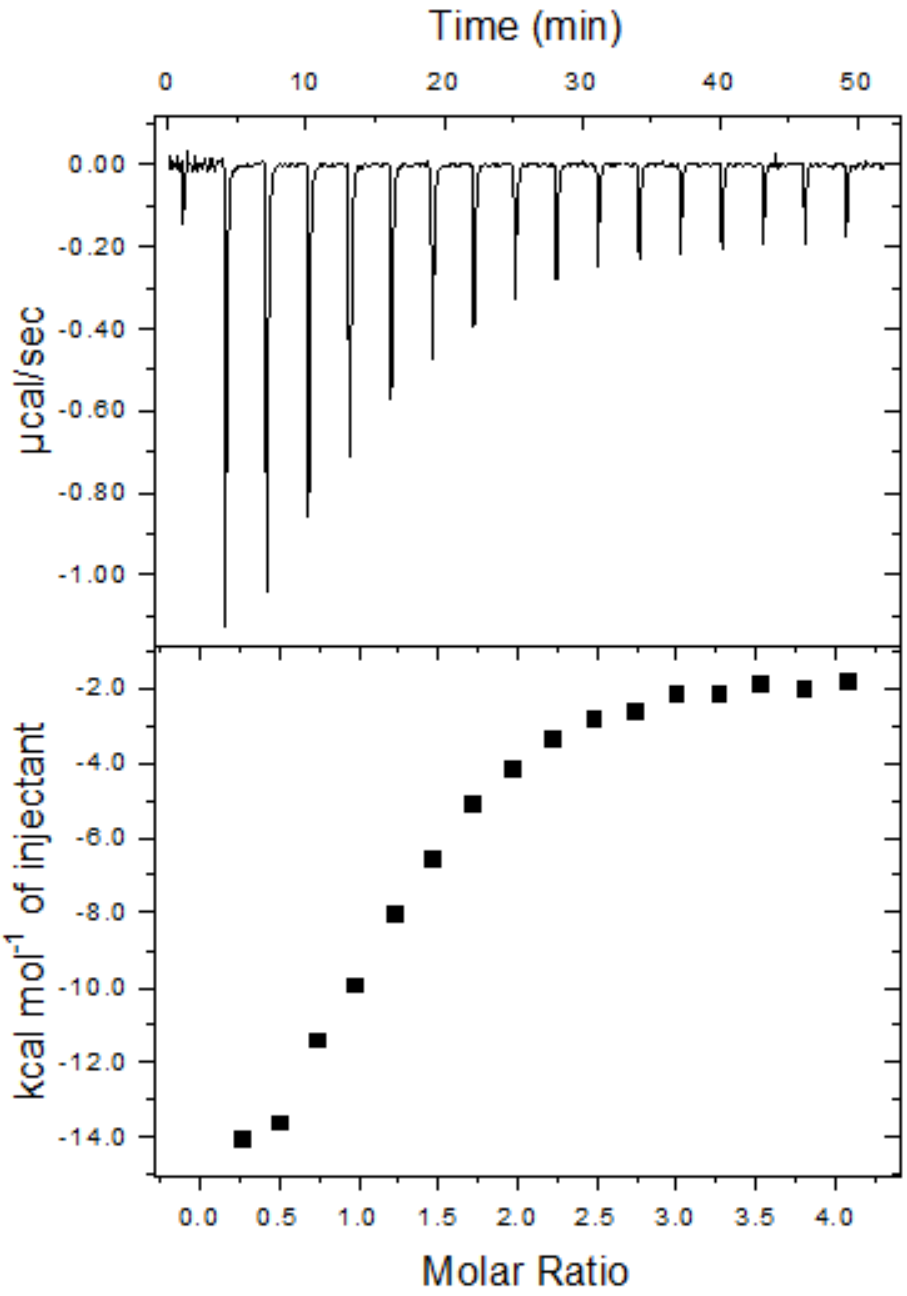


Figure 63: The ITC isotherm (upper) and integrated heat-per-injection plot for the interaction between MBP-Ufrag and His-RLD2. His-RLD2 was injected into an MBP-Ufrag sample resulting in the above trace, showing a tight interaction between the two species. The data was processed using the Origin ITC200 software to generate both the above figure and the thermodynamic features of the reaction.

The calculated stoichiometry of the interaction (n) was 1.20 ±0.0955, the K_a value was $8.34x10^4$ ±$1.54x10^4$ M^{-1},(Kd = 11.99 µM) the ΔH was determined as -$2.353x10^4$ ±2521 cal/mol, and the resulting ΔS was -56.4 cal/mol/deg. The calculated stoichiometry for the MBP-Ufrag+RLD2 interaction is slightly higher than the 1:1 ratio observed for UBE3A+RLD2 in both the ITC (section 4.2.2) and AUC (section 4.1.3) data. This could mean that it has a different mode of

binding than the full UBE3A+RLD2, or it could just mean that the data is less reliable than previous experiments. The ΔH and ΔS values are similar to those of the UBE3A+RLD2 interaction (section 4.2.2), which would sugest a similar binding mechanism. The K_d value for Ufrag+RLD2 also similar to the K_d value for UBE3A+RLD2 so the strength of the binding is similar, but it is not quite the same. The K_d of Ufrag+RLD2 is slightly higher than that of full-length UBE3A, so it is possible that although the Ufrag region is responsible for a key part of the interaction, other areas of UBE3A are also involved in coordinating binding.

In order to ensure that the observed interaction is due to binding of Ufrag to RLD2 rather than the MBP tag and RLD2, the MBP protein was expressed and purified without any conjugated products. I then repeated, as a control, the ITC experiment with just MBP and RLD2 to see if any interaction was observed (Fig. 64).

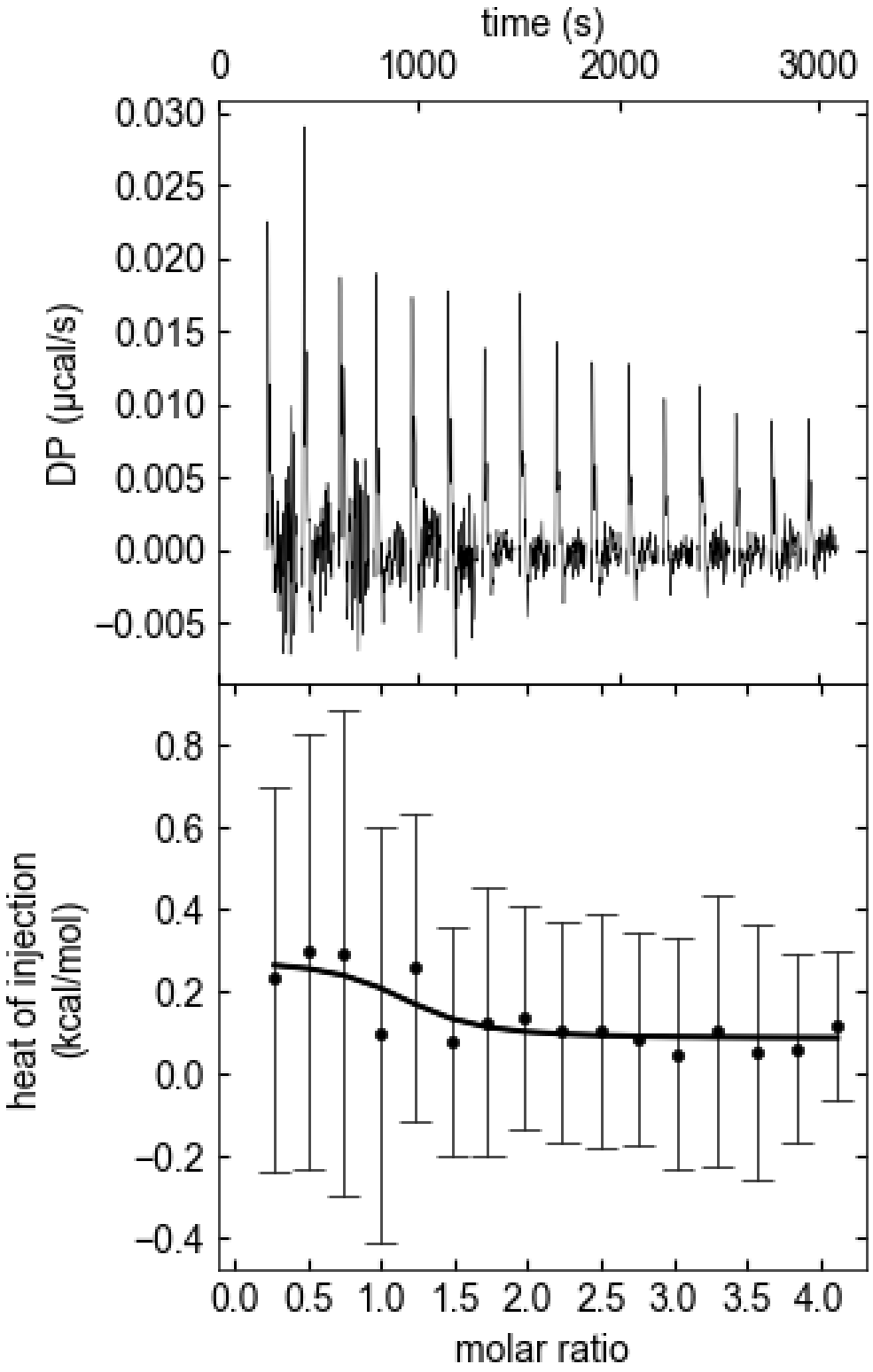


Figure 64: The ITC isotherm (upper) and integrated heat-per-injection plot for the interaction between MBP and His-RLD2. His-RLD2 was injected into an MBP sample resulting in the above trace, showing no significant interaction between the two species. The data was processed using the SedPhat software to generate both the above figure and the thermodynamic features of the reaction.

The calculated K_d value was 1.171 uM, the ΔH was determined as 0.21 kcal/mol, and the stoichiometry was 1.103. The MBP + RLD2 run did produce more of a trace than you would expect from a blank run, but the amplitude of the spikes was much lower than you would expect from a typical protein-protein interaction. The small effect observed upon mixing MBP and RLD2 was more suggestive of a dilution effect, and it does not nullify the reaction observed upon mixing of the Ufrag construct and RLD2. However, the lack of an interaction between MBP and RLD2 still does not preclude the possibility that the MBP tag prevented the full interaction between Ufrag and RLD2.

4.3 Complex Formation Through Co-Purification

Although each protein was purified separately (see chapter 3) in order to map the interactions as they occur through biophysical techniques, co-purification techniques can also be informative regarding protein interactions. If a protein interaction is strong enough to survive chromatography techniques then it is likely to be a significant interaction in cells, and on-column complex formation can be a quick way to determine if the presence of a purification tag has an effect on the interaction before significant time is spent on separate purifications.

4.3.1 UBE3A+PSMD4

UBE3A and PSMD4 were both purified separately, as described throughout chapter 3, before being mixed in a 3:1 molar excess of RLD2. The mixture was passed through an S200 column, as described in section 2.6.7, and the traces for the complex sample and the UBE3A only sample were superimposed in order to compare the profiles (Fig. 65).

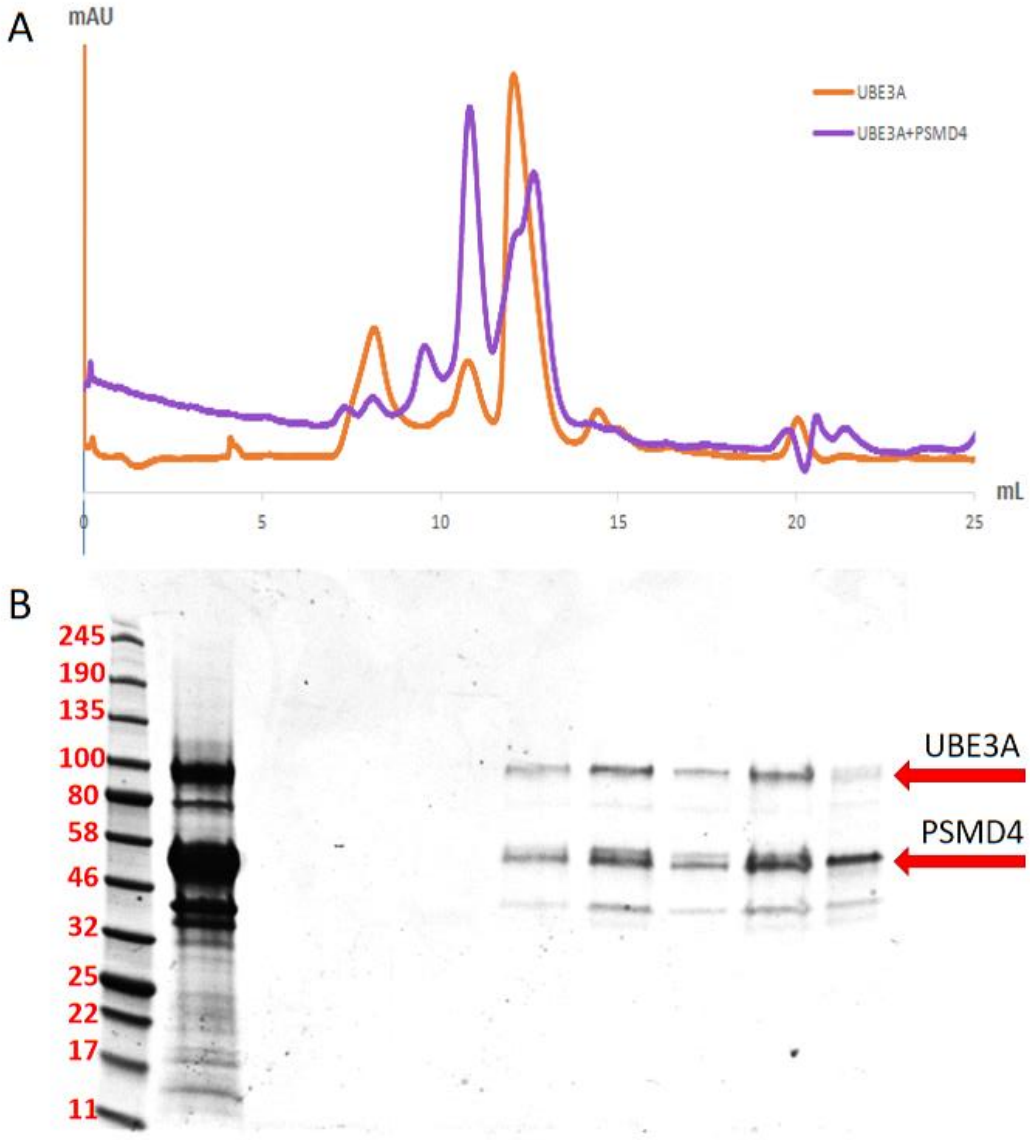


Figure 65: The association of UBE3A and PSMD4 visualised through SEC. A) The A280 trace for size exclusion of the UBE3A+PSMD4 complex, shown in purple, is superimposed over the trace for the sample of UBE3A only, shown in orange. B) The fractions from the UBE3A+PSMD4 SEC run were subjected to analysis by SDS-PAGE and staining with a Coomassie-based dye.

Upon binding to PSMD4, there is a substantial shift in the A280 profile compared to the profile for UBE3A alone on the same column. The peak for the monomeric UBE3A sample decreases with a concurrent increase in the peak just prior. The increase in this higher molecular weight peak in the complex sample but not in a sample of UBE3A at a higher concentration (data not shown) suggests that it is a result of the formation of a 150 kDa complex of UBE3A+PSMD4. SDS-PAGE analysis suggests a 1:1 ratio of PSMD4 and UBE3A in the fraction sample, which is supported by the observation of a 1:1 interaction stoichiometry upon ITC analysis (section 4.2.1). The presence of this complex following SEC suggests that the interaction is highly stable.

4.3.2 UBE3A+RLD2

UBE3A and RLD2 were also expressed and purified separately (see chapter 3) before being mixed in a molar ratio of 1:3 with excess RLD2. This mix was subjected to size exclusion as described in section 2.6.7, and the resulting fractions were run through SDS-PAGE (Fig. 66).

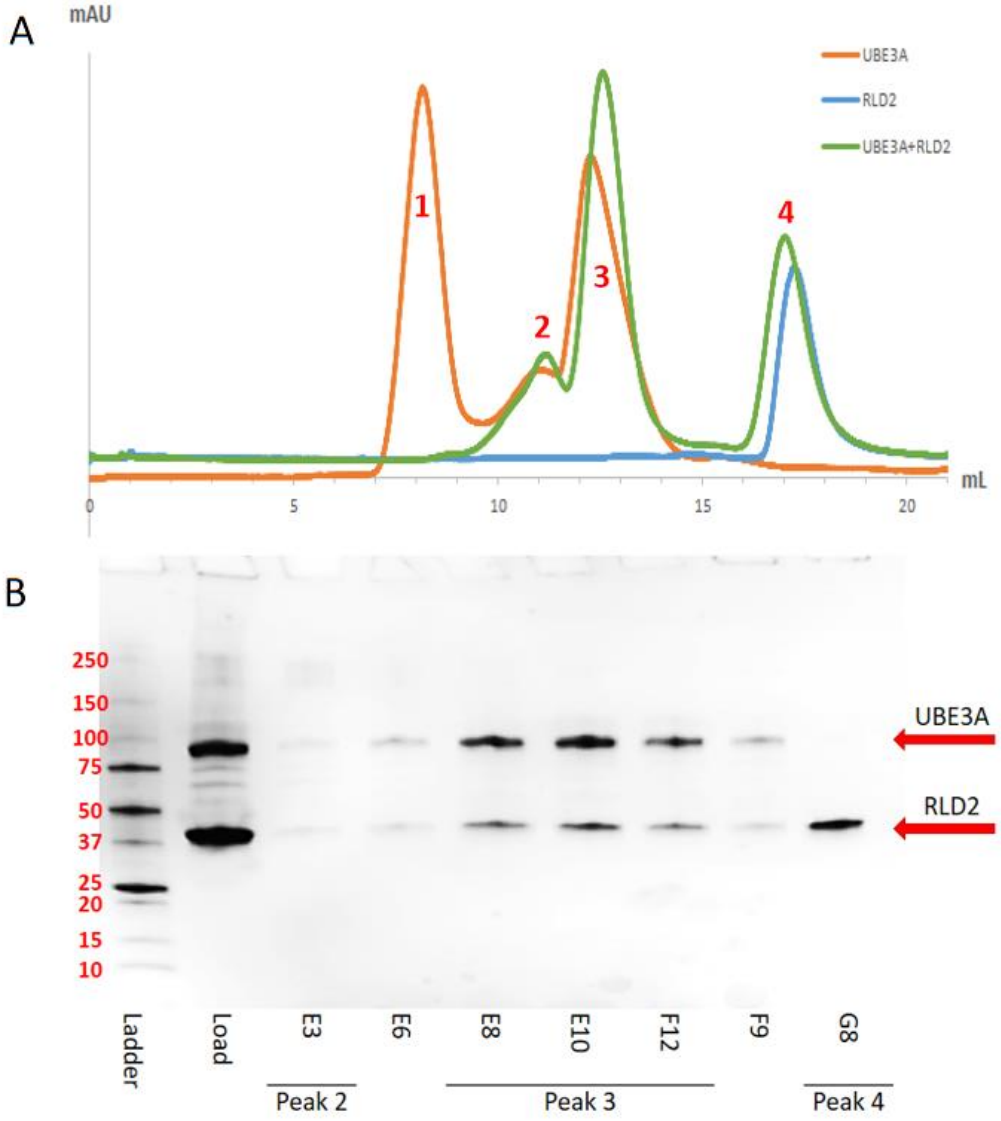


Figure 66: Size exclusion of a complex of UBE3A and RLD2. A) The absorbance at A280 for UBE3A alone (orange line), RLD2 alone (blue line), and UBE3A+RLD2 (green line). B) Fractions from the UBE3A+RLD2 size exclusion run were subjected to SDS-PAGE and stained with a Coomassie-based dye. The expected molecular weights of both UBE3A and RLD2 are shown by the red arrows to the right of the image.

Despite the extra molecular weight of RLD2, the complex of UBE3A+RLD2 appeared to elute slightly later than UBE3A alone, which would suggest a smaller molecular weight. However, when the fractions from that peak were run on SDS-PAGE it showed an equimolar ratio of both proteins. One possibility is that the binding of RLD2 to UBE3A changes the shape of the

protein to a more compact conformation, which allows it to run through the size exclusion column as if it were a much smaller species than it is.

4.3.3 Ufrag+RLD2

The identified region of UBE3A involved in the interaction with HERC2 is only 50 amino acids (Kühnle *et al.,* 2011), which results in a peptide of roughly 6 kDa. This would make it difficult to see on SDS-PAGE during the purification process. In order to be able to easily observe the interaction between the Ufrag peptide and RLD2, the MBP tag that was used to solubilise the fragment during expression was kept attached during the purifications and generation of the complex. The MBP tag could then be removed and easily separated from the preformed Ufrag+RLD2 complex.

The His-RLD2 and MBP-Ufrag fusion proteins were purified as described in 3.2.3 and 3.2.4, mixed in an equimolar ratio, and subjected to subsequent MBPTrap and HisTrap purifications and SDS-PAGE as described in sections 2.6.3, 2.6.5, and 2.7.1 (Fig. 67).

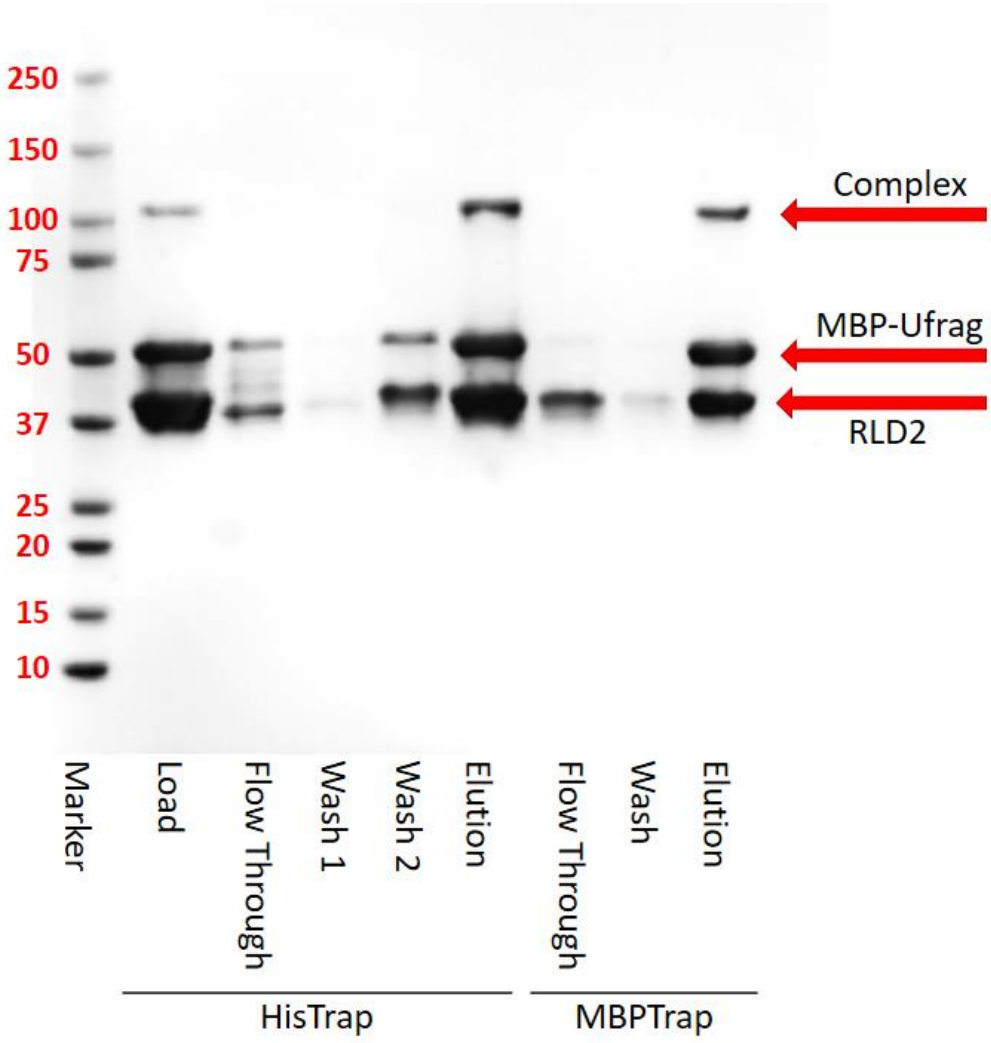


Figure 67: Purification of a complex of His-RLD2 and MBP-Ufrag through consecutive affinity chromatography purifications. The mixture was subjected to a HisTrap purification to remove any unbound MBP-Ufrag in the sample, and the elution from the HisTrap was put straight through an MBPTrap column to remove any unbound His-RLD2. The expected molecular weights for the components of the complex are shown by the red arrows on the right.

The co-purification of MBP-Ufrag and His-RLD2 shows that there is a strong interaction between the two species, but there is still the possibility of unbound constituents which necessitates the two-step affinity purification. The presence of the band around 100 kDa suggests that some of the complex has even remained intact following preparation for SDS-PAGE. This seems unlikely as the samples are heated to 95°C for several minutes in a solution containing SDS, but this is supported by a decrease in molecular weight of this band following cleavage of the MBP tag (Fig. 68)

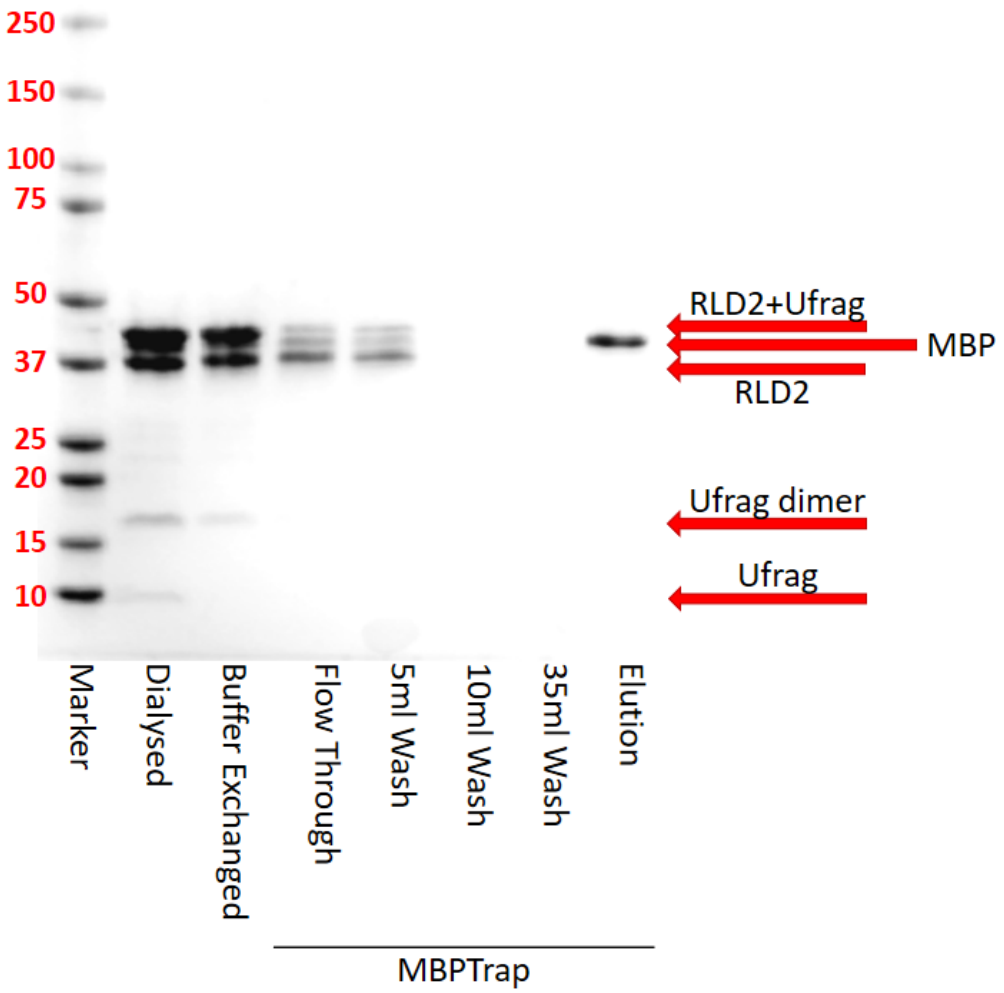


Figure 68: A reverse MBPTrap purification of the cleaved RLD2+Ufrag complex. The cleaved complex sample was subjected to buffer exchange to reduce any maltose in the sample from the initial MBPTrap, and then the sample was subjected to a reverse MBPTrap. The expected molecular weights for the various species are shown by the red arrows on the right.

The complex sample formed by a two-step affinity chromatography purification was cleaved simultaneously with dialysis as described in section 2.6.4, but due to the high affinity of maltose for the MBPTrap column it was subjected to further buffer exchange to reduce the maltose in the sample down to 1.25 µM. The sample was then subjected to a reverse MBPTrap purification in an attempt to separate the free-MBP from the cleaved complex. Despite the low level of maltose in the sample buffer, some free-MBP was still present in the flow through samples mixed in with the complex. The samples that contained both complex and free-MBP were pooled, subjected to another round of dialysis to reduce the maltose to approximately 6 nM, and then put back through the MBPTrap to isolate the RLD2+Ufrag complex.

Although the complex could be cleaved, the uncleaved sample was cleaner and produced a higher yield so was still used for various experiments. The complex of MBP-tagged Ufrag and His-tagged RLD2 following the dual affinity co-purification was concentrated and subjected to size exclusion chromatography as described in section 2.6.7 (Fig. 69).

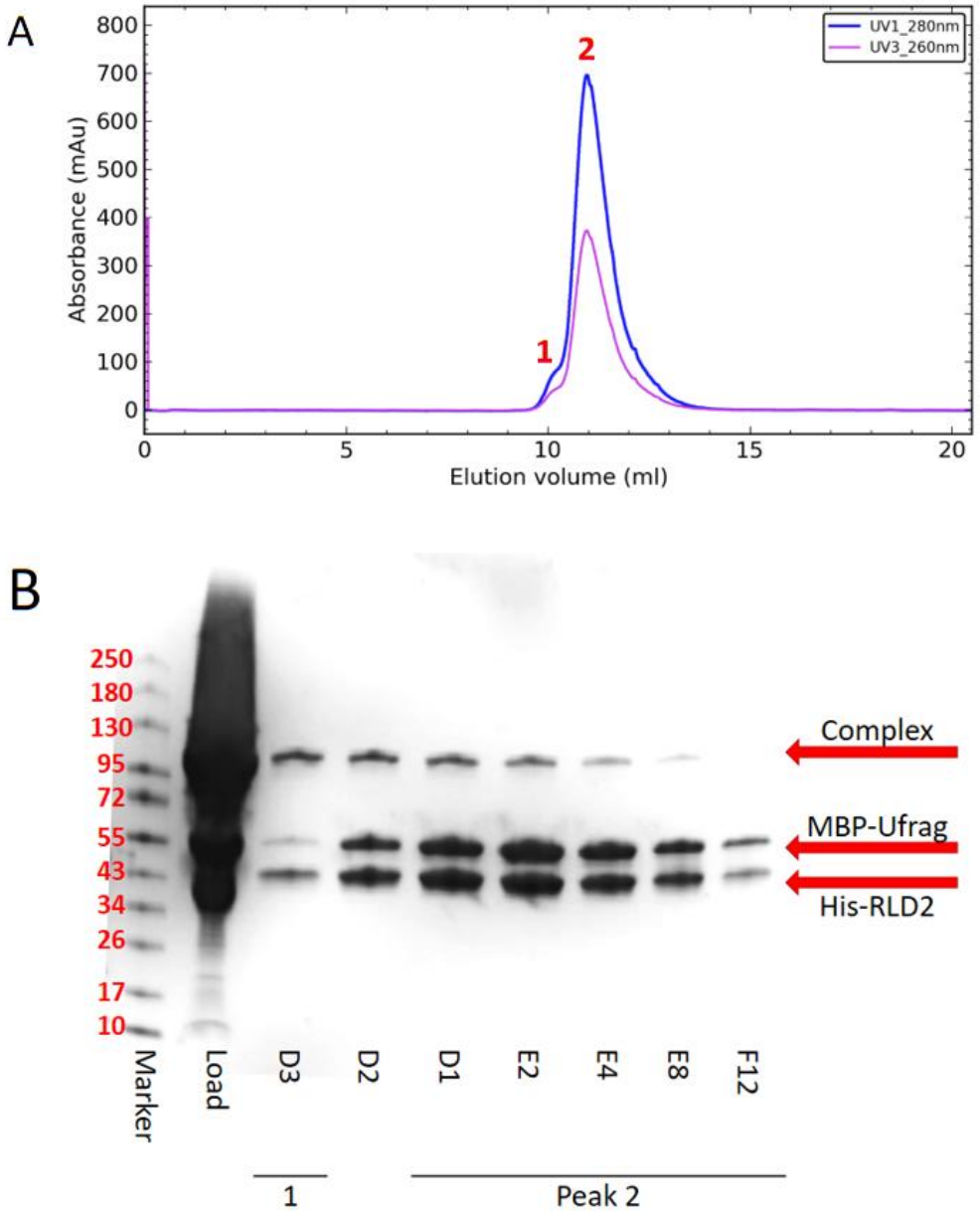


Figure 69: Size exclusion purification of a tagged RLD2+Ufrag complex. A) The absorbance profile of the complex on an S75 column. The absorbance at 280nm is shown in blue, the absorbance at 260nm is shown in pink, and the two main features of the trace are labelled in red. B) The fractions for the peak were analysed by SDS-PAGE and a Coomassie-based stain. The expected molecular weights of RLD2 and Ufrag are shown by the red arrows on the right, and the location of each fraction in the peak is labelled below the gel.

The tagged RLD2+Ufrag complex elutes almost as a single peak (Fig. 69a peak 2), but with a slight higher molecular weight shoulder peak (Fig. 69a peak 1). His-RLD2 alone eluted off the S75 column at 13 ml (see section 3.4.4, Fig. 45), whereas the His-RLD2 + MBP-Ufrag complex sample eluted at closer to 11 ml, which suggests an increase in the molecular weight as would be expected upon complex formation. The SDS-PAGE analysis of the sample within the trace (Fig. 69b) also suggests a 1:1 ratio of co-eluting His-RLD2 and MBP-Ufrag, which supports the observation of a clean 1:1 complex formation. However, the sample from the small peak 1 area of the size exclusion run shows a much higher ratio of His-RLD2 to MBP-Ufrag, and it also suggests a larger amount of the potential complex species. The size exclusion profile of His-RLD2 alone did not suggest the presence of any higher-ordered or aggregate species (section 3.4.4 Fig. 45), so it is possible that this fraction represents another complex of RLD2 and Ufrag with a different stoichiometry. This is supported by the ITC data, as the n-value for the Ufrag+RLD2 reaction is greater than 1, at 1.20, suggesting that an unequal number of each protein isinvolved in the interaction. As you cannot have 0.2 of a protein involved in an interaction, either 5 UBE3A units interact with 6 RLD2 units to form a much larger complex than expected, or there are a mix of stoichiometries within the sample, with the 1:1 ratio species comprising the largest proportion of the mix and a species with a higher ratio of His-RLD2 comprising the remainder. The SEC profile above supports this second theory, as the elution volume is indicative of a species with a molecular weight consistent with a heterodimer. The ’peak 1’ shoulder peak also supports this, as the SDS-PAGE visualisation suggests that the His-RLD2 to MBP-Ufrag ratio is much more skewed towards His-RLD2 (Fig. 69b), but the size exclusion profile of His-RLD2 alone does not suggest any oligomer or even aggregate peaks in this range (section 3.4.4, Fig, 49). This shoulder peak is also much smaller than the larger peak 2, representing the heterodimeric species, suggesting that the heterodimeric species comprises the vast majority of the sample. However, this second complex species, featuring an unequal ratio of each component, is not observed in the interaction between UBE3A and RLD2 (Fig. 66, section 4.3.2), so even if it is a true alternative state of the complex it is unlikely to be a physiologically relevant one.

4.4 Circular Dichroism

Circular dichroism (CD) works by exploiting the unequal absorbance of circularly polarised light by different secondary structure elements. B-sheets, α-helices, and disordered loop regions will each demonstrate a signature absorbance profile when subjected to differentially circularised light beams, which allows the relative abundance of each species within a larger protein structure to be determined (Greenfield *et al.,* 2007). This can be useful to predict the flexibility of a protein before subjecting it to structural studies, for example, a protein with a significant proportion of disordered regions will be

difficult to crystallise. However, another use CD is to identify protein-protein interactions. Protein complex formation will involve some alteration of the structure of each constituent, but the extent of the structural rearrangement can be observed through CD (Greenfield *et al.,* 2015).

4.4.1 UBE3A + RLD2

Samples for UBE3A, RLD2, and the UBE3A+RLD2 complex were prepared as described in in chapter 3, and then buffer exchanged as described in section 2.9.3. Due to COVID19 restrictions measurements were made by the beamline scientists at the B23 circular dichroism beamline at Diamond Light Source. The samples were measured in both the far-UV (180 – 240 nm) and near-UV absorption (240-320 nm) and CD spectra measurements, as shown in Fig. 70.

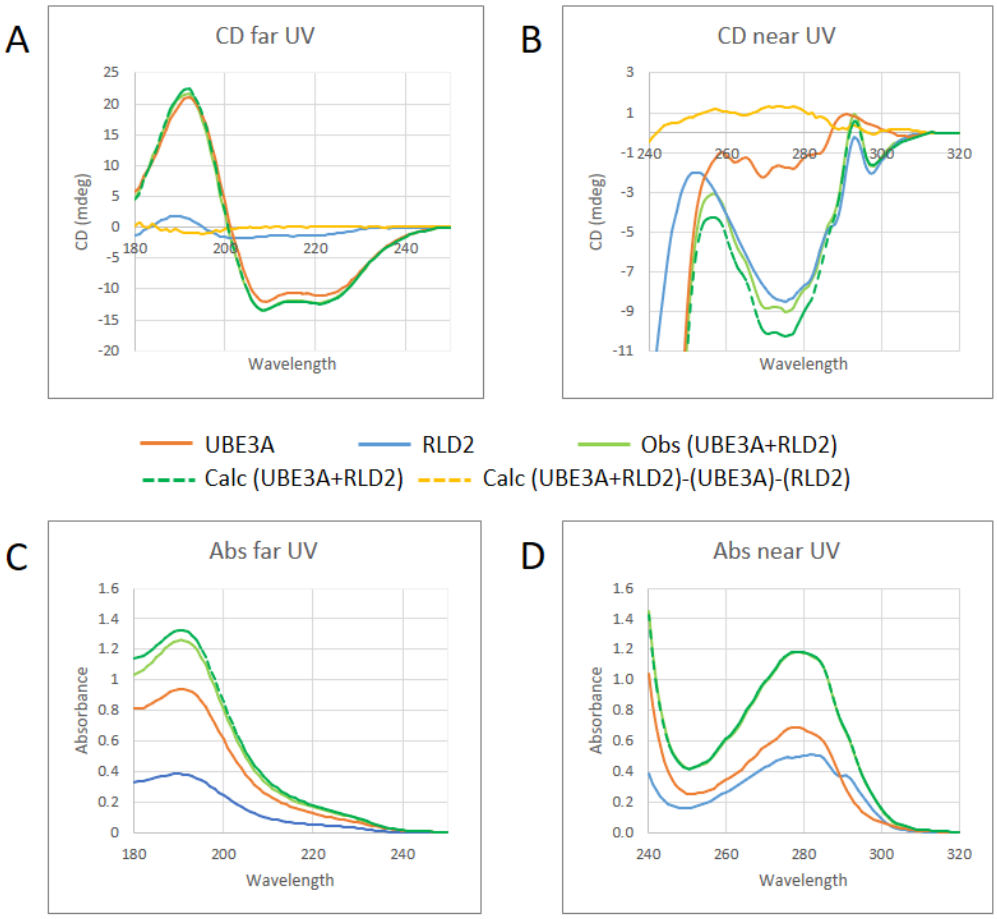


Figure 70: The interaction between UBE3A and RLD2 as studied using the circular dichroism technique. A) The CD spectra for each sample in the far-UV range. B) The CD spectra for each sample in the near-UV range. C) The absorption of each sample in the far-UV range. D) The absorption spectra of each sample in the near-UV range. The UBE3A sample is shown in orange, RLD2 in blue, the observed results for the complex are shown in green, the predicted results for the sample are shown by the dashed dark green line, and the difference between the actual results for the complex and the sum of the two components is shown by the yellow dashed line.

The spectra for the two proteins individually provides some initial information on the secondary structure elements of each protein. The CD spectrum for RLD2 alone suggests that 40% of the protein forms a beta-strand conformation and 33% is disordered, while the spectrum for UBE3A indicates a high content of alpha-helical conformation (45%) (Fig. 71). The data for each individual protein was used to calculate a predicted spectrum for a mixture of the two proteins in solution with no physical interaction, which can then be compared to the observed spectra of the complex to determine the effect of the interaction. The difference between the spectra for the UBE3A+RLD2 complex and the sum of two constituents was significant enough to confirm that a binding interaction took place, but was small enough to suggest that the changes in secondary structure upon binding are minimal. (Fig. 70, 71).

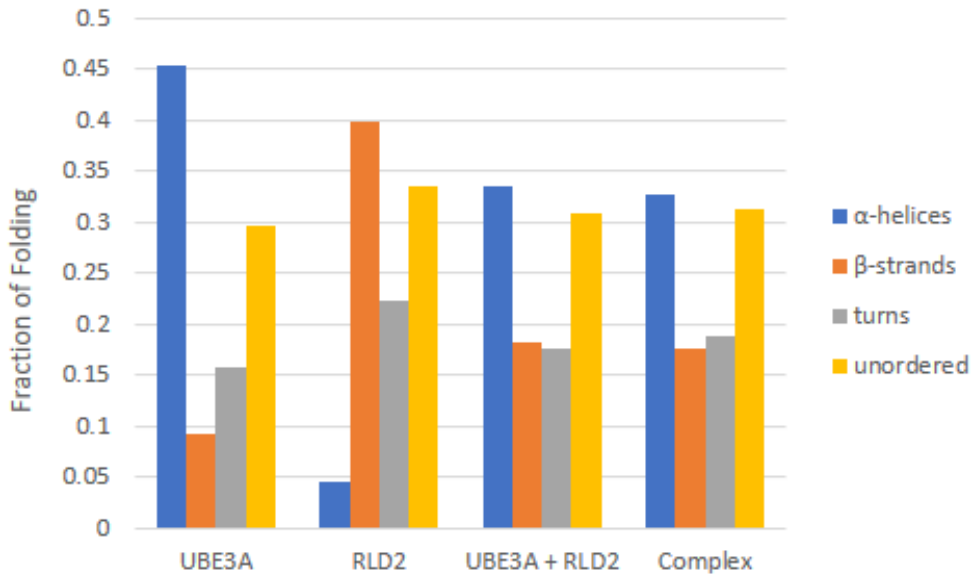


Figure 71: The fraction of each species that forms each secondary structure element. The UBE3A + RLD2 sample represents the calculated secondary structure composition of a non-interacting mixture of UBE3A and RLD2 in an equimolar ratio, while the Complex sample represents the observed data for the UBE3A+RLD2 sample provided.

The proposed binding site for RLD2 on UBE3A sits distally from the catalytic HECT domain, but the association of RLD2 and UBE3A has been shown to affect the catalytic activity of UBE3A (Kühnle *et al.,* 2011). In order for the RLD2 domain to alter the activity of the HECT domain, either the N-terminal region involved in the interaction sits closely with the HECT domain in the tertiary structure of UBE3A, or the RLD2 interaction would have to cause a large structural rearrangement within the whole UBE3A structure. The absence of a large alteration in the native-fold of either protein upon formation of the UBE3A+RLD2 complex would suggest that the initial theory is more likely.

4.5 Nano Differential Scanning Fluorimetry

Thermal melt shift measurements are an important method of determining the most stable form of a protein sample in a range of conditions. The stability of a sample is particularly important when the sample is used for structural biology techniques, such as x-ray crystallography and cryo-EM, as more flexible proteins will be difficult to determine to a high resolution, if at all. Thermal melt shift measurements were taken using the Nanotemper Prometheus instrument as described in section 2.9.4, and involved heating samples slowly and measuring the ratio of tryptophan fluorescence at 350:330 nm of the sample to determine the temperature at which it melts. In a folded protein, the tryptophans will be in a hydrophobic environment where the maximal fluorescence emission will occur at 330 nm, whereas once the protein has unfolded, the tryptophan is released into a hydrophilic environment, where the maximal fluorescence of tryptophan occurs at 350 nm. The shift from maximal fluorescence at 350 nm to 330 nm is therefore a measure of the extent of unfolding of the protein (Gao *et al.,* 2020).

4.5.1 UBE3A+RLD2

UBE3A and RLD2 were purified separately, and a UBE3A+RLD2 complex was formed as described in section 4.3.2. A small sample of each species was subjected to thermal melt analysis as described in section 2.9.4, and the melting point of each was compared (Fig. 72).

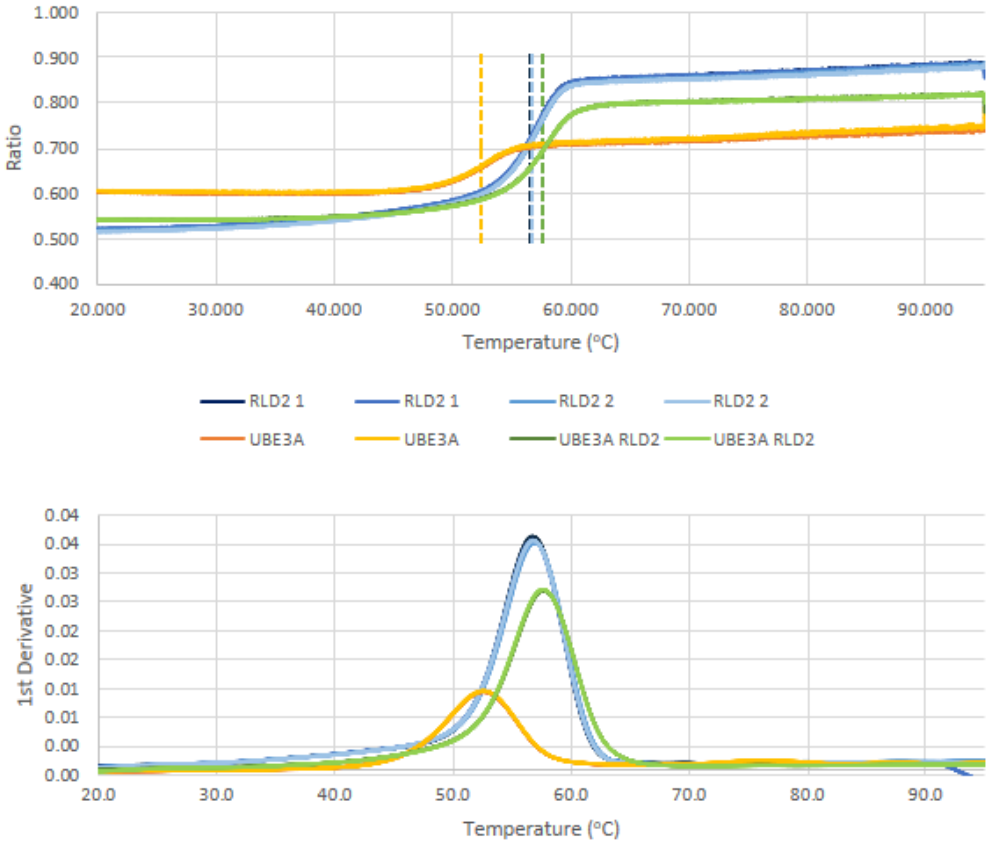


Figure 72: The thermal melt shift profile for a UBE3A+RLD2 complex compared to its individual constituents. The traces for the UBE3A replicates are shown in shades of orange, the RLD2 replicates are shown in shades of blue, and the UBE3A+RLD2 complex is shown in shades of green. The thick dashed lines in the top graph show the thermal melt point (T_m) for each sample.

The UBE3A samples have a melting temperature of 52.5°C, and the melting point of RLD2 was determined to be 56.7°C. This is representative of the predicted levels of disorder of the UBE3A and RLD2 sequences. The sample of UBE3A+RLD2, however, had a melting point of 57.6°C. The higher melting point for the complex compared to either individual constituent suggests that the formation of the UBE3A+RLD2 complex stabilised both species in some way.

4.5.2 UBE3A+PSMD4

UBE3A and PSMD4 were initially purified separately, and then a UBE3A+PSMD4 complex was formed as described in section 4.3.1, and then all three samples were subjected to thermal melt analysis as described in section 2.9.4 (Fig. 73).

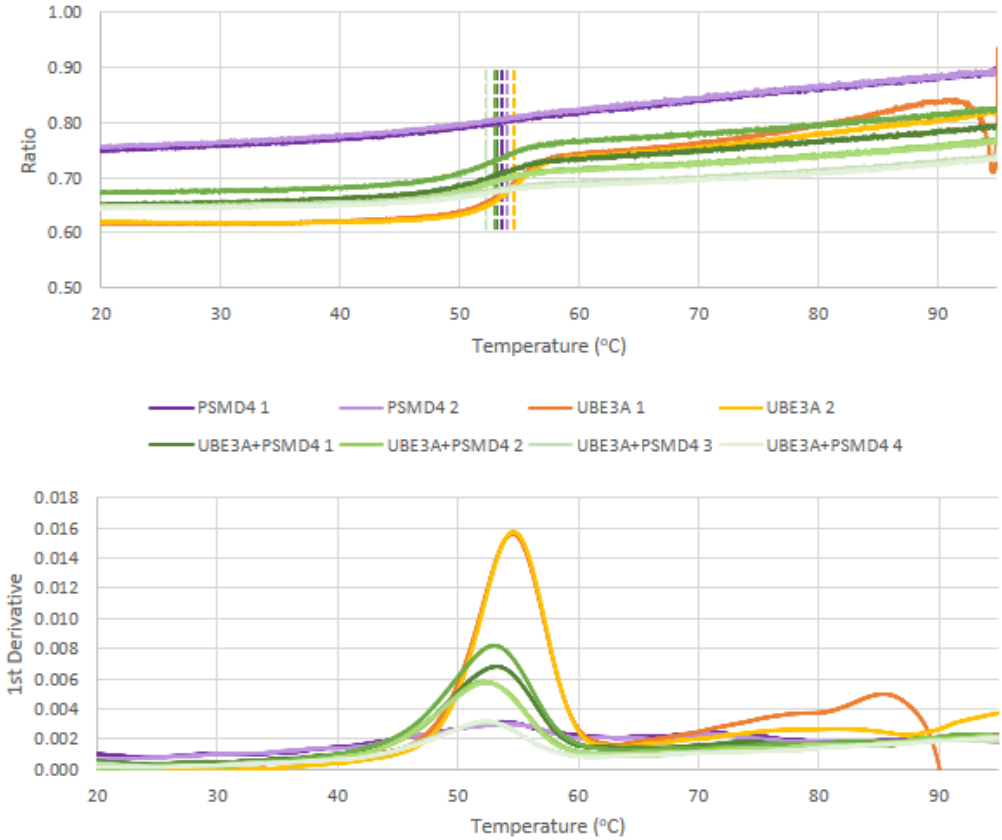


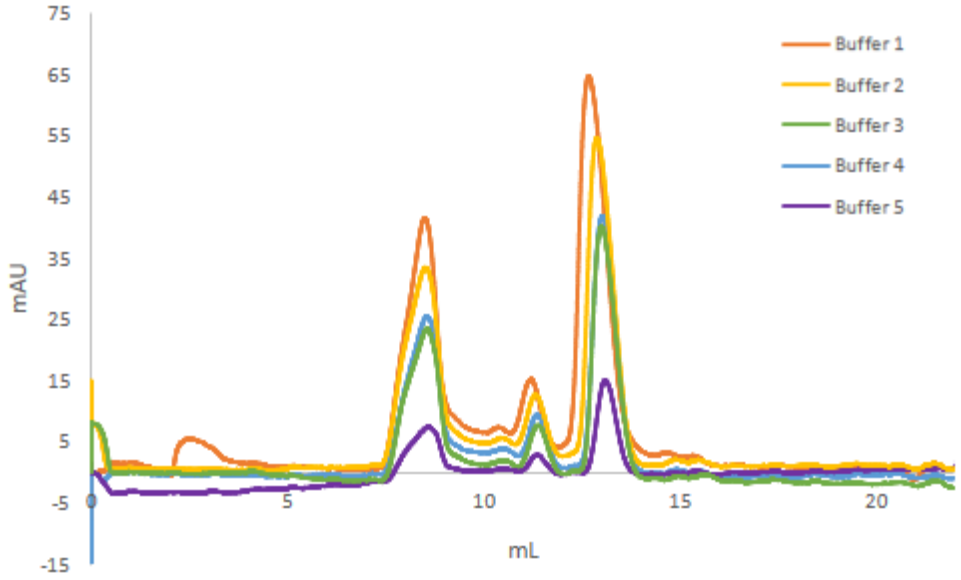
Figure 73: The thermal melt shift profile for a UBE3A+PSMD4 complex compared to its individual constituents. The traces for the UBE3A replicates are shown in orange, the PSMD4 replicates are shown in purple, and the UBE3A+PSMD4 complex is shown in shades of green. The thick dashed lines in the top graph show the precise thermal melt point for each sample.

The thermal melt data for the UBE3A+PSMD4 samples was not ideal. One issue is that PSMD4 has very little intrinsic fluorescence compared to most proteins, so the samples that contain PSMD4 have very low signal, which makes the inflection point determinations less precise. PSMD4 is also predicted to be fairly flexible and disordered, which means that there is less definition between the folded and unfolded states, as reflected in the shape of the curve for the 350 nm to 330 nm ratio (Fig. 73 top). The calculated thermal melt point for UBE3A was 54.5°C and the average calculated thermal melt point for PSMD4 was 53.7°C. This fits with secondary structure predictions based on the sequences, which suggest that PSMD4 is more disordered overall than UBE3A (Appendix 3). However, the thermal melt point for the UBE3A+PSMD4 complex was calculated as 52.5°C, which is even lower than either constituent and suggests that the complex may not be very stable.

4.5.3 UBE3A Buffer Optimisation

Although the thermal melt analyses are very useful for determining the stability of various samples under a range of conditions, the technology to carry out these experiments was not available to us initially. In order to test

the stability of UBE3A in a range of buffer conditions initially, a purified UBE3A sample was subjected to several SEC runs with buffers featuring a range of salt concentrations and pH values (Fig. 74).



Buffer 1	50 mM Tris, 150 mM NaCl, pH 8
Buffer 2	50 mM Tris, 250 mM NaCl, pH 8
Buffer 3	50 mM Tris, 500 mM NaCl, pH 8
Buffer 4	50 mM Tris, 150 mM NaCl, pH 7
Buffer 5	50 mM Tris, 150 mM NaCl, pH 8.5

Figure 74: The SEC profiles of a UBE3A sample subjected to a range of buffers. The 280nm measurement for each sample is shown, with the trace for buffer 1 shown in orange, buffer 2 in gold, buffer 3 in green, buffer 4 in light blue, and buffer 5 in purple. The constituents of each buffer are shown in the table.

Although the buffers used to purify UBE3A differ in their pH range and salt concentrations, the SEC profile for each sample looks almost identical. The only difference between each trace is a decrease in the overall concentration of the sample used for each run, but the relative ratio of each species within each run appears to be identical across all buffers used. This suggests that UBE3A is similarly stable across a variety of salt concentrations and pH values.

Following on from the initial buffer trials using SEC, the stability of UBE3A in HEPES buffer at pH 6 and pH 8 was tested using microscale thermophoresis (Fig. 75)

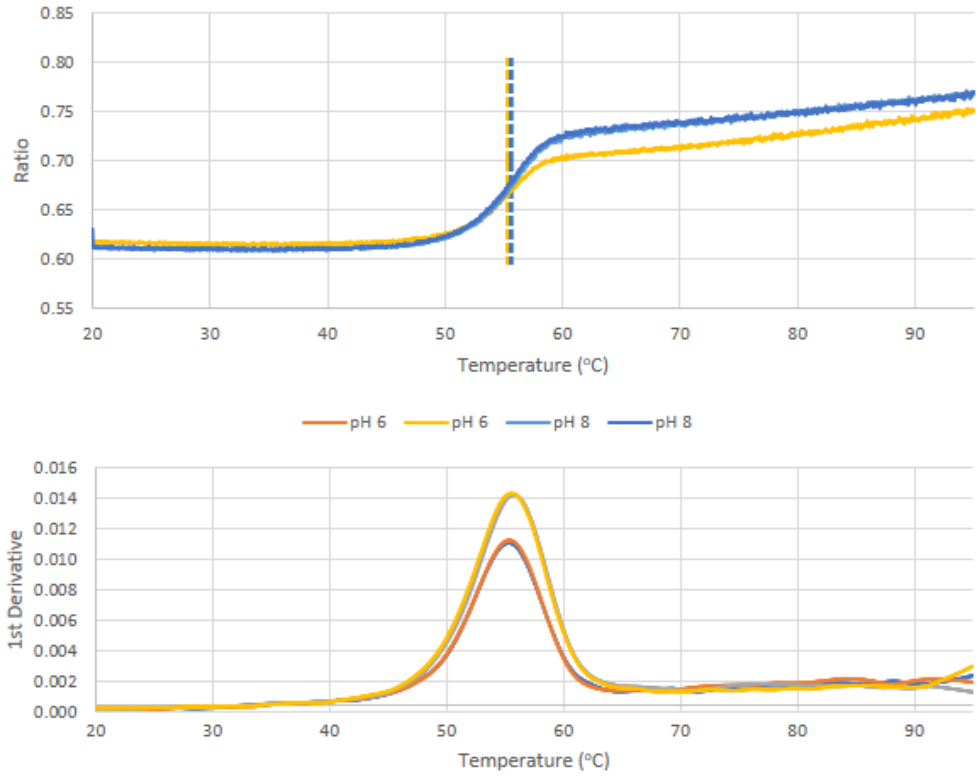


Figure 75: The thermal melt profile of UBE3A in HEPES buffers at different pH values. Samples run at pH 6 are shown in shades of orange, and samples at pH 8 are shown in shades of blue. The inflection point for the pH 8 samples is 55.6°C, and the inflection point for the pH 6 samples is 55.3°C.

The difference in melting point between the pH 6 and pH 8 samples is 0.3°C, which is minimal. This suggests that lowering the pH to pH 6, which is the suggested optimal pH value for glutaraldehyde activity, does not impact the stability of the UBE3A sample.

4.5.4 UBE3A Crosslinking

As the UBE3A sample could not be further stabilised through optimisation of the buffer alone, it was subjected to chemical crosslinking in an attempt to limit the flexibility of the protein. A sample of UBE3A was purified as previously described (see chapter 3) and subjected to glutaraldehyde treatment using a variety of conditions in order to optimise the crosslinking process (Table 8).

	UBE3A Concentration	Final % Glutaraldehyde	pH	Reaction Time
1	1 mg/ml	0.05	6	5 min
2	1 mg/ml	0.05	6	15 min
3	1 mg/ml	0.05	8	5 min
4	1 mg/ml	0.05	8	15 min
5	1 mg/ml	0.025	6	5 min
6	1 mg/ml	0.025	6	15 min
7	1 mg/ml	0.025	8	5 min
8	1 mg/ml	0.025	8	15 min
9	0.1 mg/ml	0.05	6	5 min
10	0.1 mg/ml	0.05	6	15 min
11	0.1 mg/ml	0.05	8	5 min
12	0.1 mg/ml	0.05	8	15 min
13	0.1 mg/ml	0.025	6	5 min
14	0.1 mg/ml	0.025	6	15 min
15	0.1 mg/ml	0.025	8	5 min
16	0.1 mg/ml	0.025	8	15 min

Table 8: An optimisation screen was attempted using small scale glutaraldehyde crosslinking reactions and a range of reaction conditions in order to determine the conditions that led to the most extensive UBE3A crosslinking without inducing aggregation of the sample. The parameters tested were the concentration of UBE3A in each reaction, the amount of glutaraldehyde present, the pH of the reaction buffer, and the time course of the reaction before quenching.

Each crosslinking trial reaction was carried out in an individual reaction as described in section 2.9.5. The reactions were quenched with an excess of high concentration Tris buffer, followed by heating in the presence of an SDS sample buffer. The results of the crosslinking test were visualised through SDS-PAGE (Fig. 76).

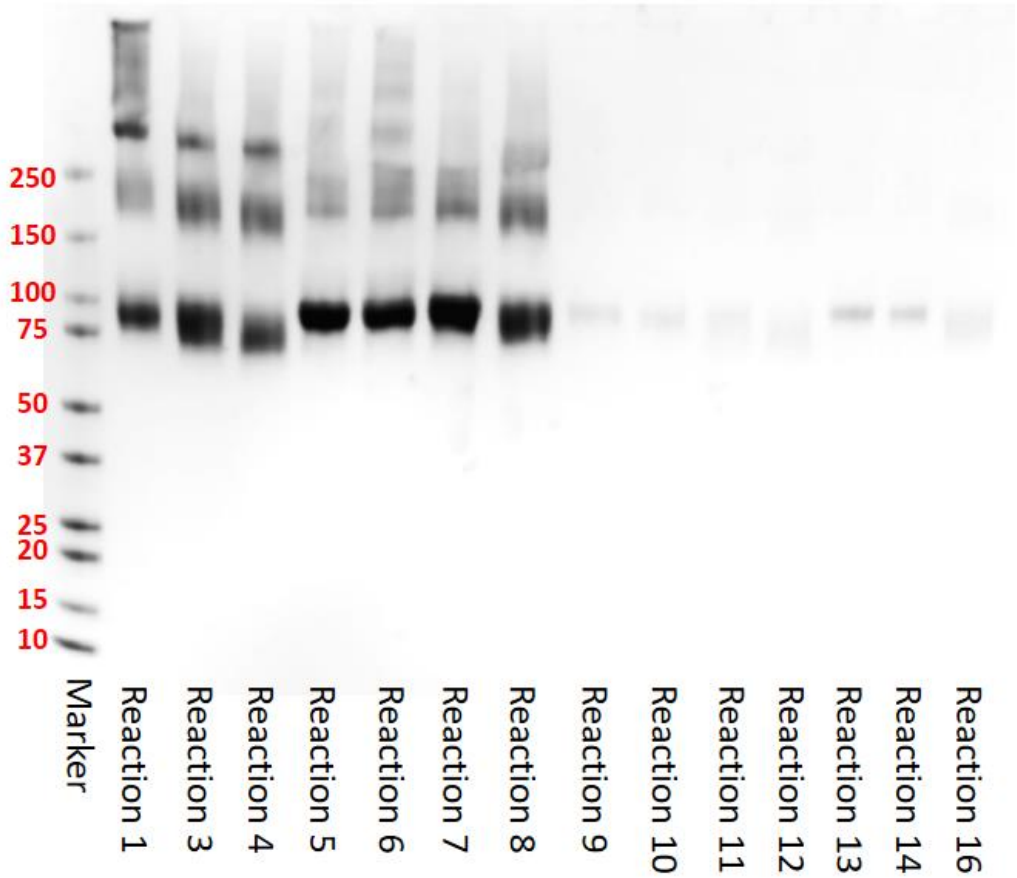


Figure 76: Initial visualisation of the UBE3A crosslinking test reactions through SDS-PAGE and a Coomassie-based dye. Two of the reactions from table 8 were left out due to space limitations on the gel.

The lower UBE3A concentration samples (reactions 9-16) appeared to be much cleaner than the higher concentration samples (reactions 1-8) in terms of higher molecular weight aggregate species, although it is not obvious whether that is just due to the contaminants being too low concentration to observe or whether they are actually cleaner samples. Although the data provided is arbitrary on this front, a decision had to be made. Theoretically, the sample is more likely to aggregate at a higher concentration, so by crosslinking at low concentration and then concentrating later if required the aggregate species would not be as extensively crosslinked. It is also possible that the crosslinked monomer may even stabilise the sample to dimmish its aggregation. Upon initial inspection, reaction 13 was presumed to produce the most stable UBE3A construct, as the band for UBE3A in that sample was the clearest of the lower concentration samples. A larger-scale crosslinking reaction for UBE3A was prepared using the reactions conditions used for reaction 13 (see section 2.9.5), and the stability of the crosslinked sample was tested using a thermal melt shift experiment (Fig. 77).

