Imperial College London
Department of Infectious Disease
Section of Synthetic and Structural Biology

Structure-function analysis of protein complexes regulating Rag signalling within the mTOR pathway and structural studies of the ΦKZ non-virion RNA polymerase

Natàlia de Martín Garrido

Supervised by Dr. C.H.S Aylett
Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy, December 2021

Abstract

FLCN mutations are associated with Birt-Hogg-Dubé (BHD) syndrome, a hereditary syndrome which, ultimately, results in kidney cancer. FLCN forms complexes with two folliculin interacting proteins (FNIPs). The FLCN complex is a GAP for RagC/D therefore being a positive modulator of mTORC1. The double nucleotide-bound setting, with the Rag GTPase heterodimer being able to simultaneously bind to both GTP and GDP, is a unique feature to the Rags amongst others GTPases, which suggests that the FLCN-GAP mechanism activity towards RagC/D is likely to be novel.

We have purified the *Spodoptera litura* FLCN-FNIP2 complex and the RagA-RagC complex. Using cryo-EM we have provided a structural 3D model for the *S.litura* FLCN-FNIP2 complex. Comparison of our 3D reconstruction to that of the human FLCN-FNIP2 complex shows that the architecture and domain organisation of the FLCN-FNIP complex is conserved in *S.litura* and that the complex retains the same conformation in the absence of the Rag GTPase heterodimer.

Bacteriophage ΦKZ is one of a few exceptional bacterial viruses to encode two multisubunit RNAPs able to transcribe all genes needed throughout the infection cycle. Previous biochemical studies have demonstrated that the ΦKZ non-virion RNAP is a 5-subunit complex able to initiate transcription from specific promoters. The complex is formed by four β/β'-like subunits and a fifth subunit with no homology to any other protein with known function but crucial for the assembly of a functional enzyme, promoter recognition and transcription initiation.

Here, we have determined for the first time the ΦKZ non-virion RNAP structure at 3.3 Å resolution with and without DNA using cryo-EM. The structures provide first insights into the organisation of the four split β/β'-like subunits and the location of gp68 within the complex. Moreover, structural comparisons show that gp68 is structurally similar and is positioned at the same location as bacterial σ factors.

Declaration

I hereby certify that all the data included in this thesis is my own work, unless otherwise stated.

Natàlia de Martín Garrido

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non Commercial Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Acknowledgements

Firstly, I would like to thank my supervisor, Dr. Christopher H.S. Aylett to give me the opportunity to work on this project. Furthermore, his mentoring and expertise have been of great value for the fulfillment of this project. I would also like to express my gratitude for his guidance through the learning of all technical and practical aspects of electron microscopy required for this project.

I would also like to thank all the members in the Dr. Yakunina's group at Peter the Great St. Petersburg Polytechnic University for providing samples and biochemical experiments which have been crucial for the fulfillment of this project, and for their expertise in the field.

I would also like to express my gratitude to Kailash Ramlaul for all his daily support and valuable discussions and for his initial guidance through the lab. He has been a great colleague both at a professional and personal level, being a great support inside and outside the lab, always willing to listen and provide advise. Additionally, I would like to thank all the other members of the Section of Structural and Synthetic Biology, in particular Ciaran McKeown and Dr. Soo Mei Chee for managing the section, Elizabeth McCormack for managing the lab, and Valerie Good for managing the Structural Biology Baculovirus expression facility at Imperial College.

I also need to thank Paul Simpson for managing the Imperial College Electron Microscopy Centre, for initial training, and for continuous technical support, and eBIC at Diamond Light Source (Didcot, UK) and the Structural Biology Platform at the Crick Institute (London, UK), managed by Nora Cronin, for allowing high-resolution cryo-EM data collection.

At a personal level, I would like to thank all my friends that have been next to me for the past three years and a half, as their support outside the lab have been crucial for keeping up during difficult moments. All my volleyball friends for keeping me fit and distracting me with amazing hours on the court, as well as the drinks that followed the training sessions and matches. And a especial mention to my Imperial College crew, Ignacio, Albert and Cesc for all the lunch breaks, coffees and after-work drinks, and Lluís, for listening to me for hours and hours, either with a beer, through voice messages or in a casual chat in our cosy kitchen. And my friends back home, specially my Picolinas and Se la lleva, my uni friends, where this journey started.

Finally, I would like to thank in Catalan my family and my partner, Marc, for their constant support, faith and patience through the fulfillment of this project:

En primer lloc, vull agrair als meus pares, Jordi i Esther, el seu constant suport, tot i estar lluny, i per sempre encoratjar-me a continuar endavant malgrat les dificultats que puguin aparèixer. Sense ells, emprendre aquest camí no hauria estat possible. Gràcies, Putxis, per fer-me la dona que soc i per donar-me l'educació i els valors que m'han fet arribar fins aquí.

També m'agradaria agrair a altres membres de la meva família que sempre han cregut en mi: el meu germà Marc, tiets i tietes, cosins i cosines i, sobretot, les meves dues àvies, Maria Teresa i Maria Rosa, dues dones exemplars, i els meus avis, Joan i Diego, que segur que estarien molt orgullosos del meu gran assoliment.

Finalment, no puc estar més agraïda a la meva parella, en Marc, que ha estat al meu costat en tot moment, recolzant-me, cuidant-me i ajudant-me en tot el possible fora del laboratori, fins i tot en els moments més difícils durant la pandèmia causada per la COVID-19. Ha estat un pilar fonamental per la realització d'aquest projecte i pel meu benestar, sempre donant-me forces i energia per continuar esforçant-me dia a dia. Gràcies, gràcies i mil gràcies.

Publications

Publications directly related to this work

de Martín Garrido N, Orekhova M, Ting Emilie Lai Wan Loong Y, et al. Structure of the Bacteriophage PhiKZ non-Virion RNA Polymerase. Nucleic Acids Research. 2021; 13, 7732-7739

de Martín Garrido N, Aylett CHS. Nutrient Signaling and Lysosome Positioning Crosstalk Through a Multifunctional Protein, Folliculin. Front Cell Dev Biol. 2020; 8, 108

Other relevant publications

de Martín Garrido N, Ramlaul K, Aylett CHS. Preparation of Sample Support Films in Transmission Electron Microscopy using a Support Floatation Block. J Vis Exp. 2021; 170, e62321

Ramlaul K, Fu W, Li H, **de Martín Garrido N**, He L, Trivedi M, Cui W, Aylett CHS, and Wu G. (2021). Architecture of the Tuberous Sclerosis Protein Complex. J. Mol. Biol. 2021; 433 (2), 166743

de Martín Garrido N, Fu W, Ramlaul K, et al. Direct transfer of electron microscopy samples to wetted carbon and graphene films via a support floatation block. J Struct Biol. 2021; 213 (1), 107677

de Martín Garrido N, Crone MA, Ramlaul K, Simpson PA, Freemont PS & Aylett CHS. Bacteriophage MS2 displays unreported capsid variability assembling T = 4 and mixed capsids. Mol Microbiol. 2019; 113, 143–152

Contents

Abstract **i**

Declaration **i**

Acknowledgements **v**

Publications **vii**

1 Introduction **1**

1.1 Folliculin, a multi-functional protein involved in nutrient signaling and lysosomal positioning 1

1.1.1 The mTORC1 signaling pathway 2

1.1.2 The Rag GTPases communicate nutrient availability to mTORC1 depending on their nucleotide binding state 5

1.1.3 The Folliculin complex and its interacting partners 9

1.1.3.1 Structural organisation of the FLCN complex 9

1.1.3.2 Folliculin acts as a GTPase activating protein to modulate the Rag GTPase nucleotide binding state 12

1.1.3.3 Other FLCN interacting partners: the AMP-activated protein kinase and the Rab GTPases . . . 13

1.2 Bacteriophage ΦKZ; a giant virus encoding its own transcriptional machinery . . 18

1.2.1 Bacteriophage biology and medical interest . . . 18

1.2.2 Bacteriophage ΦKZ . . . 20

1.2.3 The transcriptional apparatus of bacteriophage ΦKZ . . . 23

1.2.3.1 DNA-dependent RNA polymerase, the central enzyme of transcription . . . 23

1.2.3.2 Bacteriophage ΦKZ encodes its own multi-subunit RNAPs . . . 27

1.2.3.3 The ΦKZ non-virion RNA polymerase . . . 29

1.3 Objectives . . . 33

2 Materials and Methods **35**

2.1 General Procedures . . . 35

2.1.1 PCR . . . 35

2.1.2 *E.coli* transformation . . . 36

2.1.3 Golden Gate Reaction . . . 36

2.1.4 HiFi assembly . . . 37

2.1.5 Agarose Gel Electrophoresis . . . 38

2.1.6 SDS-PAGE . . . 38

2.1.7 Native PAGE and Gel-shift . . . 39

2.1.8 Mass spectrometry . . . 39

2.1.9 General Protocols for the preparation of EM grids 39

2.1.9.1 Grid washing and plasma cleaning 39

2.1.9.2 Preparation of Graphene Oxide covered grids 40

2.2 The FLCN complex and its interacting partners 41

2.2.1 Materials 41

2.2.1.1 Plasmids 41

2.2.1.2 Genes and proteins 41

2.2.2 Protein Expression 42

2.2.2.1 Virus generation and amplification 42

2.2.2.2 Establishment of protein expression conditions 43

2.2.3 Purification of the *Sl* FLCN-FNIP2 complex 43

2.2.4 Purification of the RagA-RagC complex 45

2.2.5 FLCN-FNIP2-RagA-RagC complex formation 45

2.2.5.1 Size-exclusion chromatography 45

2.2.5.2 Density gradients: GraFix and Prefixation 46

2.2.6 GTPase assays 47

2.2.6.1 Malachite green Phosphate assay kit 47

2.2.7 GTPase-Glo^{TM} Assay 48

2.2.8 Single-Particle Electron Microscopy 49

2.2.8.1 Negatively-stained specimens 49

2.2.8.2 Vitreous specimens 50

2.2.8.3 Data Collection and Image Processing 50

2.2.9 Structural model generation 51

2.3 The ΦKZ non-virion polymerase 52

2.3.1 ΦKZ non-virion RNAP expression and purification 52

2.3.1.1 Plasmids 52

2.3.1.2 Protein expression and purification 52

2.3.2 DNA-template preparation 54

2.3.3 Mass spectrometry 55

2.3.4 Electrophoretic mobility shift assay and photo-crosslinking 55

2.3.5 Native PAGE and *in vitro* transcription 56

2.3.6 Limited proteolysis 56

2.3.7 Single-Particle Electron Microscopy for the ΦKZ nvRNAP complex . . . 57

2.3.7.1 Cryo-EM sample preparation 57

2.3.7.2 Calculation of reference density 57

2.3.7.3 Data Collection and Image Processing 58

2.3.7.4 Modelling and refinement 59

2.3.8 Single particle Cryo-EM for ΦKZ nvRNAP complexes with the P119L promoter 61

2.3.8.1 Cryo-EM sample preparation 61

2.3.8.2 Screening of vitrification conditions 62

2.3.8.3 High-resolution Data Collection and Image Processing 63

2.3.8.4 Modelling and refinement 64

3 The Folliculin complex and its interacting partners 67

3.1 Structural studies of the *Sl* FLCN-FNIP2 complex 68

3.1.1 Expression and purification of the *Sl* FLCN-FNIP2 complex 68

3.1.2 Screening of vitrification conditions 73

3.1.3 High-resolution data collection using a Volta phase plate and image processing 75

3.1.4 Architecture of the *Sl* FLCN-FNIP2 complex 80

3.2 Preliminary structural studies of the *Sl*FLCN-FNIP2-RagA-RagC complex . . . 82

3.2.1 Expression and purification of the *Sl* RagA-RagC complex 82

3.2.2 Optimisation of *Sl* FLCN-FNIP2-RagA-RagC complex formation 85

3.3 Discussion 92

4 Structure of the ΦKZ non-virion RNA polymerase 99

4.1 Preparation of ΦKZ non-virion RNA polymerase complexes 100

4.2 Optimisation of vitrification conditions for the ΦKZ nvRNAP 102

4.3 High-resolution data collection, image processing and model building for the ΦKZ nvRNAP 105

4.4 Architecture of the ΦKZ nvRNA polymerase 112

4.5 Structural comparison with bacterial multisubunit RNAPs 115

4.6 Gp68 is involved in the assembly of a functional nvRNAP 119

4.7 Gp68 is structurally and functionally similar to bacterial σ factors 121

4.8 Discussion 125

5 Structure of the ΦKZ non-virion RNAP bound to its promoter P119L 130

5.1 Preparation of ΦKZ nvRNAP complexes with the late promoter P119L promoter 131

5.2 Preliminary structural analysis of the ΦKZ nvRNAP complex bound to the P119L promoter 136

5.3 High-resolution data collection and model building 139

5.4 Architecture of the ΦKZ nvRNAP bound to the KZP119L promoter 147

5.5 Organisation of the ΦKZ nvRNAP catalytic core 154

5.6 The role of Gp68 156

5.7 Structural comparison of the ΦKZ and AR9 nvRNAPs 160

5.8 Discussion 166

6 Conclusions and outlook 172

6.1 Conclusions 172

6.1.1 The Folliculin complex and its interacting partners 172

6.1.2 The ΦKZ non-virion RNAP 173

6.2 Future Work 175

6.2.1 The Folliculin complex and its interacting partners 175

6.2.2 The ΦKZ non-virion RNAP 176

Bibliography 177

Appendices 212

List of Tables

2.1 Volumes and final concentrations used for standard PCR reactions 35

2.2 Conditions used to run standard PCR reactions 36

2.3 Standard volumes used to perform Golden Gate reactions 37

2.4 List of protein constructs used for *Sl* FLCN-FNIP2 and RagA-RagC purification 42

2.5 List of buffers used for *Sl* FLCN-FNIP2 purification 44

2.6 List of buffers used for *Sl* RagA-RagC purification 45

2.7 Summary of data collection parameters for the *Sl* FLCN-FNIP2 high-resolution data using a *Volta phase plate* (VPP) 50

2.8 List of buffers used for the ΦKZ nvRNAP purification and biochemical assays . 53

2.9 List of oligonucleotides used for biochemical assays with the ΦKZ nvRNAP purification and biochemical assays 54

2.10 List of oligonucleotides based on the ΦKZ P119L promoter used for cryo-EM . 55

2.11 Summary of data collection parameters for the ΦKZ nvRNAP data collection for calculation of reference density 58

2.12 Summary of high-resolution data collection parameters for the ΦKZ nvRNAP . 59

2.13 Summary of the coverage of the ΦKZ nvRNAP structural model 60

2.14 Model refinement statistics for the ΦKZ nvRNAP structure 61

2.15 Summary of collection parameters for the ΦKZ nvRNAP complexes bound to P119L RNA/DNA (initial model) 63

2.16 Summary of high-resolution data collection parameters for the ΦKZ nvRNAP bound to the RNA/DNA scaffold P119L template 64

2.17 Summary of the coverage of the ΦKZ nvRNAP complex bound to P119L RNA/DNA structure 65

2.18 Model refinement statistics for the structure of the ΦKZ nvRNAP bound to the P119L RNA/DNA promoter 66

4.1 Buried surface area of the gp68 interaction with gp55 and gp71-73 114

4.2 Structural similarity of the ΦKZ nvRNAP subunits and bacterial msRNAPs structures 116

4.3 Comparison of the buried surface area of nvRNAP gp68 and σ factor A from *M. tuberculosis* 124

List of Figures

1.1 The mTORC1 signaling pathway is regulated by different stimuli 4

1.2 The structure of the Rag GTPase heterodimer 6

1.3 Architecture of the mTORC1 complex at the lysosome 9

1.4 Structural biology of the FLCN and the FNIP proteins 11

1.5 Lysosomal positioning in response to nutrient availability 16

1.6 Bacteriophage ΦKZ morphology and pseudo-nucleus maturation in *P.aeruginosa cells* 22

1.7 Structure of the catalytic centre of the "two-barrel" RNAPs family formed by interacting DPBBs from the β and β' subunit 25

1.8 Architecture of bacterial msRNAPs 27

1.9 Bacteriophage ΦKZ transcription strategy 29

1.10 Schematic diagram of the conserved sequence and structural regions of the ΦKZ nvRNAP subunits 30

1.11 Middle and late promoters of the ΦKZ nvRNAP subunits 31

2.1 The pG plasmid map 41

2.2 Malachite Green GTPase assay Phosphate Standards curve 48

3.1 Sequence alignment of *S.litura* and human FLCN 69

3.2 Sequence alignment of *S.litura* and human FNIP2 70

3.3 Affinity purification and mass spectrometry of the FLCN-FNIP2 complex 72

3.4 Size-exclusion chromatography of the FLCN-FNIP2 complex. 73

3.5 Negatively-stained grids of the FLCN-FNIP2 complex 74

3.6 Cryo-EM screening of FLCN-FNIP2 samples 75

3.7 High-resolution data collection using a Volta phase plate of the FLCN-FNIP2 complex 76

3.8 Data processing scheme for the cryo-EM high-resolution data collection of the FLCN-FNIP2 complex 77

3.9 Cryo-EM reconstruction of the *Sl* FLCN-FNIP2 complex 78

3.10 Structural comparison of the *Sl* FLCN-FNIP2 3D reconstruction with the human FLNC-FNIP2 structure 79

3.11 Architecture of the *Sl* FLCN-FNIP2 complex 80

3.12 Purification of the RagA-RagC heterodimer 83

3.13 FLCN-FNIP2 GAP activity towards the RagA-RagC GTPase heterodimer . . . 85

3.14 FLCN-FNIP2-RagA-RagC complex formation in the presence of GTPγS 87

3.15 Negative staining screening of the *Sl* FLCN-FNIP2-RagA-RagC complex formation in the presence of GTPγS 87

3.16 Cross-linking screening and density gradient fixation of the FLCN-FNIP2-RagA-RagC complex in the presence of GTPγS 88