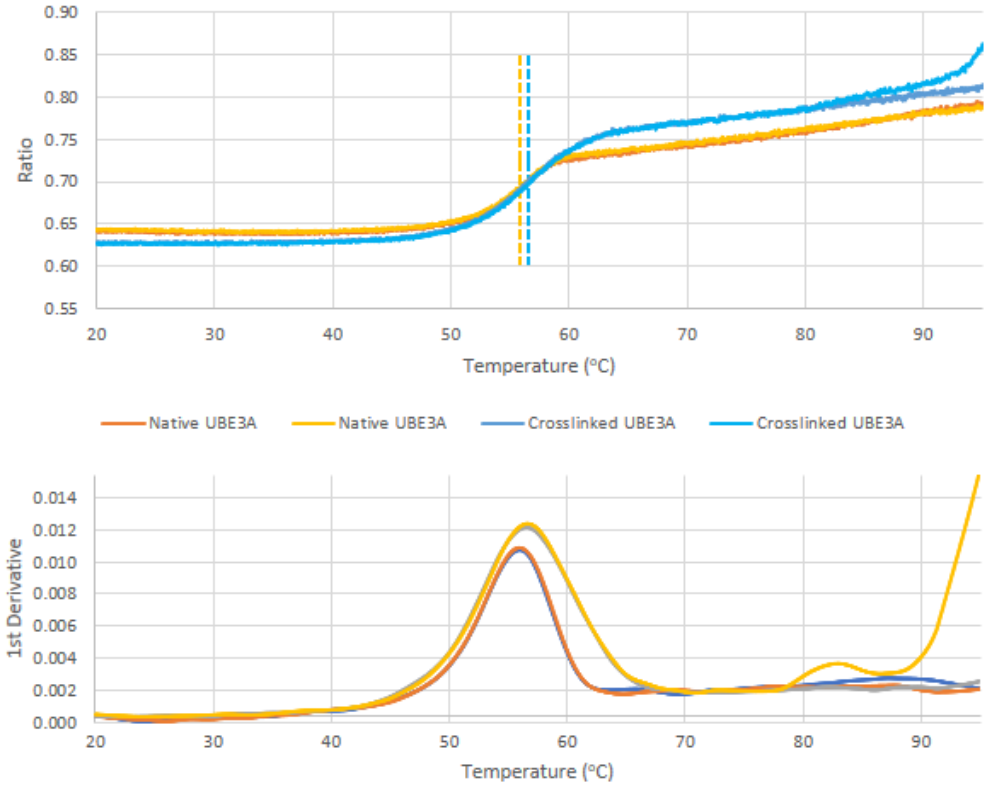


Figure 77: Thermal melt shift analysis of UBE3A after a crosslinking reaction. Samples of both native UBE3A (shown in yellow and orange) and crosslinked UBE3A (shown in shades of blue) were subjected to a thermal melt measurement to determine the melting point of each sample. The native protein had a thermal melt point of 55.9°C while the crosslinked sample had a melting point of 56.6°C.

The increase in the thermal melting point of the sample following crosslinking does suggest some stabilisation of the protein, but the effect was not as great as it could have been. A second look at the results of the initial crosslinking test (Fig. 76), this time accounting for the effects of a crosslinked sample on SDS-PAGE, suggested that sample 13 may in fact be the least effective crosslinking condition. Heavily crosslinked proteins will not be able to become fully denatured, even in the presence of SDS, so they will run less smoothly through an acrylamide gel than a native denatured sample (Griffith, 1972). With this in mind, the most crosslinked samples appeared to be reactions 9-12. Samples from the small-scale reactions 9-13 were subjected to thermal melt analysis as described in section 2.9.4 to determine the extent of the crosslinking in each sample (Fig. 78)

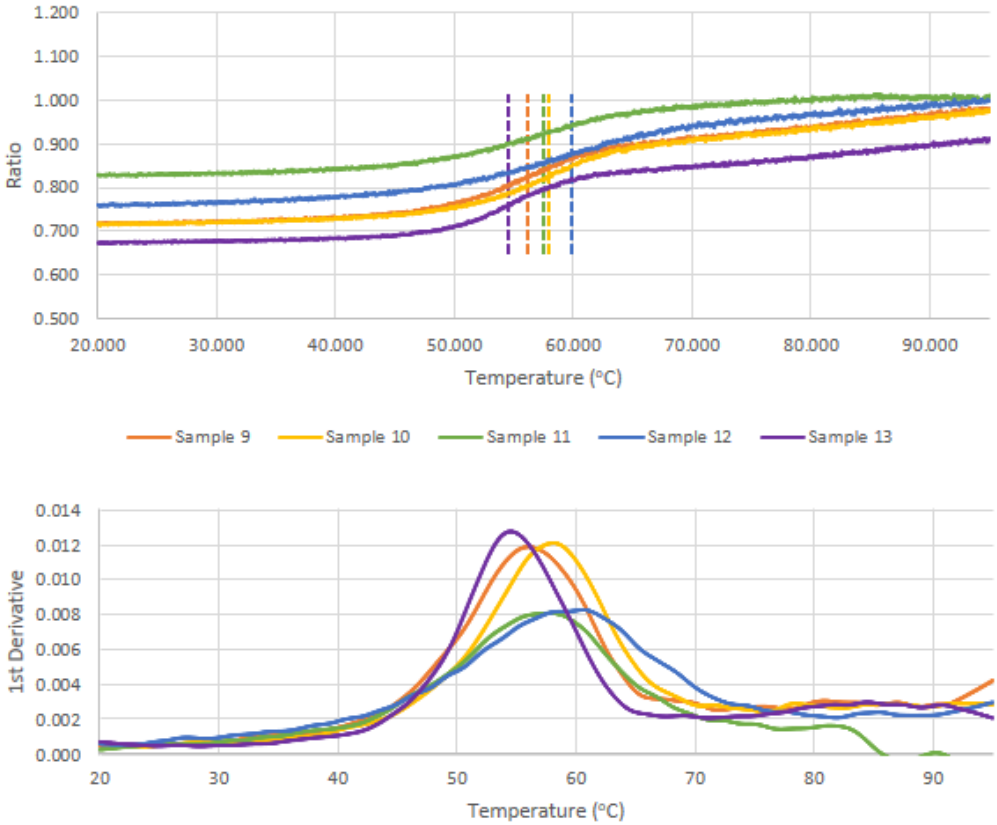


Figure 78: Thermal melt analysis of several small-scale crosslinking test reactions. The sample in reaction 9 (red line) has a thermal melt point of 56.1°C, reaction 10 (orange line) has a thermal melt point of 57.9°C, reaction 11 (yellow line) has a melting point of 57.4°C, reaction 12 (green line) has a melting point of 59.8°C, and the thermal melt point for reaction 13 (blue line) was calculated to be 54.6°C.

Reaction 13, which was carried out using the same conditions as the initial large scale UBE3A crosslinking reaction, was the least stable sample (Fig. 78). The crosslinking conditions that induced the largest stability increase in UBE3A were those used for the small-scale reaction 12. The differences in the reaction conditions between samples 12 and 13 include a higher concentration of glutaraldehyde, a longer reaction time, and a higher pH value. Glutaraldehyde is expected to be optimally active at different pH levels for different proteins (Migneault *et al.,* 2004), but both the thermal melt results and the SDS-PAGE visualisation of the small-scale reactions suggest that the sample for reaction 12 at pH 8 was more effectively crosslinked than its pH 6 counterpart, reaction 10 (Fig. 76, 78), so the conditions used for the small-scale reaction 12 (table 8) were accepted as the most efficient for crosslinking of UBE3A and used for all further samples.

A second larger-scale UBE3A crosslinking reaction was prepared using the conditions described for the small-scale reaction 12 (table 8), however both this sample and the initial weakly crosslinked sample were subjected to SEC analysis to determine the effect of harsher crosslinking on the oligomeric

state of the sample. The initial, weaker sample was referred to as ‘UBE3A crosslink 1’, while the second, more strongly crosslinked sample was referred to as ‘UBE3A crosslink 2’. Both samples were independently concentrated, subjected to SEC, and analysed through SDS-PAGE and a Coomassie-based dye (Fig. 79, 80).

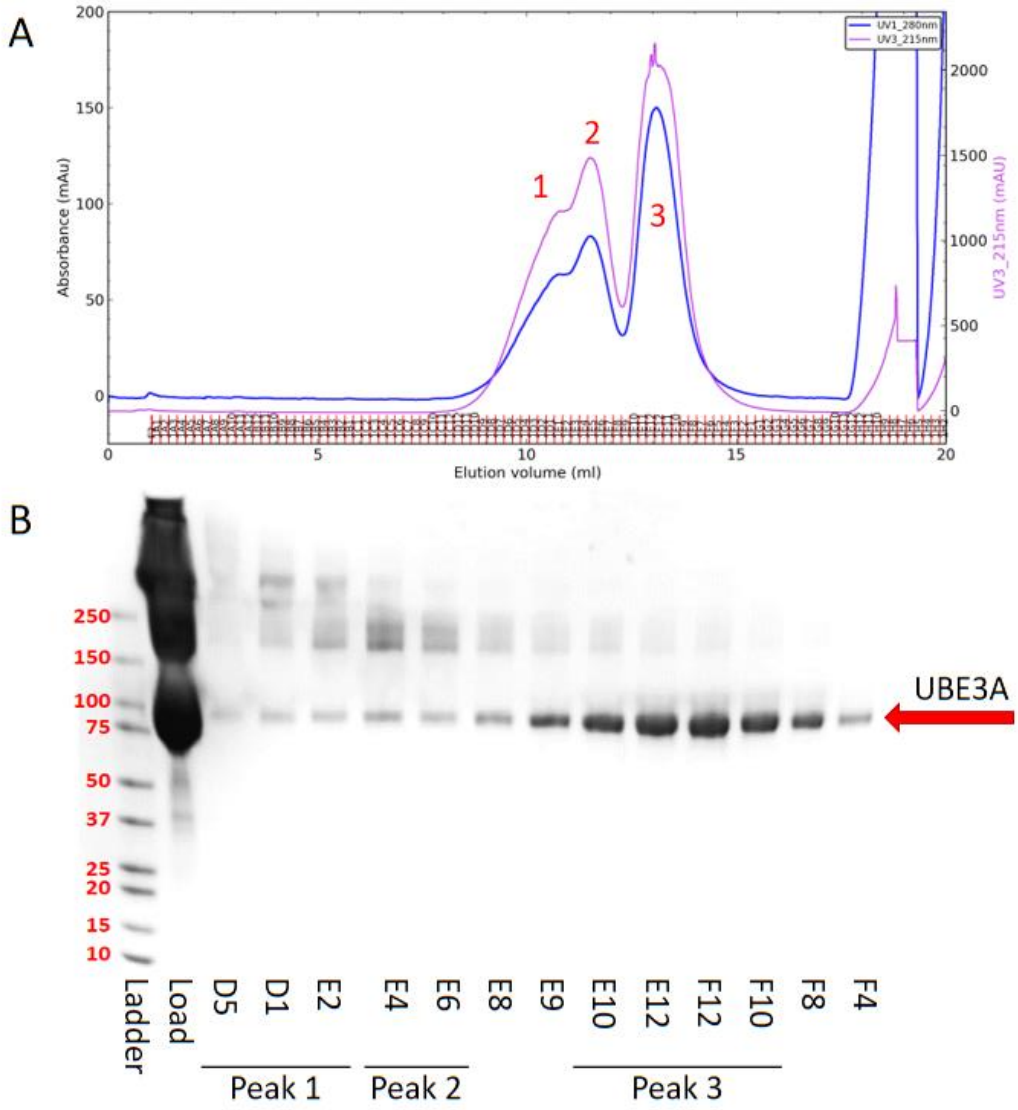


Figure 79: A) The SEC profile of the weakly crosslinked ‘UBE3A crosslink 1’ sample. The A280 trace is shown in blue, and the A260 trace is shown in red. B) The relevant fractions from the trace were subjected to SDS-PAGE and visualised with a Coomassie-based stain.

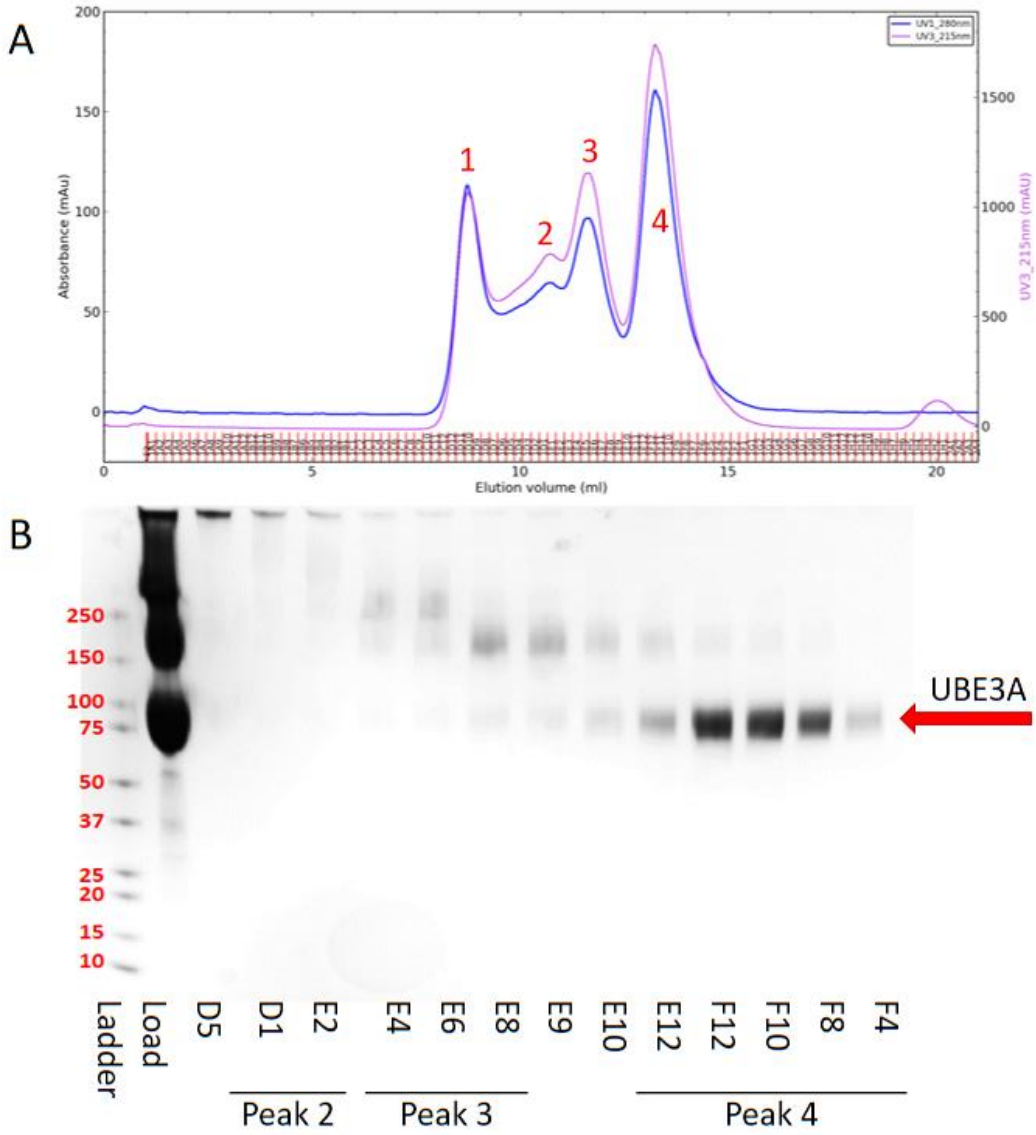


Figure 80: A) The SEC profile of the strongly crosslinked 'UBE3A crosslink 2' sample. The A280 trace is shown in blue, and the A260 trace is shown in red. B) The relevant fractions from the trace were subjected to SDS-PAGE and visualised with a Coomassie-based stain.

The traces for crosslinked UBE3A are similar to the standard SEC trace for UBE3A (Fig. 42), but with a few slight differences. The most notable thing about the 'UBE3A crosslink 1' trace (Fig. 79) is that there is no aggregate peak, although this is likely to be because the sample used for crosslinking had already been subjected to SEC to remove the aggregate species before use. The key notable difference in both crosslinked species is that the peak for the potential dimer and/or trimer is much more pronounced (Fig. 79 peaks 1 and 2, Fig. 80 peaks 2 and 3). This suggests that the multimer species are present in the uncrosslinked UBE3A sample, but they may be underrepresented in the standard UBE3A SEC trace due to the dilution effect of SEC. In both crosslinked samples, it becomes more clear that both dimeric and trimeric

UBE3A species are present in the sample, although the monomeric form is still in the clear majority.

When the two UBE3A crosslink samples are compared, the ‘crosslink 2’ condition appears to produce much cleaner samples. Other than the presence of the aggregate peak in the ‘UBE3A crosslink 2’ sample (Fig. 80 peak 1), the SEC traces for the two samples are very similar. The SDS-PAGE observations for the two samples, however, show the differences. In the SDS-PAGE for ‘UBE3A crosslink 1’ (Fig. 79b) all of the samples show multiple different bands, including a UBE3A monomer band across all of the fractions. However, in the ‘UBE3A crosslink 2’ SDS-PAGE results (Fig. 80b) the samples appear mostly clean, with more separation between the different bands. The UBE3A monomer band still appears across several other fractions, but it is at a much lower intensity than in the ‘UBE3A crosslink 1’ samples. The potential dimer and trimer species appear much more distinct, with the apparent dimer more prevalent in fractions E8 and E9 and the apparent dimer in fractions E6 and E4, although these distinctions are not reflected in the SEC trace, with all of these fractions appearing to form part of peak 3 (Fig. 80b).

4.5.5 UBE3A+PSMD4 Crosslinking

UBE3A and PSMD4 were purified separately through the affinity chromatography steps, and then they were mixed in an equimolar ratio ready for crosslinking. The UBE3A+PSMD4 complex was crosslinked as described in section 2.9.5, and then the sample was concentrated and subjected to SEC in order to observe the level of aggregation present (Fig. 81).

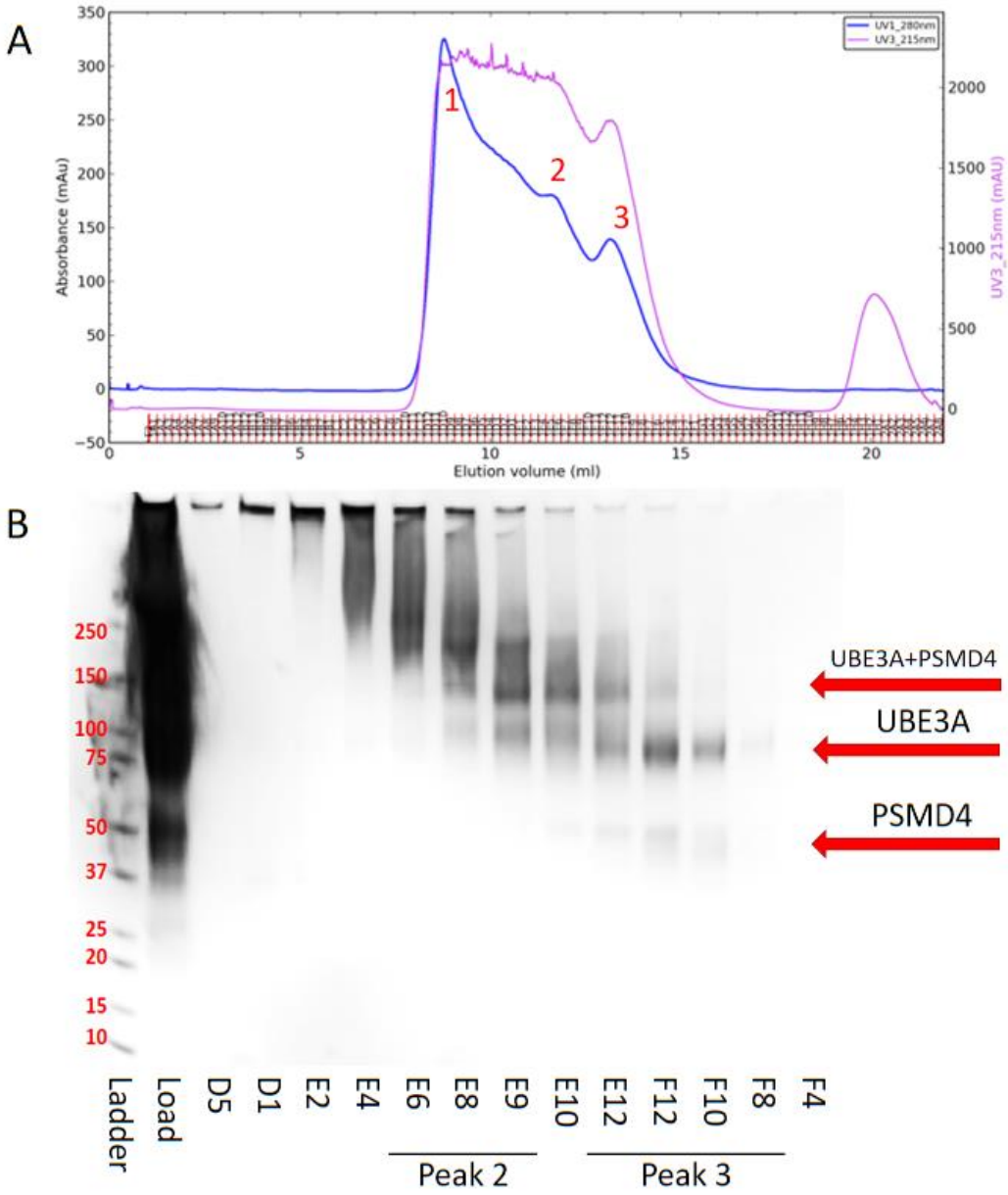


Figure 81: A crosslinked sample of UBE3A+PSMD4 after SEC. A) The SEC profile of the concentrated complex, The A280 trace is shown in blue, the A260 trace is shown in red, and the A215 trace is shown in pink. B) The fractions relating to the features of the trace were run on SDS-PAGE and stained with a Coomassie-based dye.

The trace for the crosslinked UBE3A+PSMD4 does not show any distinct features. Instead of a single peak for each species present in the sample, it appears to be all one large aggregate peak. There does appear to be a peak-like feature at around 13 ml that could be a UBE3A+PSMD4 species, but it is too intermingled with the rest of the trace to be separated. The SDS-PAGE image (Fig. 81b) also does not support the isolation a single heterodimeric UBE3A+PSMD4 species. In the lower molecular weight range, in fractions F12–F8, there are bands for both UBE3A and PSMD4, but they are not crosslinked into a complex. Fractions E8-E10 do contain a species at approximately 150

kDa, which is around the predicted molecular weight of a 1:1 UBE3A+PSMD4 complex, but it is by no means the primary species in any of these fractions. A large amount of aggregate species in varying sizes are present across most of the lanes, and no single species can be isolated from any of these fractions to sufficient purity.

4.5.6 UBE3A+RLD2 Crosslinking

UBE3A and RLD2 were purified separately through the affinity chromatography steps, and then they were mixed and subjected to SEC. The pre-formed complex was isolated and crosslinked as described in section 2.9.5, before being concentrated and subjected to SEC again to separate out any aggregate (Fig. 82).

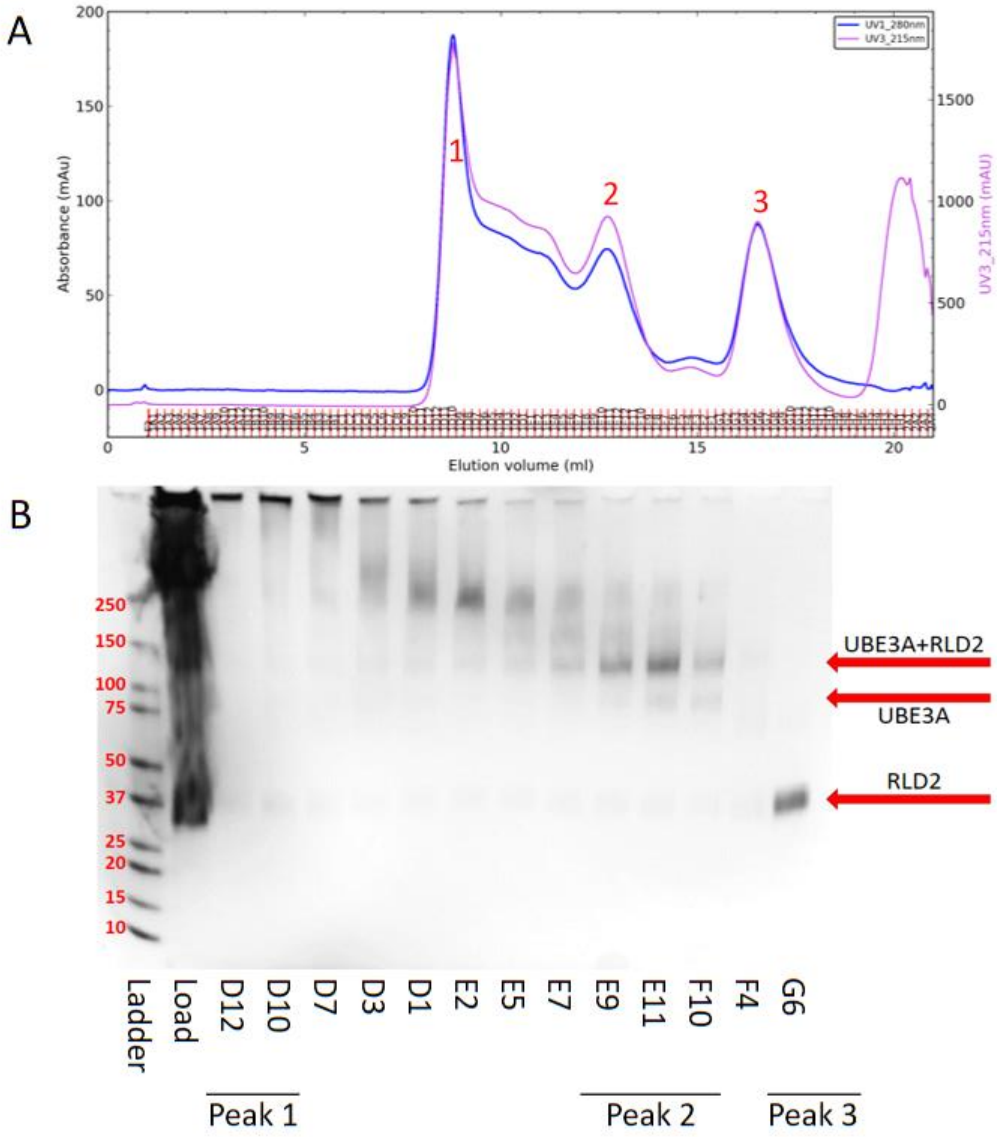


Figure 82: A crosslinked sample of UBE3A+RLD2 after SEC. A) The SEC profile of the concentrated complex, The A280 trace is shown in blue, the A260 trace is shown in red, and the A215 trace is shown in pink. B) The fractions relating to the features of the trace were run on SDS-PAGE and stained with a Coomassie-based dye. The expected molecular weights of various species are shown by the red arrows on the right.

The AKTA trace (Fig. 82a) shows a clear peak for excess unbound RLD2, but the peak for UBE3A+RLD2 has merged with a larger, broader peak for a series of aggregate species. When this trace is compared with that of the non-crosslinked UBE3A+RLD2 complex (Fig. 66), it shows a clear single peak for isolated RLD2, as can be seen in peak 3 of figure 82, and a clear peak for the complex, with a much smaller third peak representing higher molecular weight species merging to one side. The crosslinked sample however shows a much larger proportion of higher molecular weight species, with the small peak 2 from figure 66 being replaced by a much broader trace, culminating in a peak with a much higher intensity and representative of a much higher

molecular weight. This higher molecular weight peak also overlaps with the peak for the true UBE3A+RLD2 complex much more substantially, making isolation of the complex more difficult. However, SDS-PAGE of the UBE3A+RLD2 peak samples does show the presence of a clear UBE3A+RLD2 complex. Some un-bound UBE3A and RLD2 is present across various samples, along with the higher molecular weight contaminants, but the complex still forms the major component of fractions E9-F10 enough to attempt to use it for further structural work.

Although I have been able to show that UBE3A forms complexes with RLD2 and PSMD4 through AUC (section 4.1), ITC (section 4.2), co-purifications (section 4.3) and CD (section 4.4), none of these techniques are able to confirm that the complex remains intact following the plunge freezing process of making cryo-EM grids. This makes processing any cryo-EM data more difficult, as I am unable to attempt to account for any heterogeneity in the sample. This difficulty is demonstrated in the uncrosslinked UBE3A+PSMD4 cryo-EM data (section 7.2), as I was unable to determine whether the classes that were being produced were noisy UBE3A-only classes or a true UBE3A+PSMD4 complex view that just appeared similar to the UBE3A-only sample. In theory, crosslinking the complex samples should stabilise the conformation and prevent disassociation at any point in the process, allowing a more heterogeneous sample to be applied to cryo-EM grids. This appeared to work reasonably well for the UBE3A+RLD2 sample, as the SDS-PAGE image shows the presence of fairly discrete species of un-crosslinked monomers, a stabilised heterodimer, and a higher molecular weight complex sample (Fig. 82b. However, the size exclusion trace for this sample (Fig. 82a) does not show very good separation of the heterodimer and higher order species, and the SDS-PAGE image does show contaminant species present across each lane. The sample was taken forward in this instance and I was able to resolve particles of what appears to be a UBE3A+RLD2 complex (Fig. 113), although further work to improve the purity of the crosslinked sample, either through a size exclusion column with a higher resolution or through further purification steps such as ion exchange, could be beneficial and may enable generation of a higher resolution structure. The UBE3A+PSMD4 sample would also benefit from optimised separation of crosslinked species, although the resolution of the sample demonstrated here (section 4.5.5) was not as promising as the UBE3A+RLD2 sample and so was not taken forward for cryo-EM analysis. In contrast, the crosslinked UBE3A sample appeared very promising and would definitely benefit from further exploration. I was able to observe UBE3A oligomerisation in its SEC profile (section 3.2.1) and through AUC (section 4.1.1), but the monomeric state was consistently the predominant component of any sample, and this appeared to remain the case across a concentration range (Fig. 58). This made isolation of a multimeric state difficult, and I was not able to identify higher order species within any cryo-EM datasets. The

crosslinked UBE3A sample appears to show a clear separation of monomeric UBE3A, a possible UBE3A dimer, and also a possible trimeric state. The potential dimer and trimer states are not separated in the SEC trace, but the individual fraction do show separation of the states. Cryo-EM grids were set up using the trimeric and dimeric forms of crosslinked-UBE3A and were screened on a 200 kV Glacios microscope, but further data collection was not pursued due to the time constraint of the project. This would definitely be an area that could benefit from further exploration.

4.6 Discussion

UBE3A has been implicated in several different clinical contexts (Bandilovska *et al.,* 2019; Kishino *et al.,* 1997; DiStefano *et al.,* 2016; Noor *et al.,* 2015; Salminen *et al.,* 2019; Vatsa and Jana*,* 2018; Cheng *et al.,* 2019; Pyeon *et al.,* 2019; Olabarria *et al.,* 2019), but still the mechanism for its activity has not been well characterised. Ronchi *et al.,* propose a mechanism that involves a multimeric state of UBE3A, and they suggest that the physiologically relevant form is a trimer (Ronchi *et al.,* 2014). However, this observation has not been replicated by any other group and has yet to be substantiated with any biophysical or structural data beyond the initial SEC trace. SEC and AUC analysis of UBE3A purified from *E. coli* (Fig. 42 (section 3.4.1) and Fig. 58 (section 4.1.1)) suggest that the sample is primarily monomeric, although a higher oligomeric state is also present in a low quantity. However, neither SEC nor AUC were able to accurately determine the molecular weight of the higher molecular weight species to determine if it was a trimer or a dimer of UBE3A. This was due to the decreased resolution at that molecular weight range in the S200 column for SEC, and an unreliable frictional coefficient calculated from the AUC data due to the mixture of species in the sample. One observation from the AUC data (Fig. 58) is that the ratio of monomer and multimeric states of UBE3A appears to remain stable across the concentration range tested. The constant presence of the multimeric state, regardless of the concentration, suggests that it is a spontaneously forming species rather than a high concentration artefact, and neither is it a contaminant species in equilibrium with the monomer, but it also prevents isolation of it for further analysis. However, endogeneous UBE3A is subject to post translational modifications, for example phosphorylation of T485 in isoform 1 (Jason *et al.,* 2015), so I cannot say whether a more abundant multimeric form of the protein is present in cells following some form of modification. Another point to note is that the Ronchi *et al*., work (Ronchi *et al.,* 2014) was performed using the isoform 2 variant of UBE3A, whereas the work presented in this project was carried out using isoform 1. It may be possible that the extra 23 amino acids of isoform 2 enable more efficient oligomerisation of the enzyme than is suggested for isoform 1.

Once the oligomeric state of UBE3A had been assessed, the UBE3A sample was subjected to further biophysical analysis to determine if the stability of

the enzyme could be improved with different buffer conditions. An initial screen using affinity-purified sample subjected to SEC with a range of buffers suggested that the neither the multimeric state nor the propensity for aggregation of UBE3A was particularly affected by different salt concentrations or pH values (Fig. 74) This was later confirmed through the use of thermal melt measurements conducted on UBE3A in buffers at pH 6 and pH 8. Despite the regular observation of UBE3A aggregate during protein purification, its properties remained stable across a range of pH and salt concentration values. The distribution of species in the UBE3A sample was also confirmed following crosslinking of the sample. Crosslinked UBE3A was subjected to SEC and subsequent SDS-PAGE analysis of the fractions (Fig. 79, 84), and the profile reflected that of native UBE3A, where the majority of the sample was monomeric UBE3A. However, earlier fractions from the SEC profile of crosslinked UBE3A show species that could represent both dimeric and trimeric forms of UBE3A (Fig. 80). However, as these species are only apparent in the artificially crosslinked sample, it is impossible to determine whether they are both natural states of the enzyme that have been stabilised, or whether they are artefacts from the crosslinking process.

PSMD4 was identified as an interactor of UBE3A that could potentially act as both a substrate and a binding partner to influence its activity (Buel *et al.,* 2020; Kühnle *et al.,* 2018; Avagliano Trezza *et al.,* 2019). The interaction between UBE3A and PSMD4 was first demonstrated using SEC (Fig. 65), where the predominant peak from a sample of equimolar UBE3A and PSMD4 shows a shift towards a higher molecular weight elution volume. SDS-PAGE analysis of this peak shows a 1:1 ratio of UBE3A and PSMD4 present (Fig. 65b), although even in an excess of PSMD4 some UBE3A remains unbound, as indicated by the retention of the UBE3A-only peak at 12 ml. (Fig. 65a). The 1:1 ratio of the UBE3A+PSMD4 complex is supported by the ITC data (Fig. 61), and the Kd suggested by the ITC experiment also suggests a reasonably strong interaction. However, the SV-AUC data for the UBE3A+PSMD4 interaction goes against this somewhat (Fig. 59). Mixtures of UBE3A and PSMD4 in different molar ratios were subjected to SV-AUC alongside samples of UBE3A and PSMD4 alone, and the species present in each sample were compared. Given the high Kd suggested by ITC and the observation of a new species upon SEC, it was expected that this higher molecular weight species would also be observed for UBE3A+PSMD4 after AUC. However, the peaks for the complex samples show only the peaks observed in the individual protein samples. This confirms that both proteins are present in the sample, but it does not show any stable UBE3A+PSMD4 species present. It is unclear why the sample should not form a stable complex when analysed by AUC when it was observed during both ITC and SEC experiments, although one possibility is that the UBE3A+PSMD4 complex forms only under high concentration conditions. The samples used for AUC were used at a maximum concentration of 0.5 mg/ml,

whereas the samples for ITC were concentrated to 2 mg/ml and 16.4 mg/ml for UBE3A and PSMD4 respectively, and the mixed sample was concentrated to several mg/ml before being subjected to SEC. SEC does lead to a sample dilution of ~ 10x during a run, so the UBE3A+PSMD4 complex observed in the samples after SEC will not be at a particularly high concentration, but it is possible that the complex formed during the high concentration state prior to loading on SEC, and it did not dissociate through the run.

HERC2 has been identified as another binding partner of UBE3A in cells, although the implications of this interaction are not fully understood as both enzymes are HECT ligases (Kühnle *et al.,* 2011). The RLD2 domain was isolated and purified (see section 3.2.4 and 3.4.4) in order to investigate the core interaction between UBE3A and HERC2. Following initial affinity chromatography of each protein, UBE3A and RLD2 were mixed in either an equimolar ratio or an excess of RLD2, concentrated, and subjected to SEC (Fig. 66). The sample eluted in a single peak, with a slight shoulder at the higher molecular weight edge, and it appears to elute at a similar volume to UBE3A alone. This alone would suggest that the peak represents a sample of monomeric UBE3A rather than any complex, but SDS-PAGE analysis of the fractions suggests that the peak is actually comprised of both UBE3A and RLD2 in a 1:1 ratio. In order to ensure that the 1:1 ratio of UBE3A and RLD2 within the peak was not due to overlapping elution profiles of monomeric UBE3A and a potential RLD2 dimer, RLD2 alone was also applied to an S200 column and compared with the complex trace. The RLD2 only trace showed a single peak with a higher elution volume representing monomeric RLD2, and no sample was detected at the elution volume of the UBE3A+RLD2 complex. Although SEC columns are designed to separate macromolecular species based on their molecular weight, the samples are not denatured in any way prior to loading so the shape of the molecule also effects the elution profile of any sample. The slightly delayed elution of a UBE3A+RLD2 sample from an S200 column suggests that the interaction could alter the shape significantly from that of UBE3A alone.

UBE3A+RLD2 were subjected to ITC in the same way as UBE3A+PSMD4, and although it did show a significant interaction and confirm the 1:1 stoichiometry, the Kd value for the RLD2 interaction (Fig. 62) was 20 times lower than that for UBE3A+PSMD4 (Fig. 61). This contrasts heavily with the AUC data for the UBE3A+RLD2 interaction, which shows the formation of a single peak with a sedimentation value between that of monomeric UBE3A and the potential UBE3A multimer (Fig. 60). Although the multiple species present in the UBE3A-alone sample prevented accurate determination of the molecular weights of each species individually, the predicted values did suggest the same 1:1 ratio indicated by ITC and SDS-PAGE. The single peak of the complex sample also suggests that in a sample comprised of a 1:1 ratio of

UBE3A and RLD2, all of the species form a stable and spontaneous complex. No lingering monomeric forms of either protein could be detected.

Although the SEC elution profile for the UBE3A+RLD2 sample suggests that there is a significant alteration in the shape of the complex compared to UBE3A alone, the CD data for the complex does not show any significant structural rearrangements. It does confirm that the species interact, but the interaction apparently does not induce any major changes to the secondary structure composition of either enzyme. This is particularly unexpected as the key region of UBE3A involved in the interaction has been mapped to an area of 50 amino acids distal from the HECT domain that has been predicted to be disordered (Kühnle *et al.,* 2011). Biophysical analysis of RLD2 and this Ufrag region suggest that this region alone can account for a significant portion of the full-length UBE3A+RLD2 interaction, but not all of it. ITC analysis of the MBP-Ufrag binding to UBE3A suggests a Kd that is roughly half that of RLD2 into full-length UBE3A. Both values are within the same order of magnitude, suggesting that the Ufrag segment does mediate a large portion of the full UBE3A interaction, but it is still a lower Kd, which could suggest that further RLD2 interacting regions exist within UBE3A. An ITC experiment of isolated MBP into RLD2 suggested that the MBP tag does not contribute to the interaction, but as the isolated Ufrag species was too difficult to purify on its own, I are unable to determine whether the MBP tag could interfere negatively in the interaction. If the Ufrag section of UBE3A is sufficient to mediate the UBE3A+RLD2, and by extension UBE3A+HERC2 interaction, then it would be expected that the effect of RLD2 on UBE3A's catalytic activity is due to internal rearrangements of the UBE3A structure. However, the CD data suggests that this is not the case (section 4.4.1). The alternative explanation would be that the Ufrag region of UBE3A sits proximal to the catalytic site of UBE3A's HECT domain in the tertiary structure, despite its distal position in the protein sequence. As a structure of full-length UBE3A has not yet been solved, this remains a possibility.

The relative stability of the UBE3A+PSMD4 and UBE3A+RLD2 complexes is observed in the SEC and SDS-PAGE results following crosslinking of each sample. The trace for the crosslinked UBE3A+PSMD4 complex is very messy. Rather than a single peak for each species present, as observed for the native sample (Fig. 65), the crosslinked trace has devolved into a very broad peak with several indistinct subpeaks mixed in. The SDS-PAGE representation of fractions within the peak fail to show a clearly identifiable UBE3A+PSMD4 complex species (Fig. 81). Un-linked PSMD4 and UBE3A species can be observed towards the later fractions, but any 1:1 ratio complex species has co-eluted with a significant amount of non-specific aggregate species. In contrast to this, the trace for UBE3A+RLD2 is much more interpretable (Fig. 82). The SDS-PAGE profile for the trace fractions suggests that two core species, un-bound monomeric RLD2 and a 1:1 UBE3A+RLD2 complex, can be

identified and isolated to adequate purity for structural studies before the reminder of the sample devolves into indistinguishable aggregate.

5 Biochemical Activity

5.1 *in vitro* Ubiquitination Assay

As the archetypal HECT ligase, UBE3A performs its ubiquitin ligase activity as part of an E1-E2-E3 enzyme catalytic cascade. The E1 ubiquitin activating enzyme interacts with free-ubiquitin in cells to activate the thioester bond. The E2 ubiquitin carrier enzyme receives the activated ubiquitin unit from the E1 enzyme and carries it to the E3 enzyme, where the E3 ubiquitin ligase catalyses the transfer of ubiquitin onto the substrate protein (Scheffner *et al.,* 1993; see section 1.1, Fig. 2). In order to ensure that the UBE3A protein purified in chapter 3 is the active form, I performed an *in vitro* ubiquitination assay. The basic assay was prepared as described in section 2.8.1, with a commercial his-E1 enzyme catalysing the first step of the reaction. The preferred cognate E2 enzyme for UBE3A has been determined experimentally as UbcH7 (Eletr and Kuhlman, 2007), that was expressed and purified as described in 3.2.2 and 3.4.2. The assay was performed with UBE3A and also in the presence of proteins that have been identified as either potential substrates or accessory proteins of UBE3A.

5.2 UBE3A Activity

5.2.1 UBE3A Autoubiquitination

UBE3A has been shown to autoubiquitinate to regulate its activity both *in vivo* and *in vitro* (Nuber *et al.,* 1998). Due to this observation, UBE3A was initially used as its own substrate. The assay was prepared as described in section 2.8.1, and the various timepoints and controls were visualised through SDS-PAGE and a Coomassie-based stain (Fig. 83).

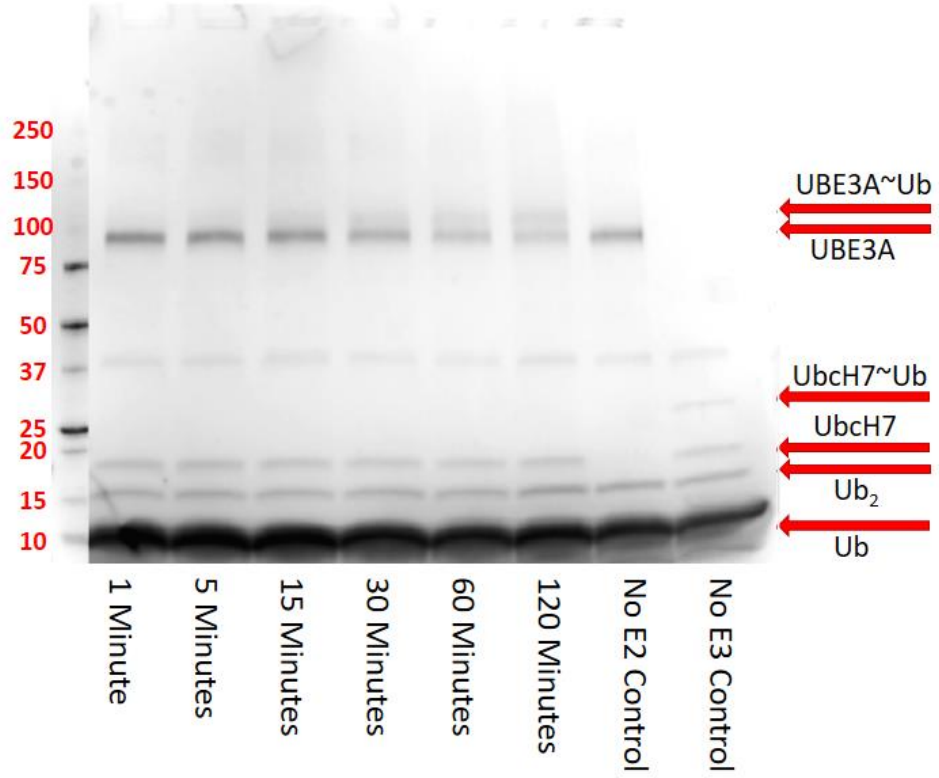


Figure 83: The in vitro autoubiquitination assay of UBE3A. The assay was prepared as described in section 2.8.1 and incubated at 37°C for a total of two h. The time points were taken in increasing intervals and quenched with SDS dye ready for SDS-PAGE. The control reactions were also left for the full two h and quenched in the same way.

UBE3A substrates are ubiquitinated following the mechanism outlined in figure 2 of section 1.1. As UBE3A is a HECT ligase, the activated ubiquitin moiety is first transferred onto the catalytic site of UBE3A, forming the thioester intermediate, before the ubiquitin is transferred to the substrate. Ideally, UBE3A activity would be measured by tracking the increase in higher molecular eight species as substrates become ubiquitinated, but this was difficult to determine using only a Coomassie stain. Instead, I attempted to use the decrease in the UBE3A band as a proxy for UBE3A activity, with the assumption that UBE3A-only band decreases and the enzyme becomes auto-ubiquitinated to form products with varying molecular weight. The absence of higher bands and the intensity of the UBE3A band in the no E2 control lane suggests that the change that occurs throughout the reaction is not an aggregation effect, although it was not possible to fully determine whether the increasing band above the UBE3A-only band represented a mono-ubiquitinated UBE3A species or the thioester intermediate state. The no E3 control lane also shows the presence of a UbcH7~Ub thioester intermediate. This species exists only transiently in the presence of a functioning E3 enzyme, but in the absence of UBE3A it can accumulate.

5.2.2 UBE3A Activity in the Presence of RLD2

The interaction between UBE3A and HERC2 has been characterised previously and the region of HERC2 involved in the interaction has been identified as the RLD2 domain, comprised of amino acids 2959-3327 (Kühnle *et al.,* 2011). The isolated RLD2 domain has been previously shown to increase the rate of UBE3A activity *in vitro* (Kühnle *et al.,* 2011), so I decided to see if I could replicate this effect and study this interaction further. RLD2 was expressed and purified (as described in 3.2.4, 3.4.4) and included in the *in vitro* ubiquitination assay as described in section 2.8.1. The time points and controls were run on SDS-PAGE and stained with a Coomassie-based stain in order to observe the effects (Fig. 84).

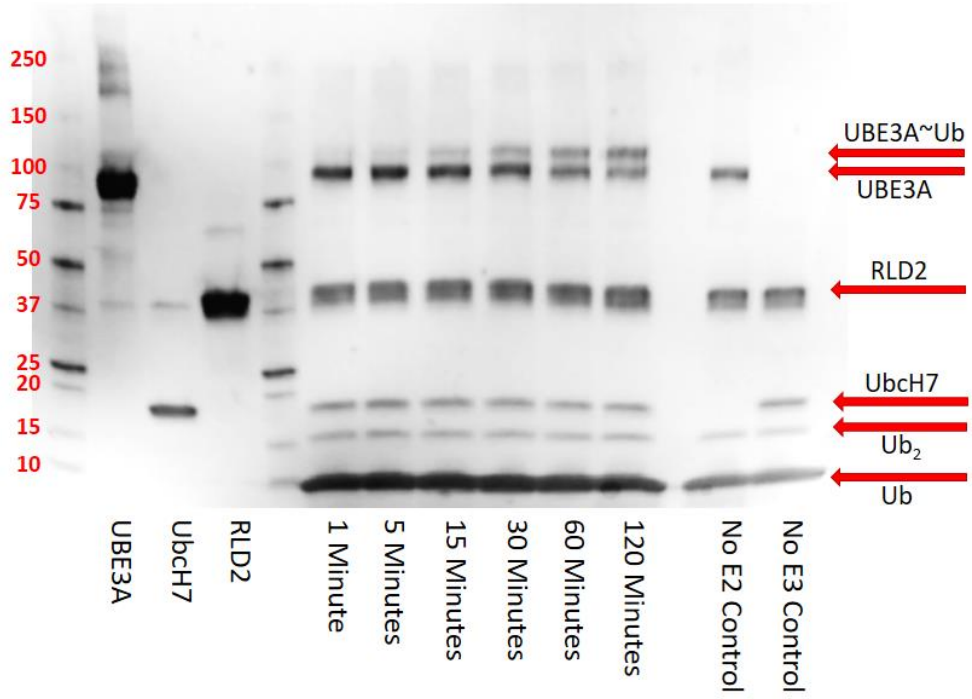


Figure 84: The in vitro ubiquitination assay of UBE3A in the presence of RLD2. The reaction was prepared as described in section 2.8.1, and time points were taken in increasing intervals and quenched with an SDS dye before visualisation with SDS-PAGE and a Coomassie-based stain. The constituent enzymes, other than the commercial E1 enzyme, were also subjected to SDS-PAGE at a higher concentration than present in the assay in order to observe any impurities that may be present in the assay samples.

The RLD2 sample appears to form two bands across all lanes, but the presence of both lanes in the control samples as well suggests that it may be a mix of his-tagged and cleaved sample left over from the purification process rather than a feature of the assay. The assay shows the decrease in intensity of the UBE3A band along with the increasing band for UBE3A~Ub that was observed in the assay without RLD2 (Fig. 83) that suggests UBE3A mono-ubiquitination. The smear above the UBE3A bands in the assay fractions may

represent longer chains of ubiquitin bound to UBE3A, but the presence of higher molecular weight species in the high concentration UBE3A sample makes it difficult to attribute that to a feature if the assay.

Along with the assay shown in figure 84, another UBE3A-only assay was run with the same samples at the same concentration to ensure as much consistency as possible between the two gels, other than the presence of RLD2. The bands for UBE3A, UBE3A~Ub, and RLD2 for each assay were then subjected to densitometry using the ImageJ software, as described in section 2.8.3, in order to determine whether the addition of RLD2 to the assay had an effect of the ubiquitination activity of UBE3A or not (Fig. 85).

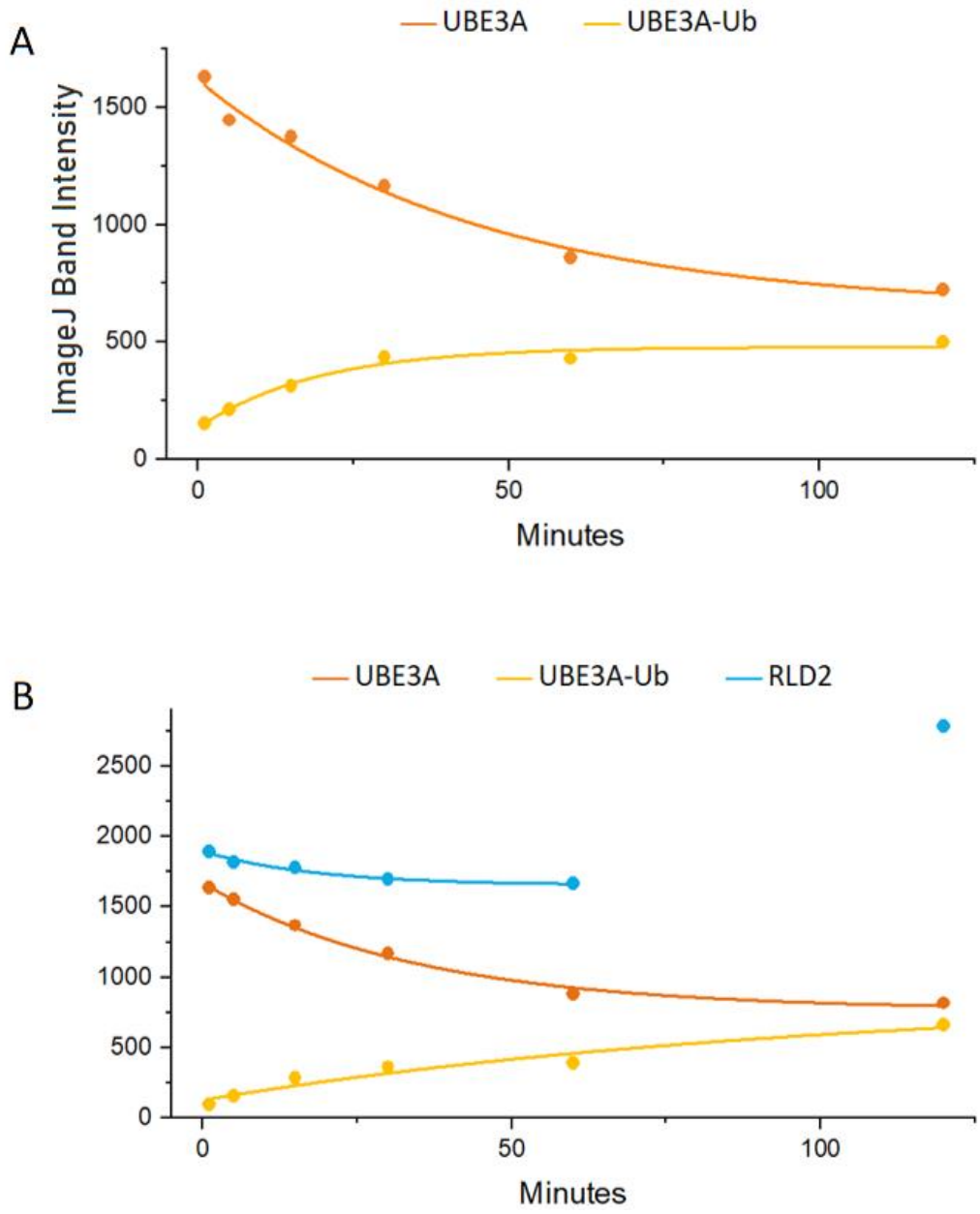


Figure 85: Densitometry analysis of the ubiquitination of UBE3A with and without RLD2. A) The arbitrary ImageJ values for the intensity of the bands for UBE3A and UBE3A~Ub across the UBE3A-only assay. B) The arbitrary ImageJ values for the intensity of the bands for UBE3A, UBE3A~Ub, and RLD2 across the UBE3A + RLD2 assay. The datapoints for each protein were fitted to a first order decay, shown as a solid line in either image, with data for UBE3A shown in orange, UBE3A-Ub in yellow, and RLD2 in blue. For the RLD2 data the final datapoint was omitted from the fit calculation to allow fitting of the data to an exponential function.

In order to compare the rate of the reaction across different assay conditions the data was fitted to a first order decay process which is defined as an exponential decay function:

$$A(t) = A_0 \exp(-k_{obs} t)$$

Where $A(t)$ is the time dependent decay of the signal (in this case the density of the band), k_{obs} is the observed rate constant and t is time. As the *in vitro* autoubiquitination assay uses UBE3A as its own substrate, the loss of un-ubiquitinated UBE3A was used as a proxy for UBE3A's HECT ligase activity. The exponential fit for the UBE3A data in each graph was used to determine the rate of loss of UBE3A for each reaction, resulting in a rate of 0.022 min^{-1} for the UBE3A-only reaction and a rate of 0.029 min^{-1} for the reaction involving RLD2. The change in the rate in the presence of RLD2 is only slight, but it does suggest an increase in activity. A key point to note from the UBE3A+RLD2 reaction (Fig. 85b) is the unexpected behaviour of the RLD2 enzyme. The RLD2 species appears to initially decrease in intensity across the time points of the reaction, and this can be fitted to a first order decay equation as with the UBE3A species, but the final datapoint at the end of the reaction shows a large increase in RLD2 presence. One possibility is that this final datapoint is an anomaly from inaccurate loading of the timepoint sample on the gel, but the other assay components appear to be present in expected quantities in the same sample, and also the same pattern can be seen in the assay involving both RLD2 and PSMD4 (see section 5.2.4, Fig 89b.) Another possibility is that RLD2 is initially involved in the reaction in some way that reduces the amount of native RLD2, maybe through an interaction with UBE3A to form a larger species that is not fully denatured upon SDS-PAGE, but it is then released back into its native form past a certain point. If RLD2 is involved in increasing the rate of UBE3A autoubiquitination through forming a tight interaction, it may then be released as the UBE3A species becomes ubiquitinated, releasing more free-RLD2 into the reaction mixture as less free-UBE3A is available for it to bind to.

5.2.3 UBE3A Activity in the Presence of PSMD4

The proteasomal shuttle protein PSMD4 has also been associated with UBE3A activity. It has been identified as a potential substrate of UBE3A *in vitro* (Lee *et al.,* 2014), but it has also been implicated in other aspects of UBE3A activity. It has been suggested that it interacts with UBE3A to carry out its function as a shuttle protein and bring UBE3A to the proteasome (Buel *et al.,* 2020). This could be to allow UBE3A to target subunits of the proteasome for degradation, as they are identified substrates of UBE3A, but it could also be to allow more efficient degradation of other cellular targets (Buel *et al.,* 2020). I expressed and purified PSMD4, as described in sections 3.2.3, 3.3,2 and 3.4.3, and included it in the assay with UBE3A as described in section 2.8.1. The assay samples were subjected to SDS-PAGE and stained with a Coomassie-based dye for visualisation (Fig. 86).

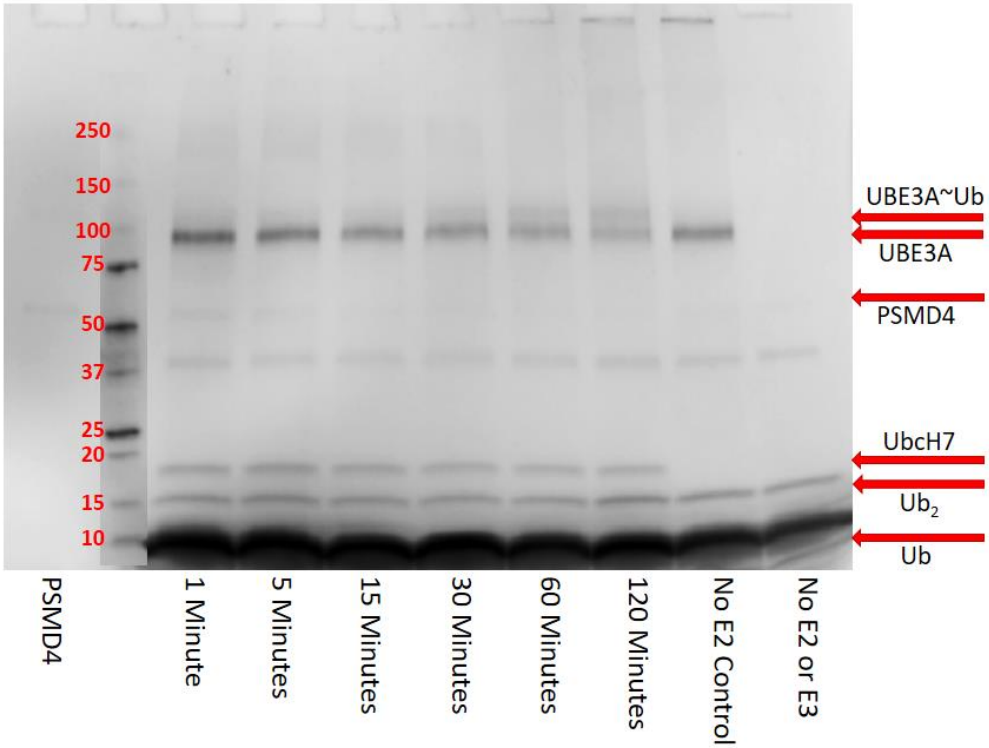


Figure 86: The <u>in vitro</u> ubiquitination of UBE3A in the presence of PSMD4. A sample of PSMD4 at the concentration it appears in the assay is shown on the left of the gel, and the band in the assay samples is labelled. A contaminant band is seen across all lanes of the assay samples, but its presence in both control lanes, including the lane without any E2 or E3 enzymes, precludes it as a product of the assay reaction. The ubiquitination of UBE3A can still be observed by the increasing presence of the UBE3A~Ub band and the concomitant decrease in intensity of the UBE3A only band.

UBE3A ubiquitination can still be observed in the presence of PSMD4, so whatever the interaction between PSMD4 and UBE3A is it does not interfere with the activity of the HECT domain. In order to determine if it is able to increase the activity of UBE3A, densitometry analysis was performed on the UBE3A and UBE3A~Ub bands of both the PSMD4 assay and a simultaneously performed UBE3A-only assay (Fig. 87). The PSMD4 band was also analysed to determine if it decreases across the reaction time as a result of becoming ubiquitinated, but the concentration of PSMD4 in the assay to begin with was fairly low so any decrease in intensity of the bands is difficult to determine.

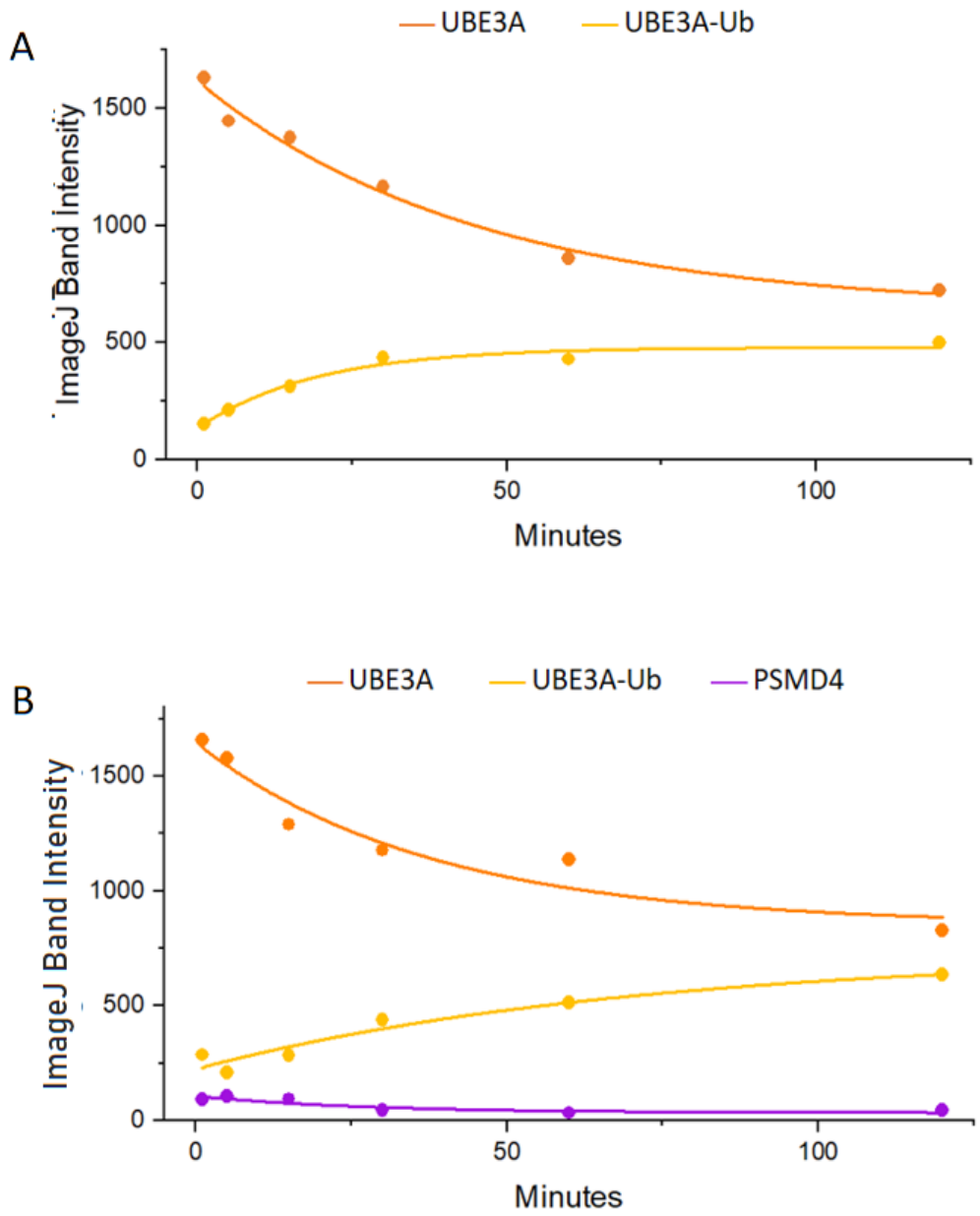


Figure 87: in vitro ubiquitination activity of UBE3A in the presence of PSMD4. A) The absolute values for the intensity of each band in the UBE3A-only assay gel, plotting over the time-course of the reaction. B) The absolute values for each band over the time-course of the UBE3A + PSMD4 assay. The datapoints for each protein were fitted to an exponential curve to allow determination of a rate of loss of free-UBE3A, as shown by the solid line in each graph. UBE3A data is shown in orange, UBE3A-Ub in yellow, and PSMD4 in purple.

The rate of the reaction, as determined by tracking the rate of loss of free-UBE3A, including PSMD4 was calculated as 0.027 min^{-1}, compared to the rate of 0.022 min^{-1} for the assay without PSMD4. This suggests that adding PSMD4 to the reaction may increase the rate of UBE3A ubiquitination. However, the exponential fit of the UBE3A data set is much less accurate for the assay in the presence of PSMD4, so the rate of the loss of UBE3A is much less reliable than the previous data sets.