3.17 Screening of FLCN-FNIP2-RagA-RagC negatively-stained grids after density gradient fixation 89

3.18 FLCN-FNIP2-RagA-RagC complex formation in the presence of GppNHp and GDP 91

3.19 Organisation of the FLCN-FNIP2 complex at the lysosomal membrane 93

3.20 The Rag GTPase heterodimer nucleotide states 95

4.1 Size-exclusion chromatography of the ΦKZ nvRNAP 100

4.2 Screening of vitrification conditions for the ΦKZ nvRNAP 103

4.3 Initial reference-free three-dimensional reconstruction of the ΦKZ nvRNAP . . . 104

4.4 High-resolution data collection of ΦKZ nvRNAP complexes 105

4.5 Data processing scheme for the cryo-EM high-resolution data collection for the ΦKZ nvRNAP complex 106

4.6 High-resolution 3D reconstruction of the ΦKZ nvRNAP core 107

4.7 High-resolution 3D reconstruction of the ΦKZ nvRNAP including the clamp . . 108

4.8 Representative high-resolution density for the ΦKZ nvRNAP 3 Å map 109

4.9 Building a molecular model for the ΦKZ nvRNAP 110

4.10 Ramachandran plot of the ΦKZ nvRNAP structure including the clamp 111

4.11 Architecture of the ΦKZ nvRNAP 112

4.12 Low resolution features of the ΦKZ nvRNAP reconstruction 113

4.13 Integration of gp68 in the ΦKZ nvRNAP complex 114

4.14 Structural comparison of the ΦKZ nvRNAP with bacterial msRNAPs. 115

4.15 The catalytic core of the ΦKZ nvRNAP 116

4.16 Architectural elements of the ΦKZ nvRNAP gp74 subunit 117

4.17 Architecture of the ΦKZ nvRNAP clamp 118

4.18 Comparison between the ΦKZ nvRNAP and the bacterial flap 118

4.19 Gp68 is necessary for the assembly of an active nvRNAP 119

4.20 Interaction of gp68 with the universally conserved motif βa15 120

4.21 Gp68 of the ΦKZ nvRNAP is involved in DNA binding 122

4.22 Gp68 is positioned identically to eubacterial σ-factor 123

4.23 Gp68 is structurally similar to eubacterial σ-factor 123

4.24 Electrostatic potential surface of the ΦKZ nvRNAP 124

5.1 Analysis of the ΦKZ nvRNAP binding to the P119L-4nt DNA template 132

5.2 Gel filtration profile of the ΦKZ nvRNAP complexes bound to the P119L-4nt DNA template. 132

5.3 Data collection of ΦKZ nvRNAP complexes bound to the P119L-4nt DNA template. 133

5.4 Data processing scheme for the ΦKZ nvRNAP complexes and P119L-4nt DNA template. 134

5.5 Analysis of the ΦKZ nvRNAP binding to the P119L RNA/DNA template . . . 135

5.6 Initial structural analysis of ΦKZ nvRNAP complexes bound to the P119L RNA/DNA template 136

5.7 Data processing scheme for initial cryo-EM data collection of ΦKZ nvRNAP complexes bound to the P119L RNA/DNA template 137

5.8 Initial comparison of the ΦKZ nvRNAP-DNA reconstruction with an *E.coli* transcribing complex 138

5.9 High-resolution data collection of the ΦKZ nvRNAP in complex with an RNA/DNA P119L template 139

5.10 Data processing scheme for the high-resolution cryo-EM data collection of the ΦKZ nvRNAP complex bound to the P119L RNA/DNA template 141

5.11 High-resolution 3D reconstruction of the ΦKZ nvRNAP bound to the P119L RNA/DNA template 142

5.12 Representative high-resolution density of the ΦKZ nvRNAP-P119L RNA/DNA template 3.3 Å map 143

5.13 Ramachandran plot of the ΦKZ nvRNAP structure bound to the P119L RNA/DNA template 144

5.14 New features of the high-resolution 3D reconstruction of the ΦKZ nvRNAP bound to the P119L RNA/DNA template 145

5.15 The upstream DNA and the RNA of the the ΦKZ nvRNAP bound to the P119L RNA/DNA template 146

5.16 Structure of the ΦKZ nvRNAP complex bound to the P119L RNA/DNA template and structural comparison to an *E.coli* transcribing complex 148

5.17 The ΦKZ nvRNAP trigger loop 149

5.18 Structural comparison of the ΦKZ nvRNAP and *E.coli* RNAP structural elements included in the clamp. 150

5.19 The ΦKZ gp123 middle domain interacts with the downstream DNA 151

5.20 The ΦKZ nvRNAP and *M.tuberculosis* RNAP $\beta 2$ domain comparison 151

5.21 Comparison of the ΦKZ nvRNAP and the *E.coli* RNAP trigger loop insertions. 152

5.22 The catalytic centre of the ΦKZ nvRNAP complex bound to the P119L RNA/DNA promoter 154

5.23 The gp68 architecture and position within the complex is similar to bacterial σ factors 156

5.24 Structural comparison of the gp68 CTD with bacterial σ^4 region 157

5.25 Structural comparison of the gp68 linker with region 3 of bacterial σ factors . . 158

5.26 Structural comparison of the gp68 NTD with bacterial σ region 2 159

5.27 Structural comparison of the ΦKZ and AR9 nvRNAPs 161

5.28 Structural comparison of the ΦKZ and AR9 nvRNAP trigger loop insertions . . 162

5.29 Sequence alignment of the ΦKZ and AR9 nvRNAP promoter recognition subunits 163

5.30 Structural comparison of the ΦKZ and AR9 nvRNAP promoter recognition subunit CTDs 164

5.31 Structural comparison of the ΦKZ and AR9 nvRNAP promoter recognition subunit NTDs 165

List of Abbreviations

4E-BP eukaryotic translation initiation factor 4E-binding protein

4s four-subunit

5s five-subunit

AMPK AMP-activated protein kinase

BHD Birt-Hogg-Dubé

BLAST Basic Local Alignment Search Tool

cryo-EM cryogenic electron microscopy

CTD C-terminal domain

CV column volumes

DDM n-Dodecyl-beta-Maltoside

DENN Differentially expressed in neoplastic versus normal cells

DEPTOR DEP domain containing mTOR interacting protein

DPBB double psi beta-barrel

dsDNA double-stranded DNA

ECF extracytoplasmic function

EM Electron Microscopy

EMSA Electrophoretic mobility shift assay

FL2 fork loop 2

FLCN folliculin

FNIP Folliculin interacting proteins

GAP GTPase activating protein

GATOR GAP activity towards Rags

GEF Guanine exchange factor

GrOx Graphene Oxide

GTA glutaraldehyde

HTH helix-turn-helix

KICSTOR KPTN IKTFG2 C12orf66 and SZT2

LAMTOR Late Endosomal/Lysosomal Adaptor MAPK and mTORC activator/regulator

LB Lysogeny broth

Lst Lethal with Sec13

min minutes

mLST8 mammalian lethal with Sec13 protein 8

MS mass-spectrometry

msRNAP multi-subunit RNAP

mTOR mammalian TOR

NBD nucleotide binding domain

NTD N-terminal domain

nvRNAP non-virion RNAP

PCRs polymerase chain reactions

PIKKs PI3K-related protein kinases

PRAS40 proline-rich Akt substrate of 40 KDa

R_h hydrodynamic radius

Rag GTPase Ras-related GTP-binding protein

Ragulator Rag and mTORC1 regulator

RAPTOR regulatory-associated protein of mTOR

Rheb Ras homolog enriched in brain

RLU relative light units

RNAP DNA-directed RNA polymerase

s seconds

S6K ribosomal protein S6 kinase

SAMTOR S-adenosylmethionine sensor upstream of mTORC1

SDS-PAGE Sodium dodecyl sulphate–polyacrylamide gel electrophoresis

SEC Size-exclusion chromatography

SGD stochastic gradient descent

ssDNA single-stranded DNA

ssRNAP single-subunit RNAP

SW3 switch 3

SWI switch region I

SWII switch region II

TEM transmission electron microscope

TFE3 Transcription factor binding to IGHM Enchancer 3

TFEB Transcription factor EB

TOR Target of Rapamycin

TSC tuberous sclerosis complex protein

TSCC TSC complex

UV ultraviolet

VPP Volta phase plate

vRNAP virion RNAP

Chapter 1

Introduction

1.1 Folliculin, a multi-functional protein involved in nutrient signaling and lysosomal positioning

The FLCN gene has been identified as responsible for *Birt-Hogg-Dubé* (BHD) syndrome, an hereditary renal cancer syndrome, which is associated with the development of renal oncocytomas, a specific histological type of kidney cancer (Toro et al., 1999). According to the Global Cancer Observatory, in 2020 there were more than 400 000 cases of kidney cancer diagnosed and over 170 000 deaths. The estimated risk of people with BHD developing kidney cancer is between 15-30 % (ttps://www.cancer.net/cancer-types/birt-hogg-dube-syndrome). FLCN has been identified as a tumour suppressor (Vocke et al., 2005; Birt et al., 1977) as most of the mutations found in the *FLCN* gene in BHD patients result in truncation of the protein (i.e. insertion/deletion, nonsense open reading frames, or the loss of proper mRNA splicing), therefore causing loss of function (Lim et al., 2010). The *FLCN* gene encodes the protein *folliculin* (FLCN). The identification of *FLCN* mutations as the cause of BHD-associated tumours led to an increase in studies investigating the loss of FLCN function in kidney cancer, revealing that mutations affecting this protein led to distinct phenotypes depending on the cellular context.

Hence, FLCN is a multi-functional protein involved in several cellular processes, such as energy and nutrient homeostasis, membrane trafficking and lysosomal biogenesis and positioning (Schmidt and Linehan, 2018).

FLCN forms a complex with two larger *Folliculin interacting proteins* (FNIP)s, FNIP1 (Baba et al., 2006) and FNIP2 (Hasumi et al., 2008b; Takagi et al., 2008). The same behaviour has been found in yeast, where respective orthologues *Lethal with Sec13* (Lst)7 and Lst4 also form a stable complex (Pacitto et al., 2015). Due to their close relationship with FLCN, FNIPs have also been suggested to act as tumour suppressors (Hasumi et al., 2015; Mo et al., 2019), however further studies are needed to support this. Structural studies have reported that both FLCN and FNIP contain both longin and *Differentially expressed in neoplastic versus normal cells* (DENN) domains (Nookala et al., 2012; Zhang et al., 2012; Pacitto et al., 2015). These protein folds are associated with the regulation of small GTPases. Consistently, functional studies have shown that the FLCN-FNIP complex is involved in the modulation of the Rag and Rab GTPases (Dodding, 2017; Schmidt and Linehan, 2018), which in turn regulate the mTORC1 signaling pathway, and membrane trafficking and lysosomal positioning, respectively.

1.1.1 The mTORC1 signaling pathway

The *Target of Rapamycin* (TOR) was first identified in *Saccharomyces cerevisiae* (Heitman et al., 1991) and few years later the mammalian homologue, *mammalian TOR* (mTOR) was identified by two different research groups (Brown et al., 1994; Sabatini et al., 1994). mTOR is the catalytic core of two distinct complexes associated with different accessory proteins that perform different functions (Kim et al., 2002; Loewith et al., 2002; Sarbassov et al., 2004; Shimobayashi and Hall, 2014; Laplante and Sabatini, 2012): mTORC2, which is associated with cell survival, metabolism and cytoskeletal functions, and mTORC1, a central regulator of metabolism and cell growth (Rabanal-Ruiz and Korolchuk, 2018; Oh and Jacinto, 2011). The central components of mTORC1 are mTOR, which provides the catalytic core, *regulatory-associated protein of mTOR* (RAPTOR) (Kim et al., 2002; Loewith et al., 2002; Hara et al., 2002), and *mammalian lethal with Sec13 protein 8* (mLST8) (Chen et al., 2003; Kim et al., 2003),

while *proline-rich Akt substrate of 40 KDa* (PRAS40) and *DEP domain containing mTOR interacting protein* (DEPTOR) complete the complex but have a mainly regulatory function (Peterson et al., 2009; Haar et al., 2007; Oshiro et al., 2007). The mTORC1 signaling pathway ensures that the cell only grows under optimal conditions, hence its dysregulation causes several metabolic disorders and cancers (Saxton and Sabatini, 2017; Mossmann et al., 2018; Guertin and Sabatini, 2007). mTORC1 is a member of the *PI3K-related protein kinases* (PIKKs) family with serine/threonine kinase activity towards target substrates. When conditions are favourable, mTORC1 promotes anabolic processes, such as cell proliferation or cell growth, through the phosphorylation of *ribosomal protein S6 kinase* (S6K) and *eukaryotic translation initiation factor 4E-binding protein* (4E-BP) (Rabanal-Ruiz and Korolchuk, 2018; Kim and Guan, 2019).

The level of activation of mTORC1 depends on two different pathways, which are regulated by three small GTPases. The current understanding is that any of the pathways is sufficient to activate mTORC1 independently: mTORC1 is recruited to the lysosome by the *Ras-related GTP-binding protein* (Rag GTPase) heterodimer where is fully activated by GTP-bound *Ras homolog enriched in brain* (Rheb) (Sancak et al., 2008; Long et al., 2005; Ramlaul and Aylett, 2018; Rabanal-Ruiz and Korolchuk, 2018). Multiple signals are integrated to finely regulate mTORC1 activation. It is believed that signals corresponding to the intracellular nutrient state, such as amino acid availability, are usually integrated through the Rag GTPase axis. On the other hand, external and survival stimuli, such as growth factors or stress, are thought to regulate mTORC1 activity through the modulation of the Rheb nucleotide state. However, it is not clear how the integration of these stimuli is balanced in different cell types and contexts. Both axis are regulated independently by different multiprotein complexes. The *TSC complex* (TSCC), composed of two large subunits, *tuberous sclerosis complex protein* (TSC) 1 and TSC2, also known as hamartin and tuberin respectively (Tomasoni and Mondino, 2011), and a smaller protein called TBC1D7 (Dibble et al., 2012), negatively modulates mTORC1 acting as a *GTPase activating protein* (GAP) towards Rheb (Tee et al., 2002; Zhang et al., 2003). On the other hand, the regulation of the Rag GTPases axis is more intricate and involves multiple protein complexes with GAP and *Guanine exchange factor* (GEF) activity: *Rag and*

mTORC1 regulator (Ragulator), a GEF for RagA/B, FLCN-FNIP, a GAP towards Rag C/D, *GAP activity towards Rags* (GATOR) 1 a GAP towards RagA/B, GATOR2, an inhibitor of GATOR1, and *KPTN IKTFG2 C12orf66 and SZT2* (KICSTOR), which is thought to regulate the GATOR1 complex (Ramlaul and Aylett, 2018; Rabanal-Ruiz and Korolchuk, 2018; Peng and Jewell, 2020). Hence, the mTORC1 signaling pathway involves several multiprotein complexes that integrate distinct signals that will ultimately output an on/off signal for the activation of mTORC1 (Figure 1.1).

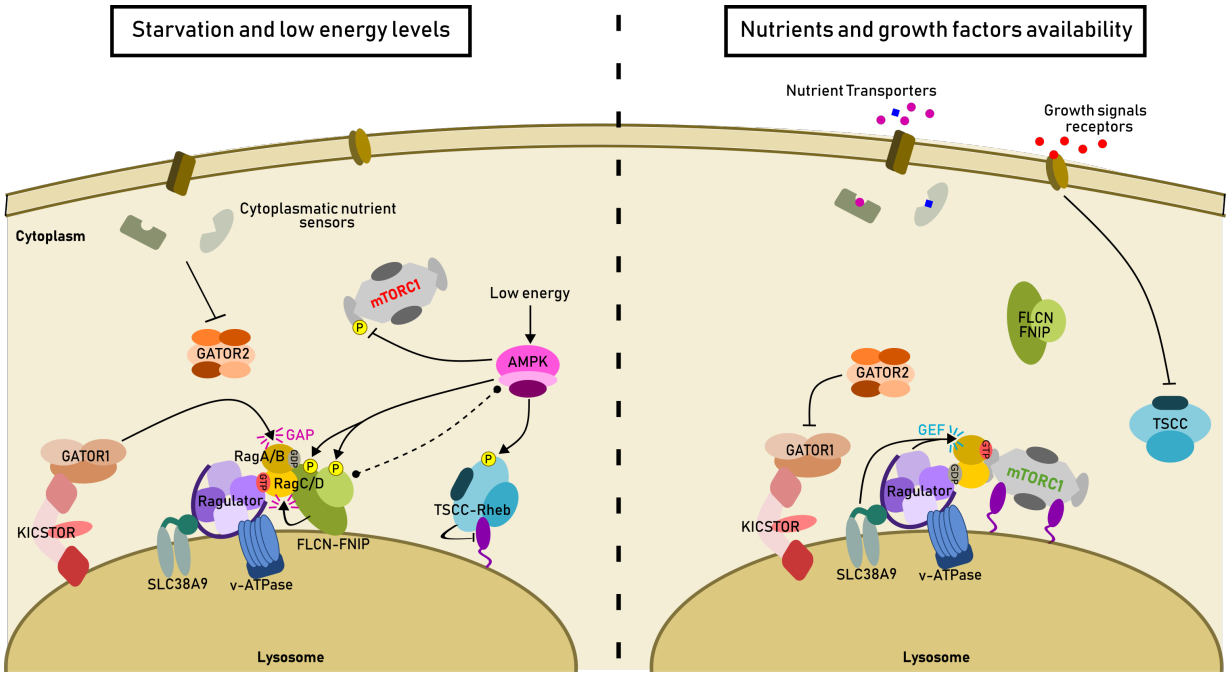


Figure 1.1: The mTORC1 signaling pathway is regulated by different stimuli. (Left) When nutrients are low mTORC1 is not recruited to lysosomes. Conversely, FLCN is recruited to the lysosome interacting with RagA/B-GDP, downstream of GATOR1 GAP activity (shown in pink). Low energy levels activate AMPK (activation is shown in black arrows) which is thought to phosphorylate Raptor and TSCC, both events having an inhibitory effect on the mTORC1 activity (inhibition is shown in black lines with a bar at the end). AMPK is also thought to bind and phosphorylate FLCN-FNIP (putative binding is indicated by a doted black line terminated with black dots). (right) In the reverse situation, when nutrients and growth factors are available, FLCN disperses from the lysosome, after exerting its GAP activity towards RagC/D, to leave space for mTORC1 recruitment, while GATOR1 is inhibited by GATOR2 and Ragulator and SLC38A9 exert their GEF activity towards RagA. The appropriate Rag GTPase nucleotide state recruits mTORC1 to the lysosome. In parallel, growth factors inhibit TSCC which in turn exposes Rheb on the lysosomal surface to allosterically activate mTORC1. Figure adapted from (de Martín Garrido and Aylett, 2020) under a Creative Commons Attribution License (CC-BY).

1.1.2 The Rag GTPases communicate nutrient availability to mTORC1 depending on their nucleotide binding state

The fact that mTORC1 depends on nutrient availability is generally accepted, however how the direct stimuli are sensed inside the cell is still unclear. Lately, a group of proteins have been reported to act as metabolic sensors for the mTORC1 signaling pathway, being able to sense the level of amino acids. Some of these metabolic sensors are cytoplasmatic nutrient sensors (Sestrins, CASTORs and *S-adenosylmethionine sensor upstream of mTORC1* (SAMTOR)), amino acid transporters (SLC38A9) and the v-ATPase, located at the lysosomal membrane (Rabanal-Ruiz and Korolchuk, 2018; Ramlaul and Aylett, 2018). Most of the intracellular signals, such as amino acid levels, will end up converging on the Rag GTPase heterodimer.

The Rag GTPases are anchored at the lysosomal membrane through the pentameric complex Ragulator (Sancak et al., 2010; Bar-Peled et al., 2012), and together, they form the platform that will recruit mTORC1 to the lysosome if there are favourable conditions for cell growth. In mammals, there are four distinct Rag GTPases: RagA and RagB, which are paralogues between them (90% sequence identity in humans) and orthologues of yeast Gtr1 (Schürmann et al., 1995), and RagC and RagD which are paralogues between them (81% sequence identity) and orthologues of yeast Gtr2 (Sekiguchi et al., 2001). Members of the Rag family include typical functional elements which are characteristic to all GTPases of the Ras family: the P-loop, guanine binding motifs, and *switch region I* (SWI) *switch region II* (SWII) (Figure 1.2) (de Martín Garrido and Aylett, 2020). In order to transmit the nutrient status to mTORC1, SWI and SWII adopt different conformations depending on their nucleotide setting. Each Rag GTPase comprises two distinct domains: the *C-terminal domain* (CTD) and the *nucleotide binding domain* (NBD). The large CTDs are responsible for the Rag GTPase heterodimerisation, while there is no direct contact between their NBDs, located on the other side of the enzyme and facing opposite directions (Figure 1.2) (de Martín Garrido and Aylett, 2020).