Whereas RLD2 did not seem to act as a typical substrate for UBE3A (Fig. 85b), the amount of free-PSMD4 in the reaction does seem to decrease as the reaction progresses, as would be expected from a more typical substrate protein (Fig. 87b, purple line). The PSMD4 intensity data could be fit to an exponential curve similarly to the UBE3A data, and the rate of loss of PSMD4 was calculated as 0.014 min^{-1}, which is less than the rate of autoubiquitination of UBE3A, but shows the negative trend associated with a transition from free-PSMD4 to variably ubiquitinated PSMD4 species. The amount of PSMD4 in the reaction mixture was difficult to quantify in every time point sample due to its low concentration (Fig. 86) so calculation of the rate of decrease of PSMD4 is not very reliable, but the general trend of the data does appear to show a decrease in the presence of that species rather than the spurious trend shown by the RLD2 sample in its concomitant assay. This is supported by previous observations that PSMD4 acts as a substrate of UBE3A (Kühnle *et al.,* 2018).

5.2.4 UBE3A Activity in the Presence of RLD2 and PSMD4

Both RLD2 and PSMD4 have been identified as potential interactors of UBE3A, but how either interacts with UBE3A is not yet fully understood. Both PSMD4 and RLD2 samples, purified as described in section 3, were included in the *in vitro* ubiquitination assay of UBE3A to see if any further effect could be observed with both proteins present, which could potentially suggest that both proteins interact with different areas of UBE3A, or whether one protein would preclude the binding of another. The assay was prepared as described in section 2.8.1, and the samples were run on SDS-PAGE and stained with a Coomassie-based dye (Fig. 88).

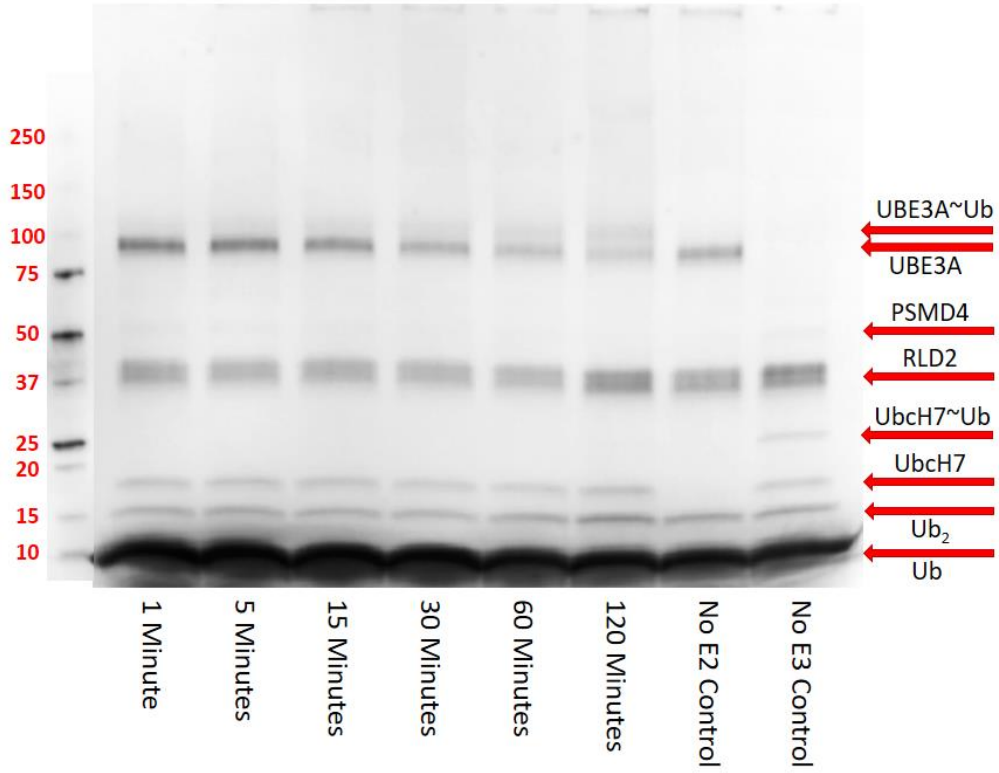
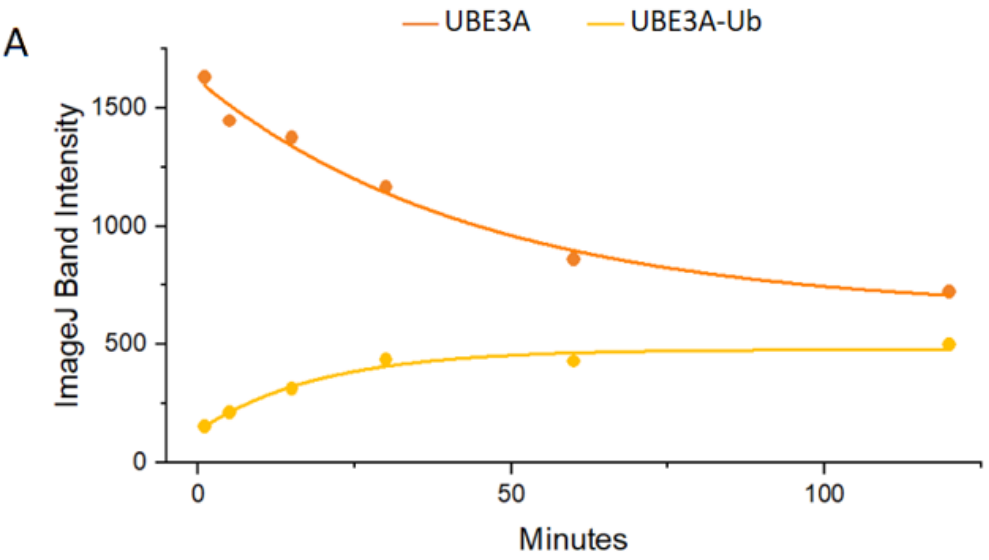
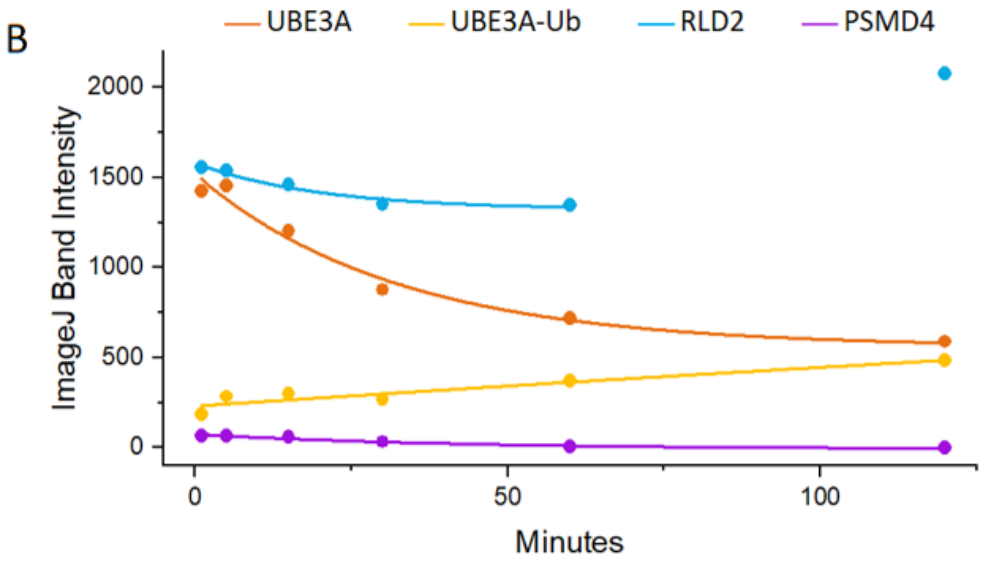


Figure 88: An in vitro ubiquitination assay involving UBE3A, RLD2, and PSMD4. The enzymes and visible ubiquitinated products are labelled on the right and the molecular weight standard is labelled on the left. The band for PSMD4 is faint across all samples, including the controls, but all other species are easily identifiable. The ubiquitination process can be followed by the increasing UBE3A~Ub band across the assay time points and a corresponding decrease in intensity of the UBE3A band.

The ubiquitination of UBE3A can be observed in the above figure so the HECT domain of the protein is still active in the presence of both proteins, but in order to make any further observations on the effect on UBE3A catalytic activity, the assay gel was subjected to densitometry analysis, as described in section 2.8.3 (Fig. 89).


A
UBE3A
UBE3A-Ub
ImageJ Band Intensity
1500
1000
500
0
0
50
100
Minutes


B
UBE3A
UBE3A-Ub
RLD2
PSMD4
ImageJ Band Intensity
2000
1500
1000
500
0
0
50
100
Minutes

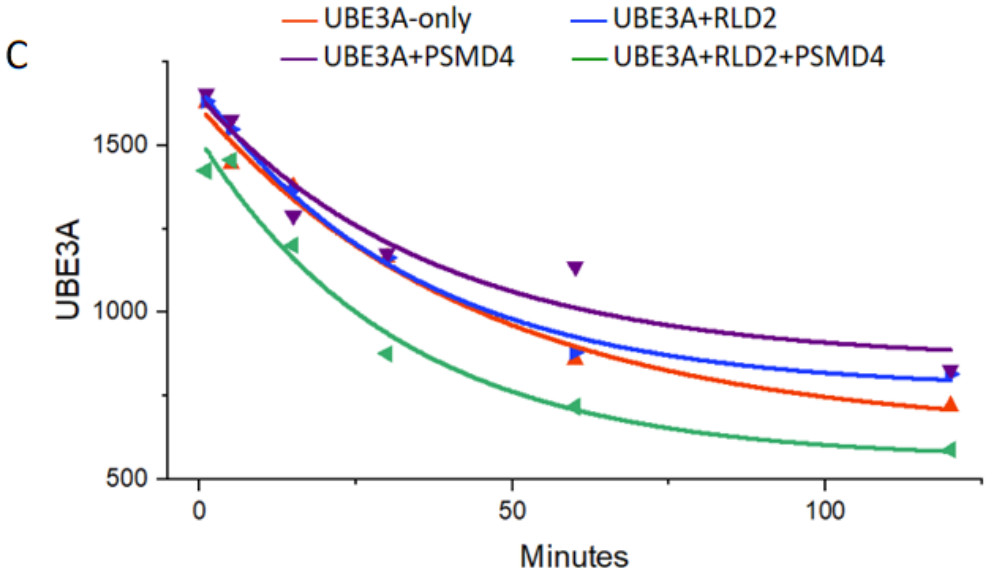

C
UBE3A-only
UBE3A+RLD2
UBE3A+PSMD4
UBE3A+RLD2+PSMD4
UBE3A
1500
1000
500
0
50
100
Minutes

Figure 89: The in vitro ubiquitination assay in the presence of both RLD2 and PSMD4. A) The arbitrary absolute values for the intensity of each band present across the UBE3A-only assay. An exponential fit for each protein is shown as a solid line. B) The arbitrary absolute values for the intensity of each band present across the UBE3A+RLD2+PSMD4 assay. An exponential fit for each protein is shown as a solid line. C) A comparison of the rate of decrease of UBE3A across each of the different assays. Each data set was fitted to a first order decay equation, with the dark orange line shows the intensity of the UBE3A band across the UBE3A-only assay, the blue line shows UBE3A presence across the assay in the presence of RLD2, the purple line shows UBE3A across the assay with PSMD4, and the green line shows the rate of UBE3A decrease across the assay containing both RLD2 and PSMD4.

Assay	**Δ UBE3A**	**Δ UBE3A-Ub**	**Δ RLD2**	**Δ PSMD4**
UBE3A-only	0.022	0.052	/	/
UBE3A+RLD2	0.029	0.01	0.0542	/
UBE3A+PSMD4	0.027	0.014	/	0.038
+RLD2+PSMD4	0.031	0.001	0.0497	0.024

Table 9: The rate constants for each of the constituents of the UBE3A in vitro ubiquitination assays. The units for each rate constant is min^{-1}. The rate constants were calculated by fitting each dataset to a first order decay reaction, as described in section 5.2.2. For the RLD2 data in both reactions the data was fitted excluding the final anomalous datapoint.

In the assay with both RLD2 and PSMD4 present, the calculated rate of decrease of UBE3A, calculated using the exponential fit of the data, was 0.031 min^{-1}. This suggests a faster rate of the reaction than in all three assays tried up to this point. In addition to the loss of UBE3A, as PSMD4 appears to act as a substrate (see section 5.2.3), the data for the PSMD4 values across the assay were also fitted to an exponential curve, and the rate of loss of PSMD4 was calculated as 0.024 min^{-1}. The increase in the rate of loss of PSMD4 in the presence of RLD2 could suggest that RLD2 may increase the rate of both autoubiquitination of UBE3A and trans-ubiquitination of substrate proteins by UBE3A, even though the two forms may be enacted through different mechanisms (Kao *et al.,* 2000; Ronchi *et al.,* 2017). Although PSMD4 appears to act as a substrate for UBE3A, it also appears to increase the activity of UBE3A, as the rate of the reaction increases with its presence, both with and without RLD2 present. PSMD4 has been suggested to interact with UBE3A *in vivo* beyond its role as a ubiquitination substrate (Buel *et al.,* 2020), but its effect of the catalytic activity of UBE3A has not been studied.

When the rate of increase of the UBE3A-Ub species is observed across all assays (Table 9), it shows a decrease in the generation of the species when the other assay components are added. This could suggest that the UBE3A-Ub

species does in fact represent the monomeric ubiquitinated UBE3A species and the decrease in its generation represents a change in the preferred substrate of UBE3A, or it could mean that the UBE3A-Ub species is the thioester intermediate, and the rate of the second step in UBE3A ubiquitination, the transfer of the ubiquitin onto the substrate, is increased when other components are present.

Ultimately, the data for the assays described here is not as robust as I would have liked. Modelling the activity of UBE3A through its mono-ubiquitination is not ideal, as it does not differentiate between actual autoubiquitinated UBE3A and UBE3A in its thioester intermediate state with ubiquitin bound to the active site ready for transfer to a substrate. A more accurate method of determining the rate of UBE3A HECT ligase activity would have been to carry out an anti-Ub western blot of the assay gels and then use the intensity of higher molecular weight ubiquitinated species as a measure of ubiquitinated substrates. Unfortunately, attempts to quantify the activity in this manner were made, but were ultimately unsuccessful. The anti-Ub Western blots carried out as part of this experiments were not able to identify even the free-Ub species present in the sample, let alone the higher molecular weight species representing variously multi-ubiquitinated products that I assume would be present in a successful *in vitro* ubiquitination assay.

Another issue with the modelling of the data here is that the assays were each carried out with only a single substrate concentration, so I cannot begin to extract any mechanistic data. The mechanism of ubiquitination by UBE3A comprises of two steps, the first is a concentration-dependent bi-molecular association between UBE3A and the activated Ub bound to UbcH7, while the second is a concentration-independent transfer of ubiquitin from the active site to another site either within UBE3A or another substrate molecule. If the assay conditions had been applied to a range of substrate concentrations I would be able to determine whether the whole reaction was concentration-dependent or not, which would provide an insight into which step of the mechanism is the rate-limiting step. As it stands, I was able to demonstrate some activity of UBE3A, and some rudimentary observations could be made, but given more time and resources these experiments could be carried out again to provide more reliable observations as well as more mechanistic insights.

A further issue with the data presented in this chapter is that each datapoint is represented as only a single datapoint, with no replicates shown. This prevents any meaningful interpretation of the data as it stands, as it prevents elucidation of the robustness of the observations. Throughout the project the assays were repeated multiple times, but inconsistencies with the experimental setup, such as varying protein concentrations or time points, prevented the compiling of all of the data into a single dataset. The patterns observed across the data shown here do suggest that there is some

biochemical activity occurring, but any meaningful interpretation of this would require a much more thorough experimental plan, even if the technical issues addressed in the preceding paragraphs could be addressed. The work presented in this chapter may not provide any conclusions on the mechanism of activity of UBE3A, either alone or with binding partners, but it does stand as a proof of concept of the assays described, and could provide the starting point for a more in depth analysis of the implication of the interactions between UBE3A and its various partner proteins.

6 EM Sample Preparation and Optimisation

Although the requirements for a cryo-EM sample are not quite as rigid as a sample for x-ray crystallography, in that a sample may work in a larger variety of buffer conditions and there is less of a requirement for absolute purity of the sample, the sample preparation process is still a fairly involved process that typically requires a lot of optimisation prior to data collection. Some proteins will have very specific requirements for ice conditions and grid types, and it is not usually possible to work out what will or will not work before trying a certain condition. A lot of the sample optimisation process for cryo-EM grids is a trial and error approach, where a lot of different conditions must be trialled in order to find one that will work. Typically, optimisation of cryo-EM samples will involve several stages, including negative staining, repeated cycles of grid preparations and grid screening, and finally optimisation of data collection parameters before a structure can be obtained (Fig. 90).

Figure 90: The process of optimising a sample for data collection involves several steps before a protein structure can be generated. Once a sample has been prepared, negative stain is used to confirm the sample quality and concentration range, and then several successive rounds of grid preparation and screening are typically required to optimise the grid conditions. Following grid optimisation, the data collection parameters must be optimised for the sample, and then the data analysis parameters also require careful optimisation.

6.1 Negative Stain

In order to confirm that homogeneous particles are present, consistent with the requirements of single particle reconstruction methods, samples were first characterised using negative staining. This involves mixing the sample with a form of stain that stains the background of the images, leaving the protein unstained. Several different negative stain solutions exist, but I used a 1% uranyl acetate solution as it provides the best mix of high contrast due to its electron density, high resolution due to its relatively fine grain size, and also easy storage (De Carlo and Harris, 2011; Scarff *et al*., 2018). Grids were prepared for UBE3A as described in section 2.10.1, and then imaged on the JEOL 2100 microscope at the Research Complex at Harwell (Fig. 91).

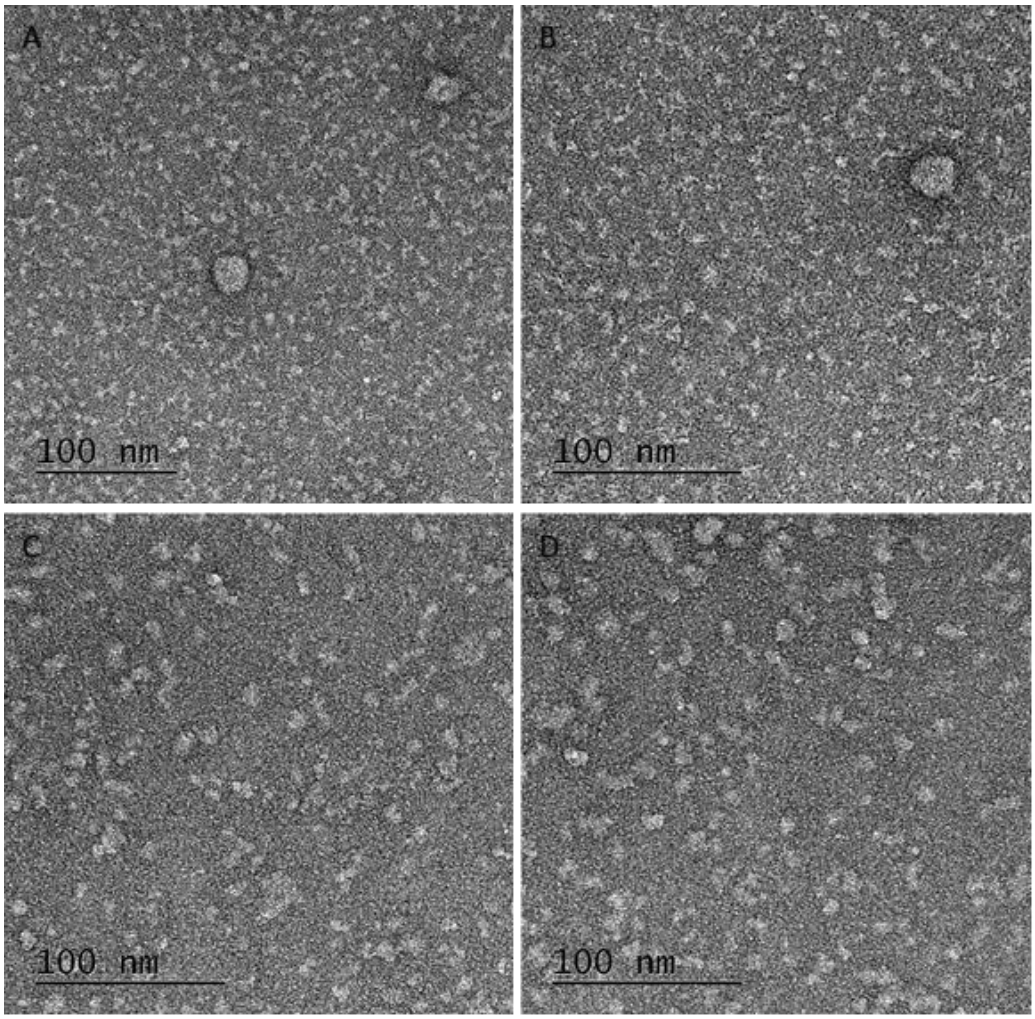


Figure 91: Negative stain images of UBE3A stained with a 1% uranyl acetate solution. A) Image taken at 60k magnification with a protein concentration of 0.025 mg/ml. B) Image takes at 80k magnification, with a protein concentration of 0.025 mg/ml. C) Image taken at 80k magnification with a protein concentration of 0.0125 mg/ml. D) A second image taken at 80k magnification with a protein concentration of 0.0125 mg/ml.

The negative stain images can sometimes be used to create a low resolution initial model to provide a starting point for cryo-EM image analysis, but they are also useful to get an idea of the rough shape and size of the protein, as well as the quality and useful concentration range of the sample. Although the images above were not processed any further, they do demonstrate that the UBE3A sample was not prone to excessive aggregation, and they suggest a concentration range that can be used as a guide for cryo-EM grids. Negative stain grids typically require a sample concentration 10-20 times lower than that for cryo-EM grids, so the negative stain dataset collected for UBE3A suggests that a concentration range between 0.1 and 0.5 mg/ml would be suitable for cryo-EM data collection.

6.2 Grid Types

One consideration when optimising sample conditions is the type of grid that is used. The most commonly used cryo-EM grids for biological samples have a regular pattern of holes within a uniform carbon film on top of a metal support grid. The standard pattern comprises holes of 2 μm diameter with 2 μm between each hole, but there are various different grids available with smaller or larger holes and different spacings. Grids with smaller holes can be useful for getting really thin ice, as the thinnest ice in the centre of the hole needs to be sustained over a much smaller surface area (Glaeser *et al.,* 2016; Fig. 92c). Lacey-carbon and holey-carbon grids also exist and can be useful for determining the ideal hole size for a sample as both contain a variety of different sized holes to allow sampling of various ice thicknesses at once. Lacey carbon grids look less like a uniform carbon film with holes in it and more like a net formed with a carbon string. They are typically more hole than carbon. Holey-carbon grids however are formed from a single uniform carbon layer that has been treated, typically by sonication, to create a number of irregular holes in the film. Unlike lacey-carbon grids these grids are typically more carbon than hole.

Adding an extra support layer over the carbon film can also help with modulating the ice thickness, as it creates a flat edge for one side of the ice, halving the effect of the ice gradient and making it easier to make fine adjustments to the ice thickness. An extra support layer also negates the air-water interface for that half of the grid, creating a larger usable space within the same thickness of ice (Drulyte *et al,* 2018; Glaeser *et al.,* 2016; Fig. 92b). This allows more orientations to be present within a thinner ice layer. The most common extra support layers are solid graphene, which is extremely thin but naturally hydrophobic and fairly difficult to apply to grids, or a graphene oxide-detergent mix, which is much easier to apply but provides a much less uniform coating (Drulyte *et al.,* 2018).

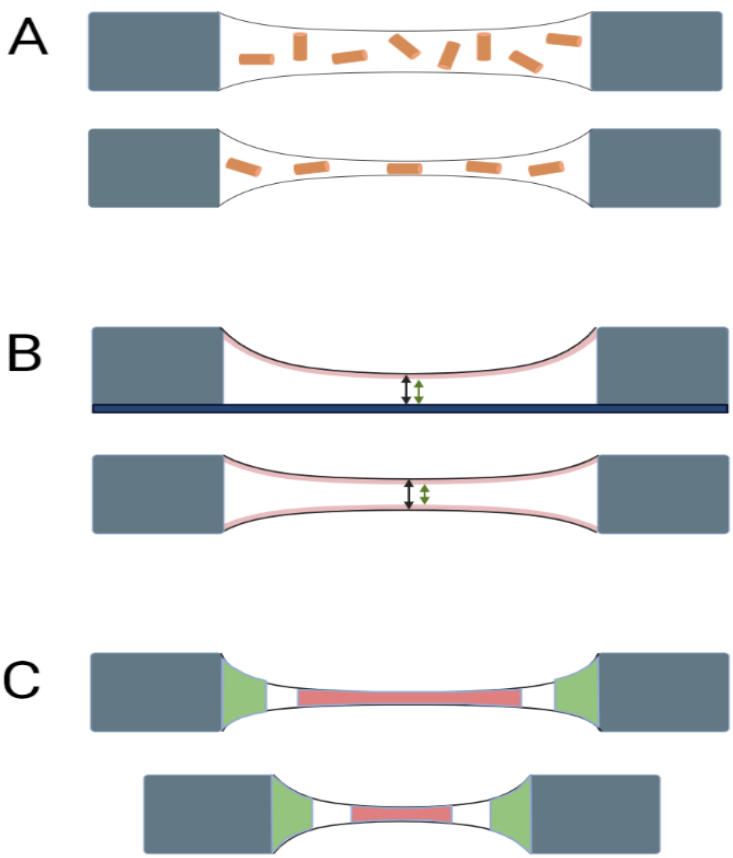


Figure 92: The effects of different grid conditions on ice thickness and particle distribution. A) A thicker ice layer within the holes of the carbon film allows objects to sit in multiple orientations within the ice while a thin layer of ice restricts the possible orientation of objects suspended in the ice, limiting the number of different views available for the 3D reconstruction of the object. B) An extra support layer such as graphene or graphene-oxide creates more usable ice within the ice thickness due to the elimination of the air-water interface on one side of the grid. C) Smaller holes support thinner ice regions due to the decreased surface area needed to support the thin ice.

UBE3A isoform 1 is 97kDa, which is quite small for a cryo-EM sample. This means that in order to see the particles clearly upon imaging the ice must be as thin as possible, without being too thin to affect the orientation sampling. This requirement led to a lot of optimisation of conditions through repeated rounds of sample preparation and screening on an FEI 200 kV Talos Arctica or Glacios microscope. Due to the small size of UBE3A, QuantiFoil R1.2/1.3 grids were used with a graphene oxide (GrOx)-DDM coating, in order to generate the thinnest ice possible (Fig. 93).

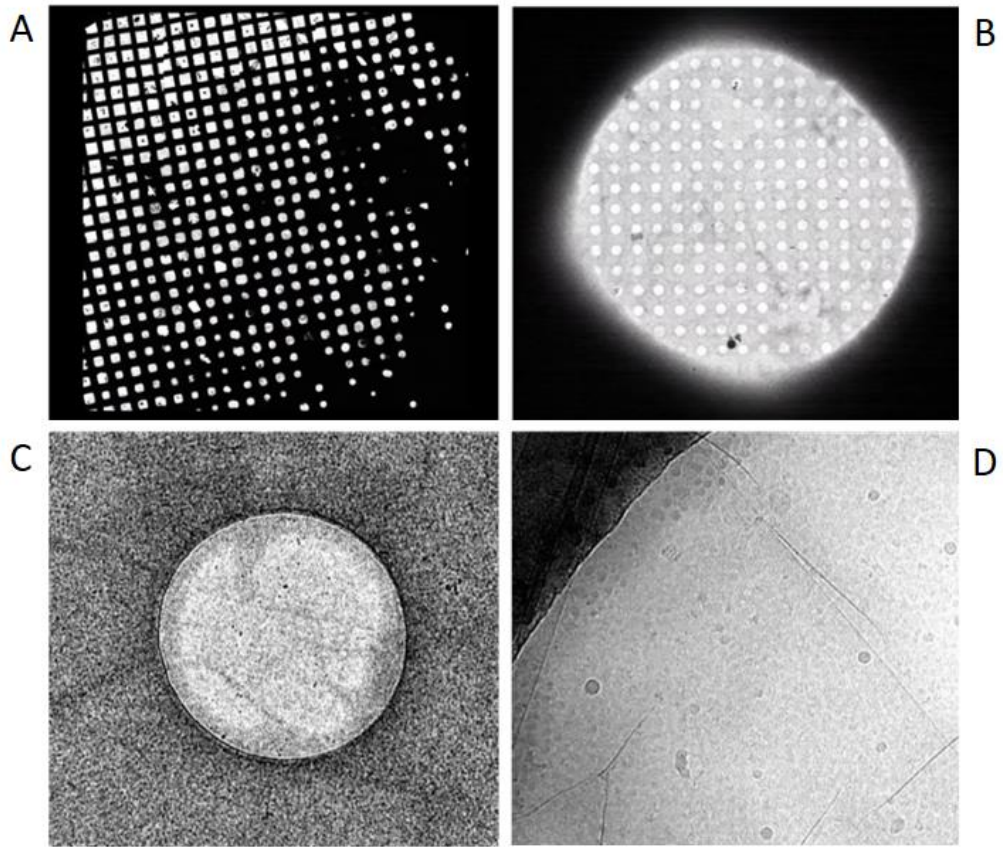


Figure 93: Representative images of a cryo-EM sample on a 300 mesh copper QuantiFoil R1.2/1.3 grid with a hand-applied GrOx-DDM coating. A) An Atlas image of the grid. This is generated by taking images at the 'Atlas' magnification setting across the grid and then joining them together to create one single image of the entire grid. At this magnification you can see the mesh size of the grid. B) An image taken at the 'GridSquare' magnification setting. At this magnification you can begin to get an idea of the ice thickness within each hole, and you can also see the graphene-oxide coating. C) An image taken at 'Hole' mag. In this image you can see the lines across the hole caused by folds or overlapping sheets of graphene-oxide. D) An image taken at 'Data Acquisition' magnification. At this magnification particles embedded in the ice are observed. The fold of graphene-oxide are also much clearer in this image.

The GrOx-DDM coating did allow us to generate very thin ice within the holes, but it also introduced artefacts into the images that made processing the data difficult. The GrOx is applied as a suspension of differently sized GrOX flakes, so it is difficult to ensure only a single layer is applied. This led to the observation of creases and overlapping areas of GrOx across the grid. This led to some contrast issues, as although the ice was thin the advantages of that were outweighed by the extra signal generated when there were several layers of GrOX. The main disadvantage to the uneven application, however, was that the processing software was not very adept at distinguishing the GrOx folds from particles, so many particles were missed in favour of the GrOx artefacts, and the software had difficulties in separating the noise from the

real particles that were picked. Because of these issues, more grids were made using the same QuantiFoil R1.2/1.3 grids but without the GrOx-DDM application (Fig. 94).

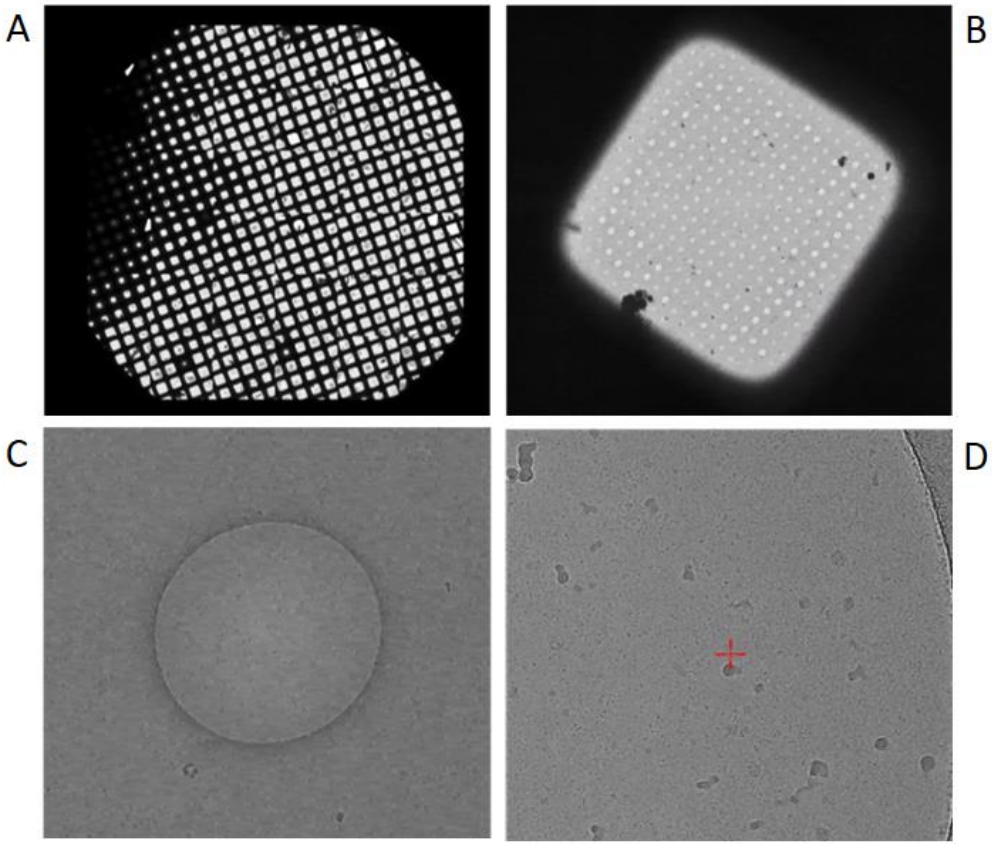


Figure 94: Representative images of a cryo-EM sample on a 300 mesh copper QuantiFoil R1.2/1.3 grid without any extra coatings. A) An atlas image showing the distribution of different ice thicknesses across the grid. B) A gridsquare image, showing the distribution of ice thicknesses within an individual gridsquare. Here you can see a mixture of dry holes and holes that contain an ice gradient. The dry holes are distinguishable by their uniform colour, while the holes that contain an ice gradient appear darker around the rim and lighter towards the middle of each hole where the ice gets thinner. C) An image of a single ice hole. D) A high magnification image of the sample within the hole. You can see the particles embedded in the ice, distinct from the ice contamination that is also present.

The R1.2/1.3 grids without the GrOx-DDM coating allowed us to be much more consistent with the grids that I made, which made it easier to optimise the other settings, such as blot time, glow discharge durations, and sample concentration. One key difference between the ice from the GrOx-DDM and the plain R1.2/1.3 grids however was the presence of an ice gradient within each hole. The GrOx-DDM grids appear much more uniform across the majority of the hole, while the non-coated grids show much thicker ice around the edges of the holes than in the middle. This is most easily seen in the gridsquare images (Fig. 93b, Fig. 94b) and is due to the effect of the GrOx

layer on the surface tension of the sample within the hole before freezing (Fig. 92b). This could cause problems as the proteins become excluded from the centre of the grid if the ice becomes too thin while still overlapping in the thicker ice around the edges, making separating particles during processing more difficult. However, the relatively large magnification size needed for small particles means that only a relatively small area of each hole will be able to be imaged, which allows specific targeting of the area in the middle of the gradient where the ice is still thin enough for contrast and thick enough for particles in all orientations.

While R1.2/1.3 grids provided good ice conditions, the protein sample had a tendency to cluster on the carbon film around the edges of the holes. One attempt to get around this was to use grids with an ultrathin layer of continuous carbon across the grid. This support layer was most easily acquired as a 3nm carbon film over LaceyCarbon grids, which makes data collection and processing less consistent, but can also allow sampling of a wider range of ice conditions (Fig. 95).

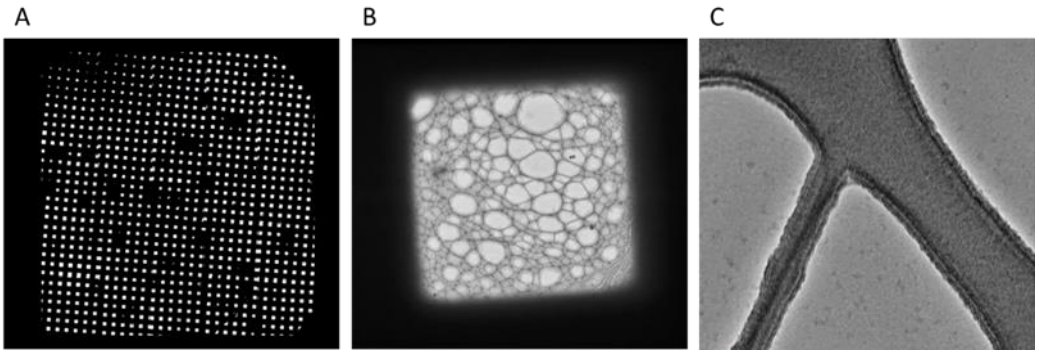


Figure 95: A 400 mesh copper grid with a LaceyCarbon film and an extra layer of super fine continuous carbon film. A) An atlas image showing the smaller grid squares relative to the 300 mesh grids due to the increased fineness of the mesh support. B) A GridSquare image showing the lack of regular repeated holes in the carbon film. The holes in LaceyCarbon grids are a mixture of sizes, which can lead to varying ice thicknesses within the same grid. C) A data acquisition image from the same grid. As LaceyCarbon grids have no regular holes there is no need to take an image at 'Hole' mag, instead a regular pattern of acquisition areas is distributed across an entire grid square to maintain an easily automated data collection strategy. This results in images with varying amounts of carbon support film in each image.

These grids provided a much better distribution of particles within the ice, and with the help of a Volta phase plate the particles were easily identified within each micrograph (Fig. 95c). However, the carbon film does increase the background noise in the images, and the altered signal to noise ratio makes it more difficult to align particles during the data processing stages. UBE3A is already a fairly challenging protein, the particles are small, asymmetrical, and appear to be more elongated than globular, so the extra challenge of the

signal to noise ratio introduced by the carbon film was not ideal. Data was collected from the LaceyCarbon continuous carbon grids multiple times to see if the increased protein yield could compensate for the difficulties with processing, but ultimately the uncoated R1.2/1.3 grids with fully optimised glow discharging and sample concentrations were the best option.

Other grid types also exist that may provide better a better option for other samples, such as the UltrAuFoil grids from QuantiFoil or the Au-flat grids from Mitegen that feature both gold supports and a gold support film containing the regularly spaced holes. These gold-on-gold grids are supposed to improve the possible resolution of a dataset by decreasing the amount of movement in each micrograph (Russo and Passmore, 2014). With the standard carbon on copper grids, or even carbon on gold, the carbon foil and the copper or gold gridbars will contract slightly differently when cooled and when warmed up by the beam, so more motion will be observed in each micrograph. For the UltrAuFoil and C-flats however, the gold support film and the gold grid bars will contract and expand at the same rate, maintaining the positioning of particles suspended in the ice (Russo and Passmore, 2014). These were briefly considered for use with the UBE3A sample, but ultimately the motion correction stage was not the limiting step of data processing for any of the UBE3A datasets, and the gold grids appeared to be more prone to contamination by miniscule gold particles from the manufacturing process. It was possible to clean the grids with detergent before use to ensure that these gold puncta were removed before applying the sample, but only once the problem had already been identified. These grids may have been worth pursuing further at other points of the project, but at the time the increased availability and promising optimisation of the carbon on copper R 1.2/1.3 QuantiFoil grids lead to their continued use over the gold-on-gold alternatives.

6.3 Blotting Conditions

When grids are prepared on an FEI Vitrobot the ice thickness can be controlled through two settings, the blot time and the blot force. The blot force determines the distance between the pads that hold the blotting paper. This parameter is different for each machine depending on how it has been calibrated, so a blot force of 2 will be different from one machine to another. Generally, this setting is optimised for each machine to determine the setting where the blotting pads just touch on the bottom edge, and then it is not changed very much. Varying the blot force can still be a useful determinant of ice thickness, especially for grids with additional support layers or if thicker ice would be beneficial.

The blot time is a much more controllable parameter, this is the amount of time that the blotting pads will spend in contact with the grid. A longer blot time will remove more of the sample solution and lead to thinner ice, but if it is blotted for too long the grid can become too dry. This setting is much more

subtle than the blot force, minor changes in the blot time will not be distinguishable from the variability between grids of even the same conditions. This makes it much more controllable parameter than the blot force, which can only be altered in more significant increments.

Another factor that was only occasionally considered during the blotting process was the wait time. This refers to the amount of time between applying the sample and blotting the grid. For typical grids the blot time was kept at 0s, as a longer wait time creates more opportunity for evaporation of the sample, although the increased humidity in the vitrobot chamber should negate that as much as possible (Glaeser *et al.,* 2016; Passmore and Russo, 2016). However, when applying a GrOx-DDM coating to grids, a wait time of ~10 s was advised. This is because the GrOx-DDM grids were not glow discharged prior to application of the GrOx-DDM solution, so I rely on the DDM on the grid and a careful application of the protein sample to encourage an even spread of protein across the grid. However, a wait between the sample application and blotting of the grid also allows the sample to dissipate, resulting in a more even spread of particles, but also allowing the protein to settle onto the GrOx-DDM surface rather than just being blotted off completely.

6.4 Sample Considerations

Once the ice conditions have been optimised, the final variable to consider is the sample itself. It is important to ensure any sample used for cryo-EM is as homogeneous as possible and is also not aggregated. While it can be relatively simple to distinguish between molecules of vastly different sizes within EM micrographs, when proteins are a similar size, or if different orientations of the protein appear to be vastly different sizes, it can be difficult to select only the protein of interest from a mix. Aggregation can also cause issues in the particle picking and subsequent data processing steps, as the software may identify particles within an aggregate and try to process them, but they may not be in a physiologically relevant state, and even if they are any particles that overlap will introduce extra signal that will confuse the alignment software. However, another issue with heterogeneous or aggregated samples is that it is harder to accurately determine the concentration of the protein of interest within the sample. The heterogeneity and presence of aggregated sample can be identified through a range of biophysical techniques, including AUC and SEC-MALS, as well as SDS-PAGE analysis.

Another consideration when preparing a sample for cryo-EM is the sample buffer. While most proteins require a significant salt concentration to stay in solution, high salt concentrations can cause issues with cryo-EM. Salt molecules within a buffer will scatter electrons, albeit to a lower extent than the proteins embedded in them (Drulyte *et al.,* 2018). This leads to a much worse signal-to-noise ratio, which makes aligning particles much more difficult, even if the software is still able to pick the individual particles.

Glycerol also has the same signal-to-noise ratio altering effects, so while salt concentrations should be kept below 300mM as a general rule, glycerol should not be used at all if possible (Thompson *et al.,* 2016; Drulyte *et al.,* 2018). Both salt and glycerol can also affect the flash freezing of the buffer in the formation of vitreous ice, so samples with high salt and glycerol concentrations are also more likely to lead to the observation of crystalline ice conditions (Drulyte *et al*., 2018).

One of the most important factors to consider when preparing a sample is the protein concentration. This factor is one of the easiest to predetermine based on knowledge of the molecular weight of the protein involved, although the effective concentration of the protein within the ice can be affected by various features of the vitreous ice formation so it still requires some trial and error to fully optimise so as to have good distributions of particles. A handy table developed by Vinothkumar and Henderson (Vinothkumar and Henderson, 2016) shows the relationship between protein size and the number of particles expected at a given concentration, this can be used as an initial guide to determine a range of protein concentrations to try. However, the final concentration of protein within the ice depends on a number of factors, including the propensity of the protein to stick to the blotting paper, the preference of the protein for the carbon support over the ice holes, and also the thickness of the ice. A higher sample concentration means more particles per image, which effectively means more data from a single data collection session. However, when the protein concentration is too high particles may begin to touch or overlap, and it becomes impossible to distinguish the boundaries of individual particles.

The final factor to consider when preparing cryo-EM grids is the volume of the sample applied to the grid. Typically volumes of 2-3µl are used, although larger volumes may increase the effective concentration of particles within the ice in situations where the sample does not withstand further concentration in solution. However, this is not very reliable and relies on the sample remaining within the grid rather than sticking preferentially to the blotting paper during the blotting process.

6.5 Screening Grids

Once the grids have been made in a range of conditions, they must be screened before they can be used for a full data collection. The purpose of screening grids is to determine whether the grids that have been made are suitable for data collection, and this involves investigating all of the features identified above, including the ice thickness, the protein concentration, the sample conditions, etc. Depending on how much optimisation is required, sometimes many rounds of sample optimisation and grid screening are required before any large datasets can be collected (Fig. 90).

Throughout this project, grids were screened on both the Talos Arctica and the Glacios microscopes housed at eBIC, Diamond light source. Both microscopes feature a 200 kV beam rather than the 300 kB beam used on a Titan Krios microscope for data collection. The Talos Arctica is equipped with a Volta Phase Plate (VPP) to allow visualisation of otherwise low-contrast datasets, while the Glacios microscope was a later addition and is such equipped with the latest Falcon IV detector. The Falcon IV detector allows for better visualisation of particles than previous model, and the 200 kV beam allows for a greater signal to noise ratio compared to a 300 kV beam to allow visualisation of mid-to-low contrast particles without the need for the phase plate (Peet *et al.,* 2019; Merk *et al.,* 2020; see section 7.2.1).

For each grid that was screened, images were taken firstly of the whole grid as a series of Atlas magnification images, then of a single gridsquare, then of a single hole, or a small group of holes depending on the size of the holes and orientation in the microscope, and then of the sample at a data acquisition magnification (as seen in Fig. 93, Fig. 94, and Fig. 95). This was repeated for several different areas of the grid, different areas within each gridsquare, and then multiple locations within the holes in order to get an idea of the sample and ice conditions across as much of the grid as possible. If the sample needs further optimisation, notes are made on which areas are good and which areas need optimising, and then further grids are made to this specification, before being screening again. If the sample looks nice across a large enough area of the grid to enable data collection, the grid is retrieved from the microscope at the end of the session and stored until a data collection session can be arranged. Screening grids is a key step in sample optimisation for cryo-EM, as it is the only way to visualise the effects of changing parameters during grid preparation.

7 Data Processing and Structures

7.1 UBE3A

7.1.1 Data Collection Considerations

As UBE3A was the primary target of the PhD project, grids of UBE3A samples were prepared and data was collected in a number of ways. Several different types of grids were trialled during the grid optimisation stages (see section 6.2), but standard QuantiFoil R1.2/1.3 grids with a 300 mesh copper support were selected as the best for the sample. This allowed for the most reproducible thin ice areas, relative to grids with larger hole sizes, while retaining the uniformity and high contrast that was lost with the various GrOx and carbon support grids that I tried. The sample was also optimised through careful consideration of the protein purification process to ensure that the final sample buffer was compatible with flash freezing, and that the protein could be subjected to grid preparation immediately following purification without any freezing, concentration, or storage stages. An attempt was also made to stabilise the isolated protein through crosslinking with glutaraldehyde (see section 4.5.4), but screening of the resulting grids suggested that the crosslinked sample still suffered from many of the same issues as the native sample, so further data collection of this grid was not pursued.

Most of the datasets for UBE3A were collected on QuantiFoil R1.2/1.3 grids with native UBE3A at a high degree of purity, but there were differences in the concentration of the sample, the ice thicknesses, and particularly the data collection conditions. As UBE3A is a particularly small particle at 98 kDa, with the added complication of it being an elongated, unsymmetrical shape, I adopted two strategies in an attempt to gain a high resolution model. One involved the use of the phase plate and very careful selection of ice areas during data collection, while the other involved the relatively recent application of the aberration-free image shift (AFIS) technique in order to collect an extremely large dataset. The increased contrast of the phase plate allowed for easier particle picking, although the elongated shape of UBE3A still provided its own problems, while the AFIS approach simply resulted in a much larger dataset, so micrograph and particle selection could be much more stringent during the data processing stages without losing the statistical power of the calculations.

Data was also collected on different Titan Krios microscopes with different settings. UBE3A data collected on the Titan Krios at the LISCB used the VPP and the Falcon III camera, while the later datasets collected on Krios 3 at eBIC used the Gatan K3 detector without the phase plate. The highest resolution model of full-length UBE3A at present was a result of the earlier phase plate dataset, although it is unclear whether that was due to better data collection conditions, better data processing technique, or simply a better biochemical preparation of the sample.

Direct Electon Detectors (DEDs) sit after the specimen field of a cryo-EM microscope and convert the scattered electron beam into a digital signal. Modern DEDs for cryo-EM are available from FEI, Gatan, and Direct Electron, but the FEI Falcon range and Gatan K2/K3 cameras are most common. Both detector types are examples of complementary metal-oxide superconductor (CMOD) devices and are comprised of many individual pixel units (McMullan *et al.,* 2016). Each pixel in these devices is comprised of multiple layers. The core principle of a CMOD DED relies on two layers of p-doped silicon, which is when impurities are introduced into the matrix of a semiconductor to change its electrical properties (McMullan *et al.,* 2016; Kuijper *et al.,* 2015). P-type doping refers to the inclusion of electron acceptor atoms into the matrix, which results in the transfer of electrons from the incident electron beam being pulled out of their valence state and into the valence layer of the semiconductor atoms. Alternatively, n-type doping refers to the inclusion of electron-donor atoms into the matrix. When the electron is pulled out of its original valence state it creates a positive energy state in the void left behind, which is referred to as a hole. The transition of the electron between the valence band and the conduction band is referred to as an electron-hole pair, and it forms the fundamental energy-carrier unit of semiconductors (Tipler and Mosca, 2008a). In CMOD devices there are two layers of p-doped silicon, the base layer is more heavily doped than the top layer to create a potential barrier between them. The top, lightly doped silicon layer is referred to as the active layer as radiation-generated electrons, i.e. those that have been scattered by interactions with the specimen, will not be able to pass through the potential barrier while the electrons from the unaltered incident beam will continue into the highly doped bottom silicon layer. Within the active layer, the trapped electrons will travel using a mixture of drift and diffusion to an active-particle layer above the active layer containing diodes that discharge the electron-hole pair and convert the beam to a digital signal. The diodes in the active particle layer are formed by the junctions between p- and n-type areas when P-wells, areas of p-doped material, are immersed in an otherwise n-doped layer (Tipler and Mosca, 2008a; McMullan *et al*., 2016). In commercial CMOD detectors, the active particle layer is also topped by a passivation layer, a layer of inert substance that allows the electron beam to pass through to the active layers while housing the electrical components of each pixel (McMullan *et al.,* 2016; Kuijper *et al.,* 2015) (Fig. 96).

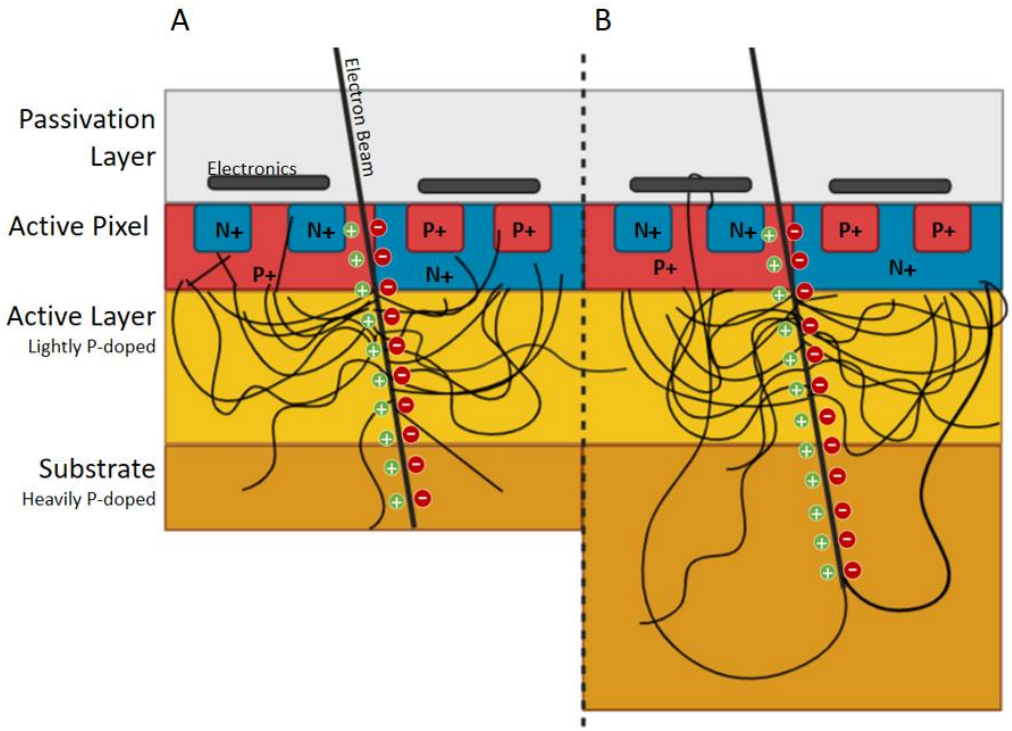


Figure 96: A diagram of the different layers in a CMOD direct electron detector. The top layer is a passivation layer that houses the electrical components. The active pixel layer is comprised of P-wells and N-wells, each comprised of P-doped or N-doped silicon and housing smaller wells of oppositely doped silicon. Below the active pixel layer is the active layer, comprised of lightly doped silicon to split the electron beam into electron-hole pairs, and below the active layer is the substrate layer of heavily doped silicon (Kuijper et al., 2015; McMullan et al., 2016). A) A back-thinned detector model with a thinner substrate layer. B) A detector that has not been back-thinned, showing the potential for electrons to re-enter the active layer in a thicker substrate layer.

While the detection of the electrons and subsequent conversion into digital data is essential for the generation of cryo-EM data, the process does introduce some extra noise into the data. The quality of a DED is indicated by its detective quantum efficiency (DQE), which is defined by the ratio of signal to noise ratios of the data before and after passing through the detector:

$$DQE = {\left(S/N\right)^2}_{out} \Big/ {\left(S/N\right)^2}_{in}$$

Where S/N is the signal to noise ratio. The DQE will always be a number below 1, with a DQE of 1 meaning that absolutely no signal has been lost or noise introduced during the detection process. The DQE can be described as a function of the spatial frequency, and will vary depending on the dose used (McMullan *et al.,* 2016; Kuijper *et al.,* 2015).

While the FEI and Gatan camera are both CMOS DEDs, there are differences in their designs and applications that allow them to have different benefits depending on the sample requirements. The Falcon III camera uses a larger pixel size compared to the Gatan K3, which means that a deeper active layer can be used, so that more of the sample-scattered electrons can be detected: this makes it well-suited for low-dose applications (Kuijper *et al.,* 2015; McMullan *et al.,* 2014). The back-thinning of the substrate layer, the highly-doped silicon layer, also allows for reduced back-transmission of unscattered electrons that may have been able to diffuse back up into the active layer while retained in a thicker substrate (McMullan *et al.,* 2014). The Falcon camera therefore has a higher DQE relative to the K3, as well as a larger image area that allows for more particles to be observed per image at the same magnification, but the Gatan K3 camera has a much faster acquisition speed, operating at 1500 fps compared to the Falcon III's 40 fps (Song *et al.,* 2019). The increased acquisition speed allows a larger dataset to be collected in the same timeframe, which can increase the resolution of the resulting model by providing more examples of rarer views, as well as just providing more signal for the more common views once the images have been averaged together during classification jobs. The higher framerate of the K3 also allows for better application of the counting mode, where each electron is detected individually rather than being integrated together to form a larger signal. The Falcon III is also able to use counting mode, but in order to identify the individual electrons with a lower shutter speed the exposure time must be increased significantly to maintain a low dose rate while increasing the overall dose (Jeong *et al.,* 2019). This makes the K3 camera much faster than the Falcon III. The increased exposure time also negatively impacts the drift in the images, as the sample will begin to move around as the ice melts, although this can be corrected for during data processing due to the splitting of each micrograph into multiple frames (Jeong *et al.,* 2019).

Ultimately, the choice between which camera to collect data with is not usually a determining factor in the resolution of the dataset. The Falcon III has a better DQE and a larger field of view so less micrographs are needed to produce a high quality dataset, while the Gatan K3 has a much faster acquisition rate at the cost of a more modest DQE and image size, meaning that the increased size of the dataset it balanced out by the need for more micrographs. The more recent Falcon 4i detector may change this, as it has a much higher framerate than the Falcon III, at 310-320 fps (ThermoFisher.com(a)) , which will allow for a much faster acquisition speed and better implementation of counting mode, while retaining the high DQE.

7.1.2 Data Processing Considerations

As UBE3A was the primary target of the project, several different data sets were collected for a UBE3A-only sample. These were collected using a variety of data collection settings, but the data was also processed using a range of

different programs and methods in order to maximise the information I could get out of each dataset.

One of the key variable steps in data processing throughout this project has been the particle picking. Relion has a reference-based autopicking process built in, which is particularly useful if a high resolution homolog structure is available, or a low resolution template of the target protein. However, although UBE3A has several homologs across different organisms, there is no structure available for any of them. To get around this, a subset of micrographs are picked and the resulting particles are subjected to a 2D classification job. From this, a few representative orientations are selected and used as 2D templates for the template-based picking program. The most precise way to pick particles from a subset of micrographs would be to manually pick them, but this can be very time consuming and may also introduce selection bias into the system.

An alternative to manual picking is the Laplacian of Gaussian (LoG) function within Relion (Zivanov *et al.,* 2018). While most autopicking programs will work by detecting certain images within the micrographs and comparing the picked particle to a defined particle image, the LoG program instead identifies areas with a steep change in contrast, thereby finding the edges of features within the image. Certain parameters are inputted by the user, such as the min and max particle dimensions, a default picking threshold, and a higher standard deviation threshold to distinguish between particles within the ice and other features with a high contrast gradient such as the edge of the hole, particles on the carbon support, or ice contamination. A LoG function is a mathematical function often used as an edge-detection technique, it involves the application of a Gaussian filter over the image, in the case the micrograph, followed by a Laplace operator to identify the steep changes in contrast around any definable objects (Jain *et al.,* 1995). The Laplace operator can be defined as:

$$\Delta \mathrm{f} = \nabla^2 = \nabla \cdot \nabla \mathrm{f}$$

Where Δf denotes the Laplace of a function, $\nabla \cdot$ the divergence, and ∇f the gradient of the function (Jain *et al.,* 1995). The Laplace operator is particularly useful for identifying objects with a somewhat gradual edge as opposed to a sharp defined edge. With a sharp edge a measure of the gradient is sufficient to identify the location, but when there is a more gradual shift in grey values, as I the case with many biological samples, the gradient operator leaves an equally broad signal. However, as the Laplace operator is a second order differential it features a zero crossing in the middle, allowing the edge to be defined by the easily identifiable zero-crossing point (Fig. 97; Jain *et al.,* 1995).

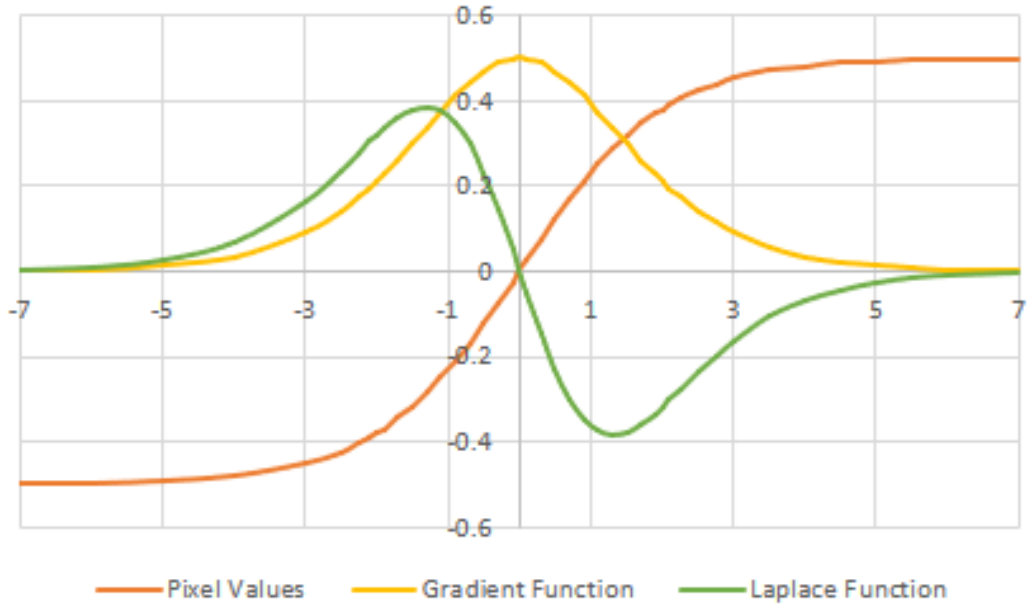


Figure 97: The Laplace operator applied to a gradual edge. The dark orange line demonstrates the gradual change in grey values at the edge of an image feature, the yellow line demonstrates the gradient operator and the green line shows the application of the Laplace function to the function of the red line. While both the gradient and Laplacian function of the line share a broad peak, the Laplace function contains a zero-crossing which can be easily identified as the mid-point of the edge.

Once a subset of particles has been picked, either by the LoG picker or manually, the 2D references are generated and used to run Relion's template-based autopicking job on the whole dataset. Relion's reference-based autopicking works using a template-based approach to particle-picking, meaning that areas in the micrograph are compared to a set of reference images and particles are picked depending on the similarity between the chosen area and the template (Scheres, 2015). Relion's autopicker tool first assumes independent Gaussian noise, but then the noise is normalised using position-dependent factors to bring the mean recorded noise in each micrograph to zero with a standard deviation of one. The position-dependent normalisation factors are particularly key due to varying contrast levels in different micrographs within a cryo-EM dataset due to the use of a range of defocus values during data collection, as well as the inherent differences in ice and sample qualities in different areas of a grid. Once a potential particle has been identified, a box is drawn around it with a user-defined length, and then a circle is drawn within the box with a user-defined radius. The area within the square but outside the circle is used to determine the background noise of the particle image, which is used to normalise the area within the circle. This results in a set of particle images with zero-mean values and uniform standard deviations, allowing accurate comparison and compiling of particles in varying ice thicknesses and taken at different defocus values (Scheres, 2015).

Once the particle image has been normalised the program calculates the probability of observing the template image in the orientation and position identified, as well as calculating the probability of identifying only noise in that location. It then divides the first value by the second to create a ratio that represents how much more likely it is that the particle image contains a true particle rather than just noise, and if the ratio is greater than 1 it determines that the potential particle in that position is a true particle. When using the Relion autopicking program the user can define a threshold for accepting particles in order to control the number of particles picked per micrograph. In order to do this, the algorithm first calculates an expected value for the previously mentioned ratio assuming to an independent Gaussian noise distribution and using the particle template provided. However, no micrographs actually show a perfect Gaussian noise distribution so the confidence in the identification of the particle in the image will be lower for the observed data than the expected value. Division of the observed value by the expected value generates a similarity metric which the user can define to select only particles that exceed a chosen confidence level (Scheres, 2015).

One limitation of template-based particle picking, beyond the requirement for a reference image, is that the reliance on the provided image can introduce bias into the process, particularly when compared to a feature-based approach (Scheres 2015). If the data is particularly good, i.e. the particles are clear, well-defined, and homogeneous, then the statistics-based approach used by the LoG function may be sufficient or even beneficial due to this. However, a benefit of the template-based approach is that it allows the detection of weaker signals (Scheres, 2015). As UBE3A is a relatively small protein by cryo-EM standards, and it is also elongated, meaning that some orientations will provide stronger signals than others, it can be difficult to optimise the parameter in the LoG picking job to select all possible views of UBE3A whilst limiting the selection of low-contrast noise areas or high-contrast ice and carbon areas. LoG was useful for speeding up the process of picking the initial subset of micrographs, but the picks were then manually curated before generating the 2D classes. The reference-based approach was then used to ensure that true UBE3A particles could be identified from ice or other contaminants.