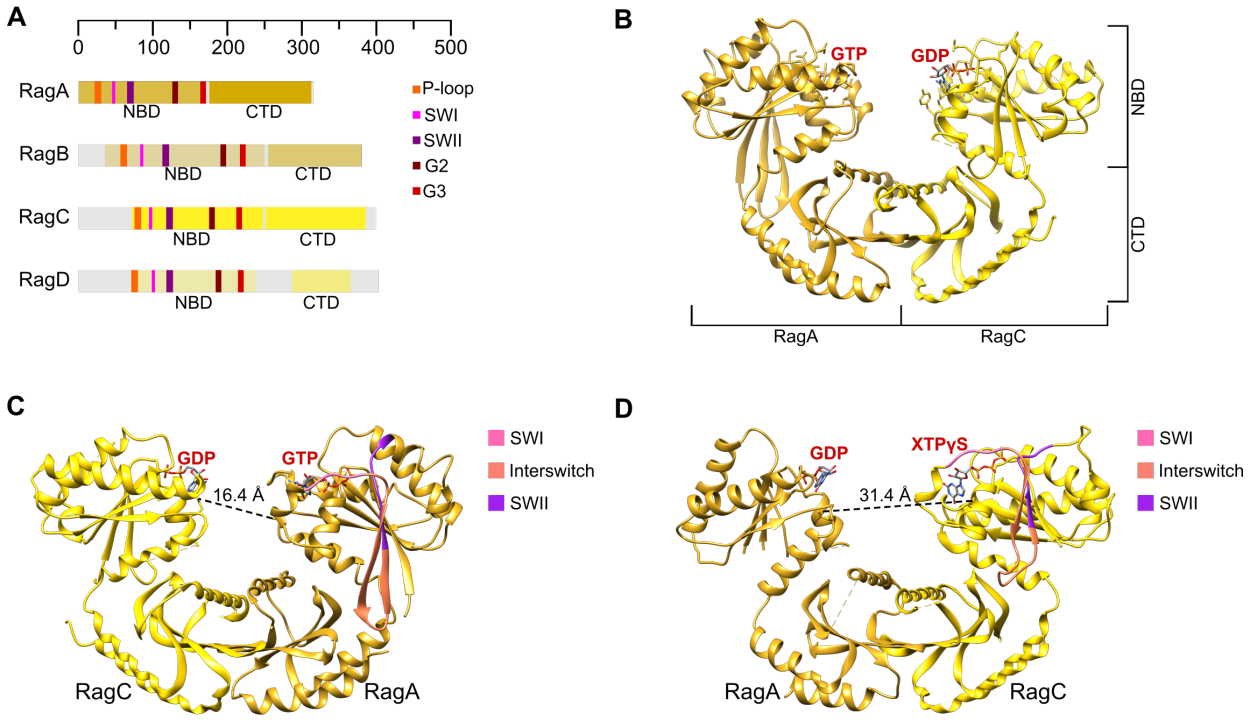


Figure 1.2: The structure of the Rag GTPase heterodimer. A) Schematic representation of the human Rag GTPases with domains coloured accordingly and typical regulatory elements from GTPases highlighted in small boxes. Top ruler represents number of amino acids. B) RagA-RagC heterodimer structure with GTP and GDP bound respectively (PDB: 6U62) (Rogala et al., 2019). Both Rag GTPases are shown in cartoon representation, with RagA shown in gold and RagC in yellow. C) Structure of the RagA-RagC heterodimer in its active nucleotide state (PDB: 6U62) (Rogala et al., 2019). Switch and interswitch regions from RagA-GTP are coloured according to the legend on the left. D) Structure of the RagA-RagC heterodimer when it forms a complex with Ragulator and FLCN-FNIP2 (PDB: 6NZD) (Lawrence et al., 2019). In this case, the Rag GTPase heterodimer is in its inactive nucleotide binding state, that is RagA-GDP and RagC-GTP (XTPγS). Switch and interswitch regions from RagC-GTP are coloured according to the legend on the right. Figure adapted from (de Martín Garrido and Aylett, 2020) under a Creative Commons Attribution License (CC-BY).

The two sub-types of Rag GTPases form heterodimers between them, RagA/B-RagC/D, with RagA/B-GTP and RagC/D-GDP being the active conformation that recruits mTORC1 to the lysosomal membrane through a direct interaction with the mTORC1 component Raptor (Sancak et al., 2008). Kinetic assays with radiollabelled nucleotides (Shen et al., 2017) have suggested that the unique heterodimeric Rag GTPase architecture allows crosstalk between them. When one subunit is bound to GTP the heterodimer adopts a conformation that impedes the binding of a second GTP or induces fast GTP hydrolysis if it happens. Furthermore, structural studies have shown how the nucleotide states of the NBD is sensed by the corresponding CTD (Anandapadamanaban et al., 2019), adopting different conformations depending on the nucleotide state of the Rag GTPase. Another distinct feature of the Rag GTPase heterodimer is

its low basal GTPase activity compared to other members of the Ras family (Shen et al., 2017; Frech et al., 1994), which indicates that they are more dependent on their GAPs (GATOR1 for Rag A/B and FLCN-FNIP2 for RagC/D) to switch between nucleotide state. Overall, the active conformation of the Rag GTPase heterodimer, bound to both GDP and GTP, and the intersubunit crosstalk are unique features that provide them with a larger number of possibilities for nucleotide binding states in comparison to other small GTPases (Mishra and Lambright, 2016), allowing them to interact specifically with a wide range of regulators to finely modulate the recruitment of mTORC1 to the lysosomal membrane.

Upon amino acid availability, the active Rag GTPase nucleotide state, RagA/B-GTP and RagC/D-GDP, is promoted, hence recruiting mTORC1 to the lysosome. When amino acid levels are low, the situation is reversed, favouring the opposite nucleotide setting, RagA/B-GDP and RagC/D-GTP, which results in the release of mTORC1 from the lysosomal membrane. Recent structural studies have shed light on the molecular basis of Rag GTPase signal transduction to mTORC1 (Rogala et al., 2019; Anandapadamanaban et al., 2019). The structure of the Ragulator-Rag-Raptor complex revealed that Raptor (mTORC1 component) firstly binds to RagA-GTP in a transient manner and is only fully attached to the Rag GTPases upon RagC binding to GDP (Rogala et al., 2019). In parallel, the structure of the mTORC1-RagA-RagC complex showed how Raptor interacts with the Rag GTPases by contacting their NBDs and that their switch regions play a crucial role in the binding, making sure that mTORC1 is only recruited at the lysosome when the appropriate Rag GTPase setting is promoted (Anandapadamanaban et al., 2019). Furthermore, the authors showed that the binding of Raptor to the Rag GTPases did not cause any conformational changes to mTORC1, hence reinforcing the hypothesis that mTORC1 needs to be allosterically activated by Rheb-GTP after being recruited to the lysosome by the Rag GTPases. Consistently, combination of the new structures with previous structural studies (de Araujo et al., 2017; Yonehara et al., 2017; Yang et al., 2017) showed that the Rag GTPases and Rheb would bind mTORC1 through different interfaces (Figure 1.3).

The Rag GTPases are anchored to the lysosome through *Late Endosomal/Lysosomal Adaptor MAPK and mTORC activator/regulator* (LAMTOR) 1, a member of the Ragulator complex,

where they recruit mTORC1 upon amino acid availability (Sancak et al., 2010; Bar-Peled et al., 2012). The five-subunit Ragulator complex (LAMTOR1 (p18), LAMTOR2 (p14), LAMTOR3 (MP1), LAMTOR4 (HBXIP) and LAMTOR5 (C7orf59)) has an elongated shape and serves as a scaffold for the Rag GTPase heterodimer, with LAMTOR1 being attached to the lysosome through myristoylation and palmitoylation sites within its N-terminus (Bar-Peled et al., 2012; Sancak et al., 2010; Nada et al., 2009). Structural studies have shown that the CTD of RagC interacts with LAMOTR2 and LAMTOR3, leaving the NBDs exposed as required for their interaction with the mTORC1 component, Raptor (Su et al., 2017; Yonehara et al., 2017; de Araujo et al., 2017). Apart from being a platform for the Rag GTPases, Ragulator is also a GEF for Rag A/B, hence favouring the active Rag GTPase nucleotide setting (GTP-bound Rag A/B) (Bar-Peled et al., 2012). Recently, the lysosomal amino acid sensor SLC38A9 has been reported to cooperate with Ragulator to perform its GEF activity through a novel mechanism (Shen and Sabatini, 2018).

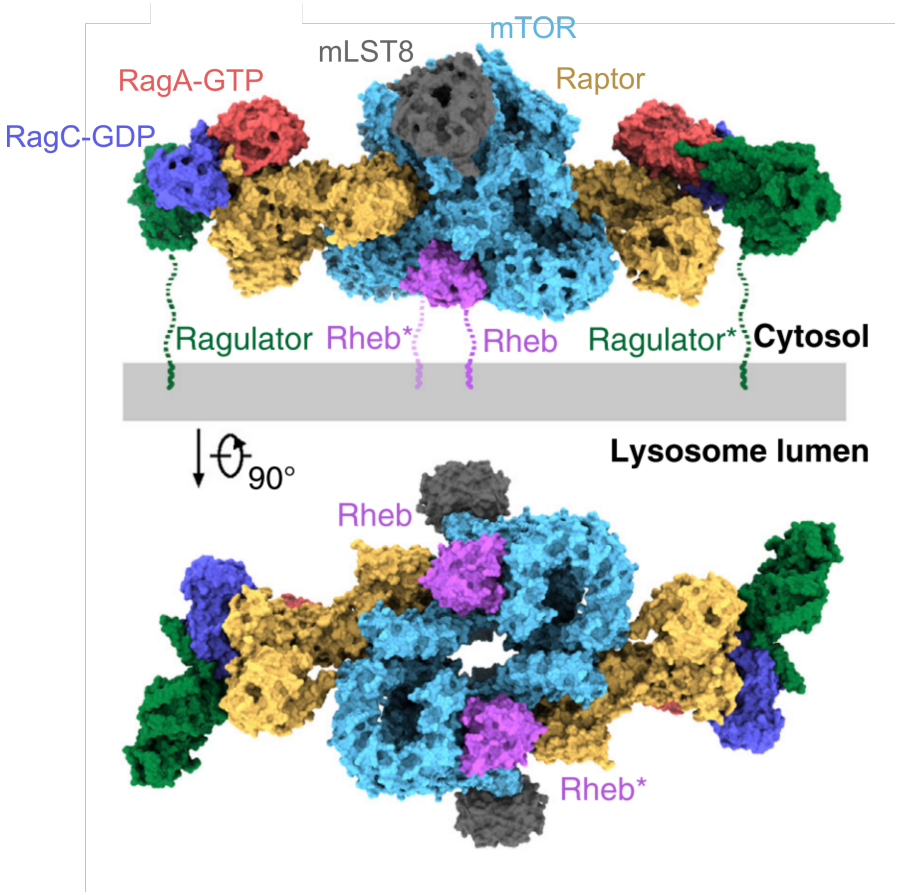


Figure 1.3: Architecture of the mTORC1 complex at the lysosome. The mTORC1–Rheb–RagA^{GTP}–RagC^{GDP}–Ragulator complex at the lysosome. The Raptor–Rag–Ragulator complex (PDB 6U62) (Rogala et al., 2019)) was superimposed with the mTORC1–Rheb complex (PDB 6BCU) (Yang et al., 2017) to generate the model. Individual complex subunits are colored and indicated by color-coded text. Figure adapted from (Peng and Jewell, 2020).

1.1.3 The Folliculin complex and its interacting partners

1.1.3.1 Structural organisation of the FLCN complex

The FLCN protein forms complexes with both FNIP1 (Baba et al., 2006) and FNIP2 (Takagi et al., 2008; Hasumi et al., 2008a), and the same behaviour has been found for their orthologues in yeast, Lst7 and Lst4 (Pacitto et al., 2015). The FLCN protein is 579 amino acids in length and has a molecular weight of 64 kDa in humans. FLCN is highly conserved across species, from mouse (92% sequence identity to human) to *Drosophila melanogaster* (28% sequence identity to human), and is expressed in a wide range of tissues, such as lung, kidney brain or heart (Nickerson et al., 2002). Many attempts to determine the structure of the FLCN-FNIP

complex failed until recently. The first structural insights into the structure of the FLCN-FNIP complex came from a crystal structure of the C-terminus (341-566) of FLCN at 2 Å resolution (Nookala et al., 2012) (PDB 3V42). The structure revealed that this fragment of the FLCN protein displayed structural similarity with the DENN family of proteins, displaying the typical α/β architecture with a β-sheet core with helices packed on each side, and a mostly helical domain (Figure 1.4). DENN domains are usually associated with N-terminal longin domains, and comparative studies with the DENN1B full domain predicted that FLCN was likely to include a longin-like domain at its N-terminus.

The FLCN interacting partners, the FNIP proteins, are much larger (130 kDa for FNIP1 and 122 kDa for FNIP2 in humans) and contain large unstructured regions, which complicates their structure determination. Both FNIPs are highly similar (74% sequence similarity in humans), and their sequence is mostly conserved from human to *C.elegans*, with the exception of the C-terminus, which is slightly divergent between species. Interestingly, FNIP1 and FNIP2 have been reported to form multimeric complexes (Hasumi et al., 2008a), however their functional relevance is still unclear. First sequence analysis revealed that, similarly to FLCN, they are likely to contain longin and DENN domains, however these domains contain large disordered regions within them, which are probably involved in phosphorylation events as they are rich in serines and threonines. The crystal structure of the N-terminus of Lst4, the FNIP yeast orthologue (Pacitto et al., 2015) (PDB 4ZY8), confirmed that this domain has a longin architecture formed by central sandwich of five β-strands with α helices packed on both sides.

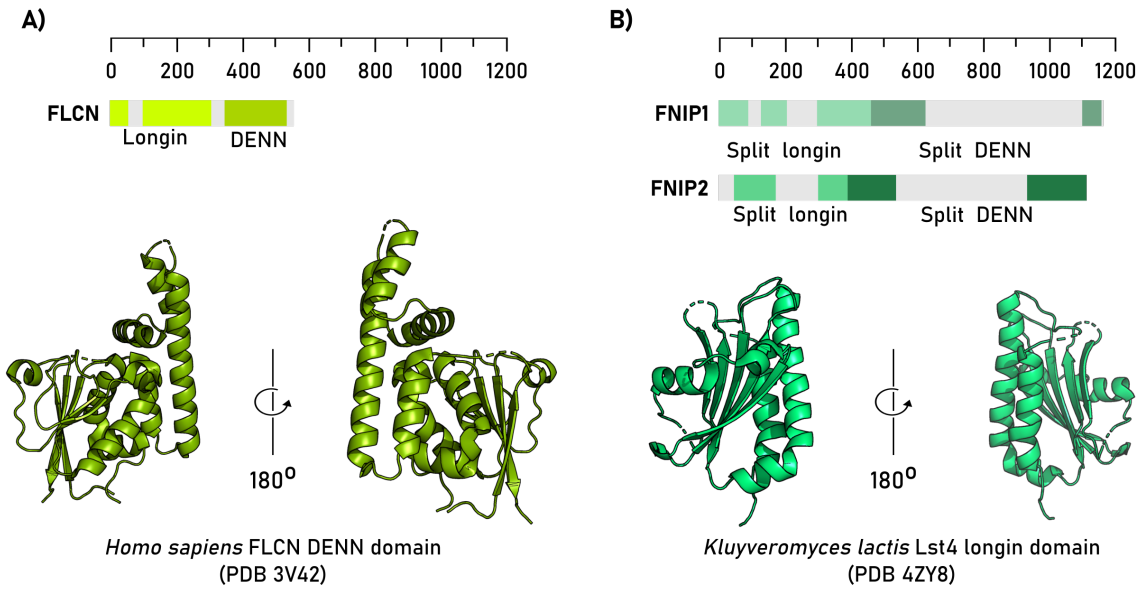


Figure 1.4: Structural biology of the FLCN and the FNIP proteins. A) The FLCN structure. (Top) Schematic diagram of the human FLCN sequence with its domain highlighted in coloured boxes. (Bottom) Crystal structure of the human FLCN DENN domain (341-566) from two different views. B) The FNIP structure. (Top) Schematic diagram of the human FNIP proteins. (Bottom) Crystal structure of the *Kluyveromyces lactis* longin domain from two different views Figure adapted from (de Martín Garrido and Aylett, 2020) under a Creative Commons Attribution License (CC-BY).

The DENN domain family is a group of proteins that have been reported to act as nucleotide exchange factors for small GTPases (Yoshimura et al., 2010). Usually, a full DENN domain is thought to comprise three subdomains: the upstream longin domain, also known as u-DENN and thought to occur independently, the DENN core and the downstream d-DENN, which are thought to occur together (Levivier et al., 2001). The longin domain is usually responsible for the interaction with the small GTPases. The fact that both the FLCN protein and the FNIPs contain DENN and longin domains suggests that they may interact with the Ras-superfamily of small GTPases, hence supporting a role for the FLCN-FNIP complex in the regulation of the Rag and Rab GTPases (Nookala et al., 2012; Petit et al., 2013b; Tsun et al., 2013; Zhao et al., 2018; Starling et al., 2016). In 2019, two *cryogenic electron microscopy* (cryo-EM) structures of the lysosomal FLCN-FNIP2 complex bound to RagA-RagC and Ragulator were published (Shen et al., 2019; Lawrence et al., 2019), revealing the architecture of the complex for the first time.

1.1.3.2 Folliculin acts as a GTPase activating protein to modulate the Rag GTPase nucleotide binding state

The FLCN-FNIP complex acts as a GAP for RagC/D, hence promoting its GDP-bound state and being a positive regulator of mTORC1 (Tsun et al., 2013). Under starved-conditions, FLCN is recruited to the lysosome interacting with RagA/B-GDP. Upon nutrient recovery, FLCN dissociates from the lysosome to allow RagC/D-GTP to recruit mTORC1 where it will be fully activated by Rheb-GTP (Figure 1.1). Coimmunoprecipitation experiments reported that both RagA and RagC coimmunoprecipitated with FLCN coexpressed with FNIP2, which suggests that both proteins of the complex are needed to perform its function (Petit et al., 2013a; Tsun et al., 2013). The same behaviour was found in yeast orthologues, Lst7 and Lst4, where the loss of either Lst7 or Lst4 decreased TORC1 activity equally (Péli-Gulli et al., 2015).

Hence FLCN is a key modulator of the recruitment of mTORC1 to the lysosomal membrane, playing a crucial role in the transduction of the nutrient signaling state. The regulation of the RagA-RagC GTPase nucleotide state is performed by multiple protein complexes which probably need to be coordinated with FLCN. Accordingly, recent studies have placed the FLCN GAP activity downstream of the GATOR1 complex (Meng and Ferguson, 2018). When amino acid and energy levels are low, GATOR1 GAP activity promotes RagA/C-GDP, which recruits FLCN-FNIP to the lysosomal membrane, while mTORC1 remains at the cytoplasm and TSCC is at the lysosome inhibiting Rheb. Upon nutrient and growth factor availability, FLCN disperses from the lysosome after performing its GAP activity towards RagC/D, which leaves space for mTORC1 to be recruited at the lysosome by RagA/B-GTP and RagC/D-GDP. In parallel, growth factors inhibit TSCC, which in turn allows the build up of Rheb-GTP that then allosterically activates mTORC1 (Figure 1.1).