However, the range of data collection settings used to collect data on UBE3A-only samples meant that the different datasets may not all be amenable to the same approaches. Some datasets were collected using relatively low concentrations of UBE3A and the phase plate, which results in a sparse distribution of high contrast particles, which was well suited for the LoG picking function. However, another later dataset was collected without the phase plate on a 300 kV microscope using a higher concentration sample, which resulted in low contrast particle images that were difficult to separate by eye. In this case, both the LoG and reference-based particle picking

functions within Relion were not able to pick the dataset accurately. For this dataset, I used the TOPAZ particle picking program. TOPAZ is a deep-learning based particle picking program, it uses deep learning techniques to train a model for a specific dataset so that the resulting picks are more accurate (Bepler *et al.,* 2019). It employs a convolutional neural network (CNN) to process the micrograph images by first considering the image as a grid of the grey-scale values for each pixel, similarly to how the Laplace operator is applied to images in LoG picking, CNNs apply a series of kernels to an image with the output of each layer acting as the input for the next, with each kernel identifying different features such as edge detection, corner detection etc. Starting with more simple kernels and increasing the complexity allows the program to determine a hierarchy of features, which allows more complex patterns to be identified (Bepler *et al.,* 2019). Different CNNs have different functional approaches, TOPAZ uses the 'sliding window' approach to try to improve the accuracy of detection, where an image is split up into several overlapping windows which the program scans individually to identify objects. The information from these individual windows is then combined through several rounds of CNN calculations in order to identify the areas in the whole image that contain objects, in this case particles. The sliding window technique can be fairly computationally expensive and can take a lot of time, so templates are typically kept as basic as possible in order to speed up the computation. However, one of the key differences between TOPAZ and previous sliding-window based CNN architectures is the use of positive-unlabelled (PU) learning, which involves assigning negative references from the unpicked areas to determine a set of negative reference images (Bepler *et al.,* 2019). One issue with typical PU models is a tendency to over-fit the model, so TOPAZ applies generalised expectation (GE) criteria to apply constraints based on provided information, i.e. parameters associated with positively labelled particles, when assigning areas as negative reference images. TOPAZ uses a minibatched stochastic gradient descent method, meaning that a subset of the provided data (minbatch) is used to apply random values to a variable until a local minimum value can be approximated. In this case, the stochastic gradient descent method is used to determine PU constraints for which the likelihood of mis-classification of an unlabelled area as a negative area is reduced. The specific method that TOPAZ uses to define these constraints is called the GE-binomial function (Bepler *et al.,* 2019). In order to ensure that the training results in a robust model without the need to fully pick each micrograph, TOPAZ employs a hybrid classifier and encoder strategy. An autoencoder is another type of artificial neural network that uses unsupervised learning, meaning that it attempts to determine aspects (coding) of the data from an unlabelled dataset, and it validates its decisions by attempting to recreate the input data from the encoded data it has generated (Theodoridis, 2020). A classifier, on the other hand, is a form of supervised learning, where labelled data is provided and the program

attempts to relate the parameters relevant to each identified class, allowing classification of unlabelled datasets once the model has been trained on the provided data. The TOPAZ software combines the two approaches to create a hybrid approach that allows a more accurate model to be generated from a relatively small reference dataset without overfitting of data (Bepler *et al.,* 2019). All of these features result in a picking model that consistently picks more particles, with a lower false-positive pick rate than other neural-network derived picking programs. The increased accuracy of picking may reduce the need for iterative rounds of 2D and 3D classification of extracted particles during downstream processing, which can be somewhat of a bottleneck in EM data processing (Bepler *et al.,* 2019).

The TOPAZ software package also include a topaz-denoise program, another CNN-based program that aims to specifically reduce noise levels of collected micrographs while retaining the high resolution features of the particles (Bepler *et al.,* 2020). Cryo-EM images typically have very low signal-to-noise ratios (SNRs), which makes particle picking and even downstream processing difficult. Some ways to increase the SNR of a dataset during data collection are to increase the defocus values, resulting in higher contrast images but causing reduction in the higher resolution information that can be found close to focus, and increasing the total dose during image acquisition, which also increases the exposure time and limits the number of images it is possible to acquire in a collection session. High- and low-pass filtering methods can also be used to attempt to reduce the SNR of micrographs during the processing stages, but this does not take into account the specific nature of cryo-EM image noise. Topaz-denoise uses a CNN approach to compare the noise between different frames of the same micrograph file in order to identify real features from structured noise artefacts without the need for a 'ground truth' reference image. The use of topaz-denoise on micrographs during processing has been shown to enable the use of shorter exposure times during collection without a reduction in the resolution of the resulting reconstruction, as well as making it easier to pick difficult, lower signal particle orientations during data processing, allowing a more complete reconstruction of the target (Bepler *et al.,* 2020).

7.1.3 UBE3A Structure

The highest resolution model of UBE3A was estimated to be 9Å, improving slightly to 8.4 Å after PostProcessing in Relion, but none of the features of a relatively high resolution structure can be seen in the current model. At roughly 8Å you should be able to begin to make out alpha helices and beta-sheet conformations, but as none one these can be seen in the UBE3A model actual resolution must be lower. However, although the model lacks sufficient detail to be able to fit the sequence, it does provide a shape to fit the previously solved HECT domain, or computationally predicted structures of full-length UBE3A into (Fig. 98). This both validates my work in this area, as

although I was not able to generate a high-resolution model I can conclude that UBE3A was present in my samples, and it validates the predicted models for UBE3A that show an overall similar shape to my experimentally-derived data.

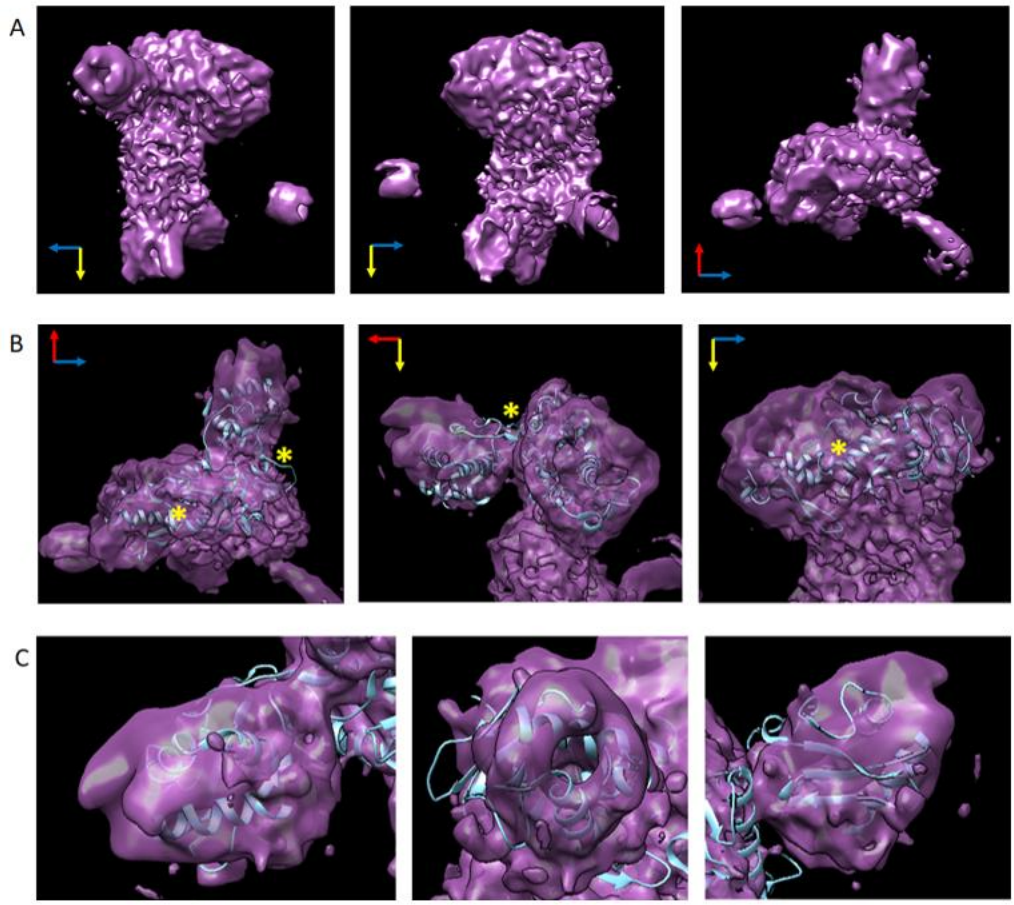


Figure 98: The low-resolution model of full-length UBE3A. A) The full-length UBE3A structure was solved to a reported resolution of 8.43Å. Different views of the model are shown, with the relative orientations indicated by the coloured arrows in the corner of each image. B) The crystal structure for the UBE3A HECT domain (1C4Z) was fit into the low-resolution model using the automated fit to map tool in chimera. The initial fit shows that the model has a realistic size and overall shape for this domain of the protein, but a different organisation of the lobes within the HECT domain could improve the fit. The locations of the hinge regions within the UBE3A HECT domain are shown by a yellow asterisk. The relative orientations of each image are shown by the coloured arrows in the corner of each image. C) The C-lobe of the HECT domain was fit into the corresponding density by hand, disregarding the fit of the rest of the HECT domain into the model.

When the HECT domain is fit into the low resolution cryo-EM model of UBE3A, it is apparent that the shape of the map corresponds closely to the shape of the HECT domain. However, HECT domains typically display some flexibility, between the three defined lobes of the structure (see section 1.5). The hinge

regions between these three domains are shown in Fig. 98b by the presence of yellow asterisks. The 1C4Z crystal structure of UBE3A HECT domain is a rigid structure, and it was formed of isolated HECT domains rather than full-length UBE3A so it may not be completely physiologically relevant. If the lobes are allowed to flex around these hinge points, the structure may fit the model even more tightly. In an attempt to demonstrate this, the C-lobe was manually arranged into the corresponding density of the low resolution model without regard for the N-terminal lobes (Fig. 98c). In this position, the C-lobe fits nicely with the low resolution model, with the secondary structure elements surrounding the defined void through the centre of the region.

The HECT domain of UBE3A is easily recognisable from the low resolution model, but unfortunately none of the rest of the protein is identifiable at this resolution. The only other fully characterised domain of UBE3A is a zinc-finger region at the N-terminus of the protein sequence (2KR1; Lemak *et al.,* 2011), and this domain is too small to be easily identifiable in a structure at this resolution. However, the rapid advancement of AI-based protein structure predictions has allowed a new way to determine different areas of the full-length model. Protein structure predictions are available for UBE3A from both AlphaFold (Jumper *et al.,* 2021) and the RoseTTA program through the Robetta server (Baek *et al.,* 2021). The AlphaFold library contains a single model for UBE3A human isoform 2, although there are no models for the other isoforms, including isoform 1 that I have been working with experimentally. However, UBE3A isoforms differ only in their extreme N-terminus, where isoform 2 contains an extra 23 amino acids. As all of the key domains will be identical between the two species, the isoform 32 model is also a good representation of the potential isoform 1 structure. The AlphaFold server is able to display the model coloured by the confidence level of each residue, with highly conserved residues shown in blue and very low confidence areas shown in orange (Fig. 99)

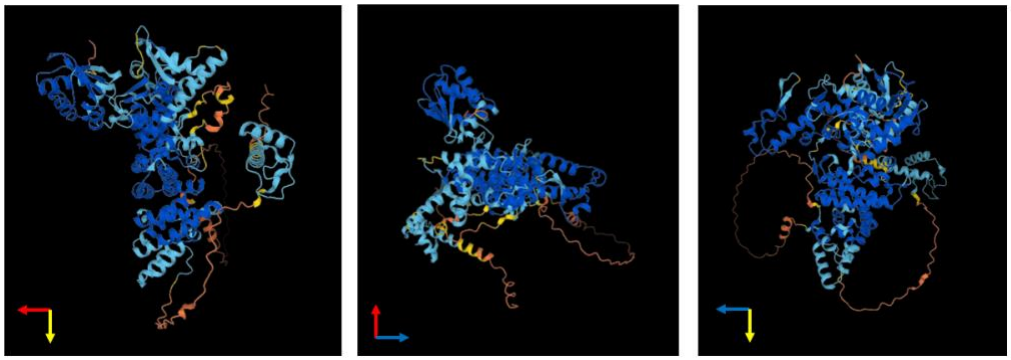


Figure 99: The AlphaFold model of UBE3A human isoform 2. Residues are coloured by their relative confidence values, residues with a confidence score of >90 are coloured blue, residues with a score >70 are coloured light blue, residues with a confidence score >50 are coloured yellow, and residues with a confidence score <50 are coloured in orange. The orientation of each view is demonstrated by the coloured arrows in the corner of each image.

The AlphaFold model of UBE3A appears to be very similar to my low-resolution map (Fig. 98a), although the handedness does not match. This is not very surprising, as cryo-EM data is collected by passing electrons through the entire sample to generate a 2D image, the handedness of the initial model is determined at random. The correct handedness is typically attributed to the model when it has reached a high enough resolution to visualise the turn of the alpha-helices within the structure. However, as my model is still too low resolution to discern individual alpha-helices, the handedness was not set. The other issue with the AlphaFold model is the presence of very long loop regions. These unstructured loop regions are inserted into AlphaFold models when the confidence values are too low to predict any structure. They do not necessarily mean that the corresponding region of the structure will be disordered, but they can indicative of a flexible region. In order to compare my low-resolution model to the AlphaFold structure, the handedness of my model was initially flipped and then the two models were superimposed (Fig. 100).

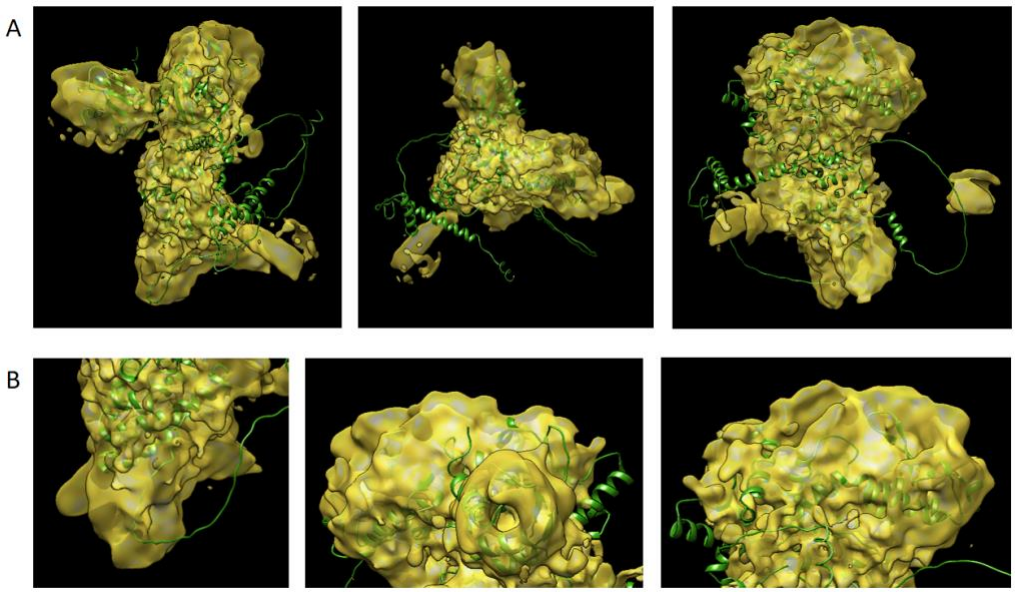


Figure 100: A comparison of the AlphaFold model with the low resolution cryo-EM model of UBE3A. A) A full view of both models to show the accuracy of the fit across the whole structure. B) Close up views of the areas within the low resolution model that do not have corresponding areas in the AlphaFold model.

The overall shape of the AlphaFold model fits the low resolution cryo-EM model very closely. The C-lobe of the HECT domain has a particularly satisfying fit (Fig. 100). Furthermore, the main body of the protein beyond the HECT domain also fits nicely into the body of the cryo-EM model. There are some areas where the AlphaFold model does not match up with the cryo-EM model, however, particularly at the bottom of the N-terminal region, and the top of the HECT domain N-lobe region. The part of the protein sequence that forms the large loop regions of the AlphaFold model is the N-terminus of the protein, so it is possible that in a true physiological state the N-terminal region is more structured than predicted, and it may fill the empty space at the bottom of the cryo-EM model. An explanation for the empty space within the HECT domain region could be similar to that used for the fit of the HECT domain crystal structure into my model (Fig. 98b), which is that the HECT domain is a flexible domain consisting of three lobes and two hinge regions between them. Although the AlphaFold model was made taking into account the full-length protein sequence rather than just the isolated HECT domain, it is still only a prediction based on minimal empiric evidence. The flexibility of the HECT domain likely makes it difficult to determine a single representative conformation, as unless an unknown factor is present to force it into a single state, different conformations around the hinge regions likely represent similar energy states. The flexibility of the HECT domain also makes it difficult to determine the structure empirically through the use of cryo-EM, particularly as UBE3A is a small protein and the differences in HECT domain conformations are likely to be very small, as it is very difficult to separate the different states during the processing stages. This means that it is currently

impossible to determine which structure, the low-resolution cryo-EM one or the AlphaFold model, is more accurate without a higher resolution empirical model.

Whereas AlphaFold only provides a single model for UBE3A, and it is only available for isoform 2, the Robetta prediction program generates multiple possible models for any given sequence. The human UBE3A isoform 1 sequence was subjected to modelling using the RoseTTAFold method through the Robetta portal and it generated five separate models (Fig. 101)

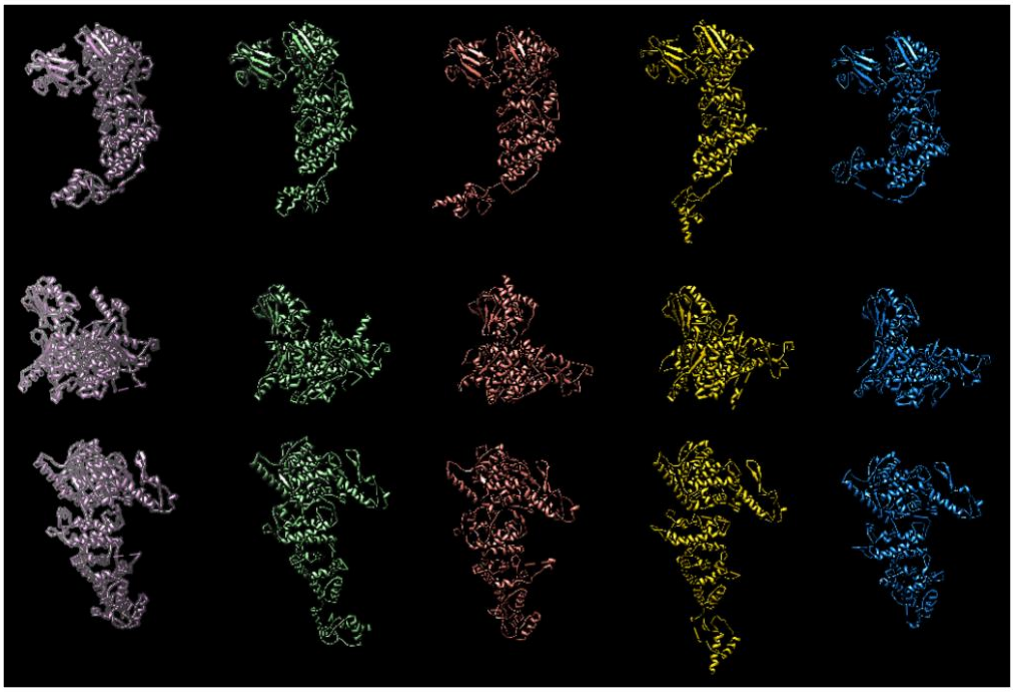


Figure 101: The Robetta deep learning predicted models for UBE3A isoform 1. Five models were generated. Model 1 is shown in pink, model 2 in green, model 3 in red, model 4 in yellow, and model 5 in blue. All models feature only the residues with an error estimate of less than 5 Å.

The models generated by Robetta were very similar is most areas, but the main difference was within the N-terminus of the structure, furthest from the HECT domain. This seemingly unconnected domain of UBE3A is also observed in the AlphaFold model (Fig. 99), where rather than appearing near the bottom of the structure it sits halfway up, connected by a longer linker than the Robetta models. The Robetta models also agree with the AlphaFold model in terms of the handedness of the structure. As all five structures were reasonably similar, only model 1 was fitted into the low resolution cryo-EM model in order to simplify the observations (Fig. 102).

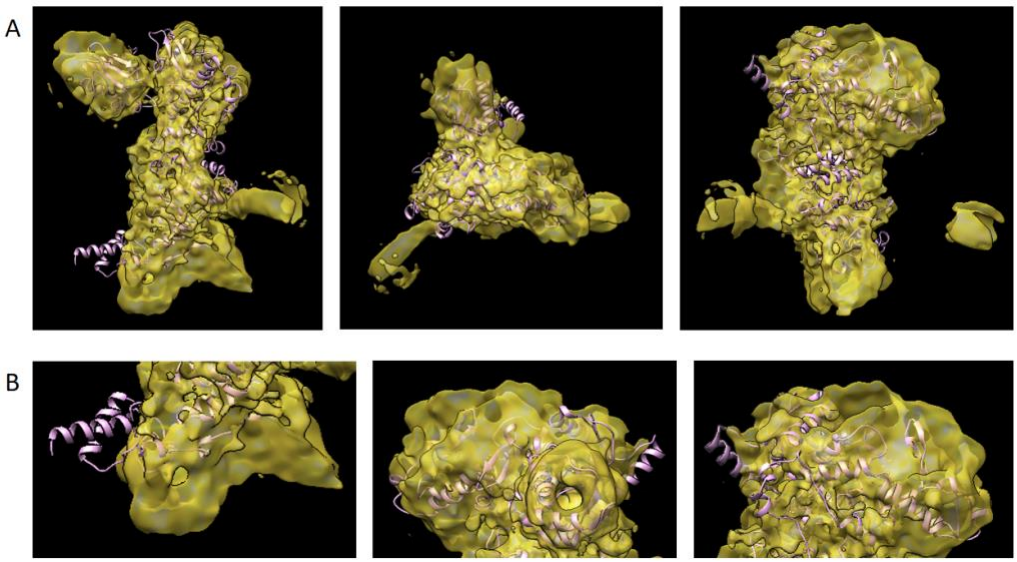


Figure 102: A comparison of the Robetta model with the low resolution cryo-EM model of UBE3A. A) A full view of both models to show the accuracy of the fit across the whole structure. B) Close up views of the areas within the low resolution model that do not have corresponding areas in the Rosetta model.

The Robetta model fits similarly well into the low-resolution model as the AlphaFold model. The more variable region at the bottom of the structure does appear to sit in a more sensible position in the Robetta model compared to the AlphaFold model, and it is easy to imagine it fitting into the density with a slight rotation. The lack of the long loop regions are also not present in the Robetta model. The Robetta model was downloaded so that residues with a confidence value of less than 5Å were not included, so even if the loop regions were modelled they wouldn't be observed here, but the initial model with all residues present still didn't contain any significant loop regions (Fig. 103b).

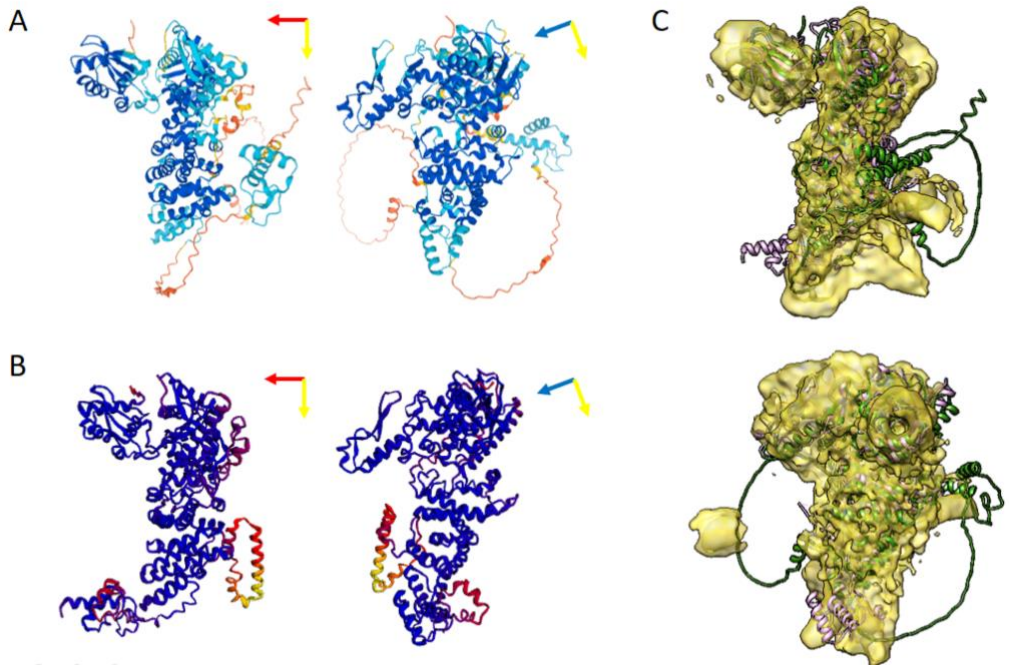


Figure 103: A comparison of the three models of UBE3A. A) Two distinct views of the AlphaFold UBE3A isoform 2 structure, coloured by the level of confidence of each residue. Darker blue areas have the highest confidence level and low confidence areas are shown in orange. B) Two views of the full Robetta UBE3A isoform 1 prediction. Residues are coloured by the error estimate value, areas with a higher error estimate are shown in red, with high confidence areas shown in blue. C) Two views of both the AlphaFold and Robetta models fitted into the low resolution cryo-EM model. The colours are uniform for each model, the AlphaFold model is shown in green, the Robetta model is shown in pink, and the cryo-EM model is shown in yellow.

The areas in which both predicted models are most confident are very similar between the two, and these areas fit into the low resolution cryo-EM model well. On the other hand, the areas with lower confidence values or higher error estimates are not so well conserved between the models, and they do not typically fit into the volume of the cryo-EM model. This is particularly noticeable for the loop regions, the weakly connected domain areas, and the tip of the helix in the large N-terminal subdomain of the HECT domain. The EM model also seems to extend further out from the predicted models in the upper and lower extremes of the structure. It is likely that the more variable areas of the predicted models would be arranged differently in order to fill much of this space, but the resolution of these areas also seems to be lower than that of the central region of UBE3A. UBE3A is a reasonably flexible protein, as suggested by the model thermal melt point calculated in section 4.5.3 and the IUPred2 prediction in appendix 3, and the cryo-EM model suggests that the extremities of the protein are more flexible than the central region. This could mean that both the well-characterised HECT domain and the less characterised N-terminal region of UBE3A function as separate

domains that are connected by the more structured alpha-helical segment in the centre of the protein.

Although the HECT domain is by far the best characterised region of UBE3A, some other areas of the protein have been identified as functional domains or regions involved in protein-protein interactions. Although the location in the protein sequence has been identified for each of these regions, the location in the full-length structure is still completely unknown. The regions involved in the association with the HPV E6 in protein are particularly intriguing, as they are comprised of four or five (depending on the strain of HPV) distinct protein regions that are scattered across the protein sequence (Fig. 19; Drews *et al.,* 2020). The AZUL domain is another reasonably well-characterised domain within UBE3A (Kühnle *et al.,* 2018; Buel *et al.,* 2020), particularly in its interaction with PSMD4 (Buel *et al.,* 2020). This domain sits at the extreme N-terminus of UBE3A isoform 1, with only the extra amino acids of isoforms 2 and 3 upstream of it. As Zinc finger domains have been identified as protein interaction domains in other proteins (Gamsjaeger *et al*., 2007), and also in UBE3A with PSMD4 (Buel *et al.,* 2020), it has been proposed that the AZUL domain is responsible for recruiting substrates for the HECT domain (Buel *et al.,* 2020). However, its location at the extreme N-terminus of the protein sequence, while the HECT domain sits at the extreme C-terminus of the protein sequences, raises questions over the mechanism of transferring a substrate between the two domains. The low resolution cryo-EM structure of UBE3A may not be high enough resolution to conclusively determine the location of each domain within the structure, but it does give us a good idea of the proximity of each region. In order to observe the proximity of each defined region or functional domain of UBE3A within the full-length protein, the Robetta and AlphaFold models were fitted into the cryo-EM volume and the relative regions within the sequence were individually coloured (Fig. 104).

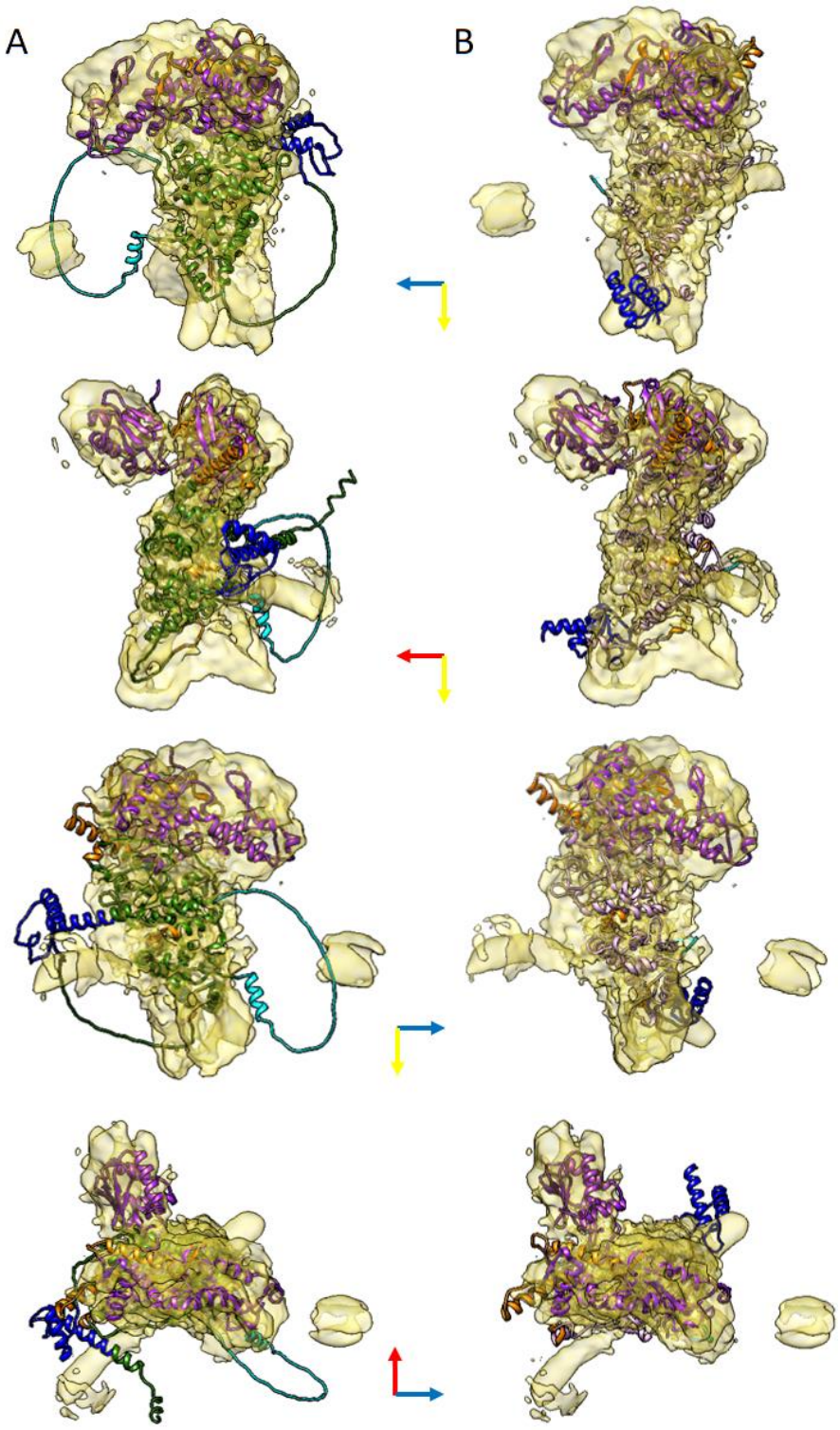


Figure 104: The low resolution cryo-EM model of full-length UBE3A with the predicted models fitted in and coloured by domain regions. A) The AlphaFold model of UBE3A isoform 2 fitted into the cryo-EM volume. B) The Robetta prediction fitted into the cryo-EM volume. The HECT domain is demonstrated by the purple regions, the AZUL domain is coloured in blue, the E6-interacting regions are coloured orange, and the region involved in the interaction with HERC2 is coloured in cyan. The orientation of each image is demonstrated by the coloured arrows down the centre of the figure.

When the domains are shown in the context of the full-length structure, some observations can be made. The LxxLL motif that is key to the E6 protein interaction sits very close to the HECT domain, despite its location roughly halfway through the protein sequence. The E6-interacting regions that sat close to the HECT domain in sequence are predictably close to the HECT domain also, but two of the most distal regions associated with E6 binding are not very close to the HECT domain at all. Both are only small protein regions, one spans residues 98-105 (isoform 1 numbering) and the other residues 287-297 (Drews *et al.,* 2020), but the corresponding small patches of orange in the predicted structures can be seen as predominantly loop regions, one in the middle of the previously undefined portion of the N-terminus, and the other at the extreme of this region furthest away from the HECT domain (Fig. 104). The E6-interacting region furthest from the HECT domain, both in the sequence and the full-length structure, is associated with an increasing binding efficiency of E6 but does not have any impact on the alteration of UBE3A activity after binding (Drews *et al.,* 2020). Its distal location could possibly be explained by a general destabilising effect of mutations in that area. However, the E6 interaction region hallway down the back face of the enzyme has been associated with changes in UBE3A activity following binding (Drews *et al.,* 2020), which is not easily explained given the distance from the catalytic site. Of the E6-interacting regions close to the HECT domain, the region closest to the catalytic site is absolutely required for binding of low-risk HPV E6 proteins, but is not measurably involved in the interaction between UBE3A and high-risk HPV E6 proteins (Drews *et al.,* 2020). Perhaps the close proximity to the catalytic site of E6 proteins in this position is a factor in the different outcomes of the high and low risk HPV infections, as a ~19 kDa globular protein docked around the catalytic site in order to reach both this domain and the LXXLL motif would physically preclude binding of many endogenous target proteins. The clinical outcomes of low-risk HPV infections do not suggest that binding of low-risk HPV E6 proteins to UBE3A prevents UBE3A activity, but in reducing the area available for substrate proteins to dock it likely alters the substrate profile of the enzyme. The other HECT-adjacent E6-interacting domains, particularly the crucial LXXLL motif, are situated around the outside edge of the large N-terminal subdomain of the domain. This probably means that E6 proteins docked here do not prevent binding of any potential substrates, but it would allow substrates of sufficient length to orientate themselves in proximity to the catalytic cysteine residue of UBE3A.

Another key domain of UBE3A is the AZUL domain at the N-terminus. The AZUL region (Fig. 104, blue) is not well conserved between the predicted UBE3A structures, with AlphaFold showing it situated halfway up the back of the protein at the end of a long linker, while the Robetta model shows it at the bottom. Neither of these models fits into the low-resolution cryo-EM

model, which suggests that neither is the correct placement. The AZUL region placement is particularly divergent across all of the initial Robetta models as well (Fig. 101), which could suggest that it is attached to the rest of the protein by a flexible linker region that makes it more flexible than the rest of the protein structure. This would explain the lower resolution at the bottom region of the cryo-EM map also, if it has a less defined position in the physiological protein.

As well as the well-defines HECT and AZUL domains, and the well-characterised E6-interacting regions, another area of UBE3A has been implicated in the interaction between UBE3A and the RLD2 domain of HERC2. The interaction region has been loosely mapped to within residues 150-200 of isoform 1 (Kühnle *et al.,* 2011), not too far from the AZUL domain and very distal to the HECT domain in terms of the protein sequence. An association between RLD2 and full-length UBE3A has been shown to increase the ubiquitination activity of UBE3A *in vitro* (Kühnle *et al.,* 2011), so the interaction must affect the HECT domain of UBE3A in some way. The sequence relating to the RLD2 interacting region of UBE3A was coloured cyan in both predicted models (Fig. 104), but unfortunately the results were inconclusive. The cyan region of the AlphaFold model spans a single alpha helix followed by a long, unphysiological loop region. The Robetta model, however, shows only a few residues of this region as a disordered loop, with the rest of the residues having been removed from the structure due to an estimated error rate of >5Å. The full Robetta model suggests that this region may form several alpha helices that protrude from the main body of the protein roughly halfway down one side (Fig. 105b), but this would also occupy a space that is not represented in the low resolution cryo-EM model. The low confidence in the predictions of this area may suggest that the area remains in a relatively flexible, disordered state until association with a binding partner induces a more stable conformation. However, CD analysis of the interaction between full-length UBE3A and the isolated RLD2 domain suggests that although an interaction definitely occurs, it does not result in any significant rearrangement of the structure of either protein (section 4.4.1). When the full Robetta model is fitted into the low resolution cryo-EM model the RLD2 interaction region still does not fit into the model, but it does sit in rough proximity to an otherwise empty are of the cryo-EM volume (Fig. 105).

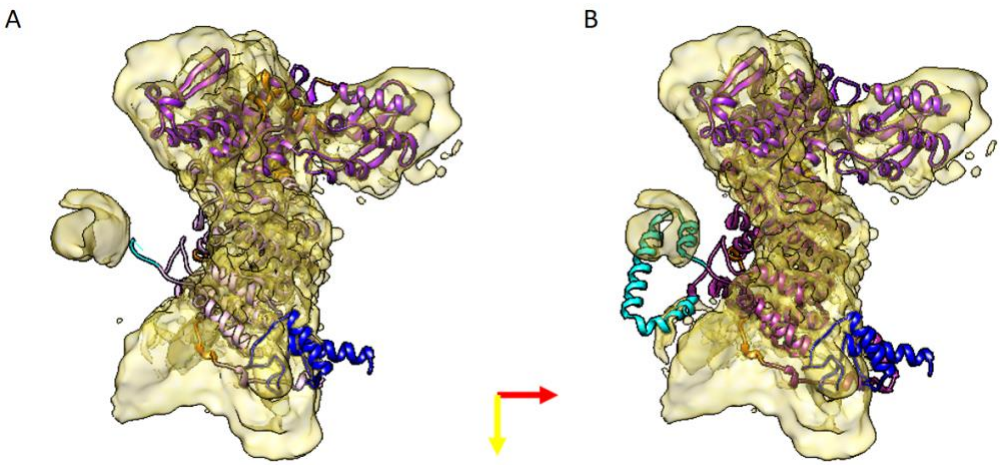


Figure 105: A comparison of the two predicted structures of UBE3A. A) The Robetta model of UBE3A truncated to remove any residues with an error estimate of >5Å. B) The full model of UBE3A predicted by Robetta. Both models are coloured as in figure 104 to show the locations of different domains across the structure.

Both the AZUL region and the HERC2 binding regions of the full UBE3A Robetta model sit outside the empirical cryo-EM volume, but both are also within proximity to empty regions of the map so that it can be easily imagined how both could fit following rearrangements around the linker regions. The AZUL domain is a well-defined domain, having been solved crystallographically twice (Lemak *et al*., 2011; Buel *et al.,* 2020), and the general shape of the domain has a consensus across all of the predicted models. The variability of the AZUL domain concerns its position relative to the rest of the protein. However, the HERC2 interacting region is much less well characterised, and a structure for this region has not yet been solved. While the AlphaFold model for UBE3A has not been able to determine any structure for this domain beyond a single alpha helix, the Robetta prediction for this domain has very high error estimates, up to 15Å at its peak (see appendix 5) so can also not be trusted. However, the Robetta model in the context of the low resolution cryo-EM model (Fig. 105b) does suggest that the domain could fit into the otherwise empty space in the electron potential map, although a higher resolution structure is required to determine the precise arrangement of the structural elements.

As well as the key functional domains of UBE3A, several point mutations have been identified for UBE3A (Sadikovic *et al.,* 2014). Several of these have been associated with Angelman Syndrome pathologies (Sadikovic *et al.,* 2014), some are associated with other UBE3A-related disorders (Yi *et al.,* 2015), while others have been identified as non-pathogenic (Sadikovic *et al.,* 2014). Most cases of AS are caused by a complete loss of UBE3A, either through a deletion of the chromosome region or an imprinting defect (Jiang *et al.,* 1999). Where AS is caused by a mutation in the *UBE3A* gene, many of these

mutations are frameshift or nonsense mutations resulting in a truncated and inactive form of the protein (Sadikovic *et al.,* 2014). However, several AS cases have also been identified where a missense mutation of a single amino acid is sufficient to inhibit UBE3A activity to a similar degree (Sadikovic *et al.,* 2014). Consideration of the location of these mutations and the associated clinical outcomes could provide key information on the mechanisms of UBE3A activity. A selection of missense mutations in UBE3A identified from a cohort of AS patients and family members (Sadikovic *et al.,* 2014) was displayed on the full-length Robetta model of UBE3A isoform 1 in at attempt to visualise how the mutations could affect the catalytic activity of UBE3A (Fig. 106).

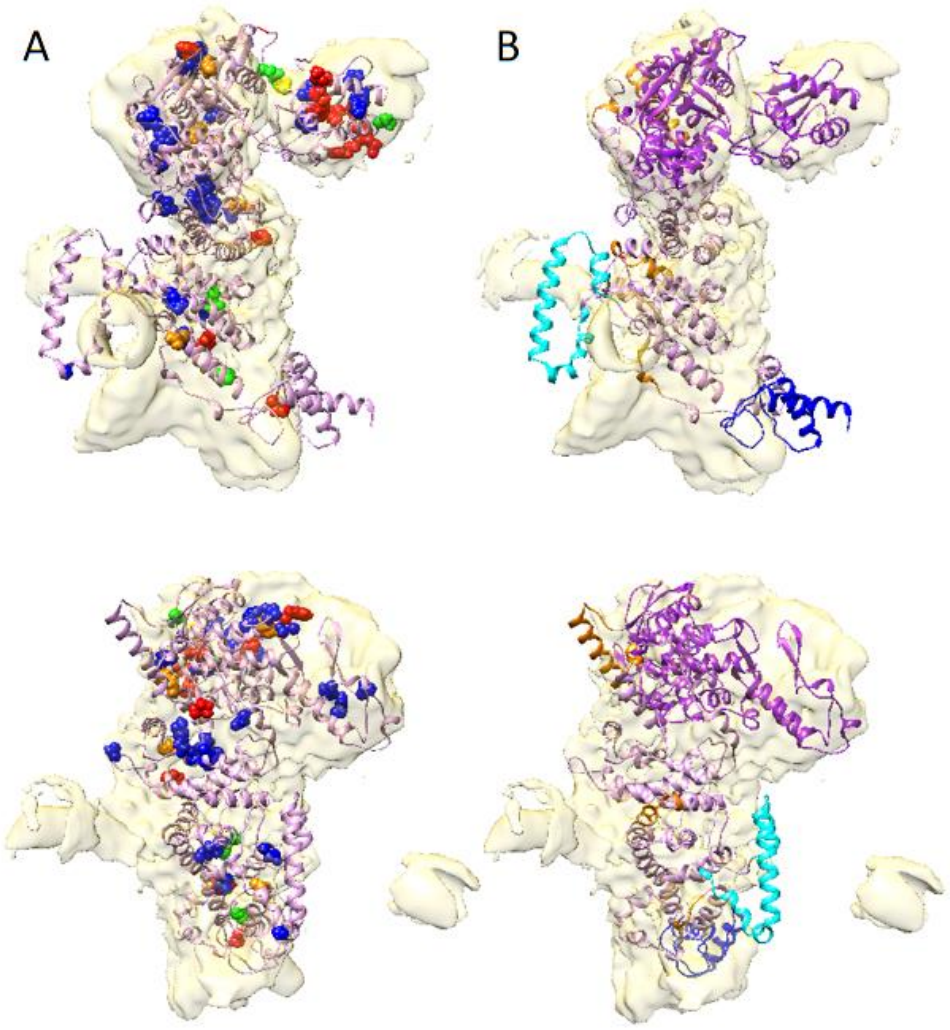


Figure 106: A comparison of the AS-related missense mutations and the domain boundaries within the Robetta prediction and low-resolution EM model of UBE3A. A) The amino acids associated with missense mutations in UBE3A identified from a cohort of AS patients and family members are shown in sphere form. Known pathogenic mutations are shown in red, mutations that are unconfirmed but likely to be pathogenic are shown in orange, de novo and other mutations of unknown consequence are shown in blue, and benign mutations are shown in green. The active site cysteine residue, itself the location of a known pathogenic mutation, is shown in yellow. B) The Robetta prediction of UBE3A isoform 1, coloured to demonstrate the identified domain regions, fitted into the low resolution EM-model of UBE3A. The domains are coloured as previously, the HECT domain is shown in purple, the E6 binding regions are shown in orange, the AZUL domain is shown in blue, and the HERC2-interacting region is shown in cyan.

Many of the AS-associated mutations cluster around the HECT domain, particularly within the C-lobe of the HECT domain. This area is close to active site cysteine residue, and many residues within the C-lobe are involved in stabilising the interdomain packing of the C- and N-lobes of the domain (Huang *et al.,* 1999). It is not surprising that mutations in this area could impact UBE3A's catalytic activity, as the HECT domain is responsible for catalysing the transfer onto ubiquitin onto substrates, which is UBE3A's main cellular role. Pathogenic mutations were also identified within the AZUL domain region, as described in work by Kühnle *et al.* (2018). The AZUL domain has been shown to be involved in UBE3A's interaction with the proteasomal substrate protein PSMD4, but as PSMD4 can act as both a substrate of UBE3A (Lee *et al.,* 2013) and a binding partner to shuttle UBE3A to the proteasome (Avagliano Trezza *et al.,* 2019; Buel *et al.*, 2020), it is unclear whether a mutation in this region would affect other UBE3A protein-protein interactions, including substrate binding, or just the association with the proteasome. UBE3A associates with the proteasome to target proteasomal subunits for degradation, but it has also been suggested that UBE3A may associate with the proteasome through the shuttle protein PSMD4 in order to ubiquitinate already ubiquitinated substrates (Buel *et al.,* 2020). This could act as a quality control check for the proteasome, to ensure that proteins designated for degradation are not missed by the proteasome during the transfer from the shuttle protein. Although the missense mutation in the AZUL domain demonstrated above has not yet been designated as pathogenic or benign, other studies have shown that mutations in this region can cause the AS phenotype (Kühnle *et al*., 2018).

Another residue was found to be mutated in the AS cohort within the HERC2-interacting region, although it is not yet determined whether this mutation is a cause of AS or not. The interaction between UBE3A and HERC2 has been characterised physically (Kühnle *et al.,* 2011), but the physiological implications are not well known. It would be interesting to see if a mutation that disrupts the interaction between UBE3A and HERC2 while leaving the catalytic area of UBE3A intact could affect the downstream effects associated with UBE3A to the extent that it would cause the AS phenotype. A neurodevelopmental disorder associated with loss of HERC2 rather than UBE3A shows a strikingly similar phenotype to AS (Harlalka *et al.,* 2013), so it is possible that the activities of the two ubiquitin ligases are interlinked to some extent.

Interestingly, a large number of the mutated residues appear to cluster within the previously uncharacterised region of UBE3A, outside of the currently identified domain boundaries. One possibility is that this core region of UBE3A is responsible for organising the interplay between the different domains, and a mutation here would destabilise the structure to abrogate UBE3A activity, or even lead to protein aggregation. However, it is also

possible that this region contains a previously uncharacterised substrate-binding domain, responsible for the substrate specificity of UBE3A. There is a region within this area of UBE3A, spanning residues 287 to 297 of isoform 1, that is involved in affecting UBE3A's ubiquitin ligase activity following HPV E6 binding, but is not necessary for the actual binding of E6 (Drews *et al.,* 2020). There are no AS-associated mutations identified within this 10 residue region, but it is possible that this helix forms part of a larger domain that coordinates substrate binding with the activity of the catalytic domain.

7.2 UBE3A+PSMD4

7.2.1 Data Collection Considerations

A sample of UBE3A+PSMD4 was purified as described in section 4.2.1, and it was subjected to cryo-EM grid preparation as described in section 2.10.2. UltraThin carbon support lacey carbon grids were used for this sample. These feature a 400 mesh copper support, a lacey-carbon film layer, and a 3nm thin layer of continuous carbon over the top of the lacey-carbon. This extra film helps to encourage the protein particles that would otherwise cluster on the carbon foil, to sit within the ice holes. Theoretically, the extra background signal introduced by the extra carbon layer can be mitigated by the use of the phase plate, but realistically, it does still impact the signal to noise ratio of the resulting micrographs (Drulyte *et al.,* 20018). The Lacey-carbon form of support film differs from the regularly spaced holes of standard quantifoil grids in that holes for the particles to sit in are not uniform. This means that the ice thickness may vary across the grid. This can make it difficult to generate a single optimal ice condition across the grid, but it does mean that a variety of ice thicknesses can be collected on within the same grid. This may be useful when the sample is prone to preferential orientation, which the UBE3A alone sample was, as it allows images with maximum contrast in thinner ice areas as well as areas with more orientations in thicker ice areas.

Although most datasets are collected on a 300 kV titan krios microscope, the 200 kV microscopes can be useful for data collection of particularly small particles. Cryo-EM works by exploiting the scattering of electrons following the irradiation of biological samples with a focused electron beam. When the focused electron beam comes into contact with the vitrified sample, the individual electrons are scattered in different ways. Electrons that interact with the outer electron shell of atoms in the specimen undergo inelastic scattering, where energy from the electron beam is absorbed by the specimen damaging it. However, electrons that come into close contact with the nucleus of specimen atoms undergo elastic scattering, where the energy of the beam is not affected but the direction and phase are (Franken *et al.,* 2020). The elastically scattered electron waves are collected through a series of electromagnetic lenses in order to focus them onto a detector, resulting in a magnified 2D image of the specimen (Franken *et al.,* 2020). However, biological samples are weak phase objects, which means that although they

do induce elastic scattering of the incident beam, the effect of the elastic scattering is not very large. In order to separate the signals caused by interference with the vitreous ice layer and the embedded biological samples, the objective aperture is taken out of focus (Cheng *et al.,* 2015). The extent of the elastic scattering depends on the thickness of the specimen, so larger, thicker proteins will produce a larger scattering effect than smaller proteins. This is why smaller protein particles have a decreased contrast compared to larger proteins. In order to increase the contrast of smaller particles, without losing the high-resolution information caused by increasing the defocus value (Cheng *et al.,* 2015), the data could be collected using either a phase plate or a lower voltage electron beam (Peet *et al.,* 2019; Danev and Baumeister, 2016).

Phase plates are a way of introducing a phase shift in the elastically scattered electrons relative to the unscattered electrons of the incidence beam. The original phase plate design, the Zernicke phase plate (ZPP), featured a thin amorphous carbon film with a hole in the centre. The hole in the centre allowed the unscattered beam to pass through unaffected, while the scattered electrons would interact with the charged film to introduce a phase shift in the resulting electron wave (Danev and Nagayama, 2001). However, the edge of the foil hole produced fringing artefacts within the images that limited the potential benefits, and they had a very short lifespan (Danev *et al.,* 2014). The most widely used phase plate in use currently is the Volta phase plate (VPP), which features a thin film of amorphous carbon without the hole in the centre. The phase contrast effect of the VPP is caused by a beam-induced generation of a Volta potential above the thin carbon film. When the film is irradiated with the unscattered electrons from the incident beam, it reacts with residual gases within the vacuum of the microscope chamber to generate a difference between the inner potential and surface potential of the film. This difference in potentials is referred to as the Volta potential (Tipler and Mosca, 2008b). When the unscattered beam comes into contact with the Volta potential above the carbon film it undergoes a negative phase shift close to the ideal value of $\pi/2$. The effect of the beam-induced Volta potential is highly localised, meaning that it only affects the electrons that pass through the beam-illuminate area. The phase plate is situated in the back focal plane of the microscope, the point where the unscattered incident beam is focused into a tight point, so that the Volta potential-induced beam shift is applied only to the unscattered electrons focused in that area and the elastically scattered electron waves remain unaffected. This results in a phase shift between the elastically scattered and unscattered signals without affecting the phase of the elastically scattered electrons themselves. A phase shift of $\pi/2$ results in a transformation of the CTF from a sinusoidal wave to a cosine waveform, which means that rather than starting with a zero point, the curve starts at its highest amplitude. The increased phase contrast introduced

by the VPP allows data acquisition much closer to focus, which reduces the effect of the CTF on dampening the high-frequency information (Danev *et al.,* 2014; Wang and Fan, 2019).

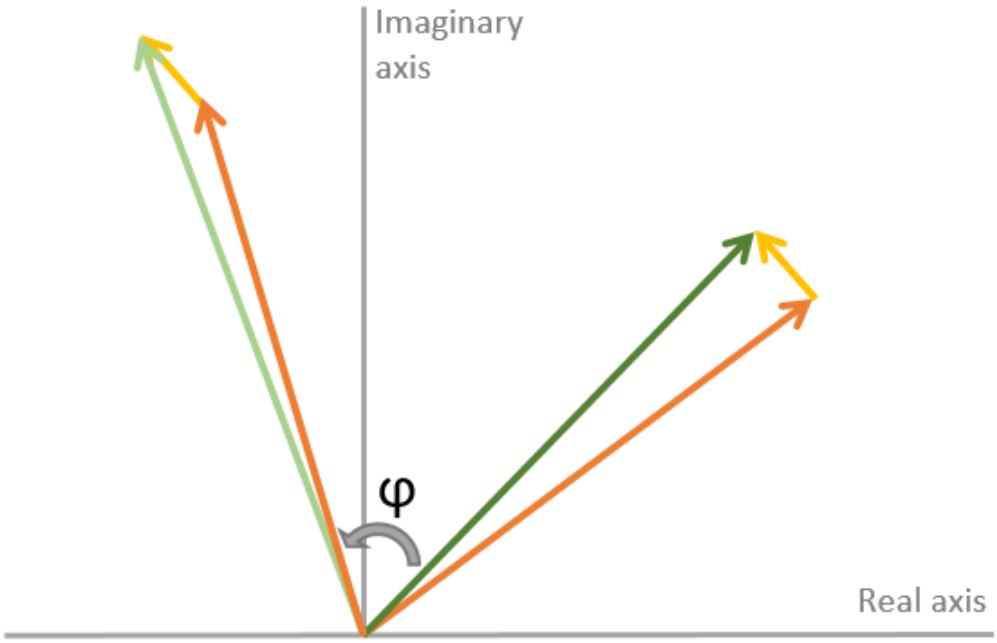


Figure 107: An Argand diagram showing the effect of the volta phase plate (VPP). The unscattered electron beam is shown in orange, the scattered electron beam is shown in yellow, and the total image is shown in dark green for the non-phase plate diagram and in light green for the phase plate representation. The VPP causes a phase shift (ϕ) of the unscattered beam, increasing the amplitude of the total signal in the final image (Wang and Fan 2019).

However, the phase plate is still not a perfect solution. Practically speaking, the VPP is difficult to configure. In order for it to function optimally it must be situated in the back focal plane, and aligning it correctly requires expertise and the process cannot be automated. This is further compounded by the lack of reproducibility in the phase shift buildup rates. Different VPP installations, or even different areas on the same VPP, may require different handling to get them to work effectively, which again limits the potential for automation of the process (Wang and Fan, 2019). Another issue with the use of a VPP during data collection is the effect it will have on the CTF. Although modern CTF estimation processes are able to take into account the phase shift when modelling the CTF of the collected data, it does introduce another variable into the equation which increases the potential for inaccurate modelling. The most serious issue with the VPP however, is that the elastically scattered electrons will also interact with the thin amorphous carbon film. They will not interact with the Volta potential in the region charged by the focused unscattered beam, but they will still come into contact with the uncharged areas of the film, leading to low levels of energy loss as some of the electrons

are both elastically and inelastically scattered by the carbon atoms. Although the energy loss is small, it still decreases the available information which can limit the overall resolution of the resulting model (Wang and Fan, 2019).

An alternative to the phase plate is the use of a lower energy microscope. Although 300 kV microscopes are the current standard for cryo-EM data collections, recent experiments suggest that the optimal energy of the electron beam for biological samples is actually 100 kV (Peet *et al.,* 2019). Decreasing the energy from 300 kV to 100 kV decreases the elastic scattering from the sample, but it decreases the inelastic scattering to a larger extent. This results in a more preferable ratio of elastic to inelastic scattering, equating to a larger signal to noise ratio. However, the main limitation in collecting data at 100 kV currently is the direct electron detectors. Although the current higher voltage microscopes can be adjusted to produce an electron beam at 100 kV rather than 300 kV, the currently available detectors are not able to capture all of the required signal from a 100 kV beam, although work in this area is ongoing (Naydenova *et al.,* 2019). In the meantime, the ratio of elastic to inelastic scattering is still better in a 200 kV microscope compared to 300 kV, and others have already demonstrated the ability to produce sub-2Å structures using a 200 kV microscope (Merk *et al.,* 2020). Collecting data on a 200 kV microscope rather than a 300 kV microscope still comes with some limitations, the increased voltage at 300 kV allows for better penetration of the beam through thicker specimens, and the reduced inelastic scattering allows for less blurring of images (Herzik *et al.,* 2019; Henderson 1995). However, the increased contrast in the lower energy microscope allows the data to be collected without the phase plate, which simplifies the data processing strategies.

7.2.2 Data Processing Considerations

Samples of a UBE3A+PSMD4 complex were collected using both approaches, but the datasets were processed to different extents due to issues with the samples used and the clipping and loading process. Two samples of UBE3A+PSMD4 on QuantiFoil R1.2/1.3 grids were collected on the 200 kV Glacios microscope, fitted with a Falcon IV camera, over 48 hours each. Neither of these datasets were processed further than 2D classes due to imperfect sample conditions, but the particles were easily visible without the use of a phase plate (Fig. 108).

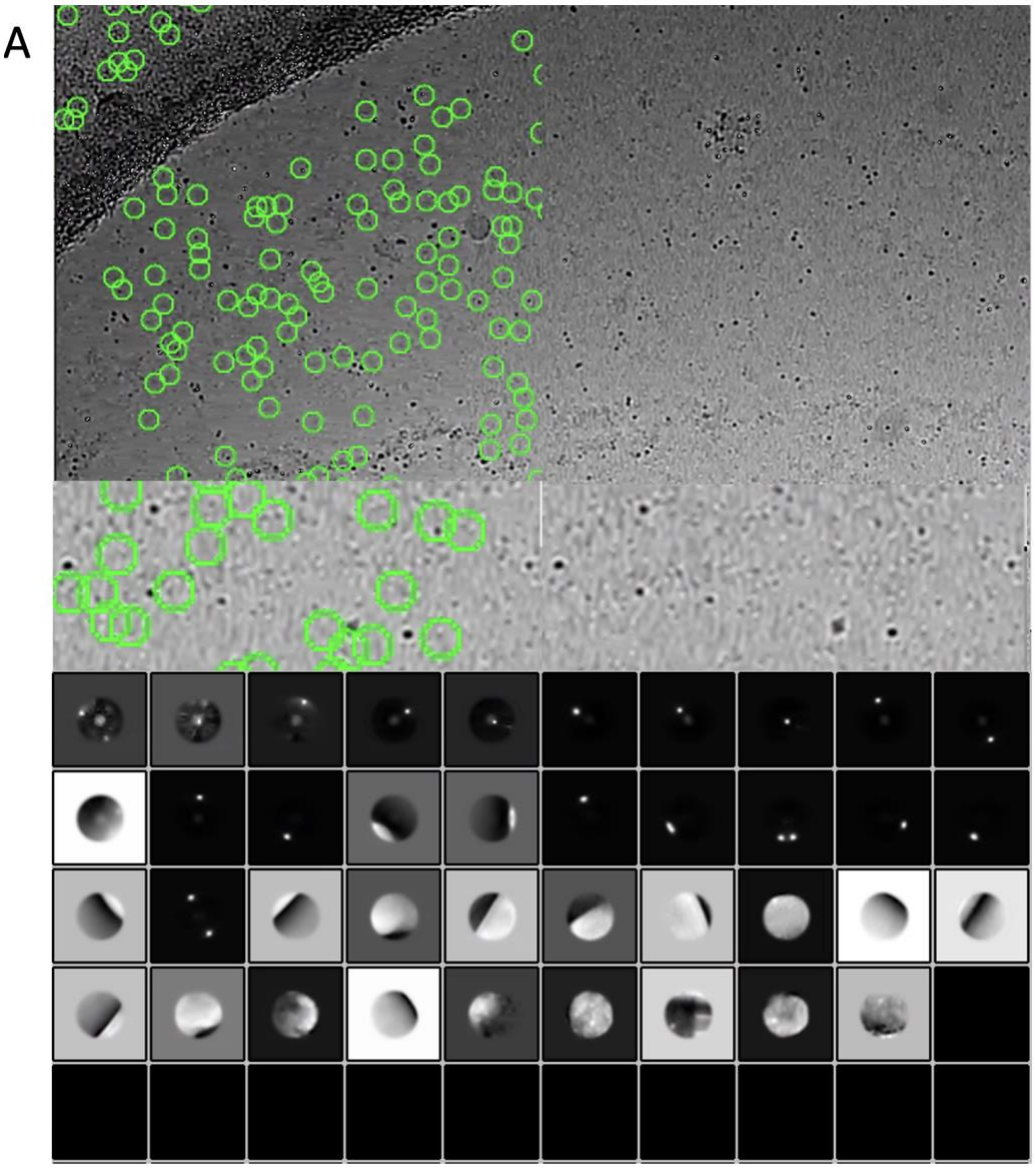


Figure 108: UBE3A+PSMD4 data collection on a 200 kV Glacios microscope equipped with a Falcon IV detector, dataset 1. Dataset 1 suffered from some form of contamination that prevented effective processing of the data. The contrast is low in the motion corrected images but the particles can still be seen by eye. Unfortunately, even after careful particle picking to avoid the contaminated areas as much as possible, the increased contrast of the unknown contaminant compared to the real particles interfered with the 2D classifications of the extracted particles.

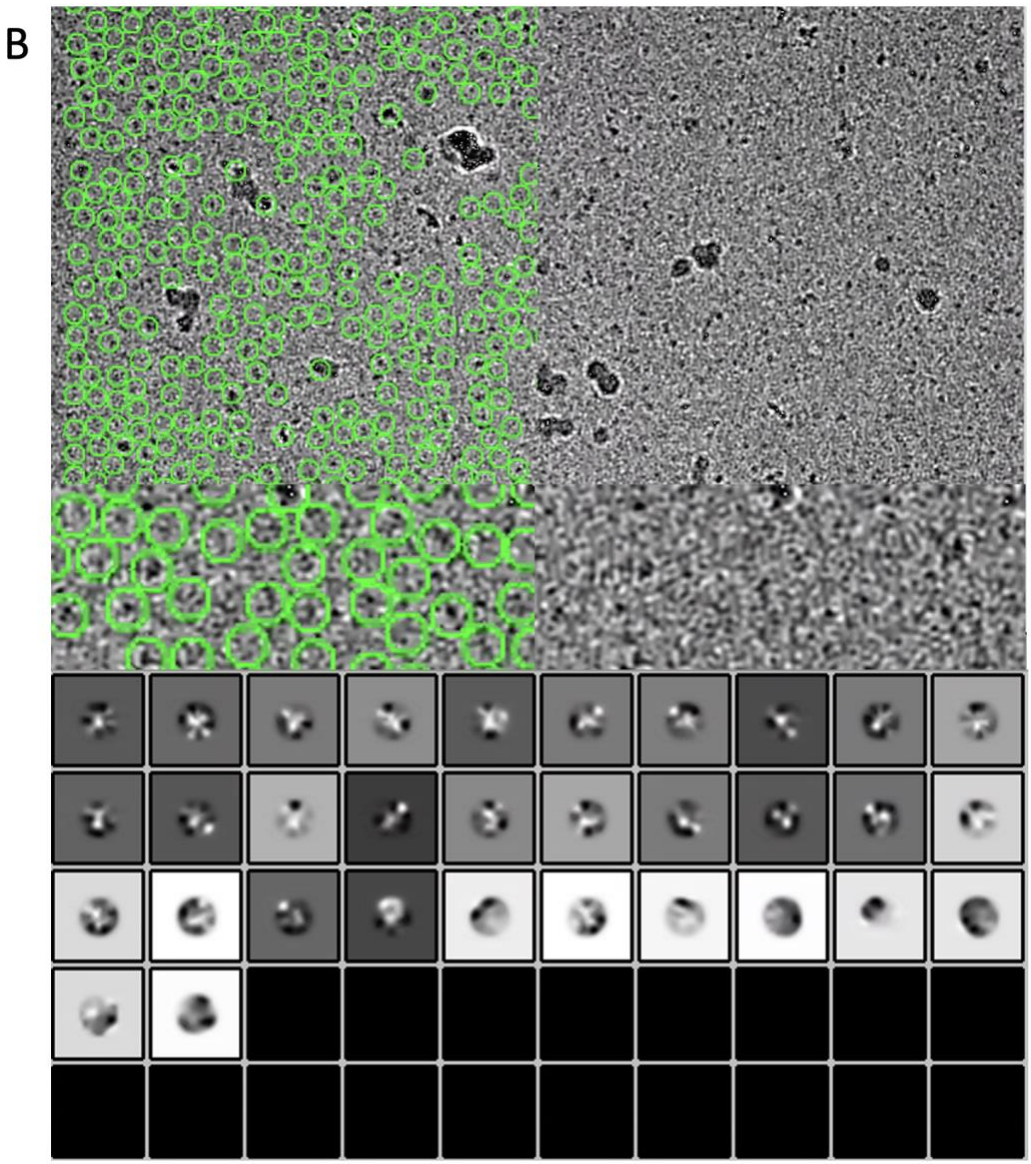


Figure 109: UBE3A+PSMD4 data collection on a 200 kV Glacios microscope equipped with a Falcon IV detector, dataset 2. Dataset 2 showed an increased level of contrast in the motion corrected micrographs so particles are clearly visible. However, the sample concentration was too high in this dataset so the 2D classes were unable to distinguish he central particle from overlapping secondary particles during the 2D classification stages, even with a tight circular mask.

Another UBE3A+PSMD4 dataset was collected on a 300 kV Titan Krios microscope with a VPP and a Gatan K3 detector. This sample was applied to the Lacey carbon grids with an ultrathin continuous carbon support (see 7.2.1) (Fig. 110).

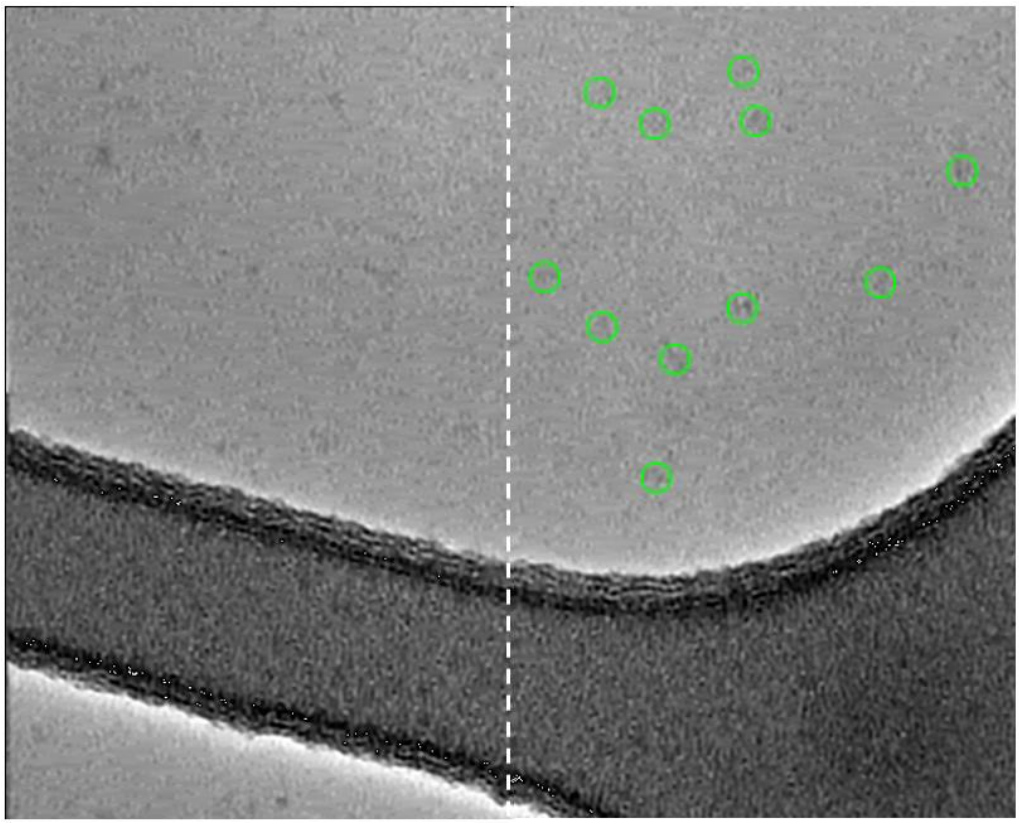


Figure 110: UBE3A+PSMD4 data collection on a 300 kV microscope with the phase plate. The sample concentration is low so there are not very many particles per image, but the particles are clearly visible in the motion corrected micrographs. The representative micrograph image has been split in two with identified particles encircled in green on the right, and the raw image shown on the left, to allow observation of the particles within context of the un-picked micrograph but also to easily demonstrate the density and distribution of picked particles within the image.

Although the particles are easily visible by eye, Relion's automated picking program still had difficulty. In order to improve the number of particles picked, I used the crYOLO program to pick particles using a model trained on my data. crYOLO was useful for this dataset because it was developed specifically to work better for small, elongated particles with low contrast (Wagner *et al.,* 2019). Relion's LoG system requires a minimum and maximum particle diameter in order to identify possible particles from other features such as ice contamination or carbon foil edges. However, the UBE3A+PSMD4 particles appear to be irregular in dimension. UBE3A alone appears to have an elongated shape, with a top profile that is much smaller in diameter than its side profile. Although the interaction between UBE3A and PSMD4 is not well characterised and the resulting shape of the complex is unknown, the particle shapes observed in the micrographs suggests that a UBE3A+PSMD4 complex is similarly elongated (Fig. 109, 110). This means that the boundaries set in the LoG picking parameters have to be fairly broad, which does not allow for very specific particle selection. By training a model in crYOLO based on my motion

corrected micrographs, the particle picking algorithm can look for the specific irregular shapes observed within a manually selected particle selection, allowing specific targeting of possible particles. LoG picking also relies on a steep contrast between the particles and the background noise, which makes particle determination difficult when the contrast in the particles is low. crYOLO's YOLO object detection method will still perform better with higher contrast images as all of the features will be more defined, but it does not rely on an observation of the steep contrast change like the LoG method does so it can handle lower contrast images much better (Wagner *et al.,* 2019).

crYOLO uses deep learning techniques to train a model for a specific dataset so that the resulting picks are more accurate (Wagner *et al.,* 2019). crYOLO employs a convolutional neural network (CNN) to process the micrograph images by first considering the image as a grid of the grey-scale values for each pixel. Similarly to how the Laplace operator is applied to images in LoG picking, CNNs apply a series of kernels to an image with the output of each layer acting as the input for the next, with each kernel identifying different features such as edge detection, corner detection etc. Starting with more simple kernels and increasing the complexity allows the program to determine a hierarchy of features, which allows more complex patterns to be identified (Yamashita *et al.,* 2018). However, different types of CNNs can detect images in different ways. The standard image detection model is the 'sliding window' model, where an image is split up into several overlapping windows which the program scans individually to identify objects. The information from these individual windows is then combined through several rounds of CNN calculations in order to identify the areas in the whole image that contain objects, in this case particles (Glumov *et al.,* 1995). This technique is fairly computationally expensive and can take a lot time so templates are typically as basic as possible in order to speed up the computation. One reason for the high computational cost of this technique is that the windows used are a set size, so if there are differently sized objects within the image the image must be downsampled many times in order to sample enough different sized objects (Wagner *et al.,* 2019). However, crYOLO uses a different CNN model for object detection known as 'YOLO' for You Only Look Once (Redmon *et al.,* 2016). The YOLO method involves looking at the entire image as a whole object rather than splitting it up into overlapping windows. It does still apply a grid system to the image, but it identifies boxes of different sizes across the image where it sees an object and then uses the grid to define the center of the object. The YOLO method is much more computationally efficient as it involves a CNN calculation on only a single image rather than several overlapping windows (Wagner *et al.,* 2019).

crYOLO doesn't use template images to identify particles, but manually picking several micrographs can be useful to ensure optimal picking. A general model is included in the program which should work on most datasets, which

would mean that several of the advantages of crYOLO could be retained without needing to manually pick micrographs. However, one of the downfalls of crYOLO's YOLO method is that the grid used to define particle locations is relatively wide, meaning that it has difficulty identifying small particles and smaller particles occupying the same grid area may not be identified separately (Wagner *et al*., 2019). Unfortunately for this project, PSMD4 is an inherently disordered protein (see Appendix 3) and UBE3A is a particularly small particle at only 98 kDa, and the non-globular nature of the structure makes some orientations appear smaller than others. In order to improve the picking accuracy for more difficult particles crYOLO includes a pipeline to manually pick a few micrographs from the difficult dataset in order to train a model for the specific sample (Wagner *et al.,* 2019). Another feature of the YOLO method compared to the sliding window method of object detection is that the sliding window technique requires examples of positive particles as well as examples of empty background areas, whereas YOLO's single pass method requires only positive particle images and it will assume that everything else constitutes background levels. This has the advantage of allowing a more complicated set of values to ignore, so the model will ignore empty grid hole areas, crystalline ice contamination areas, and areas on the carbon support equally without needing all types of negative areas to be defined. However, it does mean that in order to produce an effective model the manually picked input micrographs must be picked to completion as much as possible, although some levels of error are expected and compensated for in the algorithm (Wagner *et al.,* 2019).

Other than the particle picking using the crYOLO software, the rest of the processing for the UBE3A+PSMD4 dataset was carried out using Relion (see section 2.10.8). The data was collected across two separate sessions as the initial collection suggested that more data would be useful, but the data collection settings were kept the same. The two datasets were processed up to the CTF estimation stage separately and then merged for particle picking and further processing. Good particles were selected for through several 2D classification steps, and then an initial model was generated. This model was used to assign particles through multiple rounds of 3D classification jobs with several classes, before the final best few classes were combined into a consensus model by running them through a 3D classification job resulting in a single 3D class. This consensus model was taken forward for CTF refinements, particle polishing, and postprocessing, resulting in the final low resolution model discussed in 7.2.3.

7.2.3 UBE3A+PSMD4 Structure

Despite the various attempts at UBE3A+PSMD4 sample preparation, data collection and data processing, the best structure that I was able to obtain for UBE3A+PSMD4 is a very low-resolution model, estimated at 17.15 Å (Fig. 111).

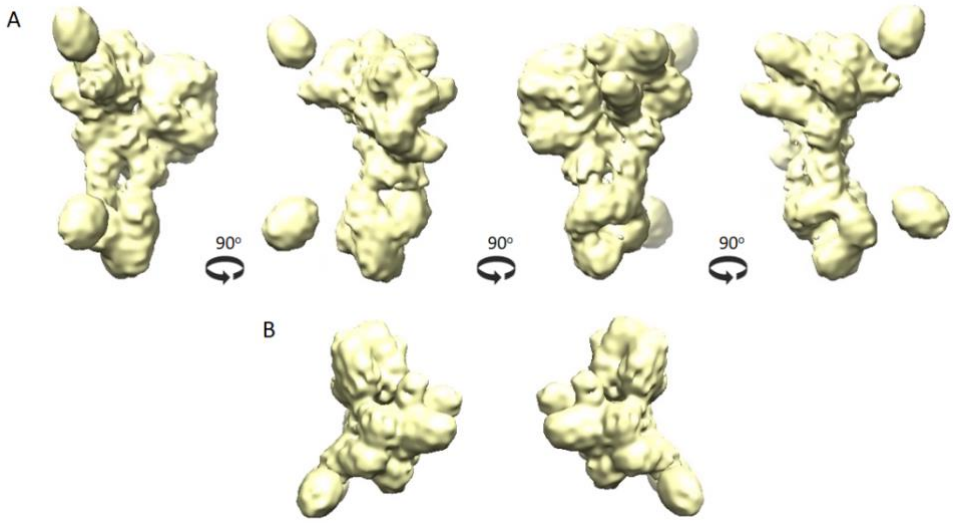


Figure 111: The low-resolution model of a UBE3A+PSMD4 sample solved by cryo-EM. A) Different side views of the calculated structure demonstrate the overall shape of the model. B) Top views of both the calculated model (right) and the flipped model (left) show the difficulty in assigning a handedness to the object.

The overall shape of the model bears a resemblance to the UBE3A-only model (see section 7.1.3), but it is also distinct enough to suggest that it could represent UBE3A in an altered conformation due to PSMD4 binding. There is no full-length structure of PSMD4, other than a low-resolution section of a structure of the full 26S proteasome (Huang *et al.,* 2016), so it would be difficult to predict how UBE3A and PSMD4 might interact, and what affect the interaction would have on the structure of UBE3A. Although PSMD4 is almost half the size of UBE3A, so would theoretically increase the size of the model by 50%, the elongated shape of UBE3A and the unknown nature of the interaction makes identification of both components difficult. It is possible that the UBE3A+PSMD4 sample has dissociated at some point during the sample and grid preparation stages, and the model above only shows a noisy representation of UBE3A alone, but it is also possible that PSMD4 could bind to UBE3A in a way that makes it difficult to identify within the complex structure. It could mimic the elongated shape of UBE3A and sit tight against the core region, or the two proteins might interact in a way that creates a more compact structure despite the increased mass. The Robetta and AlphaFold predicted structures for UBE3A were fitted into the low-resolution electron potential map for UBE3A+PSMD4 in an attempt to observe whether the unknown PSMD4 volume was necessary to fill the volume (Fig. 112).

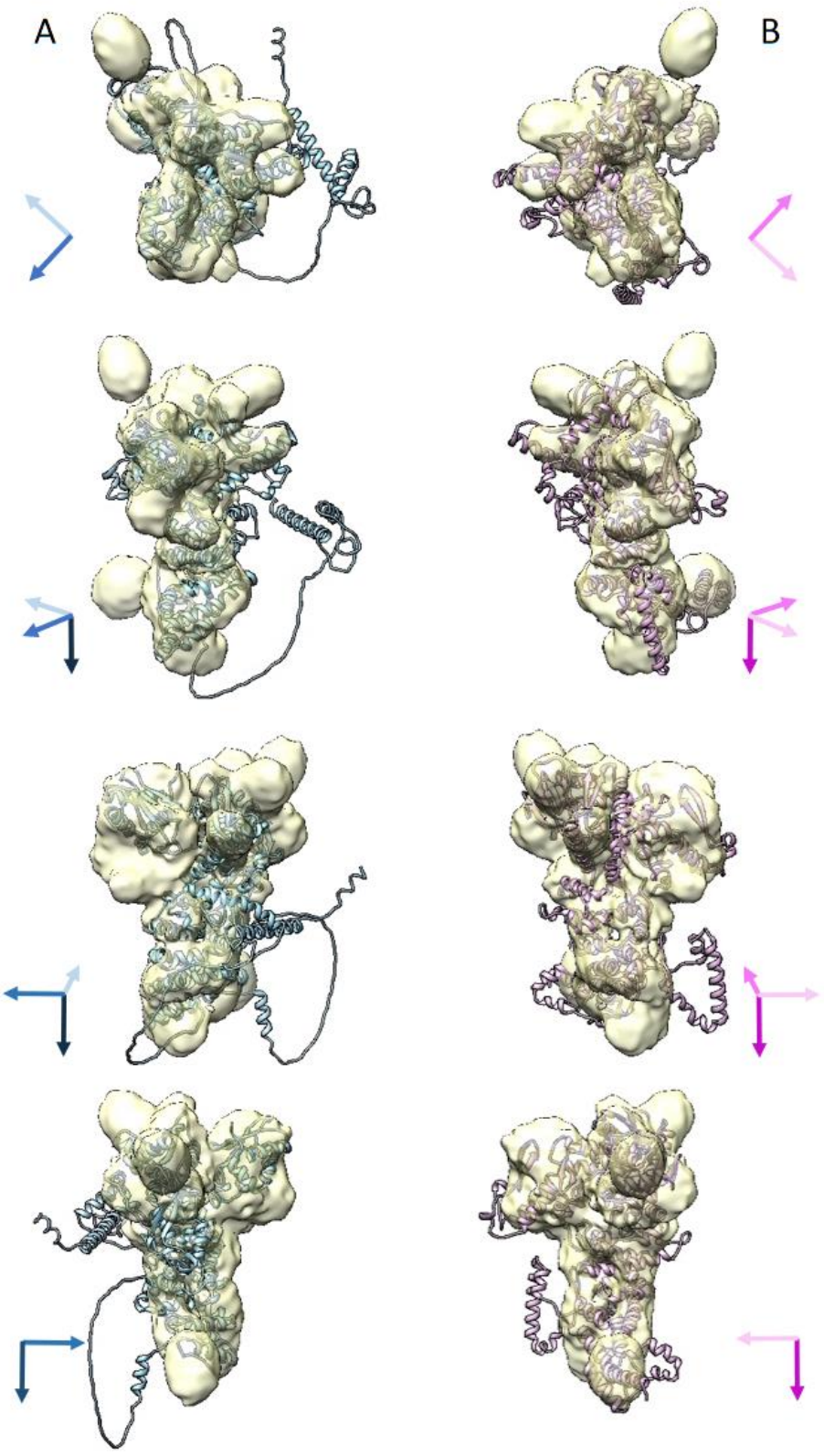


Figure 112: A comparison of the low-resolution UBE3A+PSMD4 cryo-EM model with the predicted structures of UBE3A. A) The AlphaFold structure fitted into the flipped cryo-EM model in various orientations.B) A RoseTTA predicted structure for UBE3A fitted into the unflipped cryo-EM model.

The UBE3A predicted structures do take up the majority of the cryo-EM volume, so it is unlikely that PSMD4 is binding down the length of the UBE3A core region. However, the HECT domain of UBE3A, which is the most well characterised region of UBE3A, looks abnormal. If the cryo-EM model does represent a stable UBE3A+PSMD4, it is most likely interacting with areas in the HECT domain, as well as nearby residues in the undefined N-terminal region. Either way, the current model is too noisy and too low resolution to make any conclusions about the UBE3A+PSMD4 interaction, more work in this area would be needed.

7.3 UBE3A+RLD2

7.3.1 Data Collection Considerations

The interaction between UBE3A and HERC2 has been characterised to some extent physically (Kühnle *et al.,* 2011), but the relevance of this interaction is still unknown. Since HERC2 is a giant human protein that is difficult to express and purify *in vitro*, its RLD2 domain, the domain that has been shown to interact with UBE3A, was isolated and purified instead (see sections 2.2.6, 3.2.4, and 3.4.4). A structure of the UBE3A+RLD2 complex would provide key insights into UBE3A's role in the UBE3A+HERC2 interaction, but it could also lead to a better understanding of how UBE3A interacts with accessory proteins in general.

Expression and purification of the UBE3A+RLD2 sample was optimised through several iterations, resulting in the protocol described in section 4.3.2. This sample was then applied to grids with further optimisation of the grid preparation parameters, as described in chapter 6. The most promising sample to date comprised of a UBE3A+RLD2 complex that was stabilised with glutaraldehyde crosslinking, as described in section 4.5.6 to ensure that the complex remained intact during the plunge freezing process. The single particle cryo-EM data for this sample was collected on the Titan Krios microscope at the LISCB using the K3 detector in counting mode, without the use of the VPP. For this data collection session the microscope parameters were set to the calibrated optimal settings for the microscope, as set by Dr. Christos Savva at the LISCB, with only minor adjustments made for consideration of the specific sample. This meant that parameters such as the dose rate, magnification, and exposure time were not altered from the predetermined optimal microscope settings (see section 2.10.9), but the placement of the acquisition areas and were carefully determined to acquire the optimal ice thickness preferred by the sample. The data was collected without the phase plate but using the AFIS implementation to allow rapid image acquisition, resulting in a large dataset after 45 hours of data collection.

7.3.2 Data Processing Considerations

The UBE3A+RLD2 micrographs were processed concurrently through both Relion and CryoSparc to identify if one program had any advantage over the other. Both programs were able to correct the motion and model the CTF

without issue, but the blob picker in the CryoSparc package provided better particle picking than both the LoG picker and the template-based picking program in Relion, without the need for as much hands-on curation of the picking or involvement in the process. The results from CryoSparc's blob picker were also sufficient to negate the use of either the crYOLO or TOPAZ programs, which require much more user input and manual curation of picked micrographs, as well as being more computationally expensive (Fig. 113)

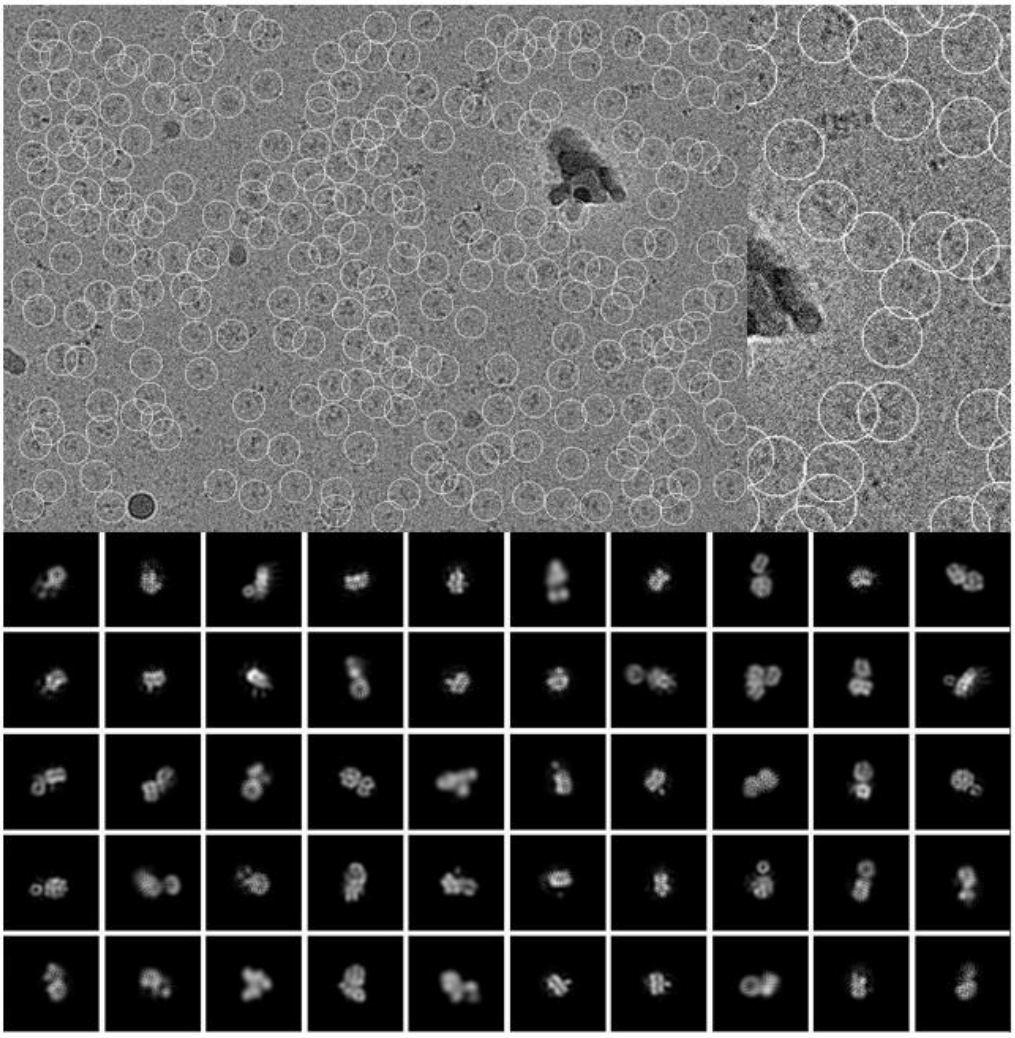


Figure 113: An example of a micrograph picked using the blob-picker job in CryoSparc. A view of the overall micrograph shows the particle density across the image, and the closer view of several picked particles around an area of ice contamination shows the avoidance of the non-ideal areas. The final round of 2D classes resulting from this set of particles suggested a good orientational representation of the sample in the dataset.

Initially, following particle picking and subsequent particle extraction, the CryoSparc particles were subjected to limited rounds of 2D classification before moving on to *Ab initio* 3D model generation. This was attempted first as 2D classification is known to be difficult for small particles with a low signal to noise ratio, and too much reliance on 2D classes can introduce more

preferential orientation issues into the dataset if some orientations of the protein show a decreased contrast in the ice relative to others. This could be the case for UBE3A, as it is a more elongated than globular structure views from the top and bottom will result in particles with an increased contrast but smaller diameter, whereas side views would result in larger particle images but a decreased sample thickness that would mean a decreased contrast. As the orientation of the UBE3A+RLD2 complex was previously unknown, only two rounds of 2D classes were used to eliminate the most egregiously bad particle images whilst retaining the full complement of productive images. This resulted in promising initial models in which the RLD2 propeller shape could easily be identified, but further refinements revealed a lot of noise without allowing any further structure identification. In order to reduce the noise in the data in an attempt to increase the resolution of the 3D models, I returned to a more thorough curation of the particles through several successive rounds of 2D classifications in CryoSparc. Following this, several *ab initio* jobs were run iteratively, generating multiple classes per job and selecting the classes that most resembled each other for continuation to the next round. This was continued until there was a clear distinction between classes containing proteins and a class of just noise. The final *ab initio* model was then taken forward to homogeneous refinement, and the particles involved in the model were re-extracted without any binning. The new particle set was used for a second homogeneous refinement job to attempt to increase the resolution of the model, but unfortunately the map was still too low resolution to make out any structural features of either constituent protein. It was decided that further attempts to refine the model would not enable reconstruction of a high-resolution map from the given data, so a final homogeneous refinement job was run and the processing was stopped.

7.3.3 UBE3A+RLD2 Structure

The crosslinked UBE3A+RLD2 sample produced a large amount of particles that resulted in initially promising 2D classes (Fig. 113), but was only able to produce a low-resolution structure (Fig. 114).

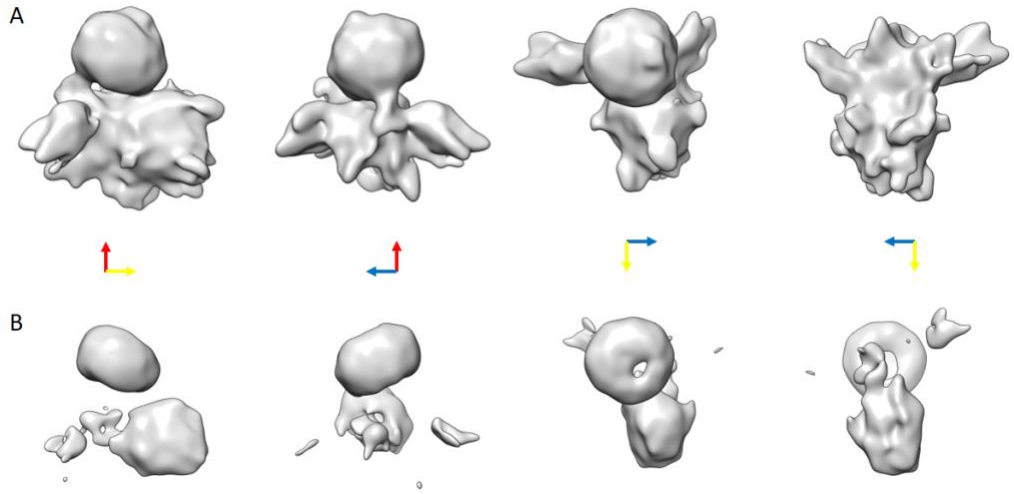


Figure 114: The low-resolution cryo-EM model for a crosslinked complex of UBE3A and the RLD2 domain of HERC2. The orientations of the different views are shown by the coloured arrows between the two rows. A) The cryo-EM map is shown with a volume level that shows the maximum area before disconnected noise artefacts begin to appear. B) The same cryo-EM map is shown with a decreased volume to level to show more detail in the RLD2 domain region.

The UBE3A+RLD2 map is too low resolution to be able to model a sequence into the density, but it does show an overall shape of the interaction. The 'puck' shape of the RLD2 domain can be clearly seen at the top of the densities shown in figure 114, while the remaining density seems to be much noisier but it does resemble the rough shape and size of a UBE3A monomer (see section 7.1.3). At a low volume level the pore region of the RLD2 is visible, with a larger hole in the face closest to the UBE3A region and a smaller hole on the outer face. The high-resolution crystal structure of the RLD2 domain (Fig. 117) shows that the 'propellers' of the seven-bladed β-propeller are angled so that the central pore forms a 'V' shape, with a larger pore on one side than the other. This is a characteristic feature of a β-propeller structure (Chen *et al.,* 2011). The UBE3A region is not as well-defined, and a higher volume level is required to see the full area. However, at the higher volume level a UBE3A model can be fitted into the density, showing a reasonable similarity between the model and the map (Fig. 115).

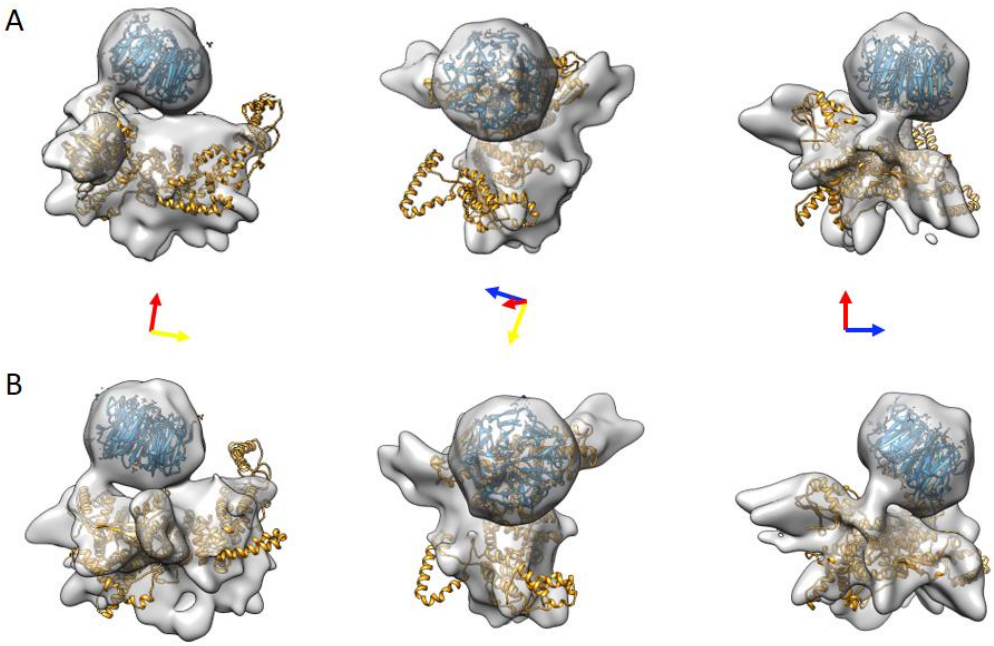


Figure 115: Model of the UBE3A and RLD2 monomers fitted into the low-resolution UBE3A+RLD2 cryo-EM map. The RLD2 map derived from x-ray crystallography is shown in blue, and the Rosetta model 1 for UBE3A is shown in orange. A) The original cryo-EM map generated by cryosparc. B) The cryo-EM map was flipped in chimera to examine the handedness.

Although the cryo-EM map is very low-resolution and particularly noisy in the UBE3A region, the volume is roughly the right size and shape for a believable complex structure and so it gives an idea of the potential interaction interfaces. However, the low resolution of the map makes it impossible to accurately determine the handedness. In an attempt to overcome this, the volume was flipped in chimera to reverse the handedness and the UBE3A and RLD2 models were fitted into both models to allow a direct comparison of the fit (Fig. 115). The flipped map possibly fits the models slightly better, particularly when considering the angle of the C-lobe and N-lobe of the HECT domain (Fig. 115a and b, right). However, the resolution is too low and the noise contribution is too high to confidently determine the true handedness of the structure.

The RLD2 structure has reasonably distinctive faces that can be discerned in the low-resolution map, but the seven blades of the β-propeller show a pseudo rotational symmetry that makes accurate determination of the residues involved in the interaction impossible. However, the UBE3A structure has a less symmetrical structure that allows an approximation of the areas in proximity to RLD2 in the complex structure more possible. The identified domains of UBE3A can be mapped onto the Rosetta model as it is fitted into the complex map, and the interaction interface can be narrowed down the broad areas of UBE3A's surface (Fig. 116).

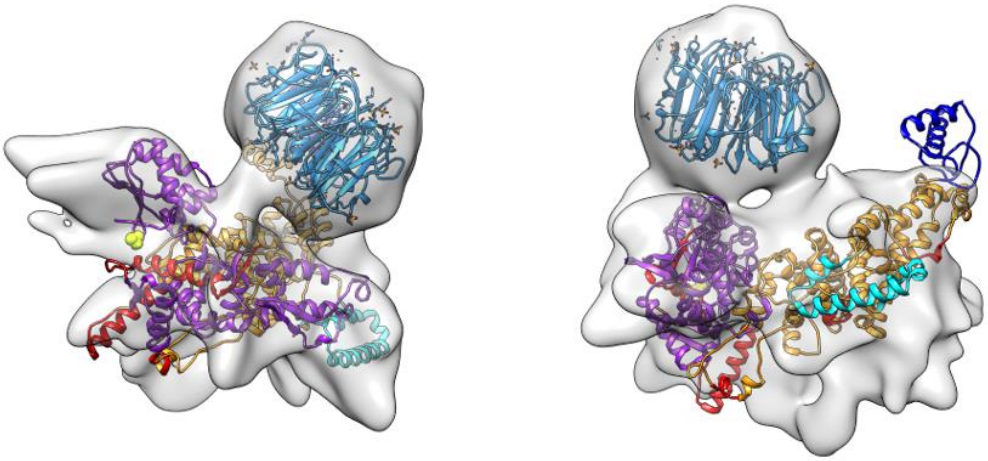


Figure 116: The UBE3A and RLD2 models fitted into the cryo-EM map of the complex, with the UBE3A model coloured to show the location of known domains. The HECT domain is shown in purple, the E6 binding regions are shown in red, the AZUL domain is shown is dark blue, the HERC2-interacting region as defined by Kühnle et al., 2011 is shown in cyan, the catalytic cysteine residue is shown in a spherical form in yellow, and uncharacterised regions are shown in orange.

The areas of density that connect the UBE3A and RLD2 regions, as suggested by the low-resolution cryo-EM model presented, appear to cluster in three points. One interaction bridges between the open face of RLD2 and the catalytic cleft of UBE3A's HECT domain, another connection point bridges the edge of the open face of RLD2 and the large N-terminal subdomain of UBE3A, and a final interface connects the opposite edge of the same RLD2 face with a region midway down the uncharacterised central region of UBE3A. Kühnle *et al.,* (2011) used truncation mutants to define the region of UBE3A involved in the interaction as the area shown in cyan (Fig. 105, 116). However, this appears to form a flexible domain within the UBE3A structure, as suggested by the inability of both AlphaFold and the Robetta server to place it in relation to the rest of the structure with confidence. Given the proposed flexibility of this region, it is possible that it undergoes a conformational change upon RLD2 binding to bring it up into position between the uncharacterised core of UBE3A and the RLD2 unit. CD data for the interaction between UBE3A and RLD2 suggested that there was no large conformational rearrangement upon binding, but if the HERC2-binding region is connected to the rest of the UBE3A structure by a flexible linker, it may be possible to significantly alter the orientation of the domain relative to the rest of the structure without inducing a change in the secondary structure elements. One of the AS-associated UBE3A mutations displayed in figure 104 (section 7.1.3) sits on the front face of this core region of UBE3A where the RLD2 domain may connect, so it is possible that this single residue (I329 from isoform 1) is responsible for holding the region identified by Kühnle *et al.,* in place on the core region of

UBE3A in the presence of RLD2. It would be interesting to see if a mutation of this residue would affect the binding interaction with RLD2.

Ultimately, the experimentally derived map is too low resolution to make any conclusions about the complex and the residues involved. The crosslinking process also adds a level of uncertainty to the structure, as it may artificially introduce connections that would not be present in a physiological complex. The RLD2 domain also only represents a small portion of the whole HERC2 protein, so the RLD2 region may sit within HERC2 in a way that limits the available interaction interface with UBE3A. However, the low-resolution map presented here does provide a potential initial insight into the interaction, and could hopefully aid in understanding any future findings.

7.4 RLD2

7.4.1 Data Collection Considerations

HERC2 is a key interacting partner protein of UBE3A. The interaction between the two proteins has been studied to some extent (Kühnle *et al*., 2011; Martinez-Noël *et al.,* 2012), but the physiological significance of the interaction is still unclear, and there is no structural information available for the complex beyond the identification of approximate regions of interaction within each protein. The HERC protein family is named for the presence of both HECT and RLD (RCC1-like domain) domains within each family member (García-Cano *et al.,* 2019). HERC2 is one of the giant HERC proteins, 4838 amino acids in length and with a molecular weight of 527 kDa, it contains three RLD domains and a single HECT domain (García-Cano *et al.,* 2019). The large size of HERC2 makes it a promising target for cryo-EM structure determination, but unfortunately its size and complexity makes it extremely difficult to express and purify. Attempts were made to purify full-length HERC2 during this project, and considering the absence of any existing literature suggesting that this feat has been accomplished before, the results have been promising (see section 3.6). However, due to time constraints of this project I also looked for ways to simplify the process. The interacting regions of UBE3A and HERC2 were mapped to the RLD2 domain of HERC2, and a region spanning amino acids 150-200 of UBE3A isoform 1 (Kühnle *et al.,* 2011), so constructs were created and the products were expressed and purified for each region (see section2 4.3.2 and 4.3.3). Although full-length HERC2 is a great size for cryo-EM, the isolated RLD2 domain is less than 50 kDa and so would be extremely small for cryo-EM techniques. RLD domains are also predicted to have a very stable structure (Hadjebi *et al.,* 2008), so the RLD2 domain from HERC2 was instead investigated using X-ray crystallography.

The expression and purification of RLD2 was relatively straightforward, although the removal of the N-terminal His-tag proved challenging. Although the cleaved product could be generated, this was only achieved in small quantities, therefore for much of the work involving RLD2 in this project the

tag was retained. Crystal screens were prepared with both tagged and untagged RLD2, both to maximise the chance of obtaining any crystal of the target, and also to see if the presence of the tag had an effect on the resulting structure. Both constructs produced robust crystals in many different conditions, and all of the crystal diffraction data was collected at the Diamond Light Source (beamline I04). The data was collected to high resolution as described in section 2.11.3, and the space group and resolution estimate were calculated using the DIALS program (Winter *et al.*, 2018).

7.4.2 RLD2 Structure

The crystal structure of the RLD2 domain of HERC2 was solved to 1.28Å, which allowed us to observe that it forms the seven-bladed β-propeller structure typical of RLD domains. The structure was a high enough resolution to determine the orientations of all the side chain residues, and even to resolve the hole at the centre of the aromatic residues (Fig. 117b).

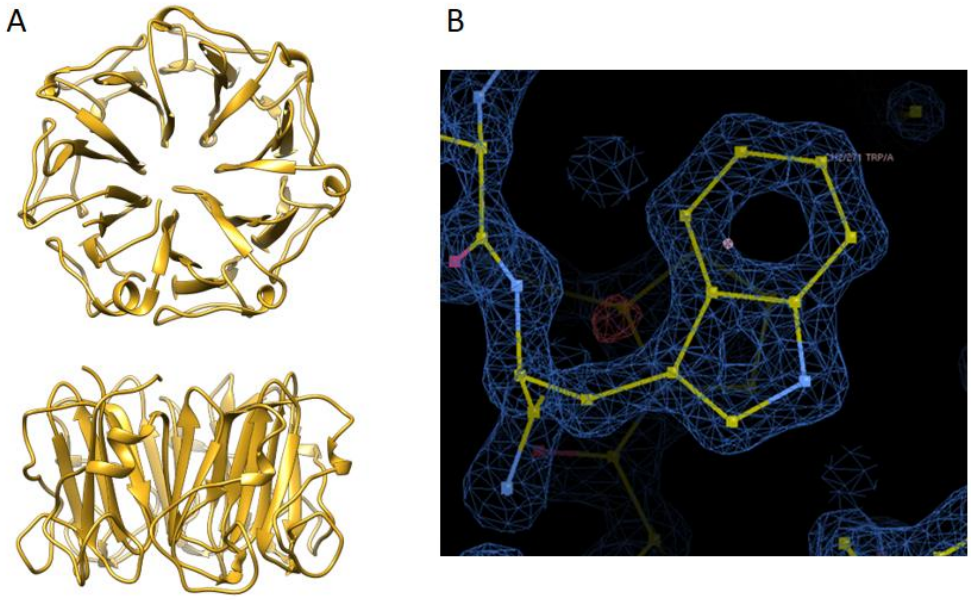


Figure 117: The crystal structure of RLD2 solved by X-ray crystallography. A) A ribbon diagram of the crystal structure shows the seven-bladed β-propeller structure of the domain. B) The electron density map for residue W271 shows the holes within aromatic rings.

Space group	$P2_1$
Unit cell dimensions (Å)	a=57.232, b=39.532, c=71.923 β=96.812°
Resolution (Å)	42.1 – 1.28 (1.31-1.28)
Total reflections	538,983 (21,051)
Unique reflections	77,630 (3,542)
Completeness (%)	96.3 (87.5)
Multiplicity	6.9 (5.9)
$\langle I/\sigma \rangle$	8.6 (0.4)
R_{merge} (%)	10.6 (189)
R_{pim} (%)	4.3 (82.9)
$CC_{1/2}$	0.998 (0.304)
Refinement	
R_{work}/R_{free} (%)	14.5/18.8
No. of atoms	14,623
Macromolecule	2,797
Solvent	236
Average B-factors	
Macromolecule	18.2
Solvent	25.2
RMSD bond lengths (Å)	0.013
RMSD angles (°)	1.76
Ramachandran plot favoured / outliers	99.6 % / 0 %
Clashscore	1.94

Table 10: The associated crystallography statistics for the RLD2 structure. The values in parentheses represent the highest resolution shell.

HERC2 contains three RLD domains, two of which already have structures deposited in the PDB (RLD1 – 4l1m, RLD3 – 3kci). The sequences of the three RLD domains of HERC2 show around 50% identity, and alignment shows that some areas show a much higher degree of conservation than others (Fig. 118).

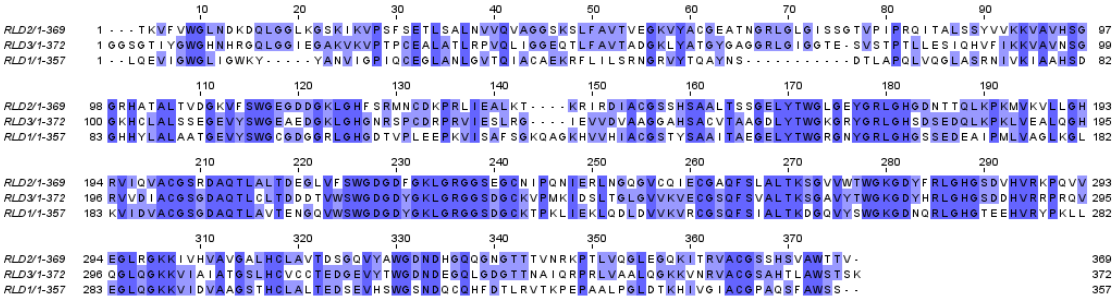


Figure 118: A multiple sequence alignment for the three RLD domains within the HERC2 protein. The domain boundaries were determined using the SUPERFAMILY domain designations, and the sequences were aligned using the ClustalO implementation within the JalView program (Waterhouse et al., 2009). The sequences were coloured by the percentage identity across the three sequences, showing areas with a higher level of conservation with a darker blue colour, and the residues with no conservation uncoloured.

As the alignment of the three domains suggests that some areas may be more conserved than others, a superposition of the structures of the three RLD domains was used to determine whether regions of sequence conservation could be identified in the protein structure. (Fig. 119).

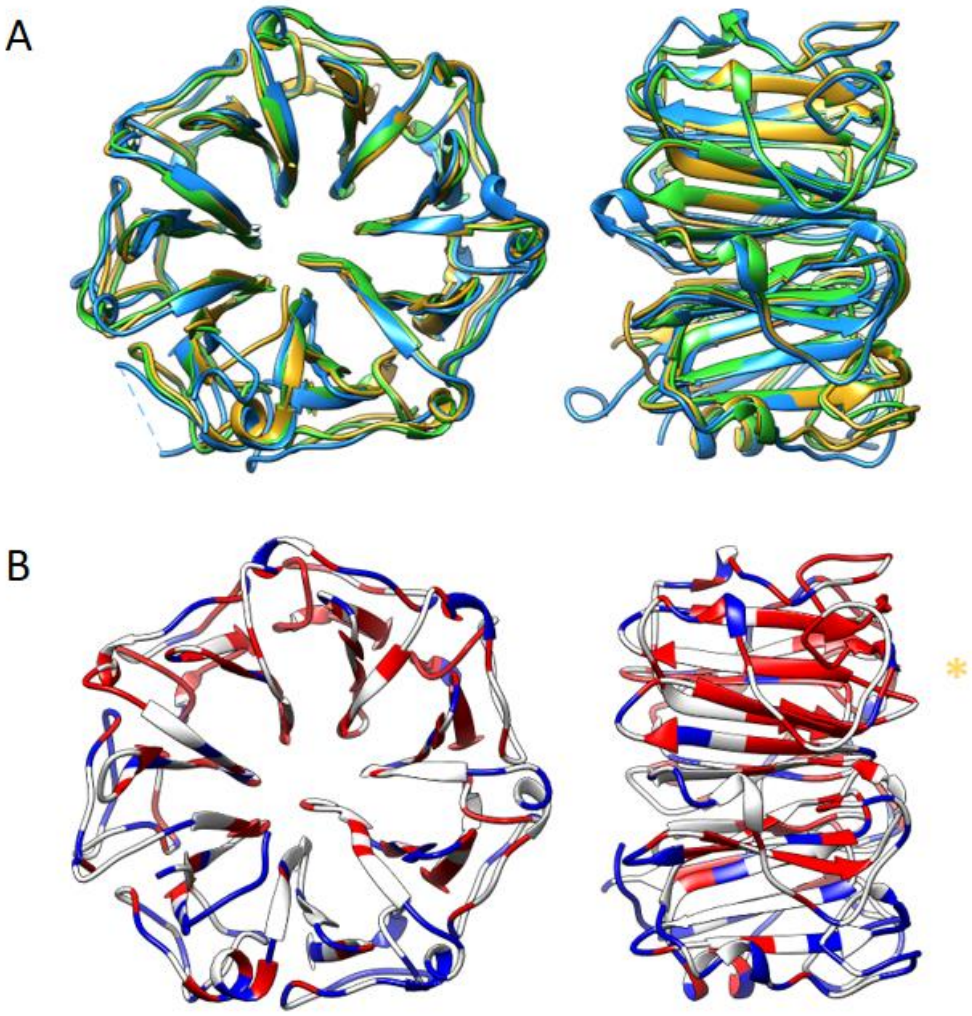


Figure 119: A comparison of the three RLD domains of HERC2. A) The three crystal structures for the RLD domains of HERC2 are superposed, showing the high level of structural conservation between the domains, even amongst the less-structured loop regions. The RLD1 structure (4l1m) is shown in blue, the RLD2 structure determined in this work (not yet deposited) is shown in gold, and the RLD3 structure (3kci) is shown in green. B) A ribbon display of the RLD2 structure coloured by the level of conservation of each residue across the three domains. Highly conserved residues are coloured red, residues with low conservation are coloured blue, and white residues are somewhere in between. The more conserved face of the domain is indicated with the yellow asterisk. The conservation threshold was determind by the relative abundance of each residue across all three sequences, so resides that are conserved across all three sequences are shown in red, residues that are conserved across two sequences are shown in white, and residues that differe between all three sequences are displayed in blue.

When all three HERC2 RLD structures are overlain in Chimera (Fig. 119a) they show a striking level of structural similarity. The core RMSD for the RLD2 and RLD3 (3kci) structures is 0.6352, and the RMSD for RLD2 and RLD1 (4l1m) is

1.0711. Even many of the less-ordered loop regions align precisely. The key differences between the structures are within the side chain residues that protrude towards the centre of the propeller. Within Chimera, the sequences for each domain were aligned based on the structural alignment, and the residues within the structure were coloured based on the level of conservation in that alignment (Fig. 119b). There are conserved residues spread throughout the structure, but one area, towards the top right of both views, shows a higher level of conservation than the rest. It appears to map more towards one face than the other as well, indicated by the asterisk in Fig. 119b, which could suggest a conserved role for this region of the domain.

One interesting observation regarding this high structural similarity is that the sequences for the RLD1, RLD2, and RLD3 domains are not expected to be quite so similar. Hidden Markov models (HMMs) are a tool used in computational biology to define a sequence region. The HMM of a certain domain will show the probability of any possible residue occurring at any point in the protein sequence while retaining the core features of the defined domain. A given sequence is compared to the HMM to determine if it matches the core domain architecture sufficiently to be defined as that domain (Yoon *et al.,* 2009). While the Superfamily database defines the RLD region by a single overarching HMM, PFam defines the domain a series of seven identifiable repeats, one for each blade of the propeller. Although the sequence for each blade is similar, reflected by the conserved structure needed to form the seven-bladed propeller, the sequence similarity between each blade in a single domain tends to decrease from the N-terminus to C-terminus of the domain. The RLD2 region for HERC2 contains all seven of these RCC1-like repeat sequences, but only four of these repeats are identified within the RLD1 domain, and only 6 within RLD3, hence the use of the Superfamily domain boundaries when isolating the RLD sequences for the sequence alignments. This suggests that although the structural similarity is retained across all three domains, there is more variation in these areas at the sidechain level, potentially mediating the different roles of the three domains within HERC2. RLD domains are a defining feature of HERC proteins (García-Cano *et al.,* 2019), and they most likely perform the same functions within them. In order to see if the structural similarity and highly conserved region observed in the RLDs within HERC2 could also be observed in the RLDs of other HERC proteins, other members of the HERC protein family were identified using the HMMER program (Finn *et al.,* 2011) and the HMM motif defined by Superfamily, and a multiple sequence alignment was performed using the ClustalO implementation in JalView (Fig. 120).

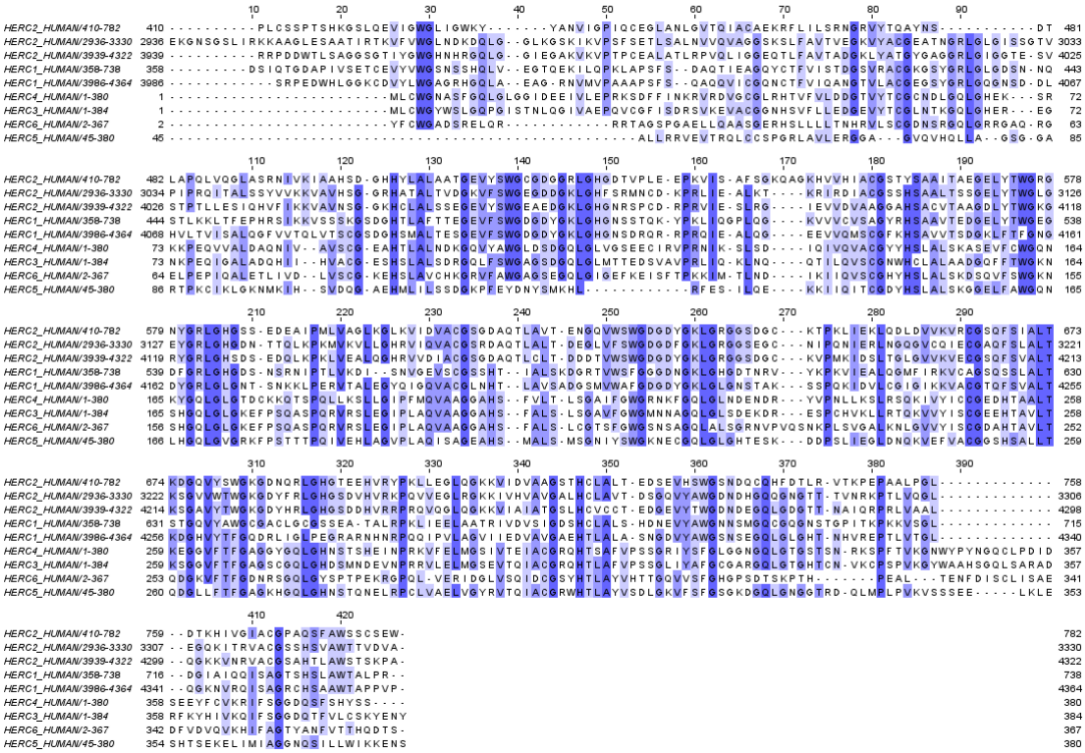


Figure 120: A multiple sequence alignment for the RLD domains within all HERC proteins. The sequences were aligned using the ClustalO module within JalView, and residues were coloured based on the percentage sequence identity of the alignment. Residues with a higher percentage sequence identity are coloured in darker shades of blue.

The sequences for all RLDs within all HERC proteins do not show quite the same level of similarity as the RLDs from HERC2 only, but that is not unexpected. The domains do still show areas with a level of conservation across all sequences (Fig. 120). In order to get a better understanding of the relationships between the different sequences a phylogenetic tree was created for the RLD domains within all HERC proteins (Fig. 121).

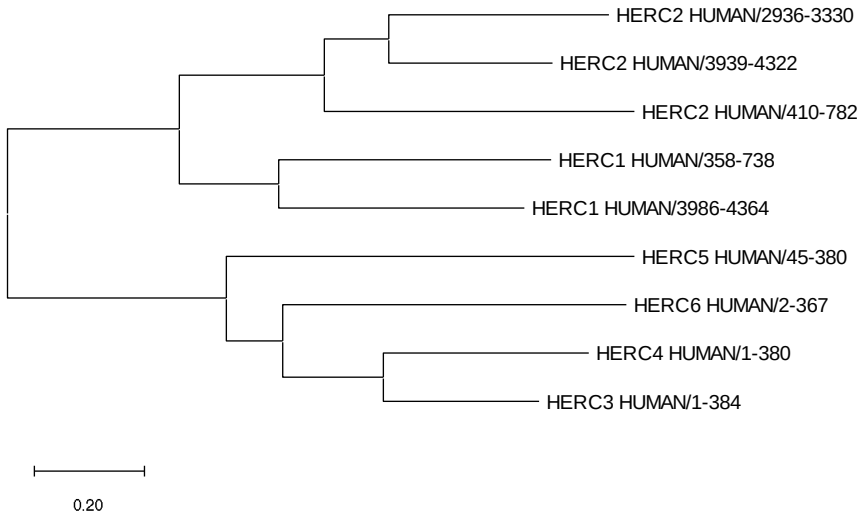


Figure 121: Evolutionary analysis of the RLD sequences within HERC proteins. The evolutionary history was inferred by using the Maximum Likelihood method and JTT matrix-based model (Jones et al., 1992). The tree with the highest log likelihood (-7139.32) is shown. Initial tree(s) for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using the JTT model. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. This analysis involved 9 amino acid sequences. There were a total of 427 positions in the final dataset. Evolutionary analyses were conducted in MEGA X (Kumar et al., 2018)

The phylogenetic tree of the HERC RLD domains shows that the RLD domains of the giant HERC proteins (HER2 and HERC1) are more similar than those of the smaller HERC proteins (HERC3, HERC4, HERC5, and HERC6), and the individual RLDs within each of the giant HERC proteins are more similar to each other. This could suggest that the multiple RLDs within the giant HERC ligases have occurred through gene duplication events, rather than through the fusion of multiple genes (Vogel *et al.,* 2005). Of the nine RLDs within the HERC proteins, experimentally determined structures are available for four. These are the three HERC2 domains (kl1m, 3kci, RLD2 – not yet deposited) and the third RLD of HERC1 (4o2w). When the structures of these domains are overlaid, they show the same structural similarity as the three RLDs within HERC2 only, but with the addition of an alpha helix in one of the loop regions (Fig. 122).

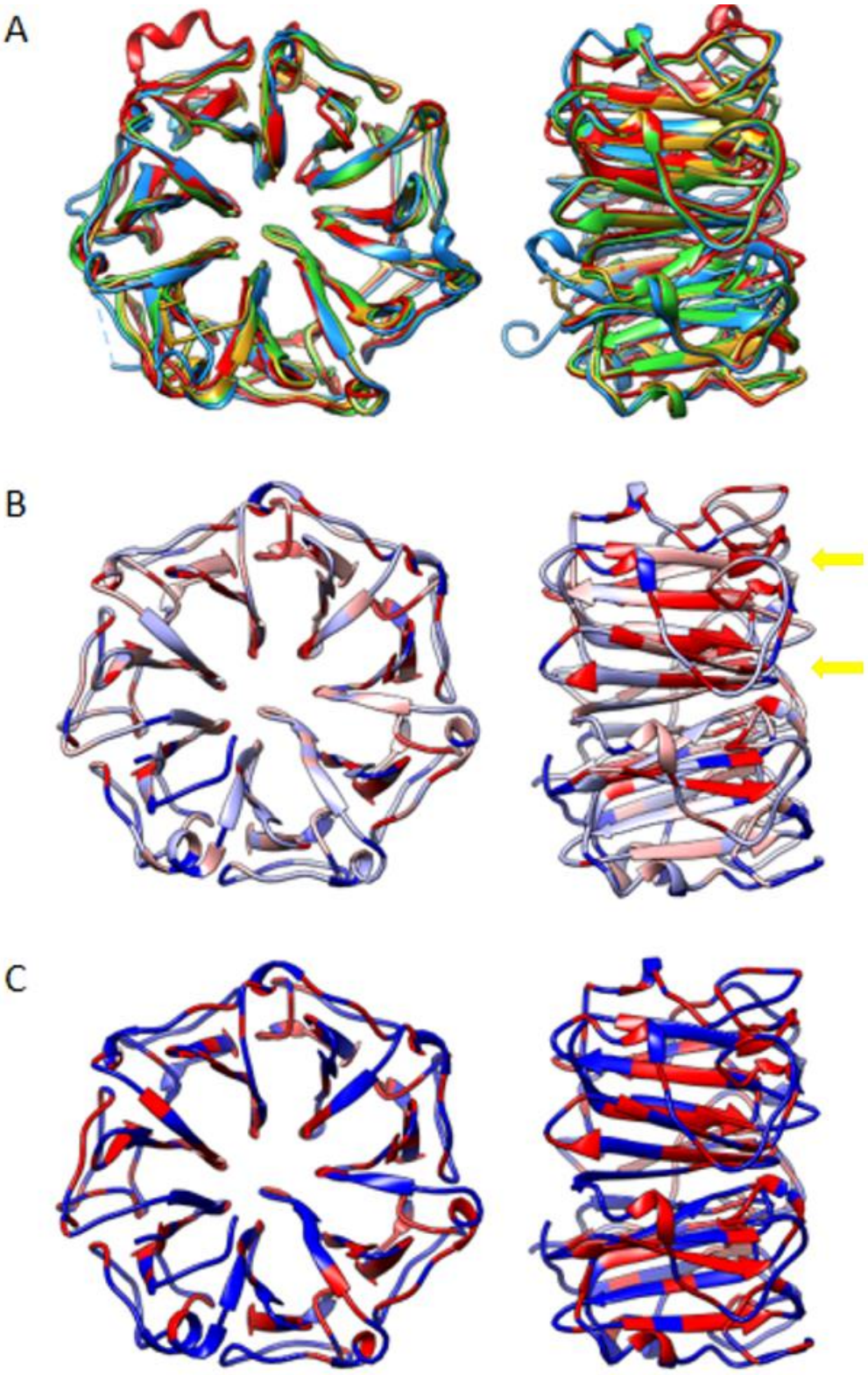

A
B
C

Figure 122: A comparison of the four known structures of RLD domains within HERC proteins. A) All four structures have been superimposed in Chimera, with RLD1 of HERC2 (4l1m) in blue, RLD2 of HERC2 in gold, RLD3 of HERC2 (3kci) in green, and RLD3 of HERC1 (4o2w) in red. B) The structure of HERC2 RLD2 alone, with the residues coloured by the level of conservation between all four structures. Yellow arrows indicate the region of he structure that appears to be the most conserved. C) The structure of HERC2 RLD2 alone, with the residues coloured by the conservation between HERC2 RLD2 and HERC1 RLD3.Highly conserved residues are shown in red, while less conserved residues are shown in blue.

The RMSD of HERC1 RLD3 and HERC2 RLD2 is 1.0740, which is only slightly higher than the RMSDs between the other HERC2 RLDs and RLD2 (1.0711 for RLD1 and 0.6354 for RLD3). However, the RMSD between HERC1 RLD3 and HERC2 RLD1 (the HERC2 RLD with the least similarity to RLD2) is 1.2716, which suggests that the RLDs within HERC2 still show a higher level of structural conservation than that of all RLDs within the HERC family. The region with the highest level of conservation still seems to map to the same region of the structure, with a slight preference for one face (Fig. 122b). Interestingly, the key difference between the HERC1 and HERC2 RLD structures, the extra alpha-helix, occurs within the region that was the most conserved amongst the HERC2 RLD domains (Fig. 119b). However, the conservation across all four domains is skewed by the fact that three of them are more closely related than the other. In order to more accurately represent the conservation of the RLDs between HERC1 and HERC2, I took the HERC2 RLD2 structure as a representative of HERC2, and the HERC1 RLD3 structure as a representative of HERC1, and rendered the HERC2 RLD2 structure by the level of conservation between the two (Fig. 122c). In this diagram, the more conserved areas are distributed more randomly throughout the structure.

Although RLD domains are one of the defining features of HERC proteins, they are not only found in the HERC family. As well as the six HERC proteins, 20 other known proteins within the human proteome contain the sequence for the RLD domain. A multiple sequence alignment for the RLD sequences of the 26 RLD-containing human proteins (Appendix 4) shows the locations of large loop regions within several of the individual RLDs, but the large blocks of similar length regions show areas of high conservation. When the isolated RLD sequences from each protein are formed into a phylogenetic tree using MEGAX (Kumar *et al.,* 2018) the evolutionary relationships between the domain sequences can be observed (Fig. 123).

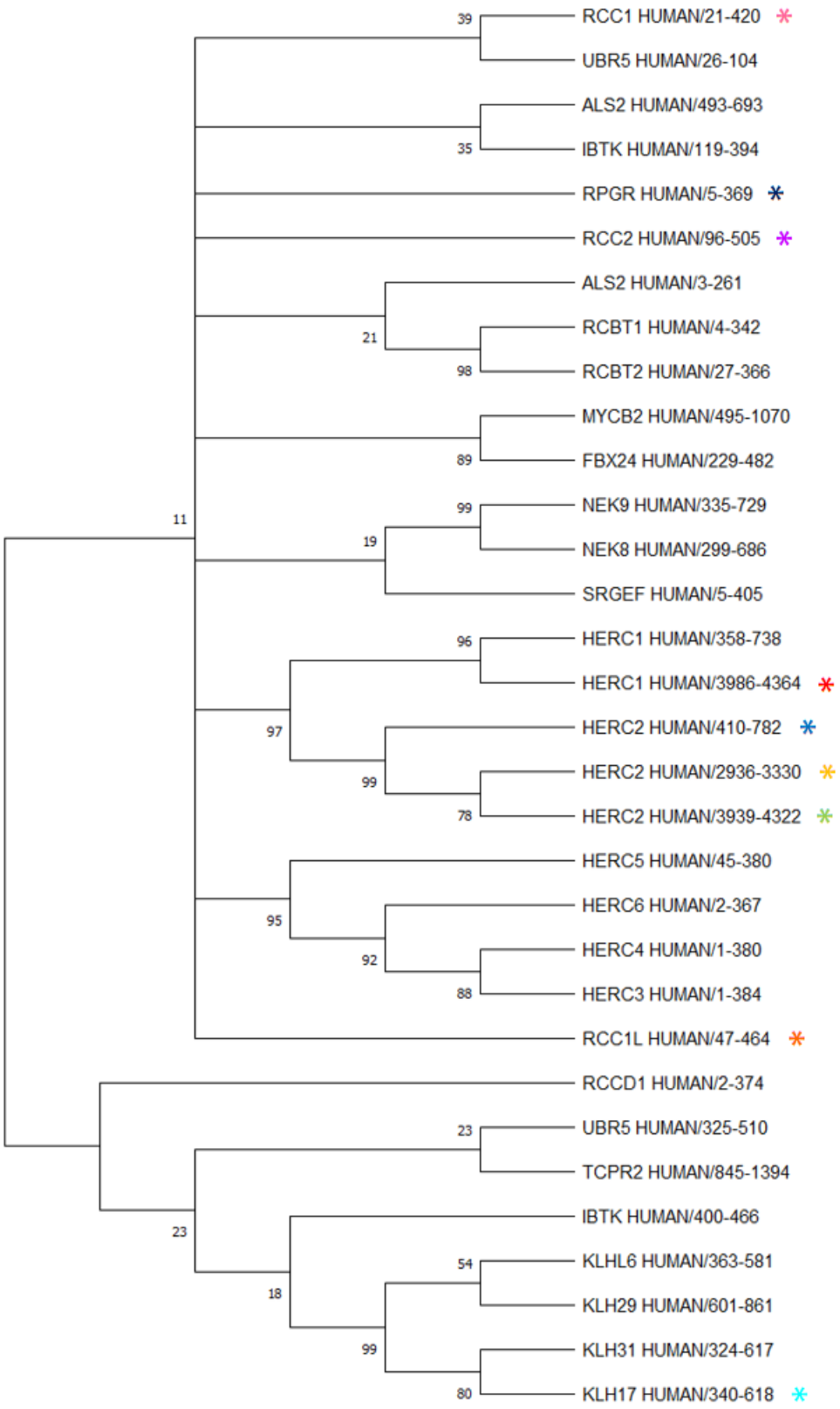

RCC1 HUMAN/21-420
UBR5 HUMAN/26-104
ALS2 HUMAN/493-693
IBTK HUMAN/119-394
RPGR HUMAN/5-369
RCC2 HUMAN/96-505
ALS2 HUMAN/3-261
RCBT1 HUMAN/4-342
RCBT2 HUMAN/27-366
MYCB2 HUMAN/495-1070
FBX24 HUMAN/229-482
NEK9 HUMAN/335-729
NEK8 HUMAN/299-686
SRGEF HUMAN/5-405
HERC1 HUMAN/358-738
HERC1 HUMAN/3986-4364
HERC2 HUMAN/410-782
HERC2 HUMAN/2936-3330
HERC2 HUMAN/3939-4322
HERC5 HUMAN/45-380
HERC6 HUMAN/2-367
HERC4 HUMAN/1-380
HERC3 HUMAN/1-384
RCC1L HUMAN/47-464
RCCD1 HUMAN/2-374
UBR5 HUMAN/325-510
TCPR2 HUMAN/845-1394
IBTK HUMAN/400-466
KLHL6 HUMAN/363-581
KLH29 HUMAN/601-861
KLH31 HUMAN/324-617
KLH17 HUMAN/340-618
39
35
21
98
89
99
19
11
96
97
99
78
95
92
88
23
23
54
18
99
80

Figure 123: Evolutionary analysis of the RLD domains within all human proteins. The evolutionary history was inferred by using the Maximum Likelihood method and JTT matrix-based model (Jones et al., 1992). The tree with the highest log likelihood (-25243.18) is shown. Initial tree(s) for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using the JTT model The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches (Felsenstein, 1985). Branches corresponding to partitions reproduced in less than 10% bootstrap replicates are collapsed. This analysis involved 32 amino acid sequences. There were a total of 860 positions in the final dataset. Evolutionary analyses were conducted in MEGA X (Kumar et al., 2018). RLD domains with a structure available in the PDB are shown by an asterisk.