Considering that both the FLCN and FNIP2 proteins are required to promote RagC/D-GDP and that they are recruited to the lysosomal membrane by interacting with the other GTPase counterpart RagA/B-GDP, the mechanism by which the FLCN-FNIP2 complex performs its GAP activity is likely to be novel. Moreover, while it is currently accepted that FLCN-FNIP is a positive modulator of mTORC1, BHD-associated renal tumours containing germline muta-

tions in the *FLCN* gene were reported to have an increase in mTORC1 activity and substrate phosphorylation levels (Baba et al., 2008; Hasumi et al., 2009; Khabibullin et al., 2014). These discrepancies suggest that the tumor suppressor activity of FLCN must be associated, not only with the mTORC1 activation, but also with other signaling pathways, supporting the notion of FLCN as a multifunctional protein involved in several cellular processes.

1.1.3.3 Other FLCN interacting partners: the AMP-activated protein kinase and the Rab GTPases

Consistent with the apparently disparate phenotypes caused by the tumour suppressor FLCN, other interacting partners have been reported for the FLCN-FNIP complex, apart from the Rag GTPase heterodimer. Another central regulator of metabolism is the *AMP-activated protein kinase* (AMPK), which is allosterically activated by high AMP/ATP ratios, an indicator of low energy levels (Carling, 2004; Shackelford and Shaw, 2009). When AMPK is activated it promotes ATP production through the activation of catabolic processes, acting in an antagonistic manner to mTORC1. AMPK inactivates mTORC1 through the phosphorylation of Raptor and TSC2 (Inoki et al., 2002; Gwinn et al., 2008), which provides a mechanism for the crosstalk between these two metabolic pathways.

The coordination between mTORC1 and AMPK provides a more sophisticated mechanism by which cell growth adapts to different conditions. Accordingly, mTORC1 inactivation by AMPK is not the only relationship between these two pathways. FLCN has been reported to bind to AMPK through the FNIPs, and both FLCN and FNIP phosphorylation levels are affected by AMPK signaling and amino acid starvation, which in turn negatively regulate mTORC1 (Baba et al., 2006; Hasumi et al., 2008a; Takagi et al., 2008). Furthermore, depletion of FLCN has been found to constitutively activate AMPK (Possik et al., 2014, 2015; Yan et al., 2014; El-Houjeiri et al., 2019), and FNIP1 mutations have been associated with an increase in the AMPK activity (Siggs et al., 2016), suggesting that both FLCN and FNIP are involved in negative regulation pathways for AMPK. Finally, recent studies have shown that the increased FLCN expression causes *Transcription factor EB* (TFEB) translocation to the nucleus in a manner-independent

of mTORC1, through AMPK-dependent dephosphorylation (Collodet et al., 2019). All these observations point at FLCN-FNIP coordinating metabolism not only through mTORC1, but also through AMPK signaling, however the functional relevance of the FLCN-FNIP interaction with AMPK remains unclear.

The lysosome is the location where mTORC1 activation occurs, however this is not the only function of this organelle in cell signaling. The lysosomal function is dependent on their bidirectional movement between the peripheral and centrosomal regions of the cell through microtubular filaments (Matteoni and Kreis, 1987). Interestingly, the intracellular distribution of lysosomes has been correlated with nutrient availability in HeLa cells (Korolchuk et al., 2011). Low amino acid levels result in the clustering of lysosomes around the centrosomal region and, in the reverse situation, upon amino acid sufficiency, lysosomes move closer to the plasma membrane. The clustering of lysosomes around the plasma membrane has been suggested as a mechanism to put in closer proximity the mTORC1 activation site with growth factor signaling (Korolchuk et al., 2011). Moreover, depletion of microtubule motors has been reported to disrupt the Rag GTPase-dependent mTORC1 activation, supporting a relationship between lysosomal positioning and mTORC1 activation. During autophagy, vesicles are moved towards the centrosomal region of the cell where they are fused to the lysosomes which, ultimately, will degrade their contents. During starvation, the translocation of TFEB and *Transcription factor binding to IGHM Enchancer 3* (TFE3) to the nucleus promotes lysosomal biogenesis and autophagy (Sardiello et al., 2009; Settembre et al., 2011), hence explaining the requirement for lysosomal clustering around the centrosomal region of the cell during starvation. This links the inhibition of mTORC1 activity to activation of catabolic processes through the lysosome (Rabanal-Ruiz and Korolchuk, 2018; Kim and Guan, 2019). Hence, the relationship between mTORC1 and the lysosome is crucial to associate the lysosomal function with nutrient availability and to coordinate the appropriate cellular response.

In previous sections, how FLCN promotes mTORC1 recruitment to the lysosome through the Rag GTPases, when nutrient levels are sufficient, has been extensively covered. However, the FLCN-Rag GTPase interaction is not the only lysosome-related function of this multifunctional protein. Pull-down assays have recently shown how FLCN interacts with Rab7 and Rab34

through their common partner RILP, a component of the lysosomal positioning machinery (Starling et al., 2016). This interaction identifies a connection between lysosomal positioning machinery and mTORC1 pathway, which could be associated with the centrosomal clustering of lysosomes during starvation (Figure 1.5). Both FLCN and FNIP2 have been reported to include DENN domains (Zhang et al., 2012), which are typical found in proteins acting as GEF for Rab GTPases (Levivier et al., 2001). FLCN seems to interact with RILP through its C-terminal DENN domain, a different region than that interacting with the Rag GTPases. Although no GEF or GAP activity has been reported for Rab34, FLCN has been identified as a GAP for Rab7, involved in EGFR trafficking (Laviolette et al., 2017). Conversely, other studies support that FLCN acts as a GEF for other Rab GTPases, which are also involved in vesicular trafficking (Nookala et al., 2012). Currently, the FLCN interaction with the Rab-RILP complex seems to be involved in the promotion of lysosomal clustering around the centrosomal region of the cell during starvation, however, further studies are required to provide a molecular mechanism for this interaction and to clarify the role of FLCN in membrane trafficking.

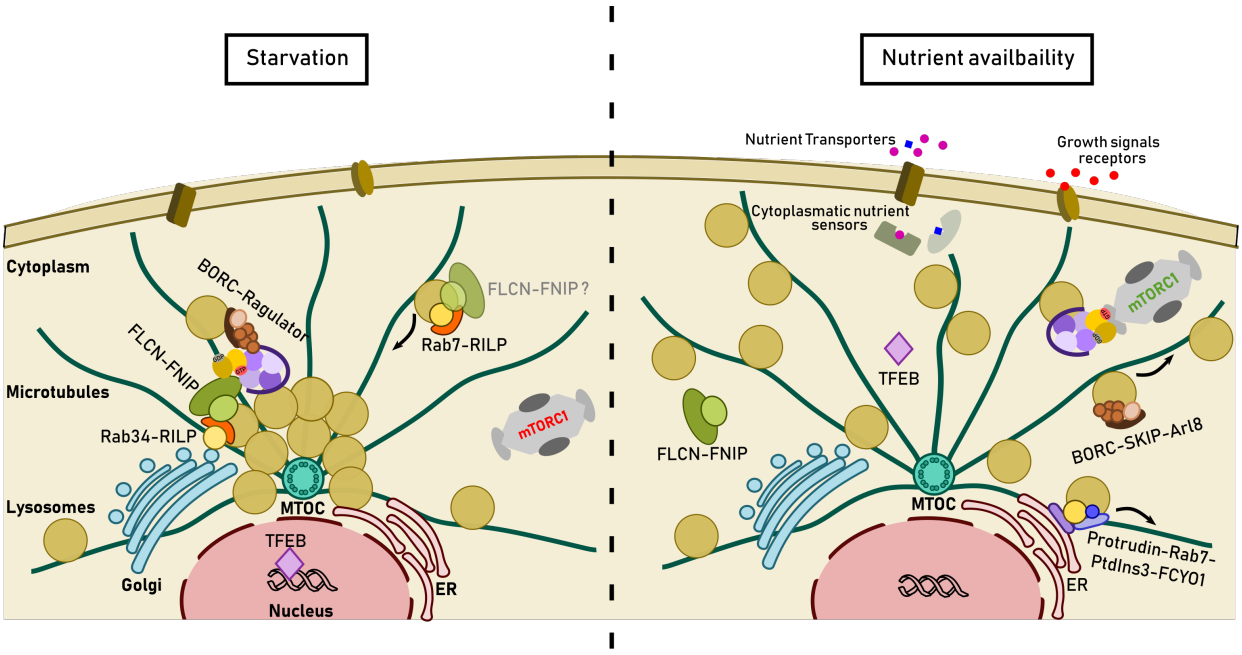


Figure 1.5: Lysosomal positioning in response to nutrient availability. Under starvation conditions, mTORC1 activity is abolished and lysosomes cluster in the perinuclear region to favor catabolic processes such as autophagy. Lysosomal clustering around the nucleus is favored by FLCN-RILP-Rab contacts on the Golgi and the inhibitory BORC-Ragulator interaction. At the same time, TFEB is relocalized to the nucleus to favor lysosomal biogenesis. When nutrient levels are recovered, FLCN disperses from the lysosome and the BORC-Ragulator interaction is broken. Kinesin dependent pathways, such as PtdIns3-Rab7-Kinesin and BORC-Arl8-Kinesin, are promoted to favor dispersal of the lysosomes. In parallel, TFEB relocalizes to the cytoplasm, where it is non-functional. Nutrients promote the recruitment of mTORC1 to the lysosome through the Rag GTPases, which occurs closer to the plasma membrane allowing full activation by growth factors through the Rheb axis. Figure adapted from (de Martín Garrido and Aylett, 2020) under a Creative Commons Attribution License (CC-BY).