The low confidence values of the earlier branches of the tree suggest that all of the RLD sequences show a reasonable degree of similarity, however, they do seem to split into two groups. Interestingly, the groups defined by the phylogeny analysis do not match up with the separate functional categories of RLD-containing proteins (Hadjebi *et al.*, 2008). In order to further investigate the similarity between the RLD/RCC structures, all RLD or RCC structures currently available in the PDB were examined in chimera (Fig. 124). The proteins with available structures are indicated by an asterisk next to the entry in figure 123.

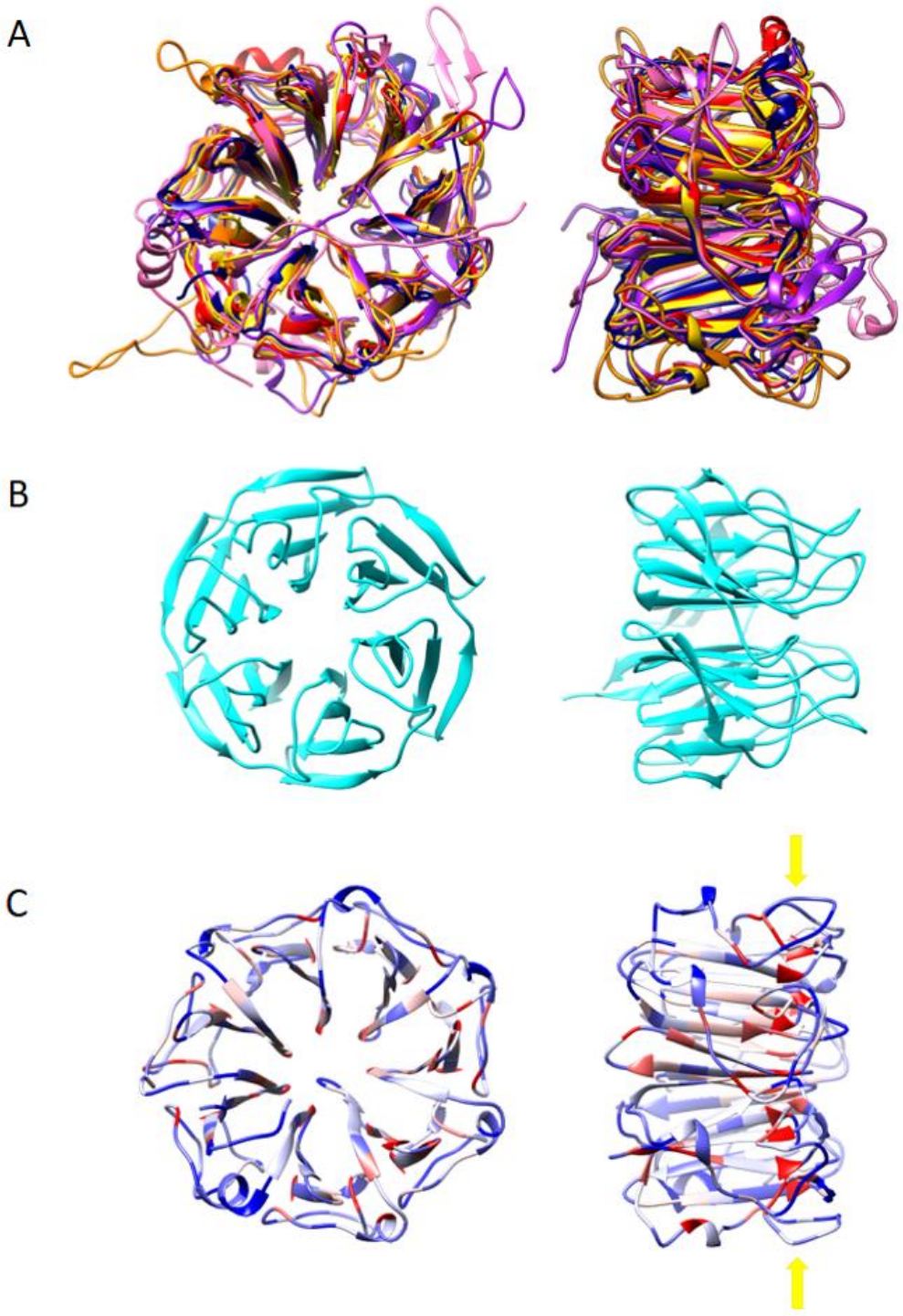

A
B
C

Figure 124: A structural comparison of all human RLD or RCC domains structures within the PDB. A) The structure for each domain, other than KHL17, was superposed in chimera. Only the RLD2 domain of HERC2 was included so as to not skew the conservation observations towards the HERC2 structures. Full-length RCC1 (5gwn) is shown in pink, RCC2 (1a12)) is shown in purple, RPGR (4jhn) is shown in navy, HERC1 RLD2 (4o2w) is shown in red, HERC2 RLD2 (not yet deposited)) is shown in yellow, RCC1L (5xgs) is shown in orange. B) The 6-bladed β-propeller structure of KHL17 (6hrl) is shown is cyan. C) The RLD2 domain of HERC2 is shown alone, with each residue coloured by the conservation of every residue shown in part a. The yellow arrows demonstrate the most conserved region of the structures.

The KLH17 structure is a clear outlier compared to the other RLD or RCC1-like structures, as the KLH17 structure forms a 6-bladed β-propeller structure (Fig. 124b) rather than the 7-bladed structure observed in every other example (Fig. 124a). The KLH17 structure also appears to include shorter β-sheets, with extended loop regions on one face making up half of the depth of the propeller structure (Fig. 124b). The KLH17 entry in the MSA does sit in a different clade to the rest of the RLD/RCC entries with available structures (Fig. 123), so it is possible that that separate clade represents a separate group within the RLD/RCC structural family, but without any structures for any other members of that group it is difficult to say for sure. When the remaining structures are superposed, (Fig. 124a), they show more structural variation than previous comparisons, as expected, but the seven-bladed β-propeller fold is still obviously retained across all structures. This observation matches the relationships between the sequences observed in the phylogenetic tree (Fig. 123).

Although the RLD domains are structurally very similar, they perform a diverse range of actions within their respective proteins (Hadjebi *et al.,* 2008). The RCC1 structure performs its GTPase activity through a β-wedge fold that intercalates with the seven-bladed propeller (Makde *et al.,* 2011). Given the range of functions for a reasonably well-conserved motif, it was theorised that the other RLD proteins would contain similar structural protrusions that would modify the domain enough to carry out the proposed activities. However, an overlay of the RCC1 structure and the RLD2 structure shows that this is not necessarily the case (Fig. 125).

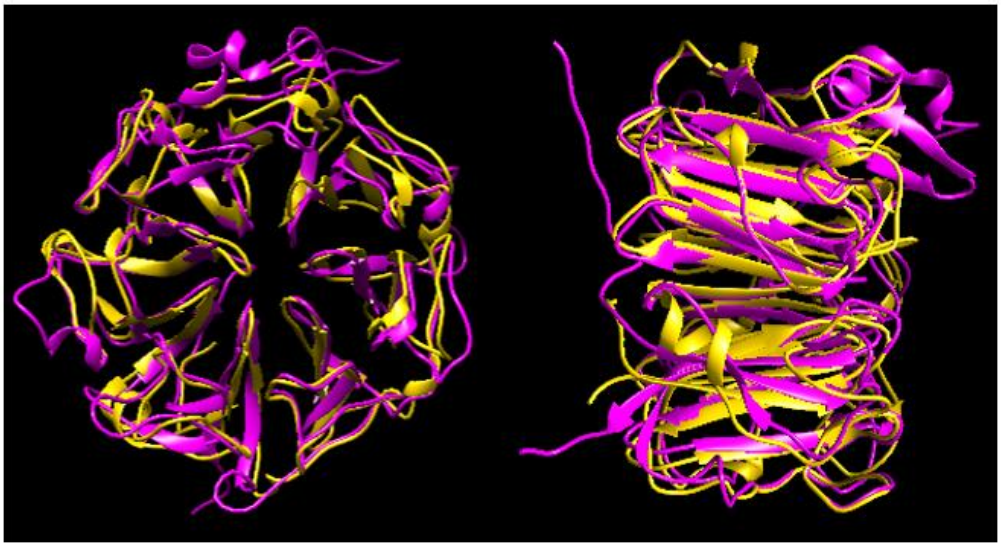


Figure 125: A ribbon diagram comparison of the RCC1 and HERC2 RLD2 structures. RCC1 is shown in pink while HERC2 RLD2 is shown is yellow. The β-wedge section of RCC1 is visible at the top of both views as a structured region in what would otherwise be expected to be an unstructured loop region between blades of the β-propeller structure.

The β-wedge section of RCC1 is clearly visible as a protrusion off the top of the β-propeller structure, while the same region in RLD2 is an unstructured loop, similar to the junction between every other blade region. There are no obvious alternative features elsewhere on the domain either to explain its differing activity. However, when you look at the comparison of all RLD structures (Fig. 124a), there are several additional elements in the loop regions of the domain that differ between proteins. Some easily identifiable examples are the extra α-helices in the outer loop region of adjacent blades in HERC1 RLD2 (red) and RPGR (navy), the extra loop density on the back face of the propeller in KLH17 (cyan), the elongated loop region of RCC1L (orange), and a structured region reminiscent of the RCC1 β-wedge on the edge of RCC2 (purple). Perhaps the core activity of the RLD seven-bladed propeller domain is a protein interaction interface, and the extra features in some of the RLD proteins confer additional activities to the protein (Hadjebi *et al.,* 2008), such as the guanine nucleotide exchange factor (GEF) activity of RCC1 or the adenyl cyclase inhibitory activity of the RPGR RLD (Hadjebi *et al.,* 2008). This is further supported by observations of β-propeller structures as a whole beyond the RLD subgroup. β-propellers are found in a wide range of proteins with diverse cellular activities, although they are mostly found within multi-domain proteins, and the interaction interfaces of the various β-propeller folds are found within the more variable loop regions between the blades of the propeller, or within extraneous structures nestled amongst the propeller structures (Pons *et al.,* 2003). The lack of extra structural elements in the loop regions of the HERC2 RLD2 structure may mean that the RLD simply acts as a binding platform for UBE3A, but other regions of the full-length HERC2 or

UBE3A structures may be involved in mediating the effects of the interaction. A previous study suggested that the isolated RLD2 domain was sufficient to influence UBE3A ubiquitin ligase activity *in vitro* (Kühnle *et al.,* 2011), which may suggest that the interaction with RLD2 causes a structural rearrangement within UBE3A to confer the change in catalytic activity, but CD analysis of the isolated RLD2-UBE3A interaction suggested that neither species undergoes a significant rearrangement in secondary structure upon binding (see section 4.4). One possibility here is that UBE3A does undergo a conformational change upon interaction with RLD2, but in such a way that doesn't affect the composition of secondary structure elements, but further work in this area is needed to clarify the mechanism and structural aspects of the interaction.

The conservation of each of the residues across the available RLD structures (Fig. 124b) is lowest around the loop regions, while the core region of the structure is more conserved The most highly conserved regions are the tip of the β-strands on one face of the β-propeller structure, which could just be a structural feature of the β-folds, but the increased conservation on one face over the other could also suggest a conserved function for this region. In order to see if this could be a protein interaction interface of sorts, all current structures of RLD/RCC domains in complex with binding partners were overlaid (Fig. 126).

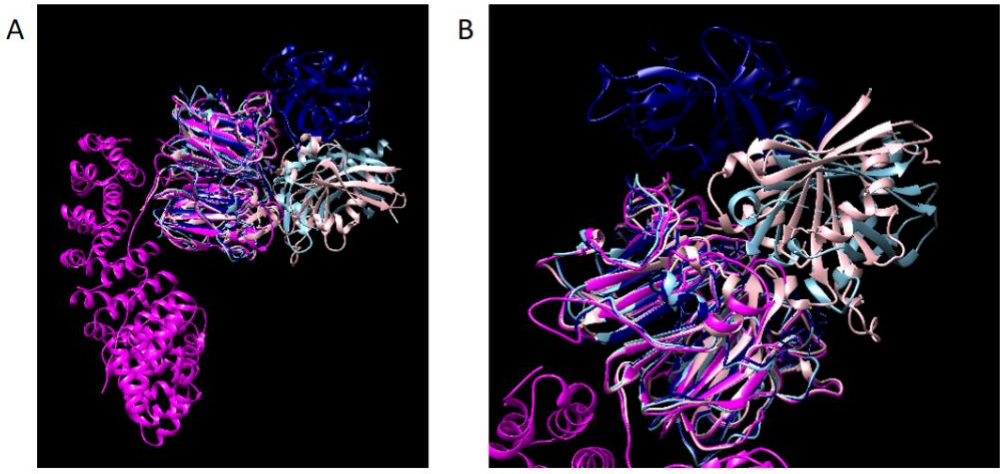


Figure 126: RLD structures in complex with binding partners. A) A full overview of all structures. B) A close-up view of the main interacting regions. RCC1 in complex with importin a3 (5tbk) is shown in magenta, RCC1 in complex with the RAN protein (1i2m) is shown in pale pink, RPGR RLD in complex with the RPGR-interacting domain of RPGRIP1 (4qam) is shown in navy, and RPGR in complex with PDE6D (4jhp) is shown in sky blue.

Of the four structures of RCC1/RLD complexes, two of them interact with the core RLD structure in a similar location, but there is no consensus site amongst all four. The importin a3 protein interacts with the opposite face of the structure compared to the other interactors, and the RPGRIP1 domain interacts on the same face as the other two, but it appears to interact more

with the distal edge of that face, rather than clustering in a definable area. The conserved residues at the tip of each β-strand are most likely to be involved in maintaining the β-propeller fold that is conserved across all family members (Hadjebi *et al.,* 2008). However, although these four structures do not suggest the presence of a single, definable protein binding site, the observation that 75% of the interactions occur on a single face of the structure, the same face that shows the increased conservation between terminal β-sheet regions, could suggest that this face is favoured for protein binding. The interaction between the RLD and its binding partners will be influenced by the position of the region within the full-length structure of the protein containing it. Since the RCC1 protein is comprised solely of the β-propeller structure with the addition of the β-wedge, interacting proteins can access the domain from either face. Although there is no full-length structure for the RGPR protein, the RGPR RLD region comprises less than 40% of the overall protein, so it is likely that one face of the β-propeller is precluded within the structure, resulting in the two observed interacting regions mapping to the same face of the RLD domain. As the RLD2 domain of HERC2 is also only a small part of the full protein, it is likely that at least one face will also be precluded. Without further information on the full-length structure of HERC2, or the specific residues of either protein involved in the interaction, identification of the interaction interface between HERC2 RLD2 and UBE3A is not possible

7.4.3 RLD2 Complex Crystals

Following the success of the RLD2 crystallisation attempt, further efforts were made to attempt to solve UBE3A+RLD2 complexes using x-ray crystallography. The full-length UBE3A protein has proven resistant to crystallisation since the HECT domain structure was solved in 1999, but it could be possible that the interaction with RLD2 could stabilise UBE3A to allow elucidation of the complex. In order to explore this possibility, a UBE3A+RLD2 sample was obtained as described in chapter 3 and subjected to a pre-crystallisation trial (PCT) (section 2.11.1). However, the CD data for the UBE3A+RLD2 interaction suggested that there was no significant re-arrangement of UBE3A upon binding to RLD2, so the likelihood of the full-length enzyme becoming stabilised enough for crystallisation was low. The ITC data (section 4.2.3) and co-purification results (section 4.3.3) for the interaction between RLD2 and the Ufrag region of UBE3A suggested that it may be possible to recapitulate at least a large part of the interaction between UBE3A and RLD2 using the Ufrag construct in the place of full-length UBE3A, so pre-crystallisation trials were also set up for a cleaved Ufrag+RLD2 sample, and an uncleaved complex with the MBP tag remaining.

Although in the case of RLD2 crystals had already begun to form in the PCT plate, generally the PCT is not able to determine whether a sample will definitely be able to form crystals. The purpose of the PCT is to determine a

concentration range at which the sample is most likely to crystallise rather than form amorphous precipitate. For the UBE3A+RLD2 sample this was demonstrated to be 5 mg/ml, and for the Ufrag+RLD2 samples 6 mg/ml was used. The samples at these concentrations were subjected to the same crystal screens as the RLD2 sample (section 2.11.2). The crystal screens for the tagged Ufrag+RLD2 complex, the cleaved Ufrag+RLD2 complex, and the full-length UBE3A+RLD2 complex resulted in a few conditions that produced potential crystals, but they were all too thin or irregularly clustered to collect data from. The Ufrag+RLD2 constructs created very dirty looking crystal clusters that did not look ideal, but they were looped and shot on beam I24 at Diamond to show that they were protein crystals. The diffraction spots were minimal and only reached the 15-20 Å level, but it did confirm that protein was present in the crystals. This suggested that the crystals could potentially be optimised through more buffer screens, or more careful tailoring of the constructs. The UBE3A+RLD2 potential crystals were not shot as they were too small, but optimised crystal screens were configured nonetheless.

Optimised crystal screens were set up, using the conditions from the original crystal screen drops that produced the potential crystals as a base to vary the salt concentrations, PEG percentages, and pH as required (section 2.11.2). Unfortunately, the optimised screens were not able to produce any higher quality crystals than the original conditions within the time frame of the project, so these samples were not taken any further. Although the original Ufrag+RLD2 crystals produced diffraction spots, suggesting that they were in fact protein crystals, further optimisation of the buffer conditions may not be the most promising approach to solve these structures. The presence of protein in the low quality crystals does suggest that the complex could be amenable to crystallisation, it would be beneficial to put more effort into altering the construct boundaries in order to define a minimal unit for Ufrag that is still able to interact meaningfully with RLD2, while reducing the number of extraneous residues that may be preventing successful crystallisation of the product. As the full-length UBE3A+RLD2 complex has proved difficult to solve with cryo-EM and unamenable to crystallisation, a combination of x-ray crystallography of the minimal Ufrag+RLD2 construct alongside a lower resolution structure of the full-length complex through cryo-EM may be the most promising approach to determining the mechanisms of interaction of the two proteins.

8 Discussion

8.1 Oligomeric States of UBE3A

UBE3A has been identified in several clinical contexts, particularly Angelman Syndrome (Kishino *et al.*, 1997) and HPV-associated cervical cancer (Scheffner *et al.,* 1993; Beaudenon and Huibregtse, 2008). It has also been implicated in several neurodevelopmental disorders (DiStefano *et al.,* 2016; Noor *et al.,* 2015; Vatsa and Jana, 2018), schizoaffective disorders (Salminen *et al.,* 2019), cardiopathologies (Cheng *et al.,* 2019), the immune response to HIV infection (Pyeon *et al.,* 2019), Alzheimer's (Olabarria *et al.,* 2019), and several types of cancer (Bandilovska *et al.,* 2019). Despite this, the physiological roles of UBE3A remain enigmatic and there is still no full-length structure available for UBE3A, often seen as integral to understanding disease processes at the molecular level.

The first piece of structural information for UBE3A was a crystal structure of the isolated HECT domain in complex with the UbcH7 E2 enzyme (Huang *et al.,* 1999) (Fig. 12). However, this structure showed a 40 Å gap between the active sites of the two enzymes, which raised questions as to how the activated ubiquitin moiety could be transferred to the active site of UBE3A. A potential answer to this came over a decade later when Ronchi *et al.,* suggested that UBE3A harbours two different E2-binding sites in its HECT domain, and that the protein functions as a trimer (Ronchi *et al.,* 2013; Ronchi *et al.,* 2014; Ronchi *et al.,* 2017). These observations came from kinetics experiments and molecular docking simulations, although no structural studies were performed to confirm this. However, the trimer idea is relatively controversial within the field of UBE3A research, as the identification of a trimer is based on the interpretation of a single size exclusion trace that has not been replicated in any subsequent work. In this project I attempted to observe the oligomeric state of UBE3A using a range of techniques to attempt to determine conclusively whether UBE3A acts primarily in a monomeric or multimeric form.

The first observation came in the form of a size exclusion trace during purification of the protein (see section 3.4.1). This suggested that the majority of the sample eluted as a monomer. However, the presence of lower abundance higher molecular weight species was consistently seen in all of the size exclusion traces generated throughout the project. The resolution range of the superdex 200 columns used to purify UBE3A are great for isolating the monomeric UBE3A species at ~98 kDa, but they are not as good at separating potential dimeric or trimeric forms, at ~200 kDa and 300 kDa respectively. This meant that a SEC profile alone was insufficient to categorise the higher molecular weight species as dimers or trimers of UBE3A. Following this, UBE3A samples were subjected to SEC-MALS (multi-angle light scattering) experiments in an attempt to derive a more accurate molecular weight for the higher molecular weight species, but the UBE3A sample was consistently

unresponsive to the technique and produced no discernible trace regardless of protein concentrations or buffer compositions used (data not shown).

I next used SV-AUC to further investigate the multimeric states within the UBE3A sample (see section 4.1.1). This resulted in a similar profile to that from SEC, with the majority of the sample forming a monomer peak, and higher molecular weight species appearing at much lower abundance. An interesting point about the UBE3A AUC trace is that the distribution of oligomeric states is not affected by the change in concentration (Fig. 58). This is typically indicative that the species are not in a dynamic equilibrium, however, isolating either peak after SEC and remeasuring with each of these techniques results in the same distribution of species. This implies that they are in a dynamic equilibrium, however as they separate in all the techniques used, the re-equilibration is slow on the timescale of transport. The major peak corresponds well to that expected for a monomer, however the minor peak is harder to characterise. It may be that this is the trimer previously observed, however molecular weight estimates are closer to a dimer. It is possible that this is a weak aggregate species of indefinite size and conformation, and thus hard to characterise.

It might be that UBE3A multimerisation is not spontaneous in cells but rather is triggered by cellular signals, such as post-translational modifications (PTMs) or accessory protein binding. Various regions of UBE3A are predicted to be flexible in the isolated monomer sequence (see appendix 3), so it is likely that in a solution UBE3A exists in many different conformations simultaneously at any time. If a small portion of the UBE3A molecules in the sample adopts the conformation required to reveal the interaction interface, and then a subset of these molecules are able to come into contact, then a small subset of the total sample will form a multimeric state. However, unless the multimerisation conformation is a particularly energetically favourable state for the monomer, the majority of the sample will not form spontaneous multimers, resulting in the profile suggested by SEC and SV-AUC. As this interaction would depend on the spontaneous adoption of a specific, energetically unfavourable conformation of UBE3A, as well as a physical interaction between different molecules with the correct conformation and orientation, an increase in concentration would not necessarily result in a linear increase in the number of productive interactions.

Further evidence of the primarily monomeric state of UBE3A also comes from EM images, as in both negative stain (Fig. 91, section 6.1) and cryo-EM methods the particles appeared to be fairly uniform in their monomeric distribution. It is possible that multimeric states were present in the sample and misinterpreted as monomers due to the elongated nature of the UBE3A particles, but dimeric or trimeric states of UBE3A were not observed at the 2D classes stage of any dataset. Further to this, complexes involving UBE3A and other species seemed to result in complexes with a 1:1 stoichiometry (see

section 4.2) and a molecular weight and S value suggestive of hetero-dimers rather than a dimer of multimers (see sections 4.3 and 4.1.3). However, as these experiments were conducted *in vitro* with a bacterially expressed protein rather than in a native cell environment, it cannot be ruled out that UBE3A may form a trimeric or otherwise multimeric state upon induction by a post-translational modification event triggered by cellular pressures. As the ubiquitination E1-E2-E3 cascade allows for a large degree of redundancy within signalling processes (Pickart *et al*., 2001), it would make sense for further regulation of ubiquitin signalling through post-translational modifications. Sequence analysis of UBE3A with the PhosphoSite Plus server (Hornbeck *et al*., 2011) reveals many potential PTM sites, including phosphorylation, acetylation, and ubiquitination acceptor sites (see Appendix 7). A phosphorylation site at T485 of UBE3A isoform 1 has been identified in the regulation of UBE3A catalytic activity (Yi *et al.,* 2015), and regulation of ubiquitination through phosphorylation of E3 enzymes has also been identified in other contexts (Gao and Karin, 2005), so it is possible that phosphorylation at another site within UBE3A enables its oligomerisation. Ronchi *et al.* (2014) suggest that the HECT domain within UBE3A is able to trimerise in the absence of a helix region immediately upstream of the catalytic domain (residues 474 – 490 of isoform 1), and PhosphoSite Plus identifies three possible phosphorylation sites in this region (T508, Y511, and S512 using isoform 2 numbering, which correlates to residues T485, Y488, and S489 in isoform 1), so it is possible that phosphorylation of one of these sites allows for a rearrangement of the helix and subsequent trimer formation. One of these is the T485 site that has already been shown to regulate UBE3A activity (Yi *et al.,* 2015), although UBE3A was shown to be more active in the unphosphorylated state. As my UBE3A sample was expressed in *E. coli* rather than eukaryotic cells it is more likely that the *in vitro* purified sample represents the unphosphorylated form, so it is unlikely that phosphorylation of T485 would lead to formation of a more catalytically active trimer. The α-helix in question resides within a defined E6-binding region (Drews *et al.,* 2020), and binding of E6 is suggested to increase UBE3A activity through increasing its oligomerisation (Ronchi *et al.,* 2014), so it is possible that phosphorylation of sites Y488 or S489 of UBE3A isoform 1 may recruit other cellular binding partners to induce oligomerisation.

However, another possible explanation for the discrepancy between my results and those of Ronchi *et al.,* (2014) is the difference in UBE3A isoforms. Throughout this work I used the shorter isoform 1 form, whereas the Ronchi *et al.* work uses the slightly longer isoform 2 form of the protein. The difference is only in the N-terminal region of the protein, where isoform 2 contains an extra 23 amino acids. Due to the spatial separation of the N-terminal residues and the HECT domain it is unlikely that the extra residues alone are able to significantly alter the oligomeric state as co-ordinated by

interactions between HECT domains, but as I have not explored any differences between the different isoforms I cannot rule it out.

8.2 Substrate Binding Interfaces in UBE3A

Another unexplained aspect of UBE3A's physiological significance is its mechanism of substrate recognition. UBE3A has been shown to regulate a wide range of targets in cells (Martínez-Noël *et al*., 2018), and alteration of its substrate specificity through binding a viral protein is an oncogenic mechanism of multiple types of cancer (Beaudenon and Huibregtse, 2008; Munakata *et al*., 2005). Despite this, a single substrate binding region of UBE3A has not been identified in existing literature. Potential interaction regions have been identified within the UBE3A sequence, for example Kühnle *et al.* (2011) used varying truncation mutants to determine a core region of 50 amino acids spanning residues 150 – 200 of UBE3A isoform 1 involved in the interaction with HERC2's RLD2 domain. Another study shows a crystal structure of a UBE3A + PSMD4 interaction involving the AZUL domain of UBE3A (Buel *et al.,* 2020), and a third study identified the regions involved in HPV E6 protein binding (Drews *et al.,* 2020). PSMD4 and RLD2 have both been identified as potential substrates of UBE3A (Lee *et al*., 2013; Wang *et al*., 2017; Kühnle *et al*., 2011), but they have also been implicated in UBE3A's catalytic activity beyond that of a typical substrate. RLD2 has been proposed to increase the catalytic activity of UBE3A (Kühnle *et al.,* 2011), and PSMD4 has been suggested to bridge between UBE3A and the proteasome to target ubiquitinated proteins for further ubiquitination to ensure their degradation (Buel *et al.,* 2020). The E6 protein has also been shown to be ubiquitinated by UBE3A (Li *et al*., 2019), but its core role is to alter the substrate specificity of UBE3A to target several tumour suppressor proteins for degradation upon HPV infection (Ebner *et al*., 2020).

All of these proteins have been identified as potential substrates, but their interaction regions are not very close in the protein sequence (Fig. 19, section 1.5) so it is unknown which, if any, of these sites represent the canonical substrate binding domain of UBE3A. Fitting of the AlphaFold and Robetta predicted models of UBE3A into the low resolution cryo-EM map of UBE3A determined in this project enabled identification of the location of these protein interaction regions in relation to the catalytic HECT domain (Fig. 104, section 7.1.3). Both the AZUL and RLD2-interacting domains were located in areas with the lowest confidence values in the predicted models, and the region that appears to correlate with them in the cryo-EM map is particularly poorly resolved. This suggests that these regions are more flexible than other areas of the protein. This makes sense when the full UBE3A-only model is examined, as both the AZUL and RLD2-interacting regions appear to sit at the most distal end of the elongated protein structure relative to the HECT domain. In order for substrates to bind at this distal point and become ubiquitinated by the HECT domain, the interacting regions must sit on a

flexible linker that allows relocation upon substrate binding to allow the target protein to come into proximity with the HECT domain catalytic site (Fig. 127).

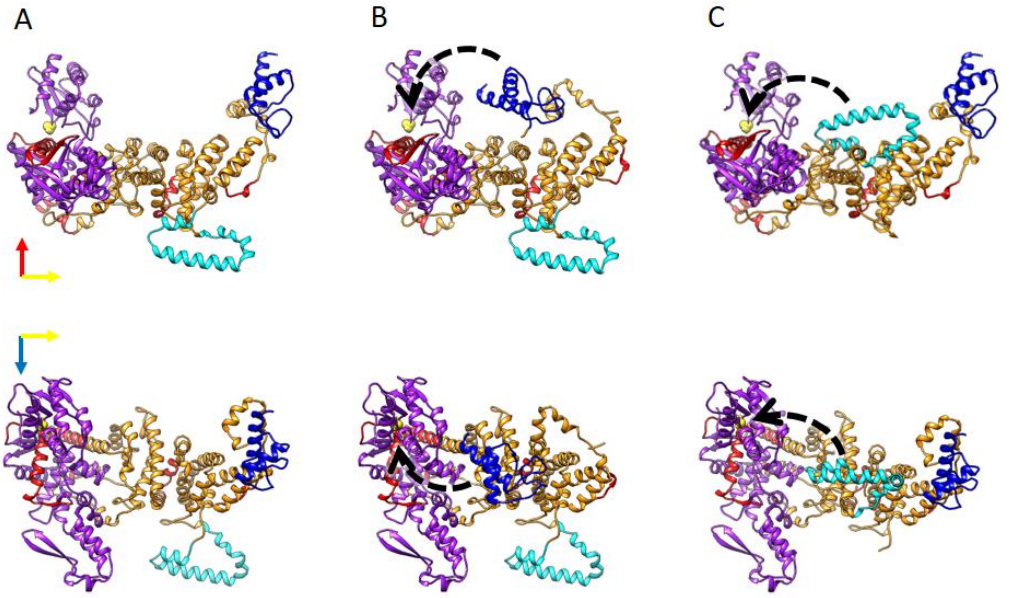


Figure 127: Rearrangements of the Robetta predicted model of UBE3A show how movements around the flexible linker regions could allow bridging between the potential substrate binding regions and the HECT domain. A) The Robetta prediction for UBE3A B) Rearrangements of the AZUL domain and the associated linker region would bring a bound substrate to a close proximity to the catalytic site of the HECT domain. C) Rearrangements of the HERC2-interacting region would allow a bound protein to interact with the HECT domain.

Although the UBE3A+RLD2 reconstruction developed in this project is low resolution and subject to a lot of noise (Fig. 114, section 7.3.3), it does suggest that the RLD2 binding region of UBE3A, as identified by Kühnle *et al*., (2011) may situate itself above the central region of UBE3A, bridging the distal point and the HECT domain (Fig. 115). This relocation of the RLD2 binding region is suggested both by a region of density suggesting a connection between the RLD2 domain and the UBE3A protein outside of either the HECT domain or the distal point, and also a truncation along the elongated dimension of the UBE3A density region. In the UBE3A-only model the density extended beyond the area of the UBE3A models with high confidence values, and it was suggested that the apparently flexible AZUL and RLD2-binding domains would orient themselves on their flexible linkers to fill this density (section 7.1.3, Fig. 104). In the UBE3A+RLD2 structure however, the cryo-EM density map extends no further than the stable alpha-helical region of the UBE3A core domain, and the AZUL domain and RLD2-binding domain must be otherwise orientated in order to fit the density (Fig. 116). The CD data for the isolated UBE3A+RLD2 interaction does suggest that no significant structural rearrangements take place upon binding (see section 4.4), but it may be that

the unstructured flexible linker remains an unstructured flexible linker in both states, and the placement of the re-orientated RLD2-binding domain may be coordinated by interactions between existing secondary structure elements.

An involvement of the core region of UBE3A in orientating flexible regions for substrate binding is also suggested by the identification of several regions involved in E6 binding that are interspersed throughout the UBE3A sequence (Drews *et al.,* 2020). When the E6-interacting regions are plotted onto the Robetta predicted model and fitted into the cryo-EM map of full-length UBE3A, they appear to form multiple discrete regions dispersed along the elongated dimension of UBE3A rather than clustering into a single location near to the HECT domain (Fig. 104, section 7.1.3). Although the E6-interacting regions appear to be distributed across the opposite face to that involved in RLD2 binding, their locations within that core region of UBE3A, despite the relatively compact shape of an E6 dimer, suggests that areas of the core region of UBE3A may be involved in a structural rearrangement, either in order to allow E6 binding or in response to interacting with an accessory protein (either E6 or RLD2).

The structures generated during this project, for both full-length UBE3A and the UBE3A complexes, were not able to reach a high resolution to enable accurate determination of interaction interfaces and the molecular mechanisms involved. However, the low-resolution structures presented can be interpreted to reveal potential new interaction mechanisms that could be explored with more targeted biophysical and kinetic experiments while further work is put into generation of a high-resolution complex.

8.3 Future Work

Although the structures produced during this project were not able to reach a high resolution, the work carried out across all areas of this project has provided a fresh insight into several areas of UBE3A protein-protein interactions, and it has identified several new avenues of future work in this area. One such avenue is the UBE3A-HERC2 interaction. The isolation of the RLD2 complex from HERC2 for structural studies of the UBE3A+RLD2 interaction is a promising idea for further studies; HERC2 is officially classified as a giant ubiquitin ligase (García-Cano *et al*., 2019) with a molecular weight of 527 kDa, which makes exogenous expression of the full-length enzyme exceedingly difficult. The size and complexity of the enzyme makes expression in anything but a mammalian cell expression system unlikely, but even in mammalian cells the size would exert significant pressure on the cellular protein expression machinery. With no examples of exogenous expression of a full-length HERC2 protein for *in vitro* purification reported in the literature to date, a significant portion of this project was dedicated to attempting to express and purify GFP-tagged HERC2 protein from HEK293 cells (see section 3.6). Although I was ultimately unable to present a clean and conclusive purified HERC2 sample, the initial observations of over-expressed very large

molecular weight species bearing GFP-tags suggests that I have been successful in expressing some level of GFP-tagged HERC2. Although it was not possible within the time frame of this project, optimisation of the expression system and subsequent purification strategies may result in significant increase in yields with potentially limited effort.

Further to the expression and purification of an *in vitro* HERC2 product, characterisation of the full-length UBE3A+HERC2 interaction will be much more informative than that of the isolated RLD2 domain. The RLD2 region encompasses only a small part of the full HERC2 protein (Kühnle *et al*., 2011), and without any structural information for the whole protein, the orientation of the RLD2 domain within it is unclear. Occlusion of several faces of the RLD2 structure may preclude some of the potential binding interfaces allowed by a free-RLD2 construct macromolecule. The use of the isolated RLD2 domain was thought to improve the possibility of generating structural information for the interaction, as the RLD2 domain is much easier to express and purify in abundance (see section 3.2.4 and 3.4.4; Kühnle *et al.,* 2011). However, despite recent advances in detector and energy filter technology available for cryo-EM (see section 7.1.2), the small size of the UBE3A+RLD2 complex combined with the flexibility of UBE3A and the asymmetrical nature of the molecule, still renders the UBE3A+RLD2 a difficult target for cryo-EM. If the HERC2 protein could be expressed and purified in order to generate an isolated UBE3A+HERC2 complex *in vitro*, the UBE3A+HERC2 complex would be an excellent target for cryo-EM due to the size of the HERC2 particles. The UBE3A+HERC2 complex is also thought to interact with further cellular proteins, NEURL4 and MAPK6, to form an even larger complex with a potentially more stabilised structure (Martínez-Noël *et al*., 2018), so even if the resolution of a UBE3A+HERC2 structure is limited by any potential flexibility of the complex, strategies for stabilisation of the complex are already clear.

However, in the absence of an exogenously overexpressed HERC2 product, the UBE3A+RLD2 complex may still represent a potential target for X-ray crystallography. Crystal trays were prepared for a UBE3A+RLD2 complex, and crystals were produced that could demonstrate diffraction spots to 20 Å. Optimisation screens were configured using variations around the original crystal condition, but no crystals of any quality formed within the remaining time frame of this project. Further optimisation of conditions may help, but a better approach would be to design a new UBE3A construct using the AlphaFold model as a guide to truncate the most flexible regions of the protein. Stabilisation of the sample used for crosslinking would also increase the crystallisation capability of the complex with potentially minimal effort. A crosslinked UBEA+RLD2 complex was generated for use in cryo-EM experiments, but this crosslinked sample was not subjected to crystallisation

screens. This is an area of work that may produce exciting results with very little time commitment required.

Another aspect of this project that I was unable to explore to the extent I intended is the work involving the UBE3A+E6+p53 complex. A structure already exists for a complex of UBE3A+E6+p53 derived through x-ray crystallography that allows some insight into the interaction (Martinez-Zapien *et al*., 2016), but this complex involves only a single domain of p53 and only a 12 amino acid fragment of UBE3A. In order to gain a full understanding of how E6 causes UBE3A to target p53, a full-length structure of the complex would be required. The existing structure was solved using the truncations described due to the observed flexibility of UBE3A preventing crystallisation of a full protein construct, but cryo-EM does not have the same rigidity constraints. Masuda *et al.,* (2019) were able to demonstrate a high yield co-expression and purification of full-length UBE3A, E6 and p53 proteins, and the complex would be a promising size for a cryo-EM target. As p53 is known to tetramerise and E6 has been shown to dimerise, it is unclear what the stoichiometry of a UBE3A+E6+p53 complex would be, but even a heterotrimer with a 1:1:1 stoichiometry would bring the size up to ~ 200 kDa, which is within the range of samples that are regularly solved to high resolutions. The flexibility of UBE3A and the corresponding complex may still be an issue for cryo-EM reconstructions, but the continuing developments in terms of data collection technology and data processing techniques could help to overcome this. Newer detectors are continuously being developed with smaller pixel sizes and faster acquisition speeds to enable higher resolution of smaller samples with less averaging required, better motion correction strategies, and larger datasets in a shorter data collection time (ThermoFisher.com(a)). Post-column filters, such as the GIF (Gubbens *et al.,* 2010) and the newer Selectris-X filters also improve the SNR of micrographs by blocking inelastically scattered electrons (ThermoFisher.com(b)), and fringe-free imaging allows more acquisitions to be taken per foilhole by limiting the area of radiation damage around each collection spot (ThermoFisher.com(c)). Newer data processing strategies further supplement these new hardware improvements by increasing the efficiency of algorithms (Punjani *et al.,* 2017) and simplifying the processes to allow easier processing by non-expert users (Li *et al*., 2020). My issues with this work were due to unforeseen issues with my existing cell stocks and subsequent COVID19-related delays in obtaining new resources. If these issues could be overcome, this is a promising area of future work with a potential for significant clinical impact.

A further aspect of the work that could be developed to produce the desired results is the *in vitro* ubiquitination assays. One improvement would be to measure the concentration of the PSMD4 sample more accurately before inclusion in the assay mixture to ensure that its presence in each timepoint sample can be clearly observed. PSMD4 is a difficult protein to quantify as it

contains no aromatic residues, but the addition of a strep tag to introduce a tryptophan residue, or potentially a labelled antibody, may allow a more accurate concentration determination. More precise quantification of the sample subjected to the *in vitro* ubiquitination assay would be instrumental in allowing observation of its decay. I also attempted to perform a Western blot with an anti-Ub primary antibody to allow detection of ubiquitinated species, rather than relying on measuring the decay of potential substrates as a proxy for UBE3A activity. My attempts at this were unsuccessful, but a thorough optimisation process to improve this would allow for much more informative observations of UBE3A activity in a range of conditions.

Another area that could benefit from further exploration is the role of post-translational modifications in UBE3A activity. Yi *et al.,* (2015) show that phosphorylation of the T485 residue inhibits the catalytic activity of the enzyme, while a non-phosphorylatable mutation of this site causes constitutive UBE3A activity *in vivo*. While attempts were made during this project to generate phosphomimetic and phospho-null mutants of this residue, the cloning process proved to be much more difficult than expected. Despite extensive attempts at troubleshooting using various cloning methods, the T485 mutants remained elusive, so *de novo* synthesis of the desired products would provide the most viable solution. With the synthesised constructs, the effects of phosphorylation, or the phospho-mimetic mutation, of this residue on the structure of UBE3A could be measured using various biophysical techniques, including CD and SAXS to observe whether a significant structural re-arrangement has occurred, or ITC with known interacting proteins to see if alteration of this residue affects its protein-protein interactions. This could be carried out alongside structural studies, particularly using cryo-EM, to attempt to visualise the mutation in high-resolution, and also *in vitro* ubiquitination assays to determine whether the effects of the phosphorylation on UBE3A activity could be recovered with the introduction of RLD2 or E6 proteins. *In cellulo* studies would also provide a valuable insight into the role of post translational modifications in UBE3A activity, although that was far beyond the scope of this project.

Similarly, many single residue mutations have been identified within the UBE3A sequence that represent pathogenic or possible pathogenic states of the enzyme (Sadikovic *et al*., 2014). Further exploration of the physical impact of each of these mutations relative to the native enzyme could potentially enable elucidation of mechanisms to rescue the effects of mutated UBE3A in Angelman Syndrome patients, as well as enabling a better understanding of the mechanism of UBE3A activity as a whole. Although the ultimate downstream effects of some point mutations are well known, such as a mutation in the active site cysteine residue (C820A) that prevents any interaction between UBE3A and ubiquitin, and the phospho-null mutation of the T485 residue (T485A) that prevents inhibition of UBE3A activity, many of

these mutations remain completely uncharacterised. Generation of many of these mutant constructs, either through targeted PCR mutagenesis of the wild-type sequence or *de novo* synthesis of each construct, would enable each to be subjected to the range of biophysical techniques and structural analysis described in this project. This would enable observation of the effects of different mutations on UBE3A stability, oligomeric state distribution, binding affinities for partner proteins, and if any of these effects enable sufficient stabilisation or oligomerisation for x-ray crystallography or cryo-EM, a high-resolution structure may be possible.

9 References

Akutsu, M., Dikic, I. and Bremm, A., 2016. Ubiquitin chain diversity at a glance. *Journal of Cell Science*, 129(5), pp.875-880.

Andrade, L., D'Oliveira, A., Melo, R., De Souza, E., Silva, C. and Parana, R., 2009. Association between hepatitis C and hepatocellular carcinoma. *Journal of Global Infectious Diseases*, 1(1), p.33.

Avagliano Trezza, R., Sonzogni, M., Bossuyt, S., Zampeta, F., Punt, A., van den Berg, M., Rotaru, D., Koene, L., Munshi, S., Stedehouder, J., Kros, J., Williams, M., Heussler, H., de Vrij, F., Mientjes, E., van Woerden, G., Kushner, S., Distel, B. and Elgersma, Y., 2019. Loss of nuclear UBE3A causes electrophysiological and behavioral deficits in mice and is associated with Angelman syndrome. *Nature Neuroscience*, 22(8), pp.1235-1247.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G., Wang, J., Cong, Q., Kinch, L., Schaeffer, R., Millán, C., Park, H., Adams, C., Glassman, C., DeGiovanni, A., Pereira, J., Rodrigues, A., van Dijk, A., Ebrecht, A., Opperman, D., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M., Dalwadi, U., Yip, C., Burke, J., Garcia, K., Grishin, N., Adams, P., Read, R. and Baker, D., 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557), pp.871-876.

Band, V., De Caprio, J., Delmolino, L., Kulesa, V. and Sager, R., 1991. Loss of p53 protein in human papillomavirus type 16 E6-immortalized human mammary epithelial cells. *Journal of Virology*, 65(12), pp.6671-6676.

Bandilovska, I., Keam, S., Gamell, C., Machicado, C., Haupt, S. and Haupt, Y., 2019. E6AP goes viral: the role of E6AP in viral- and non-viral-related cancers. *Carcinogenesis*, 40(6), pp.707-714.

Bassett, A., 2011. Parental Origin, DNA Structure, and the Schizophrenia Spectrum. *American Journal of Psychiatry*, 168(4), pp.350-353.

Beaudenon, S. and Huibregtse, J., 2008. HPV E6, E6AP and cervical cancer. *BMC Biochemistry*, 9(Suppl 1), p.S4.

Bepler, T., Kelley, K., Noble, A. and Berger, B., 2020. Topaz-Denoise: general deep denoising models for cryoEM and cryoET. *Nature Communications*, 11(1).

Bepler, T., Morin, A., Rapp, M., Brasch, J., Shapiro, L., Noble, A. and Berger, B., 2019. Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nature Methods*, 16(11), pp.1153-1160.

Berman, T. and Schiller, J., 2017. Human papillomavirus in cervical cancer and oropharyngeal cancer: One cause, two diseases. *Cancer*, 123(12), pp.2219-2229.

Bi, X., Sun, J. and Baudry, M., 2015. Yin-and-Yang of mTORC1/C2 in Angelman syndrome mice. *Oncotarget*, 6(16), pp.13844-13845.

Birch, S., Kench, J., Takano, E., Chan, P., Chan, A., Chiam, K., Veillard, A., Stricker, P., Haupt, S., Haupt, Y., Horvath, L. and Fox, S., 2014. Expression of E6AP and PML predicts for prostate cancer progression and cancer-specific death. *Annals of Oncology*, 25(12), pp.2392-2397.

Bonifacino, J., 2002. Electrophoresis and Immunoblotting. *Current Protocols in Cell Biology*, 15(1).

Bouvard, V., Baan, R., Straif, K., Grosse, Y., Secretan, B., Ghissassi, F., Benbrahim-Tallaa, L., Guha, N., Freeman, C., Galichet, L. and Cogliano, V., 2009. A review of human carcinogens—Part B: biological agents. *The Lancet Oncology*, 10(4), pp.321-322.

Brown, P. and Schuck, P., 2006. Macromolecular Size-and-Shape Distributions by Sedimentation Velocity Analytical Ultracentrifugation. *Biophysical Journal*, 90(12), pp.4651-4661.

Buel, G., Chen, X., Chari, R., O'Neill, M., Ebelle, D., Jenkins, C., Sridharan, V., Tarasov, S., Tarasova, N., Andresson, T. and Walters, K., 2020. Structure of E3 ligase E6AP with a proteasome-binding site provided by substrate receptor hRpn10. *Nature Communications*, 11(1).

Buiting, K., Williams, C. and Horsthemke, B., 2016. Angelman syndrome — insights into a rare neurogenetic disorder. *Nature Reviews Neurology*, 12(10), pp.584-593.

Bundle, D. and Sigurskjold, B., 2022. Determination of accurate thermodynamics of binding by titration microcalorimetry. *Methods Enzymol*, 247, pp.288-305.

Burette, A., Judson, M., Li, A., Chang, E., Seeley, W., Philpot, B. and Weinberg, R., 2018. Subcellular organization of UBE3A in human cerebral cortex. *Molecular Autism*, 9(1).

Bzhalava, D., Eklund, C. and Dillner, J., 2015. International standardization and classification of human papillomavirus types. *Virology*, 476, pp.341-344.

Chaturvedi, A., Engels, E., Pfeiffer, R., Hernandez, B., Xiao, W., Kim, E., Jiang, B., Goodman, M., Sibug-Saber, M., Cozen, W., Liu, L., Lynch, C., Wentzensen, N., Jordan, R., Altekruse, S., Anderson, W., Rosenberg, P. and Gillison, M., 2011. Human Papillomavirus and Rising Oropharyngeal

Cancer Incidence in the United States. *Journal of Clinical Oncology*, 29(32), pp.4294-4301.

Chaturvedi, S., Ma, J., Brown, P., Zhao, H. and Schuck, P., 2018. Measuring macromolecular size distributions and interactions at high concentrations by sedimentation velocity. *Nature Communications*, 9(1).

Chen, C., Chan, N. and Wang, A., 2011. The many blades of the β-propeller proteins: conserved but versatile. *Trends in Biochemical Sciences*, 36(10), pp.553-561.

Chen, J., 2015. Signaling pathways in HPV-associated cancers and therapeutic implications. *Reviews in Medical Virology*, 25, pp.24-53.

Chen, J., Hong, Y., Rustamzadeh, E., Baleja, J. and Androphy, E., 1998. Identification of an α Helical Motif Sufficient for Association with Papillomavirus E6. *Journal of Biological Chemistry*, 273(22), pp.13537-13544.

Chen, J., Sachse, C., Xu, C., Mielke, T., Spahn, C. and Grigorieff, N., 2008. A dose-rate effect in single-particle electron microscopy. *Journal of Structural Biology*, 161(1), pp.92-100.

Cheng, K., Li, Y., Chang, W., Chen, Z., Cheng, J. and Tsai, C., 2019. Ubiquitin-protein ligase E3a (UBE3A) as a new biomarker of cardiac hypertrophy in cell models. *Journal of Food and Drug Analysis*, 27(1), pp.355-364.

Cheng, Y., Grigorieff, N., Penczek, P. and Walz, T., 2015. A Primer to Single-Particle Cryo-Electron Microscopy. *Cell*, 161(3), pp.438-449.

Chong, J. and Blow, J., 1996. DNA replication licensing factor. *Progress in Cell Cycle Research*, pp.83-90.

Christensen, J., Izatt, R., Hansen, L. and Partridge, J., 1966. Entropy Titration. A Calorimetric Method for the Determination of ΔG, ΔH, and ΔS from a Single Thermometric Titration. *The Journal of Physical Chemistry*, 70(6), pp.2003-2010.

Copping, N., Christian, S., Ritter, D., Islam, M., Buscher, N., Zolkowska, D., Pride, M., Berg, E., LaSalle, J., Ellegood, J., Lerch, J., Reiter, L., Silverman, J. and Dindot, S., 2017. Neuronal overexpression of Ube3a isoform 2 causes behavioral impairments and neuroanatomical pathology relevant to 15q11.2-q13.3 duplication syndrome. *Human Molecular Genetics*, 26(20), pp.3995-4010.

Cubillos-Rojas, M., Amair-Pinedo, F., Tato, I., Bartrons, R., Ventura, F. and Rosa, J., 2010. Simultaneous electrophoretic analysis of proteins of very

high and low molecular mass using Tris-acetate polyacrylamide gels. *ELECTROPHORESIS*, 31(8), pp.1318-1321.

Dagli AI, Mueller J, Williams CA. Angelman Syndrome. 1998 Sep 15 [Updated 2017 Dec 21]. In: Adam MP, Ardinger HH, Pagon RA, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2020. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1144/

Dagli, A., Buiting, K. and Williams, C., 2011. Molecular and Clinical Aspects of Angelman Syndrome. *Molecular Syndromology*,.

Danev, R. and Baumeister, W., 2016. Cryo-EM single particle analysis with the Volta phase plate. *eLife*, 5.

Danev, R., Buijsse, B., Khoshouei, M., Plitzko, J. and Baumeister, W., 2014. Volta potential phase plate for in-focus phase contrast transmission electron microscopy. *Proceedings of the National Academy of Sciences*, 111(44), pp.15635-15640.

Danev, R. and Nagayama, K., 2001. Transmission electron microscopy with Zernike phase plate. *Ultramicroscopy*, 88(4), pp.243-252.

David, Y., Ziv, T., Admon, A. and Navon, A., 2010. The E2 Ubiquitin-conjugating Enzymes Direct Polyubiquitination to Preferred Lysines. *Journal of Biological Chemistry*, 285(12), pp.8595-8604.

De Carlo, S. and Harris, J., 2011. Negative staining and cryo-negative staining of macromolecules and viruses for TEM. *Micron*, 42(2), pp.117-131.

Degenhardt, Y. and Silverstein, S., 2001. Gps2, a Protein Partner for Human Papillomavirus E6 Proteins. *Journal of Virology*, 75(1), pp.151-160.

Demeler, B., 2005. UltraScan- A comprehensive data analysis software package for analytical ultracentrifugation experiments. *Modern Analytical Ultracentrifugation: Techniques and Methods*, 10, pp.210-229.

DiStefano, C., Gulsrud, A., Huberty, S., Kasari, C., Cook, E., Reiter, L., Thibert, R. and Jeste, S., 2016. Identification of a distinct developmental and behavioral profile in children with Dup15q syndrome. *Journal of Neurodevelopmental Disorders*, 8(1).

Doorbar, J., Egawa, N., Griffin, H., Kranjec, C. and Murakami, I., 2015. Human papillomavirus molecular biology and disease association. *Reviews in Medical Virology*, 25, pp.2-23.

Drews, C., Brimer, N. and Vande Pol, S., 2020. Multiple regions of E6AP (UBE3A) contribute to interaction with papillomavirus E6 proteins and the activation of ubiquitin ligase activity. *PLOS Pathogens*, 16(1), p.e1008295.

Drulyte, I., Johnson, R., Hesketh, E., Hurdiss, D., Scarff, C., Porav, S., Ranson, N., Muench, S. and Thompson, R., 2018. Approaches to altering particle distributions in cryo-electron microscopy sample preparation. *Acta Crystallographica Section D Structural Biology*, 74(6), pp.560-571.

Ebner, F., Sailer, C., Eichbichler, D., Jansen, J., Sladewska-Marquardt, A., Stengel, F. and Scheffner, M., 2020. A ubiquitin variant-based affinity approach selectively identifies substrates of the ubiquitin ligase E6AP in complex with HPV-11 E6 or HPV-16 E6. *Journal of Biological Chemistry*, 295(44), pp.15070-15082.

Eletr, Z. and Kuhlman, B., 2007. Sequence Determinants of E2-E6AP Binding Affinity and Specificity. *Journal of Molecular Biology*, 369(2), pp.419-428.

El–Serag, H. and Rudolph, K., 2007. Hepatocellular Carcinoma: Epidemiology and Molecular Carcinogenesis. *Gastroenterology*, 132(7), pp.2557-2576.

Emsley, P., Lohkamp, B., Scott, W. and Cowtan, K., 2010. Features and development ofCoot. *Acta Crystallographica Section D Biological Crystallography*, 66(4), pp.486-501.

Felsenstein, J., 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*, 39(4), p.783.

Filippova, M., Johnson, M., Bautista, M., Filippov, V., Fodor, N., Tungteakkhun, S., Williams, K. and Duerksen-Hughes, P., 2007. The Large and Small Isoforms of Human Papillomavirus Type 16 E6 Bind to and Differentially Affect Procaspase 8 Stability and Activity. *Journal of Virology*, 81(8), pp.4116-4129.

Filippova, M., Parkhurst, L. and Duerksen-Hughes, P., 2004. The Human Papillomavirus 16 E6 Protein Binds to Fas-associated Death Domain and Protects Cells from Fas-triggered Apoptosis. *Journal of Biological Chemistry*, 279(24), pp.25729-25744.

Filippova, M., Song, H., Connolly, J., Dermody, T. and Duerksen-Hughes, P., 2002. The Human Papillomavirus 16 E6 Protein Binds to Tumor Necrosis Factor (TNF) R1 and Protects Cells from TNF-induced Apoptosis. *Journal of Biological Chemistry*, 277(24), pp.21730-21739.

Finn, R., Clements, J. and Eddy, S., 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(suppl), pp.W29-W37.

Finucane BM, Lusk L, Arkilo D, et al. 15q Duplication Syndrome and Related Disorders. 2016 Jun 16. In: Adam MP, Ardinger HH, Pagon RA, et

al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2020. Available from: https://www.ncbi.nlm.nih.gov/books/NBK367946/

Franken, L., Grünewald, K., Boekema, E. and Stuart, M., 2020. A Technical Introduction to Transmission Electron Microscopy for Soft-Matter: Imaging, Possibilities, Choices, and Technical Developments. *Small*, 16(14), p.1906198.

Frohlich, J., Miller, M., Bird, L., Garces, P., Purtell, H., Hoener, M., Philpot, B., Sidorov, M., Tan, W., Hernandez, M., Rotenberg, A., Jeste, S., Krishnan, M., Khwaja, O. and Hipp, J., 2019. Electrophysiological Phenotype in Angelman Syndrome Differs Between Genotypes. *Biological Psychiatry*, 85(9), pp.752-759.

Gamell, C., Bandilovska, I., Gulati, T., Kogan, A., Lim, S., Kovacevic, Z., Takano, E., Timpone, C., Agupitan, A., Litchfield, C., Blandino, G., Horvath, L., Fox, S., Williams, S., Russo, A., Gallo, E., Paul, P., Mitchell, C., Sandhu, S., Keam, S., Haupt, S., Richardson, D. and Haupt, Y., 2019. E6AP Promotes a Metastatic Phenotype in Prostate Cancer. *iScience*, 22, pp.1-15.

Gamell, C., Gulati, T., Levav-Cohen, Y., Young, R., Do, H., Pilling, P., Takano, E., Watkins, N., Fox, S., Russell, P., Ginsberg, D., Monahan, B., Wright, G., Dobrovic, A., Haupt, S., Solomon, B. and Haupt, Y., 2017. Reduced abundance of the E3 ubiquitin ligase E6AP contributes to decreased expression of theINK4/ARFlocus in non–small cell lung cancer. *Science Signaling*, 10(461), p.8223.

Gamsjaeger, R., Liew, C., Loughlin, F., Crossley, M. and Mackay, J., 2007. Sticky fingers: zinc-fingers as protein-recognition motifs. *Trends in Biochemical Sciences*, 32(2), pp.63-70.

Gao, K., Oerlemans, R. and Groves, M., 2020. Theory and applications of differential scanning fluorimetry in early-stage drug discovery. *Biophysical Reviews*, 12(1), pp.85-104.

Gao, M. and Karin, M., 2005. Regulating the Regulators: Control of Protein Ubiquitination and Ubiquitin-like Modifications by Extracellular Stimuli. *Molecular Cell*, 19(5), pp.581-593.

García-Cano, J., Martinez-Martinez, A., Sala-Gaston, J., Pedrazza, L. and Rosa, J., 2019. HERCing: Structural and Functional Relevance of the Large HERC Ubiquitin Ligases. *Frontiers in Physiology*, 10.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. and Bairoch, A., 2003. ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research*, 31(13), pp.3784-3788.

Gentile, J., Tan, W., Horowitz, L., Bacino, C., Skinner, S., Barbieri-Welge, R., Bauer-Carlin, A., Beaudet, A., Bichell, T., Lee, H., Sahoo, T., Waisbren, S., Bird, L. and Peters, S., 2010. A Neurodevelopmental Survey of Angelman Syndrome With Genotype-Phenotype Correlations. *Journal of Developmental & Behavioral Pediatrics*, 31(7), pp.592-601.

Gewin, L., Meyers, H., Kiyono, T. and Galloway, D., 2004. Identification of a novel telomerase repressor that interacts with the human papillomavirus type-16 E6/E6-AP complex. *Genes & Development*, 18(18), pp.2269-2282.

Glaeser, R. and Downing, K., 2007. Information Delocalization in Highly-defocused Phase-contrast images of Single-particle Specimens. *Microscopy and Microanalysis*, 13(S02).

Glaeser, R., Han, B., Csencsits, R., Killilea, A., Pulk, A. and Cate, J., 2016. Factors that Influence the Formation and Stability of Thin, Cryo-EM Specimens. *Biophysical Journal*, 110(4), pp.749-755.

Glumov, N., Kolomiyetz, E. and Sergeyev, V., 1995. Detection of objects on the image using a sliding window mode. *Optics & Laser Technology*, 27(4), pp.241-249.

Grau-Bové, X., Sebé-Pedrós, A. and Ruiz-Trillo, I., 2013. A Genomic Survey of HECT Ubiquitin Ligases in Eukaryotes Reveals Independent Expansions of the HECT System in Several Lineages. *Genome Biology and Evolution*, 5(5), pp.833-847.

Greaser, M. and Warren, C., 2012. Protein electrophoresis in agarose gels for separating high molecular weight proteins. *Methods in Molecular Biology*, 869, pp.111-118.

Greenfield, N., 2004. Circular dichroism (CD) analyses of protein-protein interactions. *Methods in Molecular Biology*, 1278, pp.239-265.

Greenfield, N., 2006. Using circular dichroism spectra to estimate protein secondary structure. *Nature Protocols*, 1(6), pp.2876-2890.

Greenfield, N. and Fasman, G., 1969. Computed circular dichroism spectra for the evaluation of protein conformation. *Biochemistry*, 8(10), pp.4108-4116.

Griffith, I., 1972. The effect of cross-links on the mobility of proteins in dodecyl sulphate–polyacrylamide gels. *Biochemical Journal*, 126(3), pp.553-560.

Guan, Y., Zhu, Q., Huang, D., Zhao, S., Jan Lo, L. and Peng, J., 2015. An equation to estimate the difference between theoretically predicted and SDS PAGE-displayed molecular weights for an acidic peptide. *Scientific Reports*, 5(1).

Gubbens, A., Barfels, M., Trevor, C., Twesten, R., Mooney, P., Thomas, P., Menon, N., Kraus, B., Mao, C. and McGinn, B., 2010. The GIF Quantum, a next generation post-column imaging energy filter. *Ultramicroscopy*, 110(8), pp.962-970.

Hadjebi, O., Casas-Terradellas, E., Garcia-Gonzalo, F. and Rosa, J., 2008. The RCC1 superfamily: From genes, to function, to disease. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1783(8), pp.1467-1479.

Hanahan, D. and Weinberg, R., 2011. Hallmarks of Cancer: The Next Generation. *Cell*, 144(5), pp.646-674.

Harlalka, G., Baple, E., Cross, H., Kühnle, S., Cubillos-Rojas, M., Matentzoglu, K., Patton, M., Wagner, K., Coblentz, R., Ford, D., Mackay, D., Chioza, B., Scheffner, M., Rosa, J. and Crosby, A., 2012. Mutation of HERC2 causes developmental delay with Angelman-like features. *Journal of Medical Genetics*, 50(2), pp.65-73.

Henderson, R., 1995. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Quarterly Reviews of Biophysics*, 28(2), pp.171-193.

Hershko, A. and Ciechanover, A., 1992. The Ubiquitin System for Protein Degradation. *Annual Review of Biochemistry*, 61(1), pp.761-807.

Herzik, M., 2020. Setting Up on the Talos Arctica for High-Resolution Data Collection. *cryoEM*, pp.125-144.

Herzik, M., Wu, M. and Lander, G., 2019. High-resolution structure determination of sub-100 kDa complexes using conventional cryo-EM. *Nature Communications*, 10(1).

Hillman, P., Christian, S., Doan, R., Cohen, N., Konganti, K., Douglas, K., Wang, X., Samollow, P. and Dindot, S., 2017. Genomic imprinting does not reduce the dosage of UBE3A in neurons. *Epigenetics & Chromatin*, 10(1).

Hornbeck, P., Kornhauser, J., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V. and Sullivan, M., 2011. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Research*, 40(D1), pp.D261-D270.

Huang, L., Kinnucan, E., Wang, G., Beaudenon, S., Howley, P., Huibregste, J. and Pavletich, N., 1999. Structure of an E6AP-UbcH7 Complex: Insights into Ubiquitination by the E2-E3 Enzyme Cascade. *Science*, 286(5443), pp.1321-1326.

Huang, X., Luan, B., Wu, J. and Shi, Y., 2016. An atomic structure of the human 26S proteasome. *Nature Structural & Molecular Biology*, 23(9), pp.778-785.

Huibregtse, J., Scheffner, M. and Howley, P., 1991. A cellular protein mediates association of p53 with the E6 oncoprotein of human papillomavirus types 16 or 18. *The EMBO Journal*, 10(13), pp.4129-4135.

Iossifov, I., O'Roak, B., Sanders, S., Ronemus, M., Krumm, N., Levy, D., Stessman, H., Witherspoon, K., Vives, L., Patterson, K., Smith, J., Paeper, B., Nickerson, D., Dea, J., Dong, S., Gonzalez, L., Mandell, J., Mane, S., Murtha, M., Sullivan, C., Walker, M., Waqar, Z., Wei, L., Willsey, A., Yamrom, B., Lee, Y., Grabowska, E., Dalkic, E., Wang, Z., Marks, S., Andrews, P., Leotta, A., Kendall, J., Hakker, I., Rosenbaum, J., Ma, B., Rodgers, L., Troge, J., Narzisi, G., Yoon, S., Schatz, M., Ye, K., McCombie, W., Shendure, J., Eichler, E., State, M. and Wigler, M., 2014. The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, 515(7526), pp.216-221.

Jain, R., Kasturi, R. and Schunck, B., 1995. *Machine Vision*. McGraw-Hill, pp.149-161.

Jeong, H., Woo, J., Kweon, H. and Ryu, B., 2019. Advantages and operational strategies of Falcon 3EC for high resolution cryo-electron microscopy of biological macromolecules. *Korean Society for Structural Biology*, 7(2), pp.29-34.

Jones, D., Taylor, W. and Thornton, J., 1992. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, 8(3), pp.275-282.

Joosten, R., Salzemann, J., Bloch, V., Stockinger, H., Berglund, A., Blanchet, C., Bongcam-Rudloff, E., Combet, C., Da Costa, A., Deleage, G., Diarena, M., Fabbretti, R., Fettahi, G., Flegel, V., Gisel, A., Kasam, V., Kervinen, T., Korpelainen, E., Mattila, K., Pagni, M., Reichstadt, M., Breton, V., Tickle, I. and Vriend, G., 2009. PDB_REDO: automated re-refinement of X-ray structure models in the PDB. *Journal of Applied Crystallography*, 42(3), pp.376-384.

Jiang, Y., He, B., Li, N., Ma, J., Gong, G. and Zhang, M., 2011. The oncogenic role of NS5A of hepatitis C virus is mediated by up-regulation of survivin gene expression in the hepatocellular cell through p53 and NF-κB pathways. *Cell Biology International*, 35(12), pp.1225-1232.

Jiang, Y., Lev-Lehman, E., Bressler, J., Tsai, T. and Beaudet, A., 1999. Genetics of Angelman Syndrome. *The American Journal of Human Genetics*, 65(1), pp.1-6.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S., Ballard, A., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A., Kavukcuoglu, K., Kohli, P. and Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), pp.583-589.

Kao, W., Beaudenon, S., Talis, A., Huibregtse, J. and Howley, P., 2000. Human Papillomavirus Type 16 E6 Induces Self-Ubiquitination of the E6AP Ubiquitin-Protein Ligase. *Journal of Virology*, 74(14), pp.6408-6417.

Keller, S., Vargas, C., Zhao, H., Piszczek, G., Brautigam, C. and Schuck, P., 2012. High-Precision Isothermal Titration Calorimetry with Automated Peak-Shape Analysis. *Analytical Chemistry*, 84(11), pp.5066-5073.

Kelley, M., Keiger, K., Lee, C. and Huibregtse, J., 2005. The Global Transcriptional Effects of the Human Papillomavirus E6 Protein in Cervical Carcinoma Cell Lines Are Mediated by the E6AP Ubiquitin Ligase. *Journal of Virology*, 79(6), pp.3737-3747.

Khan, O., Fu, G., Ismail, A., Srinivasan, S., Cao, X., Tu, Y., Lu, S. and Nawaz, Z., 2006. Multifunction Steroid Receptor Coactivator, E6-Associated Protein, Is Involved in Development of the Prostate Gland. *Molecular Endocrinology*, 20(3), pp.544-559.

Kimanius, D., Dong, L., Sharov, G., Nakane, T. and Scheres, S., 2021. New tools for automated cryo-EM single-particle analysis in RELION-4.0. *Biochemical Journal*, 478(24), pp.4169-4185.

Kishino, T., Lalande, M. and Wagstaff, J., 1997. UBE3A/E6-AP mutations cause Angelman syndrome. *Nature Genetics*, 15(1), pp.70-73.

Kliza, K. and Husnjak, K., 2020. Resolving the Complexity of Ubiquitin Networks. *Frontiers in Molecular Biosciences*, 7.

Knoll, J., Nicholls, R., Magenis, R., Graham, J., Lalande, M., Latt, S., Opitz, J. and Reynolds, J., 1989. Angelman and Prader-Willi syndromes share a common chromosome 15 deletion but differ in parental origin of the deletion. *American Journal of Medical Genetics*, 32(2), pp.285-290.

Komander, D., 2009. The emerging complexity of protein ubiquitination. *Biochemical Society Transactions*, 37(5), pp.937-953.

Kühne, C. and Banks, L., 1998. E3-Ubiquitin Ligase/E6-AP Links Multicopy Maintenance Protein 7 to the Ubiquitination Pathway by a Novel Motif, the L2G Box. *Journal of Biological Chemistry*, 273(51), pp.34302-34309.

Kühnle, S., Kogel, U., Glockzin, S., Marquardt, A., Ciechanover, A., Matentzoglu, K. and Scheffner, M., 2011. Physical and Functional Interaction of the HECT Ubiquitin-protein Ligases E6AP and HERC2. *Journal of Biological Chemistry*, 286(22), pp.19410-19416.

Kühnle, S., Martínez-Noël, G., Leclere, F., Hayes, S., Harper, J. and Howley, P., 2018. Angelman syndrome–associated point mutations in the Zn2+-binding N-terminal (AZUL) domain of UBE3A ubiquitin ligase inhibit binding to the proteasome. *Journal of Biological Chemistry*, 293(47), pp.18387-18399.

Kuijper, M., van Hoften, G., Janssen, B., Geurink, R., De Carlo, S., Vos, M., van Duinen, G., van Haeringen, B. and Storms, M., 2015. FEI's direct electron detector developments: Embarking on a revolution in cryo-TEM. *Journal of Structural Biology*, 192(2), pp.179-187.

Kumar, S., Stecher, G., Li, M., Knyaz, C. and Tamura, K., 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution*, 35(6), pp.1547-1549.

Kwak, J., Shim, J., Tiwari, I. and Jang, K., 2016. Hepatitis C virus core protein inhibits E6AP expression via DNA methylation to escape from ubiquitin-dependent proteasomal degradation. *Cancer Letters*, 380(1), pp.59-68.

Lalande, M. and Calciano, M., 2007. Molecular epigenetics of Angelman syndrome. *Cellular and Molecular Life Sciences*, 64(7-8), pp.947-960.

Lamm, O., 1926. Die Differentialgleichung der Ultrazentrifugierung. *Zeitschrift für Physikalische Chemie*, 143(2), pp.177-190.

Lane, D., 1992. p53, guardian of the genome. *Nature*, 358(6381), pp.15-16.

Langer, G., Cohen, S., Lamzin, V. and Perrakis, A., 2008. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nature Protocols*, 3(7), pp.1171-1179.

LaSalle, J., Reiter, L. and Chamberlain, S., 2015. Epigenetic regulation of UBE3A and roles in human neurodevelopmental disorders. *Epigenomics*, 7(7), pp.1213-1228.

Lee, S., Ramirez, J., Franco, M., Lectez, B., Gonzalez, M., Barrio, R. and Mayor, U., 2013. Ube3a, the E3 ubiquitin ligase causing Angelman syndrome and linked to autism, regulates protein homeostasis through the proteasomal shuttle Rpn10. *Cellular and Molecular Life Sciences*, 71(14), pp.2747-2758.

Le Fevre, A., Beygo, J., Silveira, C., Kamien, B., Clayton-Smith, J., Colley, A., Buiting, K. and Dudding-Byth, T., 2017. Atypical Angelman syndrome due to a mosaic imprinting defect: Case reports and review of the literature. *American Journal of Medical Genetics Part A*, 173(3), pp.753-757.

LeConte, B., Szaniszlo, P., Fennewald, S., Lou, D., Qiu, S., Chen, N., Lee, J. and Resto, V., 2018. Differences in the viral genome between HPV-positive cervical and oropharyngeal cancer. *PLOS ONE*, 13(8), p.e0203403.

Lemak, A., Yee, A., Bezsonova, I., Dhe-Paganon, S. and Arrowsmith, C., 2011. Zn-binding AZUL domain of human ubiquitin protein ligase Ube3A. *Journal of Biomolecular NMR*, 51(1-2), pp.185-190.

Lewis, E. and Murphy, K., 2022. Isothermal Titration Calorimetry. *Methods in Molecular Biology*, 305, pp.1-16.

Li, S., Hong, X., Wei, Z., Xie, M., Li, W., Liu, G., Guo, H., Yang, J., Wei, W. and Zhang, S., 2019. Ubiquitination of the HPV Oncoprotein E6 Is Critical for E6/E6AP-Mediated p53 Degradation. *Frontiers in Microbiology*, 10.

Li, Y., Cash, J., Tesmer, J. and Cianfrocco, M., 2020. High-Throughput Cryo-EM Enabled by User-Free Preprocessing Routines. *Structure*, 28(7), pp.858-869.e3.

Lopez, S., Segal, D. and LaSalle, J., 2019. UBE3A: An E3 Ubiquitin Ligase With Genome-Wide Impact in Neurodevelopmental Disease. *Frontiers in Molecular Neuroscience*, 11.

Louria-Hayon, I., Alsheich-Bartok, O., Levav-Cohen, Y., Silberman, I., Berger, M., Grossman, T., Matentzoglu, K., Jiang, Y., Muller, S., Scheffner, M., Haupt, S. and Haupt, Y., 2009. E6AP promotes the degradation of the PML tumor suppressor. *Cell Death & Differentiation*, 16(8), pp.1156-1166.

Makde, R., England, J., Yennawar, H. and Tan, S., 2010. Structure of RCC1 chromatin factor bound to the nucleosome core particle. *Nature*, 467(7315), pp.562-566.

Martínez-Noël, G., Galligan, J., Sowa, M., Arndt, V., Overton, T., Harper, J. and Howley, P., 2012. Identification and Proteomic Analysis of Distinct UBE3A/E6AP Protein Complexes. *Molecular and Cellular Biology*, 32(15), pp.3095-3106.

Martínez-Noël, G., Luck, K., Kühnle, S., Desbuleux, A., Szajner, P., Galligan, J., Rodriguez, D., Zheng, L., Boyland, K., Leclere, F., Zhong, Q., Hill, D., Vidal, M. and Howley, P., 2018. Network Analysis of UBE3A/E6AP-Associated Proteins Provides Connections to Several Distinct Cellular Processes. *Journal of Molecular Biology*, 430(7), pp.1024-1050.

Martinez-Zapien, D., Ruiz, F., Poirson, J., Mitschler, A., Ramirez, J., Forster, A., Cousido-Siah, A., Masson, M., Pol, S., Podjarny, A., Travé, G. and Zanier, K., 2016. Structure of the E6/E6AP/p53 complex required for HPV-mediated degradation of p53. *Nature*, 529(7587), pp.541-545.

Masuda, Y., Saeki, Y., Arai, N., Kawai, H., Kukimoto, I., Tanaka, K. and Masutani, C., 2019. Stepwise multipolyubiquitination of p53 by the E6AP-E6 ubiquitin ligase complex. *Journal of Biological Chemistry*, 294(41), pp.14860-14875.

McCoy, A., Grosse-Kunstleve, R., Adams, P., Winn, M., Storoni, L. and Read, R., 2007. Phaser crystallographic software. *Journal of Applied Crystallography*, 40(4), pp.658-674.

McMullan, G., Faruqi, A. and Henderson, R., 2016. Direct Electron Detectors. *Methods in Enzymology*, pp.1-17.

Meng, L., Person, R. and Beaudet, A., 2012. Ube3a-ATS is an atypical RNA polymerase II transcript that represses the paternal expression of Ube3a. *Human Molecular Genetics*, 21(13), pp.3001-3012.

Merk, A., Fukumura, T., Zhu, X., Darling, J., Grisshammer, R., Ognjenovic, J. and Subramaniam, S., 2020. 1.8 Å resolution structure of β-galactosidase with a 200 kV CRYO ARM electron microscope. *IUCrJ*, 7(4), pp.639-643.

Mesri, E., Feitelson, M. and Munger, K., 2014. Human Viral Oncogenesis: A Cancer Hallmarks Analysis. *Cell Host & Microbe*, 15(3), pp.266-282.

Mészáros, B., Erdős, G. and Dosztányi, Z., 2018. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Research*, 46(W1), pp.W329-W337.

Michelson, M., Eden, A., Vinkler, C., Leshinsky-Silver, E., Kremer, U., Lerman-Sagie, T. and Lev, D., 2011. Familial partial trisomy 15q11-13 presenting as intractable epilepsy in the child and schizophrenia in the mother. *European Journal of Paediatric Neurology*, 15(3), pp.230-233.

Migneault, I., Dartiguenave, C., Bertrand, M. and Waldron, K., 2004. Glutaraldehyde: behavior in aqueous solution, reaction with proteins, and application to enzyme crosslinking. *BioTechniques*, 37(5), pp.790-802.

Mukhopadhyay, D. and Riezman, H., 2007. Proteasome-Independent Functions of Ubiquitin in Endocytosis and Signaling. *Science*, 315(5809), pp.201-205.

Munakata, T., Liang, Y., Kim, S., McGivern, D., Huibregtse, J., Nomoto, A. and Lemon, S., 2007. Hepatitis C Virus Induces E6AP-Dependent Degradation of the Retinoblastoma Protein. *PLoS Pathogens*, 3(9), p.e139.

Munakata, T., Nakamura, M., Liang, Y., Li, K. and Lemon, S., 2005. Down-regulation of the retinoblastoma tumor suppressor by the hepatitis C virus NS5B RNA-dependent RNA polymerase. *Proceedings of the National Academy of Sciences*, 102(50), pp.18159-18164.

Murshudov, G., Skubák, P., Lebedev, A., Pannu, N., Steiner, R., Nicholls, R., Winn, M., Long, F. and Vagin, A., 2011. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D Biological Crystallography*, 67(4), pp.355-367.

Naydenova, K., McMullan, G., Peet, M., Lee, Y., Edwards, P., Chen, S., Leahy, E., Scotcher, S., Henderson, R. and Russo, C., 2019. CryoEM at 100 keV: a demonstration and prospects. *IUCrJ*, 6(6), pp.1086-1098.

Niebler, M., Qian, X., Höfler, D., Kogosov, V., Kaewprag, J., Kaufmann, A., Ly, R., Böhmer, G., Zawatzky, R., Rösl, F. and Rincon-Orozco, B., 2013. Post-Translational Control of IL-1β via the Human Papillomavirus Type 16 E6 Oncoprotein: A Novel Mechanism of Innate Immune Escape Mediated by the E3-Ubiquitin Ligase E6-AP and p53. *PLoS Pathogens*, 9(8), p.e1003536.

Noor, A., Dupuis, L., Mittal, K., Lionel, A., Marshall, C., Scherer, S., Stockley, T., Vincent, J., Mendoza-Londono, R. and Stavropoulos, D., 2015. 15q11.2 Duplication Encompassing Only the UBE3A Gene Is Associated with Developmental Delay and Neuropsychiatric Phenotypes. *Human Mutation*, 36(7), pp.689-693.

Nuber, U., Schwarz, S. and Scheffner, M., 1998. The ubiquitin-protein ligase E6-associated protein (E6-AP) serves as its own substrate. *European Journal of Biochemistry*, 254(3), pp.643-649.

Olabarria, M., Pasini, S., Corona, C., Robador, P., Song, C., Patel, H. and Lefort, R., 2019. Dysfunction of the ubiquitin ligase E3A Ube3A/E6-AP contributes to synaptic pathology in Alzheimer's disease. *Communications Biology*, 2(1).

OriginPro, Version 2021b (9.85). OriginLab Corporation, Northampton, MA, USA.

Pan, H. and Griep, A., 1995. Temporally distinct patterns of p53-dependent and p53-independent apoptosis during mouse lens development. *Genes & Development*, 9(17), pp.2157-2169.

Passmore, L. and Russo, C., 2016. Specimen Preparation for High-Resolution Cryo-EM. *Methods in Enzymology*, pp.51-86.

Paul, P., Raghu, D., Chan, A., Gulati, T., Lambeth, L., Takano, E., Herold, M., Hagekyriakou, J., Vessella, R., Fedele, C., Shackleton, M., Williams, E., Fox, S., Williams, S., Haupt, S., Gamell, C. and Haupt, Y., 2016. Restoration of

tumor suppression in prostate cancer by targeting the E3 ligase E6AP. *Oncogene*, 35(48), pp.6235-6245.

Peet, M., Henderson, R. and Russo, C., 2019. The energy dependence of contrast and damage in electron cryomicroscopy of biological molecules. *Ultramicroscopy*, 203, pp.125-131

Penczek, P., Zhu, J., Schröder, R. and Frank, J., 2018. THREE DIMENSIONAL RECONSTRUCTION WITH CONTRAST TRANSFER COMPENSATION FROM DEFOCUS SERIES. *Series in Structural Biology*, pp.232-239.

Pickart, C., 2001. Mechanisms Underlying Ubiquitination. *Annual Review of Biochemistry*, 70(1), pp.503-533.

Plevin, M., Mills, M. and Ikura, M., 2005. The LxxLL motif: a multifunctional binding sequence in transcriptional regulation. *Trends in Biochemical Sciences*, 30(2), pp.66-69.

Pons, T., Gómez, R., Chinea, G. and Valencia, A., 2003. Beta-propellers: Associated Functions and their Role in Human Diseases. *Current Medicinal Chemistry*, 10(6), pp.505-524.

Proteinatlas.org. 2020. *Tissue Expression Of UBE3A - Summary - The Human Protein Atlas*. [online] Available at: <https://www.proteinatlas.org/ENSG00000114062-UBE3A/tissue> [Accessed 29 April 2020].

Punjani, A., Rubinstein, J., Fleet, D. and Brubaker, M., 2017. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods*, 14(3), pp.290-296.

Pyeon, D., Rojas, V., Price, L., Kim, S., Singh, M. and Park, I., 2019. HIV-1 Impairment via UBE3A and HIV-1 Nef Interactions Utilizing the Ubiquitin Proteasome System. *Viruses*, 11(12), p.1098.

Raghu, D., Paul, P., Gulati, T., Deb, S., Khoo, C., Russo, A., Gallo, E., Blandino, G., Chan, A., Takano, E., Sandhu, S., Fox, S., Williams, S., Haupt, S., Gamell, C. and Haupt, Y., 2017. E6AP promotes prostate cancer by reducing p27 expression. *Oncotarget*, 8(26), pp.42939-42948.

Rapkins, R., Hore, T., Smithwick, M., Ager, E., Pask, A., Renfree, M., Kohn, M., Hameister, H., Nicholls, R., Deakin, J. and Graves, J., 2006. Recent Assembly of an Imprinted Domain from Non-Imprinted Components. *PLoS Genetics*, 2(10), p.e182.

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A., 2016. You Only Look Once: Unified, Real-Time Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*,.

Roelofs, J., Suppahia, A., Waite, K. and Park, S., 2018. Native Gel Approaches in Studying Proteasome Assembly and Chaperones. *Methods in Molecular Biology*, 1844, pp.237-260.

Rohou, A. and Grigorieff, N., 2015. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *Journal of Structural Biology*, 192(2), pp.216-221.

Ronchi, V., Kim, E., Summa, C., Klein, J. and Haas, A., 2017. In silico modeling of the cryptic E2∼ubiquitin–binding site of E6-associated protein (E6AP)/UBE3A reveals the mechanism of polyubiquitin chain assembly. *Journal of Biological Chemistry*, 292(44), pp.18006-18023.

Ronchi, V., Klein, J. and Haas, A., 2013. E6AP/UBE3A Ubiquitin Ligase Harbors Two E2∼ubiquitin Binding Sites. *Journal of Biological Chemistry*, 288(15), pp.10349-10360.

Ronchi, V., Klein, J., Edwards, D. and Haas, A., 2014. The Active Form of E6-associated protein (E6AP)/UBE3A Ubiquitin Ligase Is an Oligomer. *Journal of Biological Chemistry*, 289(2), pp.1033-1048.

Rosenthal, P. and Henderson, R., 2003. Optimal Determination of Particle Orientation, Absolute Hand, and Contrast Loss in Single-particle Electron Cryomicroscopy. *Journal of Molecular Biology*, 333(4), pp.721-745.

Rougeulle, C., Cardoso, C., Fontés, M., Colleaux, L. and Lalande, M., 1998. An imprinted antisense RNA overlaps UBE3A and a second maternally expressed transcript. *Nature Genetics*, 19(1), pp.15-16.

Rougeulle, C., Glatt, H. and Lalande, M., 1997. The Angelman syndrome candidate gene, UBE3A/E6-AP, is imprinted in brain. *Nature Genetics*, 17(1), pp.14-15.

Ruggieri, A., Murdolo, M., Harada, T., Miyamura, T. and Rapicetta, M., 2003. Cell cycle perturbation in a human hepatoblastoma cell line constitutively expressing Hepatitis C virus core protein. *Archives of Virology*, 149(1), pp.61-74.

Russo, C. and Passmore, L., 2014. Ultrastable gold substrates for electron cryomicroscopy. *Science*, 346(6215), pp.1377-1380.

Sadikovic, B., Fernandes, P., Zhang, V., Ward, P., Miloslavskaya, I., Rhead, W., Rosenbaum, R., Gin, R., Roa, B. and Fang, P., 2014. phisp Update for UBE3A Variants in Angelman Syndrome. *Human Mutation*, 35(12), pp.1407-1417.

Salminen, I., Read, S., Hurd, P. and Crespi, B., 2019. Genetic variation of UBE3A is associated with schizotypy in a population of typical individuals. *Psychiatry Research*, 275, pp.94-99.

Sato, M., 2017. Early Origin and Evolution of the Angelman Syndrome Ubiquitin Ligase Gene Ube3a. *Frontiers in Cellular Neuroscience*, 11.

Scarff, C., Fuller, M., Thompson, R. and Iadaza, M., 2018. Variations on Negative Stain Electron Microscopy Methods: Tools for Tackling Challenging Systems. *Journal of Visualized Experiments*, (132).

Schägger, H., 2001. Blue-native gels to isolate protein complexes from mitochondria. *Methods in Cell Biology*, 65, pp.231-244.

Scheffner, M., Huibregtse, J., Vierstra, R. and Howley, P., 1993. The HPV-16 E6 and E6-AP complex functions as a ubiquitin-protein ligase in the ubiquitination of p53. *Cell*, 75(3), pp.495-505.

Scheffner, M., Nuber, U. and Huibregtse, J., 1995. Protein ubiquitination involving an E1–E2–E3 enzyme ubiquitin thioester cascade. *Nature*, 373(6509), pp.81-83.

Scheres, S., 2012. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *Journal of Structural Biology*, 180(3), pp.519-530.

Scheres, S., 2015. Semi-automated selection of cryo-EM particles in RELION-1.3. *Journal of Structural Biology*, 189(2), pp.114-122.

Scheres, S., 2016. Processing of Structurally Heterogeneous Cryo-EM Data in RELION. *Methods in Enzymology*, 579, pp.125-157.

Scheres, S., Gao, H., Valle, M., Herman, G., Eggermont, P., Frank, J. and Carazo, J., 2007. Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nature Methods*, 4(1), pp.27-29.

Scheres, S., Valle, M., Nuñez, R., Sorzano, C., Marabini, R., Herman, G. and Carazo, J., 2005. Maximum-likelihood Multi-reference Refinement for Electron Microscopy Images. *Journal of Molecular Biology*, 348(1), pp.139-149.

Schiffman, M., Wentzensen, N., Wacholder, S., Kinney, W., Gage, J. and Castle, P., 2011. Human Papillomavirus Testing in the Prevention of Cervical Cancer. *JNCI: Journal of the National Cancer Institute*, 103(5), pp.368-383.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J., White, D., Hartenstein, V., Eliceiri, K., Tomancak, P. and Cardona, A., 2012. Fiji: an open-source platform for biological-image analysis. *Nature Methods*, 9(7), pp.676-682.

Schuck, P., 2000. Size-Distribution Analysis of Macromolecules by Sedimentation Velocity Ultracentrifugation and Lamm Equation Modeling. *Biophysical Journal*, 78(3), pp.1606-1619.

Schuck, P., 2013. Analytical ultracentrifugation as a tool for studying protein interactions. *Biophysical Reviews*, 5(2), pp.159-171.

Senisterra, G., Chau, I. and Vedadi, M., 2012. Thermal Denaturation Assays in Chemical Biology. *ASSAY and Drug Development Technologies*, 10(2), pp.128-136.

Shemer, R., Hershko, A., Perk, J., Mostoslavsky, R., Tsuberi, B., Cedar, H., Buiting, K. and Razin, A., 2000. The imprinting box of the Prader-Willi/Angelman syndrome domain. *Nature Genetics*, 26(4), pp.440-443.

Shirakura, M., Murakami, K., Ichimura, T., Suzuki, R., Shimoji, T., Fukuda, K., Abe, K., Sato, S., Fukasawa, M., Yamakawa, Y., Nishijima, M., Moriishi, K., Matsuura, Y., Wakita, T., Suzuki, T., Howley, P., Miyamura, T. and Shoji, I., 2006. E6AP Ubiquitin Ligase Mediates Ubiquitylation and Degradation of Hepatitis C Virus Core Protein. *Journal of Virology*, 81(3), pp.1174-1185.

Sigworth, F., 2015. Principles of cryo-EM single-particle image processing. *Microscopy*, 65(1), pp.57-67.

Silva-Santos, S., van Woerden, G., Bruinsma, C., Mientjes, E., Jolfaei, M., Distel, B., Kushner, S. and Elgersma, Y., 2015. Ube3a reinstatement identifies distinct developmental windows in a murine Angelman syndrome model. *Journal of Clinical Investigation*, 125(5), pp.2069-2076.

Sirois, C., Bloom, J., Fink, J., Gorka, D., Keller, S., Germain, N., Levine, E. and Chamberlain, S., 2020. Abundance and localization of human UBE3A protein isoforms.

Smith, S., Zhou, Y., Zhang, G., Jin, Z., Stoppel, D. and Anderson, M., 2011. Increased Gene Dosage of Ube3a Results in Autism Traits and Decreased Glutamate Synaptic Transmission in Mice. *Science Translational Medicine*, 3(103), pp.103ra97-103ra97.

Song, B., Lenhart, J., Flegler, V., Makbul, C., Rasmussen, T. and Böttcher, B., 2019. Capabilities of the Falcon III detector for single-particle structure determination. *Ultramicroscopy*, 203, pp.145-154.

Sonzogni, M., Hakonen, J., Bernabé Kleijn, M., Silva-Santos, S., Judson, M., Philpot, B., van Woerden, G. and Elgersma, Y., 2019. Delayed loss of UBE3A reduces the expression of Angelman syndrome-associated phenotypes. *Molecular Autism*, 10(1).

Srinivasan, S. and Nawaz, Z., 2011. E3 ubiquitin protein ligase, E6-associated protein (E6-AP) regulates PI3K-Akt signaling and prostate cell

growth. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1809(2), pp.119-127.

Steffin, D., Hsieh, E. and Rouce, R., 2019. Gene Therapy: Current Applications and Future Possibilities. *Advances in pediatrics*, 66, pp.37-54.

Svedberg, T. and Fåhraeus, R., 1926. A New Method For The Determination Of The Molecular Weight Of The Proteins. *Journal of the American Chemical Society*, 48(2), pp.430-438.

Tan, W. and Bird, L., 2016. Angelman syndrome: Current and emerging therapies in 2016. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, 172(4), pp.384-401.

Taylor, G., 2010. Introduction to phasing. *Acta Crystallographica. Section D, Biological Crystallography*, 66(4), pp.325-328.

The UniProt Consortium, 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1), pp.D506-D515.

Theodoridis, S., 2020. Neural Networks and Deep Learning. *Machine Learning*, pp.901-1038.

Thermofisher.com. n.d. *Falcon 4i Electron Detector | Life Sciences EM | Thermo Fisher Scientific - UK*. [online] Available at: <https://www.thermofisher.com/uk/en/home/electron-microscopy/products/accessories-em/falcon-detector.html> [Accessed 25 February 2022].

Thermofisher.com. n.d. *Selectris and Selectris X Imaging Filters*. [online] Available at: <https://www.thermofisher.com/order/catalog/product/SELECTRIS > [Accessed 25 February 2022].

Thermofisher.com. 2019. *Fringe-Free Imaging (FFI)*. [online] Available at: <https://assets.thermofisher.com/TFS-Assets/MSD/Datasheets/fringe-free-imaging-ds0317.pdf> [Accessed 25 February 2022].

Thomas, M. and Banks, L., 1998. Inhibition of Bak-induced apoptosis by HPV-18 E6. *Oncogene*, 17(23), pp.2943-2954.

Thompson, R., Walker, M., Siebert, C., Muench, S. and Ranson, N., 2016. An introduction to sample preparation and imaging by cryo-electron microscopy for structural biology. *Methods*, 100, pp.3-15.

Thon, F., 1966. Notizen: Zur Defokussierungsabhängigkeit des Phasenkontrastes bei der elektronenmikroskopischen Abbildung. *Zeitschrift für Naturforschung A*, 21(4), pp.476-478.

Tipler, P. and Mosca, G., 2008. *Physics for scientists and engineers*. 6th ed. W.H. Freeman and Company, pp.1299-1305.

Tipler, P. and Mosca, G., 2008. *Physics for scientists and engineers*. 6th ed. W.H. Freeman and Company, pp. 1294-1295.

Trillingsgaard, A. and Østergaard, J., 2004. Autism in Angelman Syndrome: an exploration of comorbidity. *Autism*, 8(2), pp.163-174.

Tsagkaris, C., Papakosta, V., Miranda, A., Zacharopoulou, L., Danilchenko, V., Matiashova, L. and Dhar, A., 2020. Gene Therapy for Angelman Syndrome: Contemporary Approaches and Future Endeavors. *Current Gene Therapy*, 19(6), pp.359-366.

Tungteakkhun, S. and Duerksen-Hughes, P., 2008. Cellular binding partners of the human papillomavirus E6 protein. *Archives of Virology*, 153(3), pp.397-408.

Uhlen, M., Fagerberg, L., Hallstrom, B., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C., Odeberg, J., Djureinovic, D., Takanen, J., Hober, S., Alm, T., Edqvist, P., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J. and Ponten, F., 2015. Tissue-based map of the human proteome. *Science*, 347(6220), pp.1260419-1260419.

Vaccarella, S., Lortet-Tieulent, J., Plummer, M., Franceschi, S. and Bray, F., 2013. Worldwide trends in cervical cancer incidence: Impact of screening against changes in disease risk factors. *European Journal of Cancer*, 49(15), pp.3262-3273.

Vatsa, N. and Jana, N., 2018. UBE3A and Its Link With Autism. *Frontiers in Molecular Neuroscience*, 11.

Velazquez-Campoy, A., Leavitt, S. and Freire, E., 2022. Characterization of protein-protein interactions by isothermal titration calorimetry. *Methods in Molecular Biology*, 1278, pp.183-204.

Vinothkumar, K. and Henderson, R., 2016. Single particle electron cryomicroscopy: trends, issues and future perspective. *Quarterly Reviews of Biophysics*, 49.

Vogel, C., Teichmann, S. and Pereira-Leal, J., 2005. The Relationship Between Domain Duplication and Recombination. *Journal of Molecular Biology*, 346(1), pp.355-365.

Vokes, E., Agrawal, N. and Seiwert, T., 2015. HPV-Associated Head and Neck Cancer. *Journal of the National Cancer Institute*, 107(12), p.djv344.

Wagner, T., Merino, F., Stabrin, M., Moriya, T., Antoni, C., Apelbaum, A., Hagel, P., Sitsel, O., Raisch, T., Prumbaum, D., Quentin, D., Roderer, D., Tacke, S., Siebolds, B., Schubert, E., Shaikh, T., Lill, P., Gatsogiannis, C. and Raunser, S., 2019. SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. *Communications Biology*, 2(1).

Wang, H. and Fan, X., 2019. Challenges and opportunities in cryo-EM with phase plate. *Current Opinion in Structural Biology*, 58, pp.175-182.

Wang, N., Liu, D., Parokonny, A. and Schanen, N., 2004. High-Resolution Molecular Characterization of 15q11-q13 Rearrangements by Array Comparative Genomic Hybridization (Array CGH) with Detection of Gene Dosage. *The American Journal of Human Genetics*, 75(2), pp.267-281.

Wang, N., Parokonny, A., Thatcher, K., Driscoll, J., Malone, B., Dorrani, N., Sigman, M., LaSalle, J. and Schanen, N., 2008. Multiple forms of atypical rearrangements generating supernumerary derivative chromosome 15. *BMC Genetics*, 9(1), p.2.

Wang, Y., Liu, X., Zhou, L., Duong, D., Bhuripanyo, K., Zhao, B., Zhou, H., Liu, R., Bi, Y., Kiyokawa, H. and Yin, J., 2017. Identifying the ubiquitination targets of E6AP by orthogonal ubiquitin transfer. *Nature Communications*, 8(1).

Waterhouse, A., Procter, J., Martin, D., Clamp, M. and Barton, G., 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9), pp.1189-1191.

Wheeler, A., Sacco, P. and Cabo, R., 2017. Unmet clinical needs and burden in Angelman syndrome: a review of the literature. *Orphanet Journal of Rare Diseases*, 12(1).

Williams, C., Headd, J., Moriarty, N., Prisant, M., Videau, L., Deis, L., Verma, V., Keedy, D., Hintze, B., Chen, V., Jain, S., Lewis, S., Arendall, W., Snoeyink, J., Adams, P., Lovell, S., Richardson, J. and Richardson, D., 2017. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science*, 27(1), pp.293-315.

Winter, G., Waterman, D., Parkhurst, J., Brewster, A., Gildea, R., Gerstel, M., Fuentes-Montero, L., Vollmar, M., Michels-Clark, T., Young, I., Sauter, N. and Evans, G., 2018. DIALS: implementation and evaluation of a new integration package. *Acta Crystallographica Section D Structural Biology*, 74(2), pp.85-97.

Woelk, T., Sigismund, S., Penengo, L. and Polo, S., 2007. The ubiquitination code: a signalling problem. *Cell Division*, 2(1), p.11.

Wolyniec, K., Levav-Cohen, Y., Jiang, Y., Haupt, S. and Haupt, Y., 2012. The E6AP E3 ubiquitin ligase regulates the cellular response to oxidative stress. *Oncogene*, 32(30), pp.3510-3519.

Wolyniec, K., Shortt, J., de Stanchina, E., Levav-Cohen, Y., Alsheich-Bartok, O., Louria-Hayon, I., Corneille, V., Kumar, B., Woods, S., Opat, S., Johnstone, R., Scott, C., Segal, D., Pandolfi, P., Fox, S., Strasser, A., Jiang, Y., Lowe, S., Haupt, S. and Haupt, Y., 2012. E6AP ubiquitin ligase regulates PML-induced senescence in Myc-driven lymphomagenesis. *Blood*, 120(4), pp.822-832.

Yamamoto, Y., Huibregtse, J. and Howley, P., 1997. The Human E6-AP Gene (UBE3A) Encodes Three Potential Protein Isoforms Generated by Differential Splicing. *Genomics*, 41(2), pp.263-266.

Yamasaki, K., Joh, K., Ohta, T., Masuzaki, H., Ishimaru, T., Mukai, T., Niikawa, N., Ogawa, M., Wagstaff, J. and Kishino, T., 2003. Neurons but not glial cells show reciprocal imprinting of sense and antisense transcripts of Ube3a. *Human Molecular Genetics*, 12(8), pp.837-847.

Yamashita, R., Nishio, M., Do, R. and Togashi, K., 2018. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9(4), pp.611-629.

Ye, Y., Blaser, G., Horrocks, M., Ruedas-Rama, M., Ibrahim, S., Zhukov, A., Orte, A., Klenerman, D., Jackson, S. and Komander, D., 2012. Ubiquitin chain conformation regulates recognition and activity of interacting proteins. *Nature*, 492(7428), pp.266-270.

Yi, J., Berrios, J., Newbern, J., Snider, W., Philpot, B., Hahn, K. and Zylka, M., 2015. An Autism-Linked Mutation Disables Phosphorylation Control of UBE3A. *Cell*, 162(4), pp.795-807.

Yim, E. and Park, J., 2005. The Role of HPV E6 and E7 Oncoproteins in HPV-associated Cervical Carcinogenesis. *Cancer Research and Treatment*, 37(6), p.319.

Yoon, B., 2009. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Current Genomics*, 10(6), pp.402-415.

Zanier, K., Charbonnier, S., Sidi, A., McEwen, A., Ferrario, M., Poussin-Courmontagne, P., Cura, V., Brimer, N., Babah, K., Ansari, T., Muller, I., Stote, R., Cavarelli, J., Vande Pol, S. and Trave, G., 2013. Structural Basis for Hijacking of Cellular LxxLL Motifs by Papillomavirus E6 Oncoproteins. *Science*, 339(6120), pp.694-698.

Zhao, H., Piszczek, G. and Schuck, P., 2015. SEDPHAT – A platform for global ITC analysis and global multi-method analysis of molecular interactions. *Methods*, 76, pp.137-148.

Zheng, S., Palovcak, E., Armache, J., Verba, K., Cheng, Y. and Agard, D., 2017. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nature Methods*, 14(4), pp.331-332.

Zhou, Y., Zhao, Y., Gao, Y., Hu, W., Qu, Y., Lou, N., Zhu, Y., Zhang, X. and Yang, H., 2017. Hepatitis C virus NS3 protein enhances hepatocellular carcinoma cell invasion by promoting PPM1A ubiquitination and degradation. *Journal of Experimental & Clinical Cancer Research*, 36(1).

Zhu, J., Penczek, P., Schröder, R. and Frank, J., 1997. Three-Dimensional Reconstruction with Contrast Transfer Function Correction from Energy-Filtered Cryoelectron Micrographs: Procedure and Application to the 70S Escherichia coli Ribosome. *Journal of Structural Biology*, 118(3), pp.197-219.

Zimmermann, H., Degenkolbe, R., Bernard, H. and O'Connor, M., 1999. The Human Papillomavirus Type 16 E6 Oncoprotein Can Down-Regulate p53 Activity by Targeting the Transcriptional Coactivator CBP/p300. *Journal of Virology*, 73(8), pp.6209-6219.

Zivanov, J., Nakane, T., Forsberg, B., Kimanius, D., Hagen, W., Lindahl, E. and Scheres, S., 2018. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife*, 7.

10 Appendices

10.1 Appendix 1 – UBE3A MSA

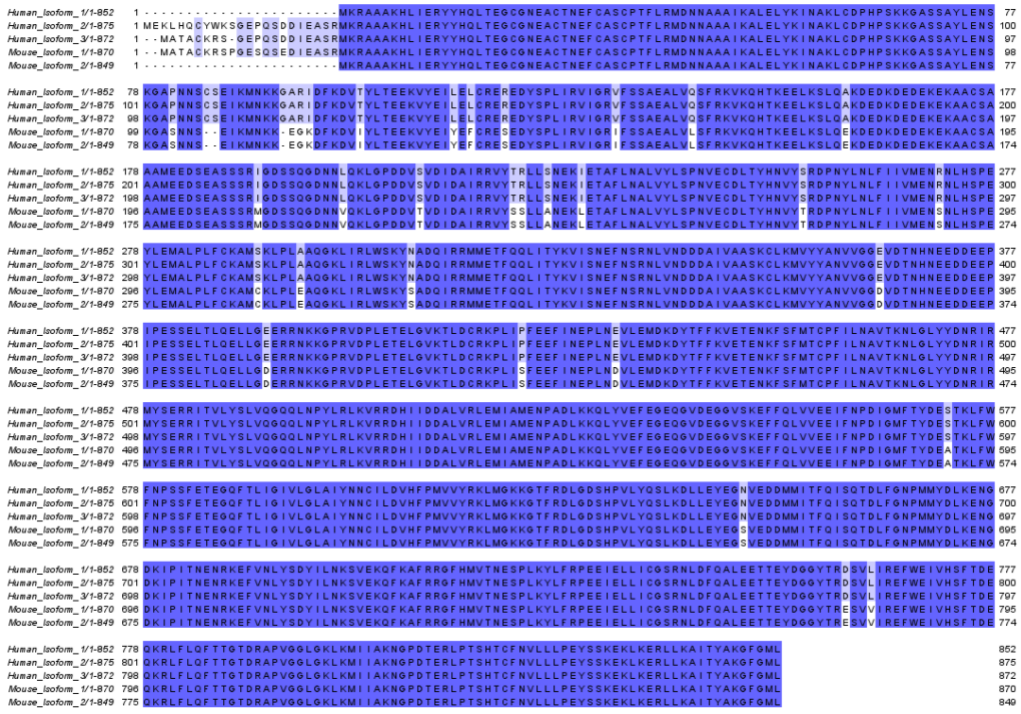


Figure 128 Full MSA of human and mouse UBE3A genes. Aligned with Clustal Omega, displayed and coloured by sequence identity with JalView 2.11.1.0

10.2 Appendix 2 – pUBE3A UBE3A Sequence

BglII...T7 promoter...XbaI...*His-tag*...E6AP...**BamHI**...*XhoI*

TCG*AGATCT*CGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAA
CAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACC**ATG**GGC*CATC*
*ATCATCATCATCACCACCATCAC*GAAAACCTGTACTTCCAGGGTGGG**ATG**AAACGTGCAG
CGGCCAAGCATCTGATTGAGCGTTATTATCACCAACTGACGGAAGGGTGCGGTAATGAA
GCGTGTACCAACGAATTCTGTGCCTCCTGCCCGACCTTTCTGCGCATGGACAACAATGC
GGCTGCAATTAAGGCTCTTGAACTGTATAAGATTAACGCTAAACTGTGCGATCCACACCC
CTCCAAAAAAGGAGCGAGCAGTGCCTACCTTGAGAATAGCAAGGGAGCACCCAATAAT
TCGTGTTCTGAGATCAAAATGAACAAGAAAGGCGCTCGTATCGACTTCAAGGATGTCAC
CTACCTGACCGAGGAAAAGGTATATGAGATCCTGGAATTATGCCGTGAACGTGAAGATTA
TAGCCCGCTGATCCGTGTGATCGGCCGTGTGTTTTCAAGCGCGGAGGCGCTGGTACAGA
GCTTCCGTAAGGTCAAACAGCACACTAAGGAAGAGCTTAAATCACTGCAAGCGAAAGA
CGAGGACAAAGATGAGGACGAAAAAGAGAAGGCGGCGTGCTCGGCTGCTGCGATGGA
AGAGGATTCGGAAGCGAGCTCATCTCGTATAGGCGACAGCTCCCAGGGCGATAACAACC
TTCAAAAACTGGGTCCGGATGACGTGTCGGTGGACATAGATGCAATTCGCCGCGTGTAT
ACTCGCCTGCTGTCGAACGAGAAAATTGAAACCGCTTTCCTGAATGCGTTAGTGTATCT
GAGCCCGAACGTGGAATGCGATCTGACCTATCACAATGTCTACAGCCGTGATCCGAACTA
TTTGAACCTGTTTATCATAGTCATGGAGAACCGGAACCTGCATTCACCCGAGTACTTGGA
AATGGCGCTTCCATTATTCTGTAAGGCAATGAGCAAATTGCCCCTGGCGGCACAAGGTA
AGCTGATACGGCTTTGGAGCAAGTATAACGCGGATCAAATACGTCGCATGATGGAAACC
TTTCAACAGCTGATTACGTACAAAGTAATTAGTAATGAGTTTAATTCGCGTAATCTGGTCA
ACGACGATGACGCGATCGTCGCGGCCTCCAAGTGCTTGAAGATGGTTTACTATGCGAAC
GTAGTCGGGGGAGAGGTGGATACCAATCATAACGAGGAAGACGACGAAGAGCCGATCC
CGGAATCGAGCGAACTGACACTGCAGGAACTTTTGGGCGAGGAACGGCGTAATAAGAA
GGGGCCACGTGTGGACCCATTAGAAACAGAACTGGGGGTAAAGACCCTTGACTGCCGT
AAACCCCTGATTCCGTTCGAGGAATTCATTAACGAACCTCTGAACGAAGTACTGGAGAT
GGATAAAGACTACACATTTTTCAAAGTCGAGACAGAGAACAAATTCTCCTTCATGACTT
GTCCGTTCATATTGAATGCTGTCACTAAGAATCTTGGTCTGTATTATGACAATCGTATTCG
GATGTACAGCGAACGCCGTATTACGGTGCTGTACTCATTGGTTCAAGGGCAACAACTGA
ATCCATACCTGCGTCTTAAAGTACGTCGGGATCATATCATTGACGACGCATTGGTACGCTT
AGAAATGATCGCAATGGAAAATCCGGCTGATCTGAAAAAACAGCTTTATGTGGAGTTTG
AGGGAGAACAGGGCGTGGACGAAGGGGGAGTCAGCAAGGAGTTTTTCCAACTGGTCG
TCGAAGAGATATTCAACCCGGATATTGGAATGTTTACGTATGACGAAAGCACCAAACTGT
TCTGGTTTAATCCTAGCTCTTTTGAGACTGAGGGCCAATTTACCCTTATTGGCATTGTACT
GGGCCTGGCGATTTACAATAACTGCATCTTGGATGTGCATTTCCCCATGGTGGTATATCGT
AAACTTATGGGTAAAAAGGGCACCTTTCGTGATCTTGGCGATAGCCACCCTGTACTGTAT
CAAAGCCTGAAGGATCTGCTGGAATATGAGGGGAATGTAGAAGATGACATGATGATTAC
CTTCCAAATTAGCCAGACCGATCTGTTTGGCAACCCGATGATGTACGACTTAAAGGAAA
ACGGGGACAAAATACCGATTACGAATGAAAACCGTAAAGAGTTTGTGAATCTGTATAGC
GACTATATTCTTAACAAGAGCGTGGAGAAACAATTTAAGGCGTTTCGGCGTGGCTTTCAT
ATGGTTACAAACGAAAGTCCACTTAAATACTTGTTCCGGCCAGAGGAAATAGAGCTGTT
AATTTGCGGGAGTCGTAACTTGGACTTCCAAGCCCTGGAAGAAACGACAGAGTACGAC
GGAGGCTATACCCGTGATAGCGTGCTGATTCGTGAATTTTGGGAAATTGTGCATAGCTTT
ACCGACGAGCAGAAACGGTTATTTCTTCAATTCACCACTGGCACCGATCGTGCGCCGGT
GGGTGGGTTGGGCAAACTGAAAATGATTATTGCGAAGAATGGGCCCGACACCGAACGG
TTGCCGACCAGCCACACCTGCTTCAACGTTCTGCTTCTGCCAGAATACTCTAGCAAAGA
AAAGTTGAAGGAACGTCTGCTTAAGGCCATTACCTATGCCAAAGGCTTCGGCATGCTGT
AATTAGTAACTCCTCGGCGCGCC**GGATCC***TCGAG*CACCACCACCACC

Figure 129: The sequence for the UBE3A gene within the pUBE3A plasmid provided by Dr. Martin Scheffner of the University of Konstanz.

10.3 Appendix 3 – Disorder Predictions

10.3.1 UBE3A

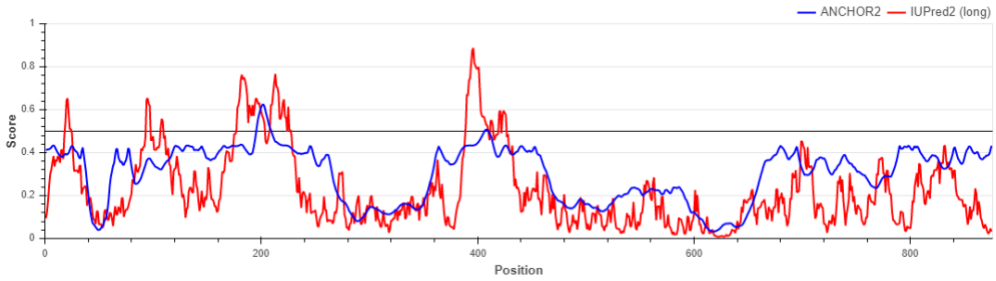


Figure 130: The disorder prediction for UBE3A predicted by IUPred2A (Mészáros et al., 2018). A score of above 0.5 denotes a disordered region.

10.3.2 PSMD4

Figure 131: The disorder prediction for UBE3A predicted by IUPred2A (Mészáros et al., 2018). A score of above 0.5 denotes a disordered region.

10.3.3 RLD2

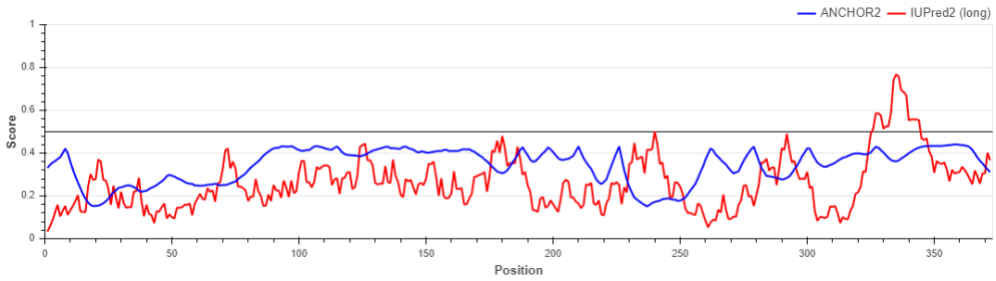
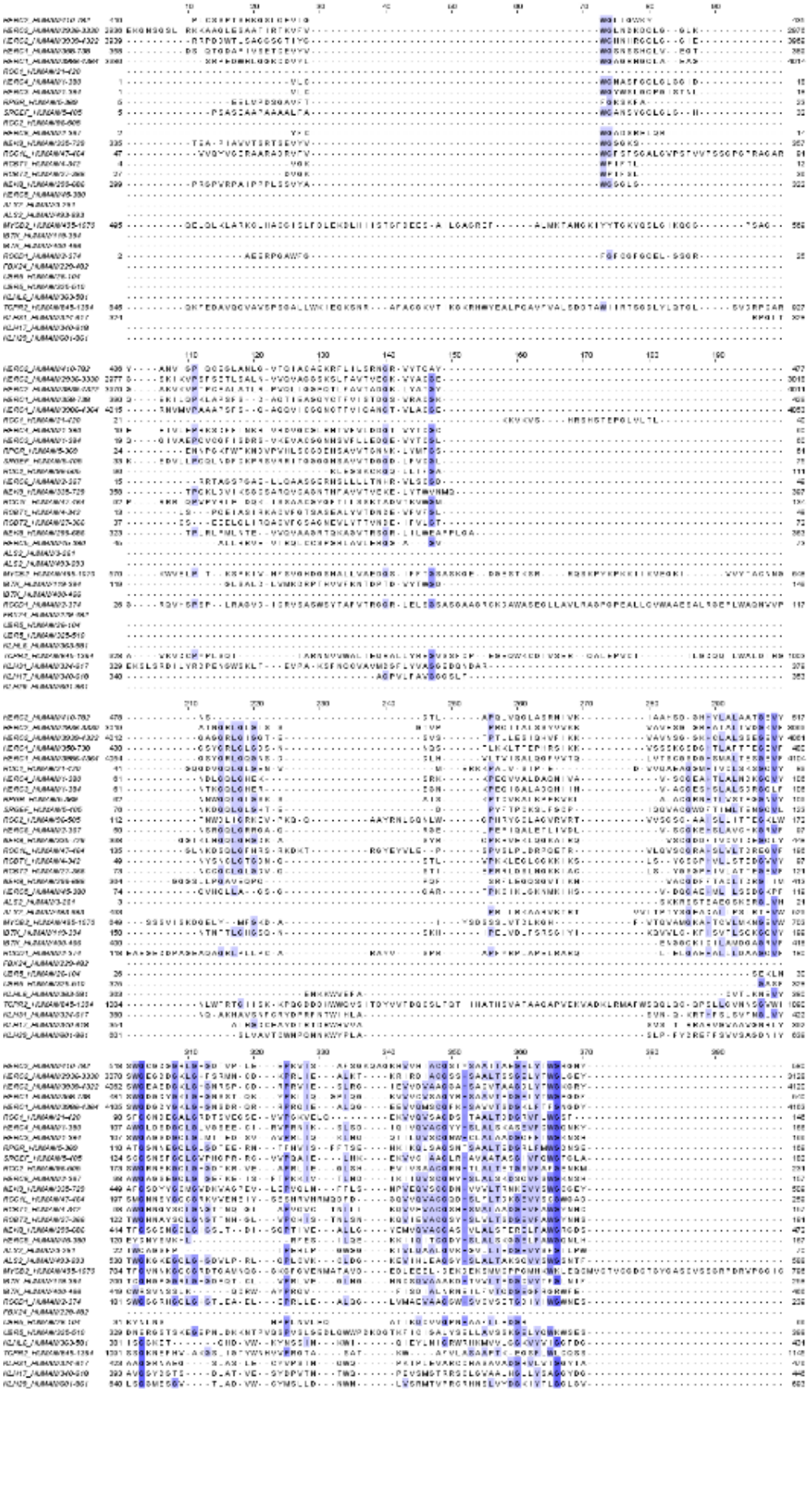
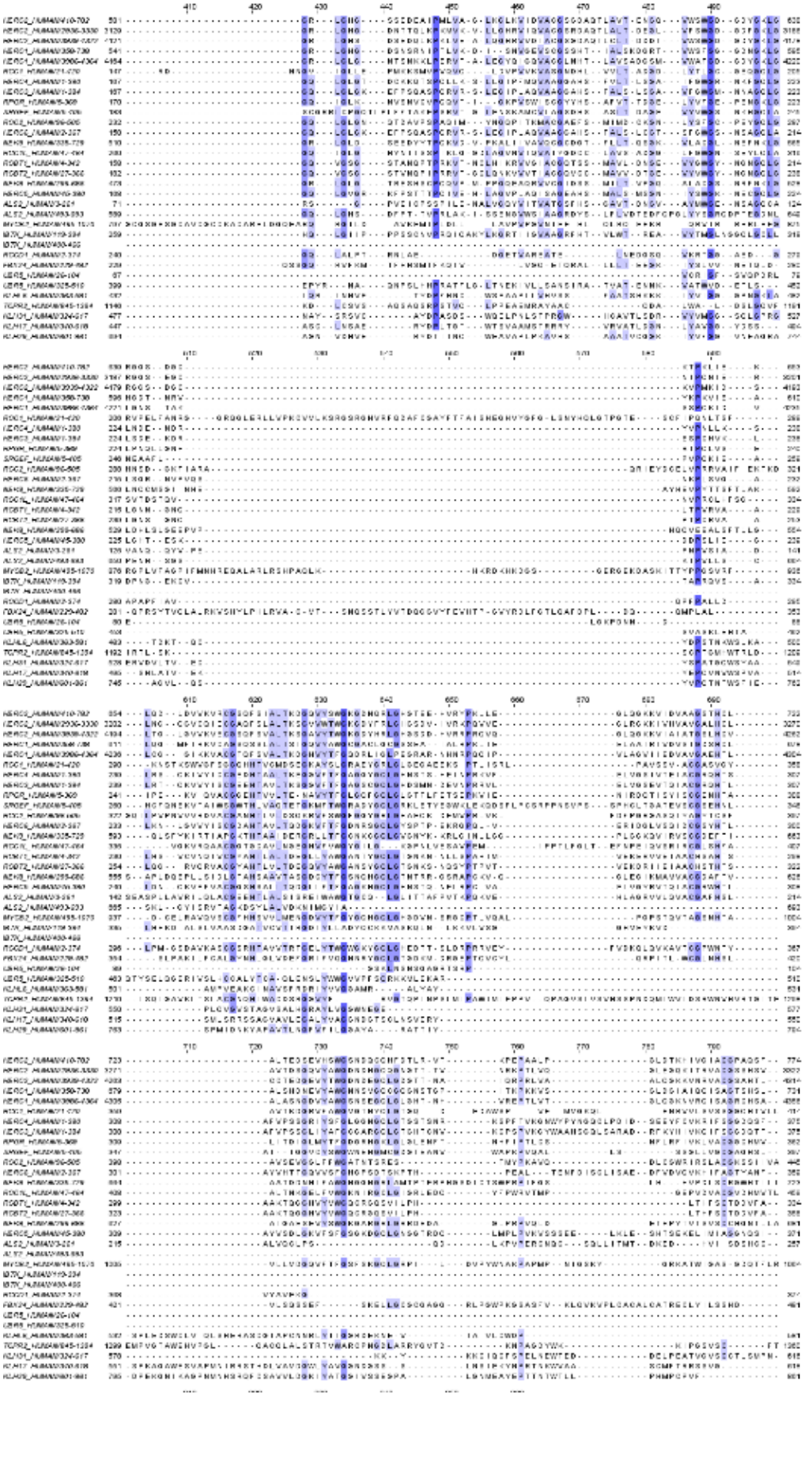


Figure 132: The disorder prediction for the RLD2 domain of HERC2 predicted by IUPred2A (Mészáros et al., 2018). A score of above 0.5 denotes a disordered region.

10.4 Appendix 4 – RLDs MSA





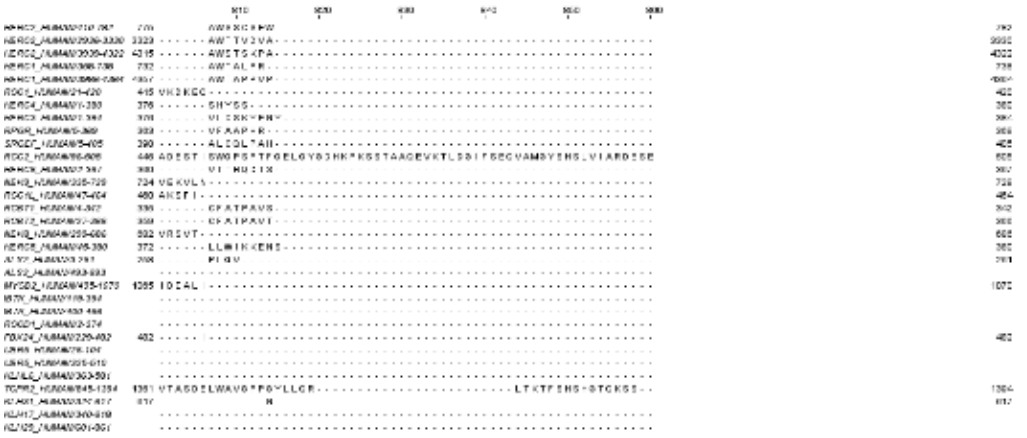


Figure 133: A multiple sequence alignments of all proteins containing an RLD domain. The sequence similarity is shown by the blue colouring, suggesting that there is a relatively low sequence similarity among the RLD domains. Long regions extraneous to the core 7-bladed β-propeller fold can be seen in different areas of the different proteins.

10.5 Appendix 5 – RobeTTA UBE3A Error Values

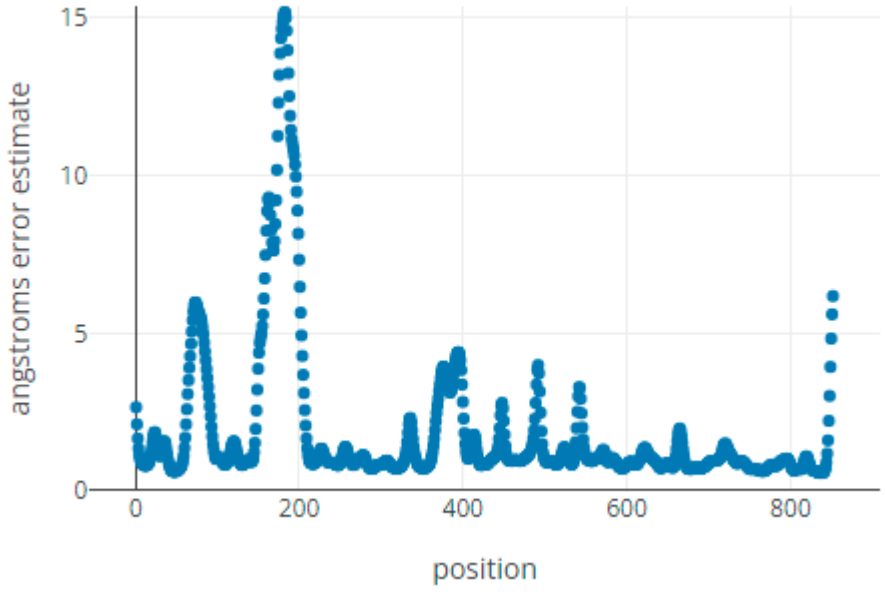


Figure 134: The error estimate for the Robetta model of UBE3A isoform 1, taken from the Robetta server website (https://robetta.bakerlab.org/ Job ID 226114). The region around residues 150 – 200, the region identified as the interacting region with RLD2 of HERC2 (Kühnle et al., 2011), shows the largest error estimate, suggesting a less rigid structure in that region.

10.6 Appendix 6 - Cryo-EM Theory

10.6.1 Screening Grids on the Glacios

Grids were screened prior to data collection using the FEI Glacios microscope belonging to the RCaH, situated at eBIC. The Glacios microscope features a 200V electron beam as opposed to the 300V beam typically used in a Titan

Krios, and the Glacis at the RCaH is fitted with a Falcon IV detector. The Glacios microscope uses the same cartridge system as the Titan Krios, where up to 12 grids can be loaded into the microscope at once. However, typically only 11 grids are loaded in a session, leaving the first slot empty in case a grid has been left on the stage from a previous session. Grids are loaded as described in 2.10.3, and once the temperatures and vacuums are at the required level, an inventory of the grids is taken. When an inventory is taken, the microscope will try to select each grid in turn to confirm that something can be retrieved from each slot. If a grid has not been loaded into the cassette properly it won't appear on the inventory, but slot that do appear on the inventory will contain a grid that can be viewed.

Once the inventory has been taken confirming that the grids were all loaded successfully, an atlas is generated for each grid. In the EPU software, in the Atlas tab, each grid to be atlased is selected from the list down the left hand side of the screen, and then the atlas session is started. Several images will be taken at the lowest magnification setting across the grid, so that each image forms a tile that are all pieced together to generate an image of the whole grid area. During this stage, the turbo pump can be left on so that the microscope is able to switch between the grids quicker, but if it is turned back to the auto-off setting then the image quality of each atlas may be slightly better. Atlasing with the turbo on typically takes ~10 min per grid, while taking atlases without the turbo pump will typically take up to 20 min per grid. When the microscope has finished collecting an atlas for a grid, the EPU software will identify the Gridsquares in each image and colour code them so that similar looking squares will be given the same colour border. The colour coding is relative to each grid, so once the optimal ice areas have been identified similar gridsquares may be easily selected, but the optimal ice areas must be determined manually for each new grid.

The atlases can give an idea of the ice quality across the hole grid, for example, it can be easy to spot particularly dry or damaged squares as well as too thick squares, but higher magnification images must be taken in order to narrow down the optimal ice areas. To do this, the stage is moved to a chosen area, by right clicking in EPU and selecting the 'move stage to here' option. The eucentric height of the GridSquare is calculated, typically using the Auto-Eucentric option in the AutoFunctions tab with the Thon Rings setting, and then an image in taken at the GridSquare mag. The Thon Rings setting is typically the same magnification as the GridSquare magnification, but the binning of the images is higher to generate more contrast between the Thon rings observed. An image taken at GridSquare magnification should show the entirety of the GridSquare without too much border around the edge, and the magnification level can be altered to meet this if needed. At the GridSquare level, the ice conditions across each square can be observed in more detail than the Atlas level. If the glow discharging was not optimal there may be

thicker ice areas in the centre of the square, like a 'fried-egg' effect. An area of the square will be selected for a closer view, again by right clicking on the area and selecting 'move stage to here', and an image will be taken at Hole/Eucentric Height magnification. For larger holes the Hole magnification setting should show an image of a single FoilHole with a minimal border around the edge, but for smaller holes, such as those in the R1.2/1.3 grids that were used frequently during this project, I preferred to set the Hole mag so that four FoilHoles and the surrounding carbon areas could be seen in a single image. Before a higher magnification image can be taken, an autofocus measurement must be taken. Using the Hole mag image, the stage is moved to the centre of the foil area between four holes. In the AutoFunctions tab of EPU, the AutoFocus option is selected, and autofocus is ran at the AutoFocus setting. The AutoFocus setting features the same magnification level used for the Data Acquisition setting, but the exposure time for each image will be shorter to speed up the focusing process. Once the AutoFocus has determined the optimal focus of the microscope, the stag is moved to a FoilHole area and an image is taken at Data Acquisition mag. This process is repeated for different areas across the grid, attempting to sample as many different areas across each FoilHole, GridSquare, and the overall atlas as is feasible within a session. Typically, in a session involving 8-12 grids, images will be taken in two or three regions of four or five different representative GridSquares, and four FoilHole images will be taken for each GridSquare area to cover the edges, centre, and inbetween region of the FoilHoles. When taking Data Acquisition level images, the exposure time and defocus values can be adjusted to try to improve the contrast of the image, and notes are taken on the visibility of the particles in different ice conditions and the concentration of the sample.

If not many particles are observed across the grid then the sample preparation needs improvement. If particles are observed on the carbon areas but not the FoilHoles, the grids either need more optimised glow discharging or some form of treatment, such as an ultrathin carbon film or GrOx-DDM application going forward. If the particles are present but difficult to see due to contrast issues, the ice thickness is probably too high. Either data can be collected in thinner areas of the same grid, or new grids must be made focussing on creating thinner ice areas. If the particles are present and the contrast is high but they appear denatured or aggregated, it is possible that the sample has denatured at the air-water interface upon flash-freezing. Some proteins are not amenable to vitrification, but optimisation of the buffer used or application of a GrOx-DDM support may be able to help with this. If particles are present, the concentration is acceptable, and they look somewhat structured, then the grid is recovered from the microscope at the end of the session and a full dataset is collected on a Titan Krios microscope.

10.6.2 Data Collection on the Titan Krios

When a grid has been screened and deemed sufficient for data collection, the recovered grid is then loaded onto a Titan Krios microscope. While up to 12 grids can be loaded at any time, for a data collection session typically two or three grids will be loaded. One will be the main grid for collection, and the other two will be a backup in case something has happened to the grid during storage, or it get damaged in the microscope, or just in case the collectable area is too small to fill the allotted time for the session so a second grid can be imaged after.

Once loaded onto the microscope stage, an atlas is taken of the grid to be collected on, and the image shift calibration is performed. This involves taking a series of images in the same position at every magnification level, starting with Data Acquisition and ending with Atlas mag, to determine if there is any shift in the stage location across different magnification settings. If there is some shift observed at any point, the ideal location of the stage is set and the image is retaken to confirm the repositioning before moving on to the next magnification level. If the image shift is not calibrated before the start of the session the microscope may have issues transitioning between magnification settings and data may be collected in areas other than those specified. Once this is set, GridSquares similar to the optimal conditions chosen during the screening session are selected to be collected from. For each square, the eucentric height is set and an image is taken. The eucentric height can be set by using the auto-eucentric option in the hole selection tab, or it can be set using the autofunction tab auto-eucentric options if the stage-tilt method is preferred. The method of determining the eucentric height seems to subject to personal preference, but either method is typically sufficicent. Once an image of the GridSquare has been taken, the holes that will be collected upon are selected. For QuantiFoil grids with regularly spaced holes, a template for a pair of holes can be adjusted to set the hole size and distance, and then a grid of potential holes will be generated. The EPU software has settings to automatically remove the holes close to the grid bars, as the extra height of the grid bars will interfere with the beam optics, and it also has an option to set an ice filter to select holes with a specific ice thickness. This ice filter method uses the histogram of light/dark values across the image, and filters are used to cut off any holes that appear to be lighter or darker than the specified levels. This can be very useful, but it is not very good at identifying between very thin ice and dry holes, particularly as the histogram is reset for each GridSquare. For the data collected as part of this project, the ice filter was used to remove the most egregiously thick or empty holes, and then the selection brush was used to manually select or deselect the remaining holes. For LaceyCarbon or HoleyCarbon grids that do not feature a regular lattice of holes, the hole selection can be done in one of two ways. The first involves the plotting of a regularly spaced lattice of holes across the hole image, and holes over large carbon areas or suboptimal ice condition areas can be

deselected using the ice filter or the selection brush tool. The second method is to manually pick areas irregularly across the GridSquare to act as FoilHoles. The second method can result in finer tuning of image acquisition areas, but it is much more time consuming.

Once the squares and holes have been selected, the acquisition areas must be set. This involves determining where the images will be taken relative to the FoilHole area, how many images will be taken per FoilHole, where the Autofocus images will be taken, and how frequently the autofocus will be run. For grids with large holes many images can be taken per hole, but for smaller holes the number of possible acquisition areas will be limited. The addition of filters such as the Gatan GIF or the Thermo Fisher Selectris and Selectris-X filters in recent years to allow fringe-free imaging has increased the number of images that can be taken in a set area, but the collectable area will still be limited by the size of the beam area. In R1.2/1.3 grids, at the start of this project only one image could be taken per hole, but the most recent dataset allowed collection of three micrographs per hole due to reduction in irradiated area sounding the acquisition area. The number of images collected per FoilHole may still be limited by the preferred location of the images to be collected. If the very centre of the hole shows the best particle distribution, only a single image can be taken, as the acquisition areas cannot be overlapping. However, if the very edge of the hole, or the region between the very edge and the very centre, is the preferred image location, more acquisition areas can be placed around the circumference of the FoilHole. The acquisition area of the FEI K3 camera is also smaller than that of the Falcon camera series, so it may be possible to take more images depending on the camera used.

The final steps before setting of a data collection session involve aligning the microscope to ensure parallel illumination. Non-parallel illumination due to inaccurately aligned microscopes can lead to local variations in magnification or defocus values, which will negatively impact the CTF estimation and the ability of the software to average the particle images, resulting in a decreased resolution limit. In a three condenser lens system, as in the Titan Krios microscopes, the crossover point of the beam between the C1 (condenser 1) and C2 (condenser 2) lenses can be varied to adjust the ‘spot size’ parameter, while the crossover point between the C2 and C3 lenses sets the beam intensity value. The C3 (condenser 3) lens allows modulation of both of these parameters while retaining the focus of the beam on the front of the objective lens, whereas in a two-condenser system, such as the Talos Arctica or Glacios microscopes, the beam intensity must be set at a constant value to ensure parallel illumination (Herzik, 2020). Although the direct alignments were attempted by myself during various screening sessions on the Glacios 200 kV microscope, the direct alignments for data collections of either the Glacios or

Titan Krios microscopes was performed by beamline staff at either eBIC or the LISCB.

10.6.3 Data Processing

10.6.3.1 RELION

RELION is a computer program that uses a Bayesian approach to determine a high resolution 3D structure from raw micrographs with fairly minimal user input required (Scheres, 2012). Bayesian means that it weighs up both the theoretical expected values and observed values of the dataset to determine the most probable interpretation of the data, rather than relying solely on the expertise of the user. The Bayesian approach allows the program to iteratively determine many of the parameters for concurrent steps in the process directly from the data used and generated by previous steps, which reduces the chance of errors introduced by inexperienced users (Scheres, 2012). RELION was not the first program developed to enable processing of cryo-EM datasets, but implementation of the Bayesian approach allowed cryo-EM data processing to become more accessible to users with less expertise in cryo-EM.

The RELION software was also the first to implement the gold-standard FSC approach. This involves separating the data into two half-sets and refining each independently. The Fourier shell correlation (FSC) between the two maps is calculated, and the point at which the FSC value = 0.143 is taken as the overall resolution. This prevents overfitting of noisy data, and allows for validation of the map generated (Rosenthal and Henderson, 2003; Scheres, 2012).

10.6.3.2 CryoSPARC

CryoSPARC is an alternative program to RELION that includes all of the processed required for full data analysis of raw cryo-EM or negative stain data. One of the key features of cryoSPARC is the increased efficiency of many of its processes relative to RELION. The cryoSPARC software was designed for effective use on commercially available computing hardware, without the need for large computing clusters. Rather than relying on GPU-acceleration of most process or other technological advances that could take many years, the focus of the cryoSPARC team was to reduce the computing cost of the processes involved by streamlining the algorithms with a particularly high computing cost (Punjani *et al.,* 2017).

CryoSPARC introduced the stochastic gradient descent (SGD) method for *ab initio* model generation, allowing a first approximation of the shape of the structure without the need for a pre-existing template model (Puinjani *et al.,* 2017). The process behind the SGD *ab initio* reconstruction method is outlined in section 10.6.3.9. The SGD *ab initio* job in cryoSPARC allowed for 3D classification of the particles without relying heavily on a 3D template. This means that cryo-EM data analysis can be carried out on completely novel proteins without homologs that have already been solved by crystallography,

but it also prevents introducing bias into the 3D reconstructions. Before a dataset has been processed to reveal the structure of the protein, the resemblance to the template provided is only a guess. *Ab initio* 3D classification allows elucidation of multiple conformations of the sample, which may not occur if the program is simply looking for the similarity of a particle to the template provided.

Another new implementation in cryoSPARC is the branch-and-bound method of structural refinements (Punjani *et al.,* 2017). The most computational expensive part of map refinements is determining the best angle to fit each 2D particle image into the 3D map. The branch and bound method introduces a lower bound, so that rather than trying to determine the probability for every possible angle, it calculates the lower bound of all poses, which is much less computationally taxing. The values outside of this lower bound are discarded, and the lower bound of this smaller subset is then calculated. This is repeated until it becomes computationally effective to calculate all absolute values for the remaining subset of possible poses, resulting in a just as accurate determination of the value with the lowest error rate, but at a much-reduced computational cost.

10.6.3.3 Motion Correction

Although cryo-EM samples are produced by immobilising purified protein samples in vitreous ice, irradiating the sample with a beam of electrons during the data acquisition process causes the ice to melt, which results in the protein particles moving around slightly during the exposure time. In order to produce the highest resolution structure possible this movement must be corrected for in data processing through a process called motion correction. Rather than taking a single 2D image of the sample, each raw micrograph is taken as a movie comprised of many frames. This allows features within each image to be tracked through the frames. The MotionCor2 program is commonly used to do this correction, it is available through RELION either by using RELION as a wrapper for the original GPU-based program, or by using RELION's own CPU-based implementation of the programme.

MotionCor2 corrects the beam-induced motion of each micrograph by first splitting each one into a 5x5 grid (Zheng *et al.,* 2017). It then maps the movement of features within each square, as well as mapping the overall movement within the micrograph as a whole. In order to track the movement of an object through the micrograph, the shift of the object between one frame and the next is described by a vector, and the total shift of the object through all the frames in the micrograph can be summarised by a series of vectors. This series of vectors can then be used to show that the total shift between any non-adjacent frames of the micrograph can be described by the summation of all adjacent vectors between them (Fig. 135).

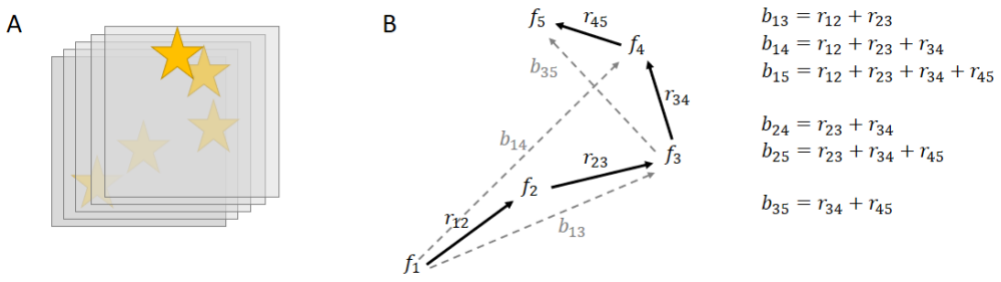


Figure 135: Motion correction of micrographs involves tracing the movement of particles across the frames of a micrograph and summarising it as a series of vectors. A) Each micrograph is split up into a defined number of frames, and the location of a defined object is determined for independently for each frame. B) The location of the particle in each frame is recorded as various points and the movement between each frame is summarised as a vector, with the total movement of the object described by a set of linear equations. r represents the relative shift between adjacent subframes, while b represents the relative shift between non-adjacent subframes.

This can be written as a set of linear equations each following the $m = (n-1)n/2$ rule, which can then be rewritten in matrix form, $Ar = b$, *where* A is an $m*(n-1)$ coefficient matrix, $r = [r_1, r_2, \ldots, r_{n-1}]^T$, and $b = [b_{12}, b_{13}, \ldots]^T$ (superscript T indicates transposition). From here, the generalised inverse matrix method is used to calculate a least-square solution of r so that $r_s = (A^TA)^{-1}A^Tb$. The residual error of each measured shift is calculated using the equation $\Delta b = ||Ar_s - b||$, and any equation with a residual error larger than 1 pixel is removed from the prior equations. After the error correction, the r_s is recalculated to get the final solution. This is then used to change the phase of each subframe's Fourier transform so that all subframes are shifted to the same origin. All subframes are then flattened into a single image in which the motion of particles during the exposure has been corrected (Zheng *et al.,* 2017).

When this theoretical framework is applied to real cryo-EM samples there are a few more factors that must be considered. The first issue is fixed pattern noise, this could be a faulty pixel in the camera or imperfect gain normalisation that leads to noise that has a fixed pattern in each subframe. If this generates a strong enough signal, particularly where the sample has been subjected to a low dose exposure, this peak may appear larger than the real cross correlation peaks and would cause the program to track the non-existent movement of the fixed artefact rather than tracking the movement of the sample itself. One way to protect against this is to apply a B-factor, a form of low pass filter that will remove the high-resolution information from the micrographs during the subframe alignment process. The high-resolution information is then restored in the corrected image so that it does not affect the final resolution of the resulting structure. Another way of protecting against fixed pattern noise is to prevent the program from aligning adjacent

subframes. This is helpful because the image shifts between adjacent subframes are likely to be fairly small so it is harder to distinguish a fixed noise artefact from a very small movement of a real object. Another consideration for real cryo-EM samples is the frame exposure time. A shorter exposure per frame will allow a finer sampling of movement, but it will also reduce the dose per image, resulting in a smaller signal to noise ratio and making identifying features within the ice more difficult (Zheng *et al.,* 2017).

10.6.3.4 CTF Estimation

After motion correction, the next step is CTF estimation. The contrast transfer function (CTF) describes the Fourier transform of the point spread function of the microscope (Thon, 1966). Cryo-EM relies on phase contrast to allow identification of features within micrographs, which has the advantage of making the observations more physiologically relevant as samples aren't subjected to fixatives or stains, but it has the disadvantage of meaning that when the images are taken with a perfect signal, i.e., a modern microscope that is perfectly and precisely aligned, the resulting image has no phase contrast and so particles are indistinguishable from noise. In order to maintain the phase contrast images are instead taken with an imperfect signal, and this imperfection is the point spread function. It is called the point spread function as each point of the image is convoluted so that rather than appearing as a distinct, precise point, each point spread into a larger, fuzzier object (Fig. 136).

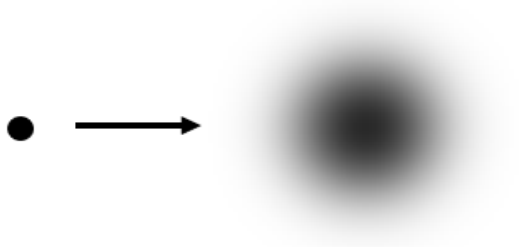


Figure 136: The point spread function causes each point to appear as a broader shape with less defined edges. This can cause some information to be lost when masks are set too tightly around particles and during image processing.

The CTF is calculated with slightly different equations depending on the software used, as different algorithms will be implemented to prioritise different aspects of the CTF estimation process. For CTFFIND4, the CTF is calculated with the equation

$$CTF(\lambda, g, \Delta f, C_s, \Delta\varphi, \omega_2) = -\sin[\mathcal{X}(\lambda, |g|, \Delta f, C_s, \Delta\varphi, \omega_2)]$$

Where

$$\mathcal{X}(\lambda, |g|, \Delta f, C_s, \Delta\varphi, \omega_2) = \pi\lambda|g|^2\left(\Delta f - \frac{1}{2}\lambda^2|g|^2 C_s\right) + \Delta\varphi + \tan^{-1}\left(\frac{\omega_2}{\sqrt{1-{\omega_2}^2}}\right)$$

Where $\mathcal{X}$ is the frequency-dependent phase shift, and is a function of λ, the electron wavelength, g, the spatial frequency vector, Δf, the objective defocus, C_s, the spherical aberration, $\Delta\varphi$, the phase shift introduced by the phase plate, and ω_2, the contribution of the amplitude contrast to the total contrast (Rohou and Grigorieff, 2019).

However, as the spherical aberration is given for a certain microscope and the wavelength is also set by the voltage of the microscope, the CTF can be considered a sine function that varies with frequency and defocus. In theory the ideal CTF function shouldn't feature a change in the amplitude of the wave, but the CTF is also affected by an envelope function caused by uncontrollable limitations in data acquisition, including limited spatial coherence and energy chromaticity of the beam, as well as specimen movement etc. (Penczek *et al.,* 1997; Cheng, 2015).

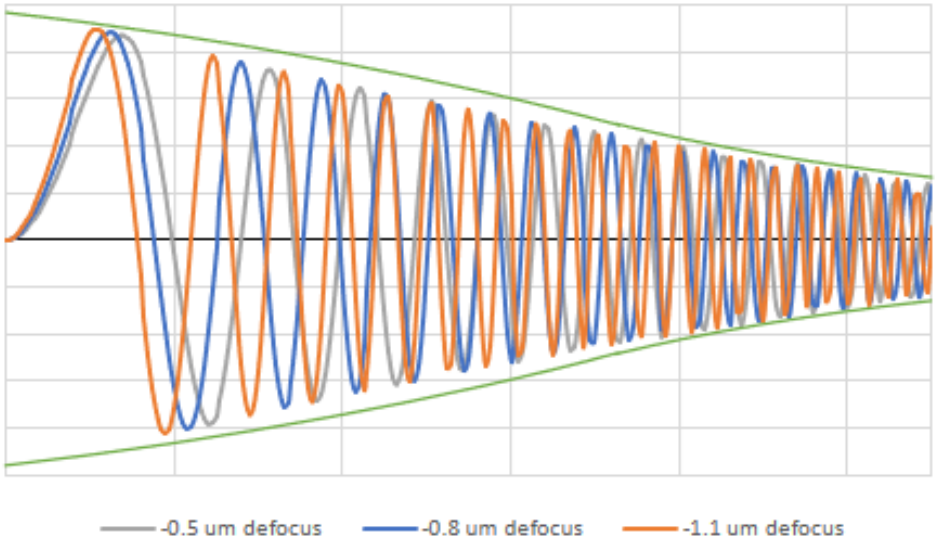


Figure 137: The effects of defocus values on the CTF. At a high defocus value the values closest to the zero crossing are much higher frequency than a low defocus image, so much more low frequency information is available. However, the number of measurable modulations is much greater than a low defocus image, which leads to greater delocalisation of electron signals, reducing the definition of the resulting images. The envelope function in represented by the green lines, and the effect of the envelope function is reflected by the decreasing amplitude of the wave.

The increase in low frequency information available for a low defocus micrograph makes the contrast of the micrograph much stronger, which allows for easier visualisation of the particles (Thon, 1966). However, the increase in modulations also leads to a greater delocalisation of the electron signals, reducing the definition of the resulting images (Glaeser and Downing, 2007). High dose images are also typically subjected to a much stronger envelope function, causing the wavelength to revert to zero at a lower frequency value than an equivalent low dose image (Chen *et al.,* 2008). A key

point to note is that both the low defocus and high defocus CTF figures have zero crossings at different points, so when collecting a cryo-EM dataset images are typically taken with a range of defocus values. This means that even though each micrograph is missing information at certain frequencies, the dataset as a whole, and so the resulting 3D structure, will contain at least some information for every frequency until the signal drops off (Cheng *et al.,* 2015; Zhu *et al.,* 1997).

10.6.3.5 Particle Picking

Once the micrographs have been imported, flattened, and contain some CTF information, the next step is identifying particles within each image. It is possible to pick particles manually, this involves selecting a box size that you think is only slightly larger than the longest dimension of your protein, and then simply clicking on features that look like particles. However, modern microscopes can take hundreds of images per hour, so a single dataset will typically contain thousands of micrographs and would take way too long to pick manually.

Relion Reference-Based Autopicking

Relion has a built-in autopicking function that uses a reference-based approach to pick particles that resemble the provided reference (Scheres, 2015). It works by first normalising the motion-corrected micrograph images by assuming an independent Gaussian noise distribution, but then applying position-dependent factors to result in micrographs with a mean recorded noise of zero and a standard deviation of one. This allows a more consistent picking pattern across micrographs that were taken at different defocus values, or with different ice conditions. The template images are then compared to the micrograph and any features that may represent a potential particle are identified. For each detected particle, a box (the size is defined by the user) is drawn around the object, and then a smaller circle is drawn within this box (the size is defined by the user). The area beyond the circle but within the square is used to determine the background noise of the image, and the area within the circle is normalised using this information. The particle image is then compared to the template image and the probability of observing the template image vs the probability of observing only noise is calculated (Scheres, 2015). This calculation results in a value that represents the confidence of each potential particle pick, and the user can define a threshold level to show only particles with a confidence level above a certain level.

Reference-based picking is particularly useful when a structure already exists for the protein of interest, or a homolog, as the pre-existing model can be used as the template to provide an accurate reference for identifying particles. If no existing structures exist templates can be generated by manually picking several micrographs and running a 2D classification job. A few different views from this job are selected to represent the key views of the molecule, and these are used as templates for Relion's reference-based

picking job. However, manually picking micrographs can be very time intensive, and picking only a few views may introduce issues with orientational sampling as rare views may not be recognised, particularly if the particles appear to be asymmetrical and elongated. Another issue with reference-based particle picking approaches is the introduction of bias into the process. If the particles do not resemble the reference image then they will not be picked, even if they are clear particles (Scheres, 2015).

Relion LoG Autopicking

An alternative to the reference-based picking within Relion is the Laplacion-of-Gaussian based picking job (Zivanov *et al.,* 2018). This involves identifying areas of steep contrast within the micrograph, which allows detection of the edges of particles without the need for any prior knowledge about the protein. The relion implementation of the LoG function allows users to input a minimum and maximum particle diameter to narrow down the objects selected, as well as filters related to the image statistics of each micrograph to attempt to avoid areas with carbon or ice contamination.
The LoG function involves the application of a Gaussian filter over the image to smooth the noise profile, followed by the application of the Laplace operator to identify the steep changes in contrast around definable objects (Jain *et al.,* 1995). The Laplace operator is a second order derivative function, meaning that it calculates the derivative of the derivative of a function. The gradient of a line is an example of a first derivative, as it describes the change in the values in the function. The second derivative then describes the change in the gradient of a line as it progresses. The Laplace operator can be defined by:

$$\Delta f = \nabla^2 = \nabla \cdot \nabla f$$

Where Δf denotes the Laplace of a function, $\nabla \cdot$ the divergence, and ∇f the gradient of the function. As the Laplace operator is a second order differential, it features a s=zero crossing at the midpoint of the change in gradient, which allows much more accurate detection of less well-defined edges compares to the first order gradient function (Jain *et al.,* 1995; See Fig. 97 in section 7.1.2).

Deep Learning Methods

As well as these more basic methods of particle picking, deep learning methods can be used to pick particles more specifically. crYOLO and Topaz are examples of deep learning particle picking programs. Both allow the user to either use a predetermined model designed to work with most protein particle presentation, or to train the program based on the real data (Wagner *et al.,* 2019; Bepler *et al.,* 2019). Training a model involves manually picking several micrographs to provide the program with a series of references for both true and false particle images True particles are those picked, false particles are generated from the areas left unpicked. Once a model has been chosen, the program will search through each micrograph image individually

to identify objects resembling the true particles. Topaz and crYOLO are both methods involving deep learning techniques, and the user experience is similar for both, but the method used in each is different. The standard object identification method within deep learning is the sliding window method, where windows of a set size are passed across the image and objects are identified within each individual window. crYOLO uses an alternative method, called YOLO for you only look once, as rather than a sliding stream of overlapping windows moving across each image, several windows of different sizes are placed over the image, with each window representing a possible particle (Wagner *et al.,* 2019). crYOLO then searches within each window to find the centre point of the object, and records that value as the particle location. Topaz, however, uses an implementation of the sliding window method known as positive unlabelling (PU) (Bepler *et al.,* 2019). The PU method speeds up computation of the sliding window implementation by allowing identification of both positive and negative areas, rather than searching through the whole image area for positive particle locations. Topaz implements this PU technique alongside a generalised expectation (GE) algorithm, that limits the user error involved in manually picking micrographs by using statistics to decide whether an unpicked area is a good example of a false particle image, or just a particle that has been missed (Bepler *et al.,* 2019).

Although both crYOLO and Topaz are highly accurate particle picking programs, particularly when they have been trained on the real data, they still allow for a picking threshold, so the picked particles can be limited to just the most likely particles, as well as various other image statistic thresholds to allow manual modulation of the picking.

10.6.3.6 Particle Extraction

Once the particles have been picked to satisfaction, the individual particle images must be extracted from the whole micrographs. Particle extraction is a fairly straightforward process without much need for optimising. One consideration though is the size of the particle files. Particles can be downscaled at this stage to reduce the computing cost of future processing steps enabling them to run quicker, but downscaling can reduce the amount of information present in the image, which can make alignments steps more difficult. With typical cryo-EM datasets you might chose to downscale the particle files to a quarter of the original file size to get through the initial alignment steps as quickly as possible, before re-extracting the refined particles without downsampling late in order to retain as much information possible in the final reconstruction. However, as UBE3A is a particularly small protein by usual cryo-EM standards, most of the processing was carried out with particles rescaled by only a half of the original size.

10.6.3.7 2D Classification

2D Classification involves combining extracted particle images together to form a discrete number of classes, each of which represents a different view of the protein in question. This combining of particle images to form a single class average image significantly improves the low signal-to-noise ratio (SNR) that is characteristic of cryo-EM images. There are several different ways of generating these 2D classes, but Relion uses a maximum-likelihood based approach that allows efficient classification of datasets with a low SNR and a heterogeneous sample distribution, while minimising any bias and requirements for *a priori* information regarding the protein structure (Scheres, *et al.,* 2005).

The 2D classification job in Relion involves two processes in one, the first involves aligning the particles to a 3D reference to determine the angular orientations of each 2D slice, and the second is the classification of the images. Relion uses a maximum-likelihood multi-reference refinement model to integrate these two processes into a single job (Scheres *et al.,* 2005). In order to determine the angular orientations of each particle image, it is assumed that each image represents a different view of a 3D reference structure, but with the addition of independent Gaussian noise. As Relion uses a multi-reference model, rather than relying on a single 3D reference it allows for the possibility of multiple underlying structures, the number of which is determined by the user. This is described mathematically as:

$$X_i(\phi_i) = A_{k_i} + R_i$$

Where $X_i \in \Re^M$ (X_i in the set of the real parts of M) is the *i*th observed image (*i*=1,…,*N*) of *M* pixels; $A_k \in \Re^M$ is an estimate for the *k*th underlying structure (*k*=1,…,*K*), and $A_{k_i}$is the correct underlying structure for image X_i; ϕ_i defines the transformation (rotation and translation) that maps X_i into A_{k_i} and $X_i(\phi_i)$ represents the transformed image; and $R_i \in \Re^M$ is the model of independent Gaussian noise, with a mean of zero and a standard deviation σ (Scheres *et al.,* 2005).

The 'maximum-likelihood' part of the model refers to the likelihood of observing the dataset ($\mathcal{X}$) given a model with a certain parameter set (Θ) ($P(\mathcal{X}|\Theta)$). Given the relationship between the observed image and an underlying structure described in the equation above, the mutual exclusivity between k_i and ϕ_i, and the fact that maximising the likelihood is the same as maximising the logarithm of the likelihood ($L(\mathcal{X}|\Theta)$), the log-likelihood of observing the entire dataset with the given model can be written as:

$$L(\mathcal{X}|\Theta) = \sum_{i=1}^{N} \log P(X_i|\Theta) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} \int_{\phi} P(X_i|k,\phi,\Theta)P(k,\phi|\Theta)\mathrm{d}\phi$$

Where $P(X_i|k, \phi, \Theta)$ is the probability of observing the image X_i given the chosen underlying structure (k), transformation (ϕ), and parameter set (Θ), and $P(k, \phi|\Theta)$ is the probability density function of k and ϕ (Scheres *et al.,* 2005).

The use of the Gaussian model of noise and subsequent probability density functions is the major advantage of the maximum-likelihood model compared to the cross-correlation (also referred to as least-squares difference) model. Whereas the cross-correlation method produces a single value that provides the highest level of correlation between the observed and predicted data, the maximum-likelihood model provides all of the possible values, weighted by the likelihood of each being correct. This is useful for structures that appear similar in different orientations, where the difference between different views could be obscured by noise (Scheres *et al.,* 2005).

The maximum-likelihood model also decreases the reliance on an initial reference structure, which decreases the bias of the resulting model as well as enabling easier identification of previously uncharacterised structures (Sigworth *et al.,* 1998). However, it does still require some form of reference in order to work. For initial 2D classes, the references are formed by averaging equally sized, random subsets of the provided particle images without attempting any alignments or classifications. This will produce a fairly featureless disc that bears very little resemblance to the final structure, but it allows a starting point for the model. The maximum-likelihood multi-reference model is iterated using an expectation-maximisation algorithm (Scheres *et al.,* 2005) so that any tiny differences in the initial starting reference will be amplified throughout the iterations to produce several distinct classes that represent the underlying structure.

Unlike later processes in Relion, the 2D classification job does not include the gold-standard FSC calculations that would allow the job to run until convergence, so the number of iterations is set by the user. However, this is typically kept at 25 iterations as that appears to be a robust balance between the computational cost of subsequent calculations and the quality of the classes produced.

10.6.3.8 3D Classification

3D Classification in Relion uses the same maximum-likelihood method as the 2D classification job to simultaneously perform alignment and classification assignments. Whereas 2D classification is limited to searching for in-plane alignments, 3D classification searches for alignment values in all orientations in order to generate a series of 3D maps (Scheres *et al.,* 2007). Although the basic principle is the same between 2D and 3D classification, the 3D job requires a much more robust initial reference model than the basic one generated within the 2D classification job. However, once a low-resolution reference map has been provided and the number of classes has been set, the

job is able to run unsupervised, allowing identification of heterogeneous groups within the sample without the requirement of prior knowledge of the different species within the sample (Scheres et al., 2007). The number of classes used for this calculation depends on the heterogeneity of the sample, which is typically not known prior to successful completion of this job, so this parameter requires a trial-and-error approach of simply running the job with different numbers of classes and seeing which works best.

A key component to single particle reconstruction is the Fourier slice theorem. This is the idea that if you take a 2D slice of a 3D object, the Fourier transform of that slice is the same as taking a 2D slice from the Fourier transform of the 3D object perpendicular to the original slice direction (Sigworth, 2016). This makes processing the data much more efficient, which is what makes high-resolution cryo-EM solutions feasible. Cryo-EM particle images are not perfect 2D representations of the 3D object in question; a lot of the signal in each particle image will be obscured by noise. This means that many particle images that represent the same section of the 3D object many not be identical, which leads to difficulty in accurately determining the correct values for the object. However, the maximum-likelihood function, as used for 2D classes, results in multiple solutions for every value that are weighted by their probability (Scheres *et al.,* 2005; Scheres *et al.,* 2007). Particle images with a higher contrast, and therefore a higher signal-to-noise ratio, will also be more heavily weighted in the calculation, leading to a more robust interpretation of potentially inconsistent data.

The main difficulty in single particle 3D reconstructions is determining the projection vector of each particle. Without knowing where the 2D slice appears in the 3D object, the Fourier transform slice cannot be accurately placed in the 3D Fourier transform. The way that projection vectors are typically determined is through a method called projection-matching, where each particle image is compared to every possible view of a provided reference map. Once the orientations of each particle image have been determined, the 2D images are merged to generate a new 3D model. This process is repeated iteratively to generate a more accurate 3D model with every iteration (Scheres *et al.,* 2007).

Random conical tilt reconstructions or subtomogram averaging can provide orientation-angle values directly using calibrated tilts of the specimen stage during data collection, but for standard single particle analysis datasets the distribution of particle orientations is completely random. If the distribution of orientations is not random, for example if the sample experiences preferential orientation due to the constraints of the ice thickness, then it can lead to an effect known as the “missing wedge”, where a section of the 3D reconstruction in real-space is missing due to the missing information in Fourier space (Sigworth, 2016).

10.6.3.9 Ab initio Model Generation

Although the 2D classification job can be run without an initial reference, the 3D classes require a more precise reference model in order to generate models with reasonably accuracy. Many proteins have homologs with already solved structures available in the PDB, and these can be used as an initial model, however, for many proteins there is no prior structural information available and *ab initio* 3D model must be generated from the data at hand. Relion uses a stochastic gradient descent (SGD) method that was developed as part of the cryoSPARC package in order to generate this (Zivanov *et al.,* 2018; Punjani *et al.,* 2017). It works similarly to the maximum-likelihood expectation maximisation iterative process used for 2D and 3D alignments, but it introduces a degree of randomness that allows the model to avoid getting trapped in local maxima. The iterative refinement method involves processing each particle image individually and using the whole set of possible orientations of all particles to determine a model at each iteration. The expectation-maximisation algorithm then iterates to re-determine all of the possible orientations of every particle using the new reference (Scheres *et al.,* 2005; Scheres *et al.,* 2007). The SGD method, however, takes a random selection of particle images from the dataset for each iteration of the algorithm and determines the optimal map for that subset of particles. This approach makes each calculation much quicker and allows a broader sampling of the data, which allows the model to escape any local maxima much easier, although it does not reach the high resolution required for structural determination of the protein so further processing steps are still required (Punjani *et al.,* 2017).

10.6.3.10 Refinements and Postprocessing

Once a 3D model has been generated through *ab initio* SGD methods and 3D classification, it can undergo further refinements and postprocessing to ensure that the model is at the highest resolution possible, and also to ensure that the data has not been over-fitted. The current gold-standard resolution estimates are taken by calculating the resolution for the point where the FSC (Fourier Shell Correlation) value is 0.143. This gold-standard threshold was proposed by Rosenthal and Henderson (Rosenthal and Henderson, 2003), but RELION was the first program to implement it routinely. The FSC value is calculated by separating the data into two random and equal halves and determining a 3D map for either. The similarity between the two maps acts as a measure of how accurate the 3D refinements are. I the two half-maps are identical, then it can safely be assumed that that is the best map for the data. If the two half-maps vary significantly, it can be assumed that the data contains a lot of noise that has been overfitted. FSC curves plot the FSC value as a function of the resolution of the map, as even very noisy maps will resemble each other at a low enough resolution, and the FSC = 0.143 value is taken as a standardised measure of the resolution of the map, independent on the size of the map. An FSC value of 0.143 was chosen as the gold-standard

as it correlates to the point that the C_{ref} value = 0.5, where the C_{ref} represents the correlation between the experimental model using the full dataset and a perfect reference model. The previous commonly used measure for resolution estimation was to take the resolution at the point where the FSC = 0.5, as that is the point where the correlation between the two half-maps is 0.5, which means that the power of the average map is comprised of 50% true signal and 50% true noise. However, that relied too much on the ability to fit each map using only half of the data. Larger datasets provide more signal, so typically result in higher resolution datasets. The FSC = 0.5 value only determined the resolution of a map generated from half the available data. The FSC = 0.143 value was determined by manipulating the equations used to calculate the correlation coefficients and introducing the perfect reference model. Although a perfect reference model does not exist, it can be introduced into the equations to rearrange them enough to calculate that the resolution that leads to C_{ref} = 0.5 also results in FSC = 0.143 (Rosenthal and Henderson, 2003). RELION was the first program to implement the FSC = 0.143 standard for resolution estimation, but it has since become the standard and is also used by other software, including cryoSPARC.

Once the FSC = 0.143 standard had been set, it allowed an automated refinement job to be created, as variables can be adjusted and the resulting resolution can be calculated to determine the limits if the variables without the need for user input to validate the adjustments (Scheres, 2012). Several further refinement and postprocessing steps can be carried out in both RELION and cryoSPARC in order to increase the resolution of the map without changing the determined locations and orientations of the particle images. The image processing process can limit the high-frequency data, particularly during the particle detection and image processing stages. The postprocessing allows sharpening of the model through a restoration of the dampened high-frequency data. The dampening is modelled using a Gaussian distribution, and generating the B-factor. A negative B-factor can then be applied to offset the original dampening to restore the information (Scheres, 2016).

10.7 Appendix 7 – PTM sites in UBE3A

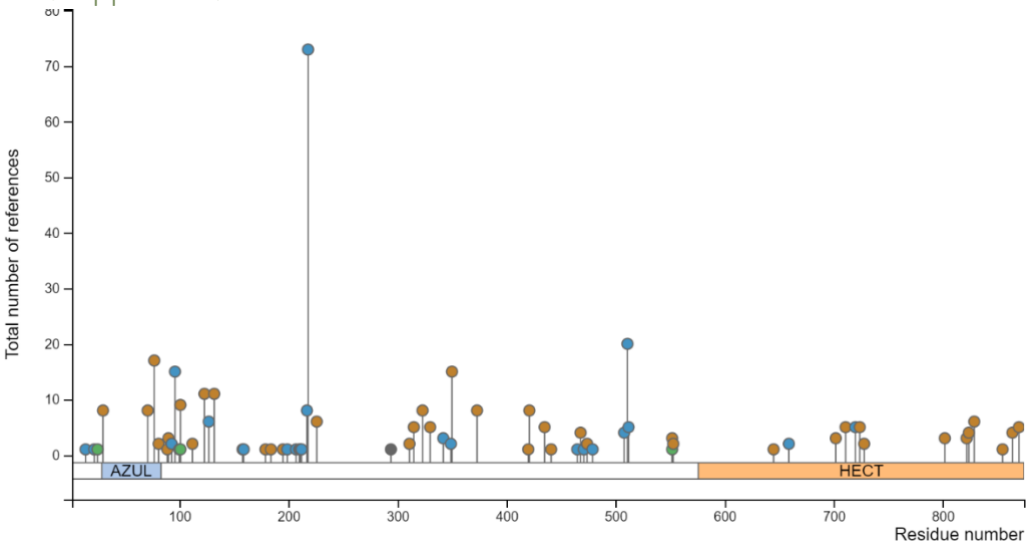


Figure 138: The possible PTM sites in the UBE3A sequence identified by the PhosphoSite Plus server (Hornbeck et al., 2011). Phosphorylation sites are shown in orange, ubiquitination sites in blue, and acetylation sites in green.

10.8 Appendix 8 – HERC2 Gel Electrophoresis

The first attempt to visualise the HERC2 expression and purification consisted of simply running the samples on a commercial 4-20% acrylamide gel with the standard tris-glycine-SDS running buffer that was used for every other protein purification during the project. The hope was that although these gels are not optimised for proteins of this size, they could still work if I subjected them to a longer run time. However, this did not turn out to be the case (Fig. 139).

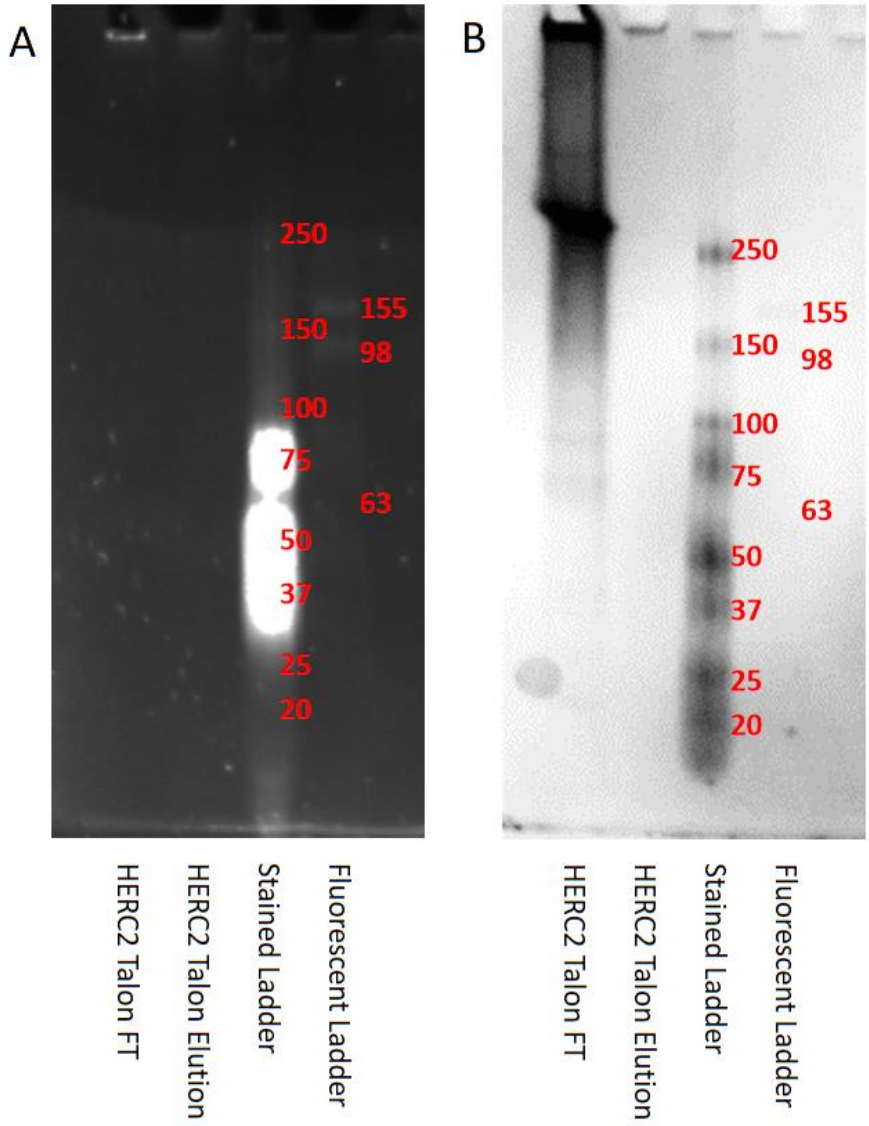


Figure 139: A small scale Talon purification of HERC2 run through PAGE under denaturing conditions. A) A UV image of the gel showing any fluorescent species present. B) Stained with a Coomassie-based dye to show all proteins present.

Although I was not able to resolve the higher molecular species on the denaturing gel, something of a high molecular weight was clearly present in one of the samples. I next attempted to use the same commercial 4-20% acrylamide gels, but with native running conditions (Fig. 140, section 2.7.2). As native gels rely on the isoelectric point of the proteins involved rather than

the size they have been shown to be better at resolving a wider range of protein sizes (Roelofs *et al.,* 2018), and they also allow for identification of GFP-tagged products as the native conditions mean that the GFP moiety is not denatured during electrophoresis.

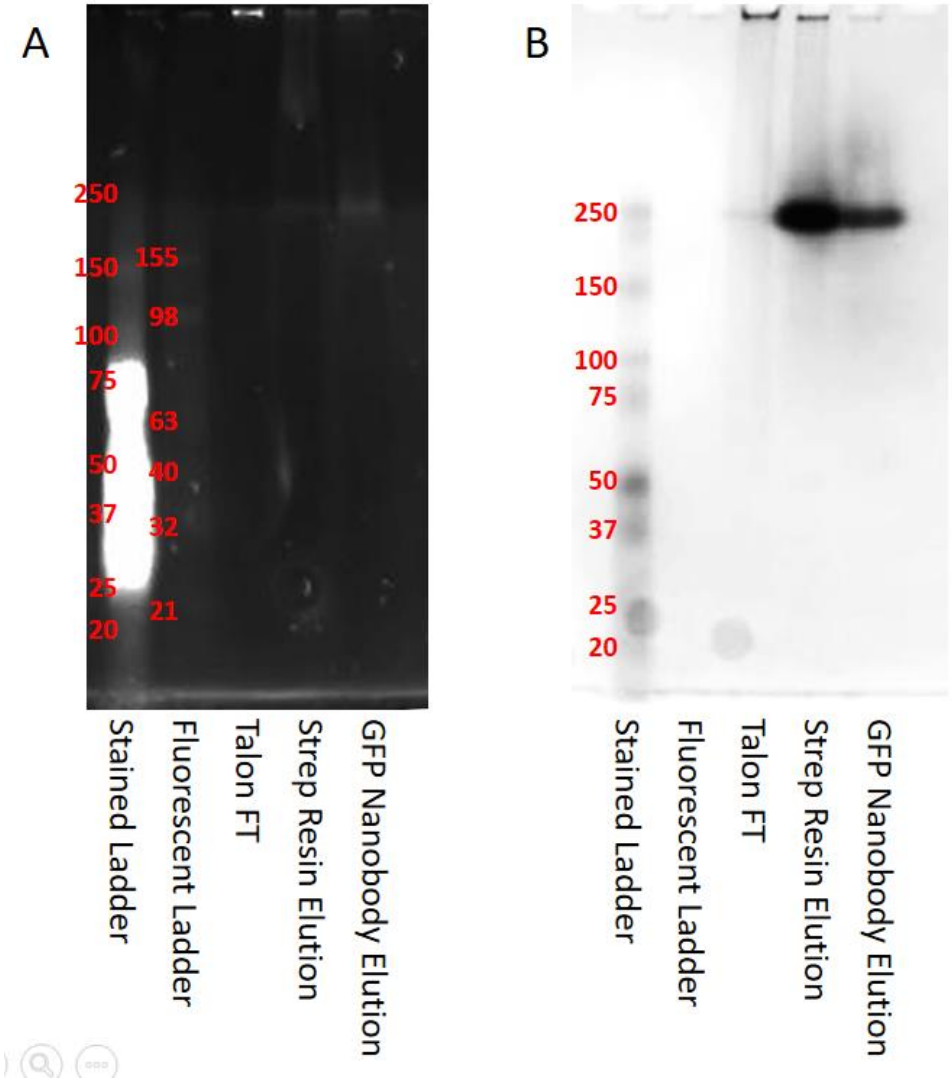


Figure 140: Small scale purifications of HERC2 run through PAGE under native conditions. A) A UV image of the gel showing any fluorescent species present. B) Stained with a Coomassie-based dye to show all proteins present. A small scale mammalian expression of HERC2 was lysed and split into three equal parts which were then purified separately with a talon resin, a strep resin, and a GFP-nanobody purification. The talon elution fraction was subjected to fluorescent size exclusion (FSEC) which did not suggest the presence of HERC2, so the flow through fraction for that purification was run on the gel along with the elutions from the other purification methods.

The native gel did not appear to show the same high molecular weight species as the denaturing gel (fig. 139), but it did show the presence of protein in the various elution samples. The bands at 250 kDa in the elution samples are the wrong size for full-length HERC2, but they are also too small for the isolated

tag region, and the tags are C-terminal in the constructs so it is unlikely that the purification tags have been expressed without the HERC2 protein. The UV image (Fig. 140a) shows a distinction between the gel above and below the 250 kDa marker that reflects the stacking gel and the resolving portion of the gel, so it is possible that the band at 250 kDa represents a species that is stuck in the stacking gel rather than a true 250 kDa sample. None of these bands are particularly fluorescent, which suggests that no GFP is present, but there is something fluorescing in the well of the talon FT sample, which could be full-length GFP-tagged HERC2 that has failed to migrate into the gel at all.

Standard native PAGE gels can be difficult to interpret as the migration of proteins through the gel depends on their shape and charge rather than the absolute size of the species', and the full structure for HERC2 is currently unknown. In an attempt to overcome this the blue-native (BN-PAGE) technique was used. This involves adding Coomassie-G to both the loading dye and the upper reservoir buffer to a typical native PAGE setup, as the slight charge of the Coomassie particles acts to mask the charge effects of the proteins in the sample (Schägger, 2001). This allows separation based on size but without the harsh denaturing effects of SDS-PAGE. Unfortunately, this method was also ultimately unsuccessful in identifying HERC2 bands from the samples provided (Fig. 141).

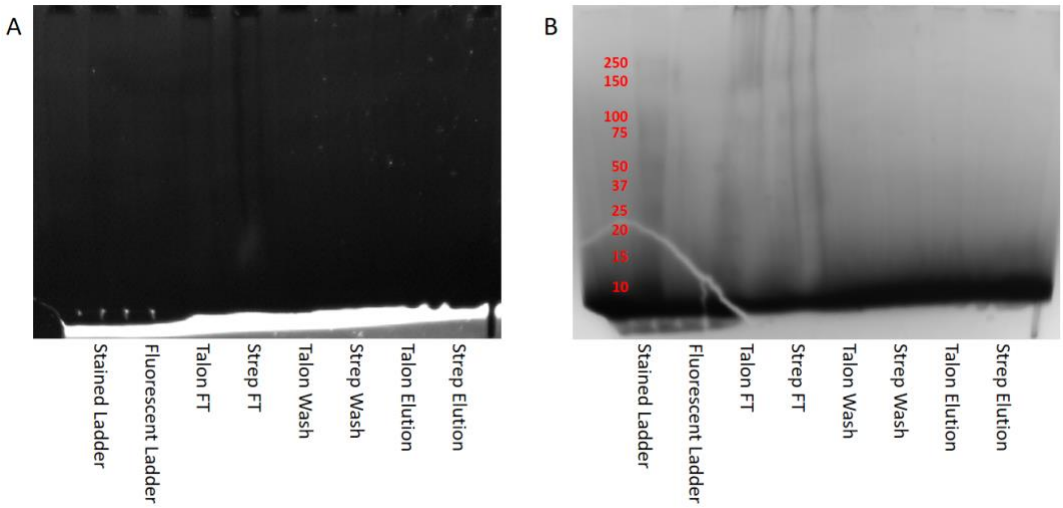


Figure 141: A small scale purification trial of HERC2 subjected to a commercial blue-native PAGE gel. A) A UV image of the gel before destaining. B) An image of the gel after destaining overnight.

The UV image of the BN-PAGE gel (Fig. 141a) does not show any of the samples, including both ladders. This is presumably due to the pervasive Coomassie dye blocking any fluorescence. However, even after destaining (Fig. 141b) nothing is very clear. The samples themselves are not clear on this gel, but the main issue with the BN-PAGE setup was that the stained ladder still does not run further down the gel, which suggests that there is no more

separation of larger molecular weight species' than In the classical native or SDS-PAGE gels.

Another method of resolving larger proteins during gel electrophoresis is a vertical agarose gel method. This involved setting a gels that are very similar to typical acrylamide gels, in terms of the gel buffer and the vertical setup, but with agarose as the gelling agent rather than acrylamide (Greaser and Warren, 2012).

The issues I encountered with this technique were mostly practical issues. Specifically, the combs were difficult to remove from the sticky agarose gel after casting, and the gels had a tendency to float out of the stands during the electrophoresis runs. However, the agarose that was available to us resulted in a fairly standard pore size, whereas separation of proteins in the size range of HERC2 would have required a specialised high pore size agarose that I was not able to acquire. This meant that even once the practical issues had been overcome, I was still not able to observe the separation that I required.

After exhausting all troubleshooting options regarding the vertical agarose gels, I attempted some horizontal agarose gels (Bonifacino *et al.,* 2002). I tried a first gel using the traditional tris-glycine electrophoresis buffers (Fig. 142a), and then further attempts used a tris-borate buffer in an attempt to increase the resolution, in both denaturing (Fig. 142b) and native (Fig. 142c-d) conditions.

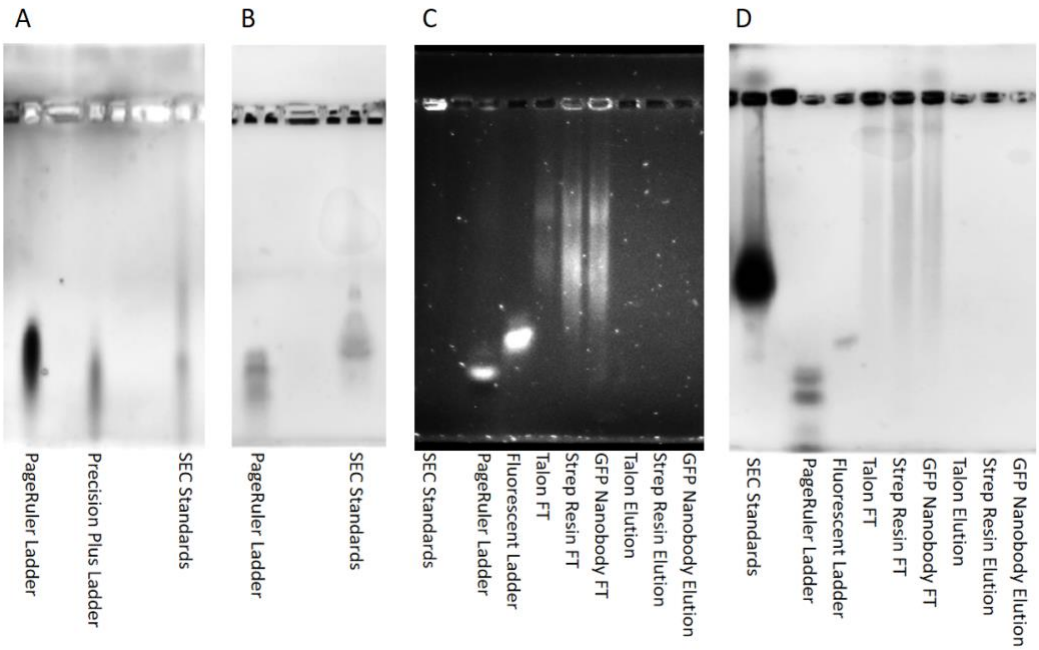


Figure 142: Horizontal agarose gels for large protein visualisation. A) A 1% agarose gel run using a tris-glycine-SDS buffer system, stained with a Coomassie-based dye. Two different stained protein ladders and a mixed SEC standards sample were run on the gel to get an idea of the molecular weight range of the gel. B) A 1% agarose gel using a tris-borate-SDS buffer system, using a stained protein ladder and the mixed SEC standards sample to show the resolving range of the gel, stained with a Coomassie-based dye. C) A 1% agarose gel using a tris-borate native buffer system, imaged using a UV light source. The mixed SEC standards, prestained protein ladder and fluorescent protein ladders are present to show the resolving range of the gel, and small scale HERC2 purification samples are also present. D) The same 1% agarose gel with a tris-borate native buffer system, stained with a Coomassie dye.

None of the protein ladders are resolved enough to show sufficient separation of the bands so it is difficult to accurately assess the resolving range of all of the gels, However, the ranges of the standards are known; the expected coverage the PageRuler ladder contains a range of bands between 10-170 kDa, the Precision Plus ladder covers a range between 11-250 kDa, the fluorescent ladder covers a range between 11-155 kDa, and the SEC standards contain proteins between 13.7 kDa and 670 kDa. The FT samples from the HERC2 purifications show some Coomassie staining (Fig. 142d) and significant fluorescence (Fig. 142c) relatively high in the gel, and it appears to be a similar height to the SEC standard stained area and significantly above any of the protein ladder stains, so it is possible that it represents some full-length GFP-tagged HERC2, even if the small scale purifications haven't worked.

The horizontal agarose gels appeared to show a promising resolving range but the resolution was very poor, so I returned to acrylamide gels, but this time utilising different buffers, acrylamide percentages, and different gel

dimensions in order to try and improve the resolution of proteins across a wider range of sizes. One attempt at finding a gel electrophoresis method capable of resolving very large proteins involved pouring gels with a tris-acetate gel buffer, which were then run using a tris-tricine running buffer. Tris-tricine systems are usually used to improve resolution of extremely small products, but when coupled with low acrylamide percentage tris-acetate gels its theoretical resolution range spanned from the very small proteins to around 600 kDa (Cubillos-Rojas *et al.,* 2010), which is approximately the size of the HERC2 construct. The lower percentage gels resulted in a larger pore size to allow migration of larger proteins, but they also had a much-decreased structural integrity that made handling difficult and resulted in more broken gels.

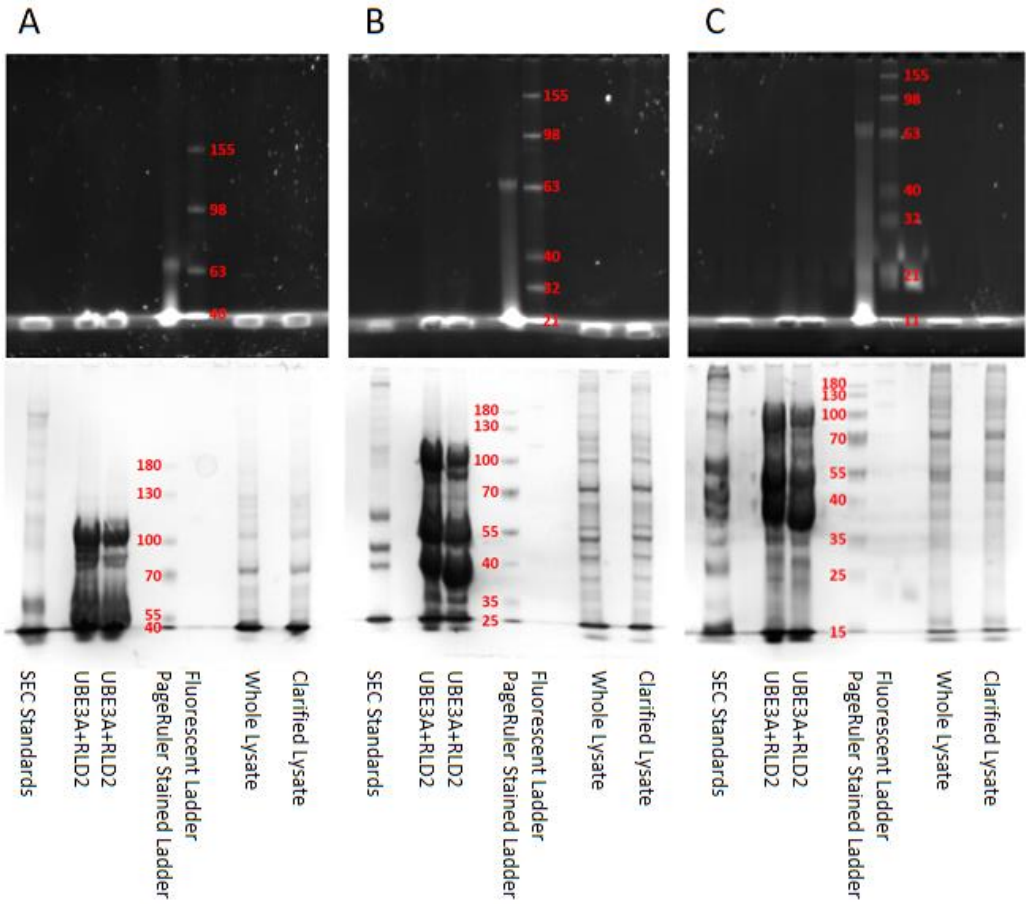


Figure 143: Hand poured BioRad mini-protean tris-acetate gels with different acrylamide percentages. A) 5%, B) 7%, C) 9%. The top row are imaged using a UV light and the bottom row are the same gels stained with a Coomassie-based dye.

Initial attempts at this gel method using the standard BioRad mini-protean II gel casting system were encouraging, with the protein ladder migrating further into the gel than previous attempts while maintaining the resolution (Fig. 143). However, I thought that a larger gel size would be advantageous in increasing the resolution in the higher molecular weight region of the gels. The large gels were first attempted with a 9% acrylamide component, even

though that had the worst migration of the smaller gels, as the larger dimensions made it difficult to handle already. The large 9% tris-acetate gel was trialled in both native and denaturing conditions (Fig. 144).

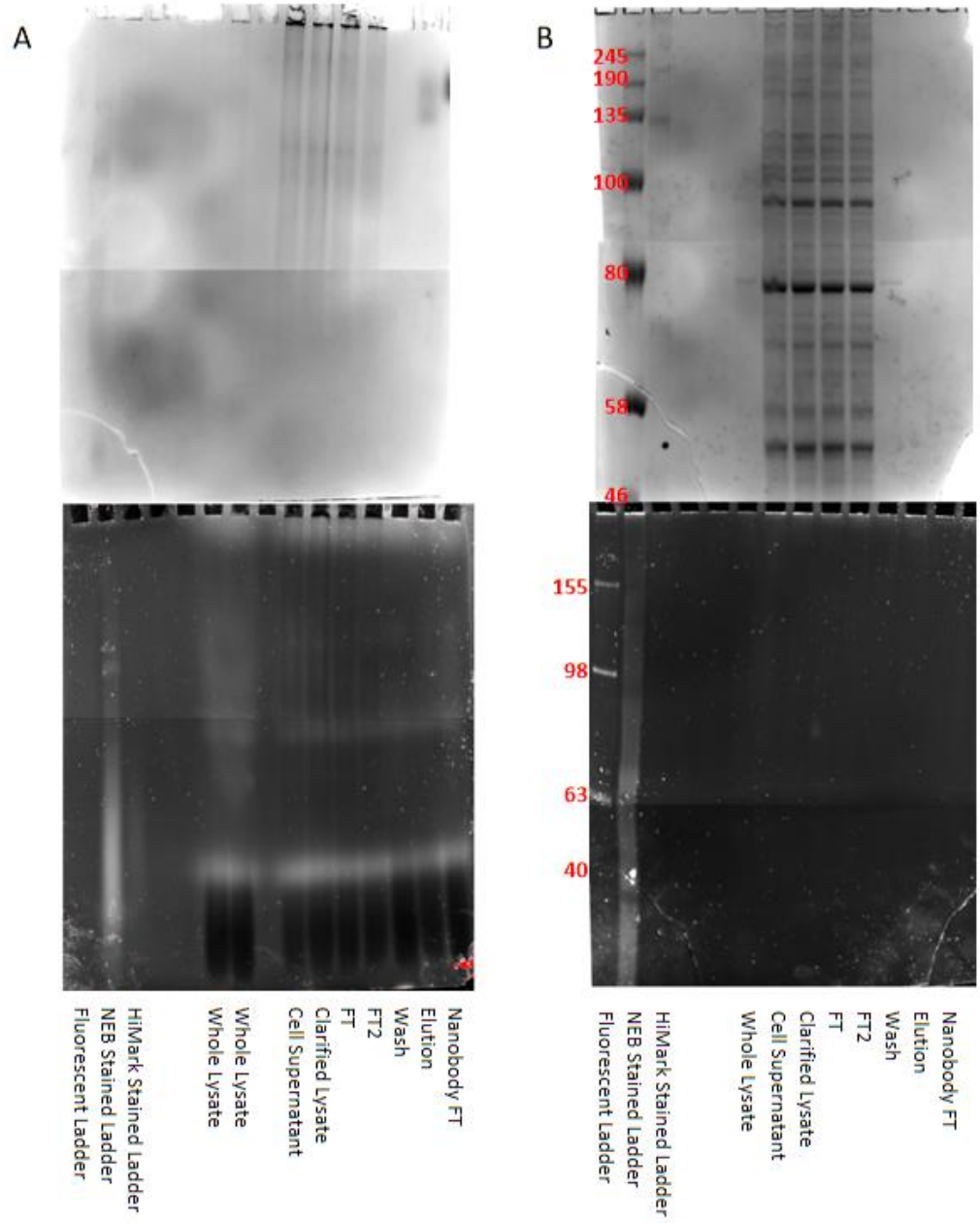


Figure 144: 9% acrylamide large tris-acetate gels with HERC2 GFP-nanobody purification samples. A) The first gel was run under native conditions. B) The second gel was run with SDS in the buffers. Both gels were imaged first with a UV source (bottom row) and then stained with a Coomassie solution and imaged using a standard light source (top row).

The Coomassie-stained native gel (Fig. 144a top) shows something on the GFP-nanobody purification lane, but nothing could be seen in the UV image for the same gel (Fig. 144a bottom). The ladders are also indistinguishable in both images, but the fluorescent ladder at least can be clearly seen in the UV

image for the denaturing gel (Fig. 144b bottom) and it shows that there is not very much migration in the higher molecular weight range. The next gel that I tried was a 5% gel of the same dimensions. Even though the 5% gel was the best of the smaller tris-acetate gels (Fig. 146), it was already fairly difficult to handle and the larger dimensions of these gels would have made it too difficult to attempt such a low acrylamide percentage. The 9% native gel did not show a high resolution so nothing was very clear, but also the samples did not fluoresce under those conditions, so the 6% gel was only attempted under denaturing conditions (Fig. 145).

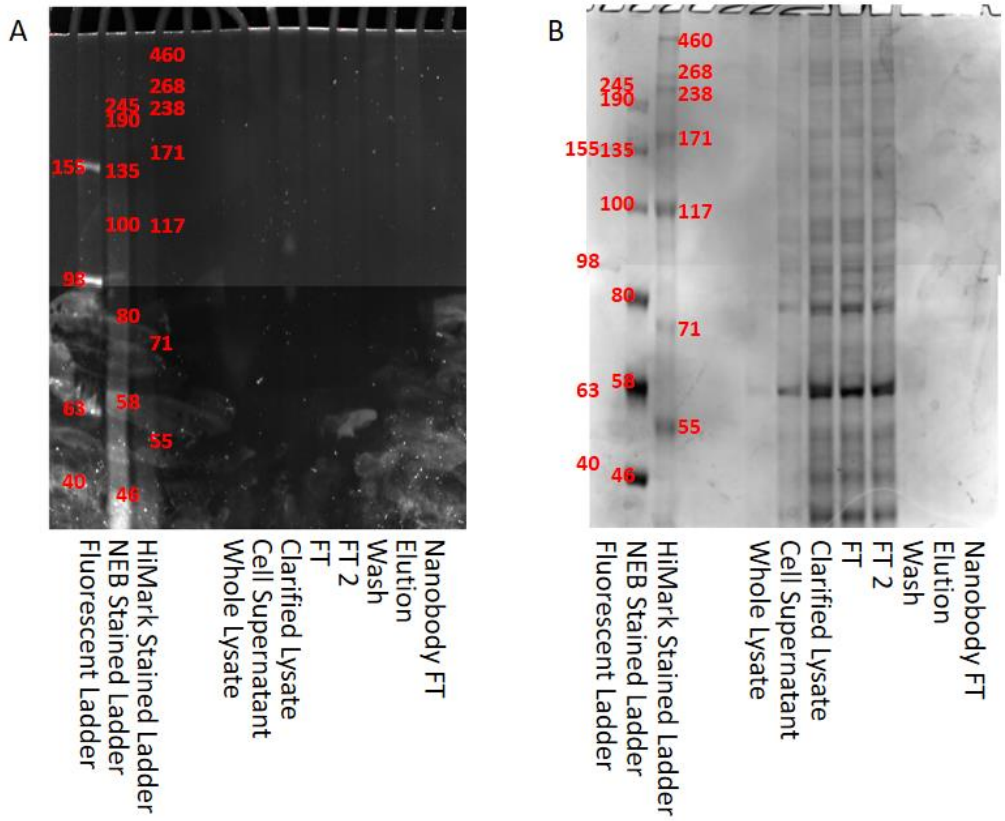


Figure 145: 6% acrylamide large tris-acetate gels with HERC2 GFP-nanobody purification samples. A) The UV image of the gel, B) A Coomassie-stained image of the gel.

None of the HERC2 purification samples produced a visible band in the UV image of the gel, but there was some fluorescence in the wells of several of the samples. This could mean that the GFP product is present in the samples, but it is difficult to determine that definitively. However, the largest band of the fluorescent ladder is still relatively high up in the gel, and there is a large size difference between the top ladder and the predicted HERC2 conjugate, so it is likely the gel is still not suitable for the HERC2 samples.