FLCN is not the only member of the mTORC1 signaling pathway that is implicated in lysosomal function. The Ragulator complex, responsible for anchoring the Rag GTPases to the lysosomal membrane, has been reported to interact with BORC, a member of the lysosomal positioning machinery, under starvation conditions (Filipek et al., 2017; Pu et al., 2017) (Figure 1.5). The interaction results in the inhibition of lysosomal transport towards the plasma membrane, hence resulting in centrosomal lysosomal clustering. Moreover, RILP has been suggested to interact with and regulate the v-ATPase (De Luca et al., 2014) and mTORC1 has also been reported to modulate v-ATPase function (Peña-Llopis et al., 2011).

All these examples provide evidence of the intimate connection between lysosomal positioning and mTORC1 signaling within nutrient signaling. FLCN is likely to play a crucial role in this crosstalk between mTORC1 and the lysosome through interactions with the Rag and Rab

GTPases. Interestingly, both FLCN and FNIP are also involved in AMPK signaling, which reinforces its crucial role in the coordination of several metabolic processes within the cell. How this pleiotropic protein coordinates all these cellular processes is still a question that remains unanswered. Structure-function studies will hopefully shed light of the functional relevance of the FLCN protein, which will hopefully contribute to a better understanding of its role in BHD-associated and other tumours.

Two structures of the FLCN-FNIP2 complex bound to the Rag GTPase heterodimer and Ragulator were published (Shen et al., 2019; Lawrence et al., 2019) while I was undertaking this project, revealing the molecular basis of FLCN-FNIP2 recruitment to the lysosome and its GAP activity towards RagC-GTP.

1.2 Bacteriophage ΦKZ; a giant virus encoding its own transcriptional machinery

1.2.1 Bacteriophage biology and medical interest

Viruses are a group of highly diverse macromolecular machines that share two common features: all viruses require a host organism to replicate and are formed by proteinacious capsids that pack and protect their genome, formed by either single- or double-stranded DNA or RNA (Prasad and Schmid, 2012). Apart from these common features, viruses exhibit great variability in their size, genome organisation, host interactions, and capsid architecture. All viral capsids have evolved to protect the encapsulated genome, and to successfully deliver it to a specific host. Viruses can infect and replicate in organisms from the three kingdoms of life, including bacteria, archaea and eukaryotes, which suggest that they probably have multiple evolutionary origins (Moreira and López-García, 2015).

Viruses that infect bacteria are known as bacteriophages. When they were discovered over a century ago (Twort, 1915; D'Herelle, 1917), their potential application to combat bacterial infectious diseases soon became apparent, however the failure of many studies and the discovery of antibiotics caused a loss of interest in the study of bacteriophages. During the 1940s, a series of studies by Max Delbrück and their collaborators on bacteriophage physiology and their relationship with their hosts led to a renewed interest in bacteriophage biology (Ellis and Delbru ck, 1939; Delbru ck, 1940a,b), with studies on bacteriophage genetics beginning in the next few years. These studies established the fundamentals of current bacteriophage biology and the foundations of molecular biology.

Bacteriophages are thought to be the most abundant organism in the planet, to an extent that their numbers can be translated to approximately a trillion of bacteriophage particles for every grain of sand in the Earth (Comeau et al., 2008). They have been isolated from multiple environments populated by bacteria, which has facilitated sequencing of their genomes. Bacte-

riophage genomes are extraordinarily diverse; they consist of either single or double stranded DNA or RNA, include sizes from 4 to 725 kb and they can be circular or linear (Ackermann, 1998). Based on the nature of their genomic material, bacteriophages have been classified into 13 families, with tailed phages of the *Myoviridae, Siphoviridae and Podoviriade* families, representing most of the currently identified bacteriophages (Sharp, 2001). Bacteriophage genomes usually show a mosaic architecture of genes that have been rearranged over years of evolution by means of horizontal gene transfer, with bacteriophages infecting the same host exhibiting great diversity within their nucleotide sequence (Hatfull et al., 2006; Kwan et al., 2005). Although the number of bacteriophage genomes that have been sequenced during recent years has dramatically increased, it is believed that they only represent a small fraction of the predicted number of bacteriophages in the biosphere, and that they probably encode multiple proteins yet to be discovered.

Over the years, bacteriophages have been used as model organisms to elucidate fundamental principles of biology, such as the discovery of mRNA as the intermediate between DNA and proteins (Brenner et al., 1961) and that nucleotides are read in groups of three, as codons, to form proteins (Crick et al., 1961). Apart from being excellent model organisms for molecular biology research, the use of bacteriophages to kill pathogenic bacteria that cause infectious diseases has been postulated as a successful alternative to antibiotics (Kakasis and Panitsa, 2019; Burrowes et al., 2011). When bacteriophages were initially discovered, the idea of bacteriophage therapy was of great interest and was implemented in some cases, however it was quickly subsided due to the use of antibiotics. Recently, antimicrobial resistance has been increasing, becoming a major global health problem that has led to a renewed interest in bacteriophage therapy. The fact that bacteria gain resistance against antibiotics does not imply that they are resistant to bacteriophage lytic mechanisms. Bacteriophages with a lytic infection cycle are those able to kill their host at the end of their cycle, in comparison to lysogenic bacteriophages that are capable of integrating part of their genome into the host genetic material.

There are several publications reporting promising results of bacteriophage therapy in animals and in humans (Kakasis and Panitsa, 2019), however, despite its potential, many studies have reported bacteriophage therapy failure due to bacteriophage selection, preparation and stor-

age (Gill and Hyman, 2010), as well as their narrower spectrum of activity. In addition, its implementation would require to put in place regulations for bacteriophage production and administration (Fauconnier, 2018), which are currently missing. Another issue may be the risk of immunological response. Although previous clinical and animal trials have not resulted in serious immunological reactions, it is crucial to test for it for every single bacteriophage. Despite its limitations, some countries have been using bacteriophage therapy for nearly a century (Reardon, 2014), and, hopefully, with recent advances in the knowledge of bacteriophage biology, bacteriophage therapy will overcome antibiotic resistance. Overall, since bacteriophages were discovered more than a century ago, their study has revealed several insights into the fundamentals of biology, such as DNA replication, gene regulation or protein folding, as well as presented these organisms as important biotechnological and medical tools. A full understanding of bacteriophage biology is therefore key to understanding their evolutionary origins, as well as providing valuable information to overcome some of the limitations of bacteriophage therapy.

1.2.2 Bacteriophage ΦKZ

Bacteriophage ΦKZ is the prototypical member of a family of giant bacterial viruses distantly related to the *Myoviridae* (Krylov and Zhazykov Zh., 1978; Krylov et al., 2007). It infects *Pseudomonas aeruginosa*, which is an intrinsically antibiotic resistant opportunistic pathogen of special concern in medical practice as it kills tens of thousands annually (Stover et al., 2000). In 2017, the World Health Organisation published a list of bacteria which are the most common causes of infections in hospitals, for which new antibiotics or alternative therapies are urgently needed (https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed). Indeed, *P.aeruginosa* was included in the list and it was identified as one of the most critical pathogens for medical practice. As bacteriophage ΦKZ can effectively combat *P.aeruginosa* infections, it is considered a potential candidate for bacteriophage therapy (Pires et al., 2015). There has been some evidence of successful bacteriophage therapy for antibiotic resistant strains (Hall et al., 2012; Med Sci et al., 2018;

Pires et al., 2015), however there are still some major concerns about the use of bacteriophage therapies in medical practice, such as their safety or their limited activity spectrum. Full understanding of the life-cycle of bacteriophage ΦKZ, as well as its evolutionary origins and the function of the proteins encoded by its large genome, may provide very valuable information to design therapies against serious *P.aeruginosa* infections.

When bacteriophage ΦKZ was discovered 40 years ago, it was the first bacteriophage classified as "giant" due to its exceptional size. ΦKZ comprises an icosahedral head encapsulating a large, densely packed DNA genome (280 334 bp), followed by a contractile long tail terminating in a baseplate (Mesyanzhinov et al., 2002)(Figure 1.6). First *Electron Microscopy* (EM) micrographs revealed further details about the morphology of the bacteriophage (Smirnova et al., 1983) and confirmed its huge size, with the head measuring 120 nm diameter and the tail being 180 nm long in its fully extended state and measuring 20 nm in diameter (Figure 1.6). The DNA inside the head is packed around a proteinacious cylinder of unknown function, called the inner body, displaying a highly ordered packed DNA (Krylov et al., 1984). Later EM studies suggested slightly larger measurements, however these differences might be due to the use of different microscopes and/or measuring techniques.

The first three-dimensional reconstruction of the ΦKZ head came with the advent of new cryo-electron microscopy methods, 30 years after the first ΦKZ micrographs were reported (Fokine et al., 2005). The 18 Å resolution reconstruction revealed that the capsid is mostly formed by copies of capsid protein, gp120, arranged into a surface lattice of hexamers with triangulation number $T = 27$. A 28 Å resolution reconstruction of the full bacteriophage revealed that the genomic DNA inside the capsid is organised into a spool-like structure around the inner body formed at a roughly ≈32°angle against the long axis of the capsid (Fokine et al., 2007), a novel organisation previously unreported in bacteriophages at that time. The same authors reported that the tail consists of a neck, a collar and a contractile sheath assembled around the tail tube, and that the baseplate has a flat hexagonal shape in comparison to the dome-shaped hexagonal aspect of T4 phages (Fokine et al., 2007). Later EM studies found that the inner body was extremely sensitive to radiation damage (Wu et al., 2012), a quality that could be used to distinguish the proteinaceous component from the surrounding DNA. The shape and

position of the inner body suggests that it may be involved in DNA packaging, however, further studies are required to confirm that. Currently, available reconstructions of ΦKZ provide an overall architecture of the bacteriophage morphology, however they lack high resolution features that would reveal further details, such as the protein composition of important parts of the bacteriophage, like the tail tube or the baseplate, as well as the role of the intriguing inner body.

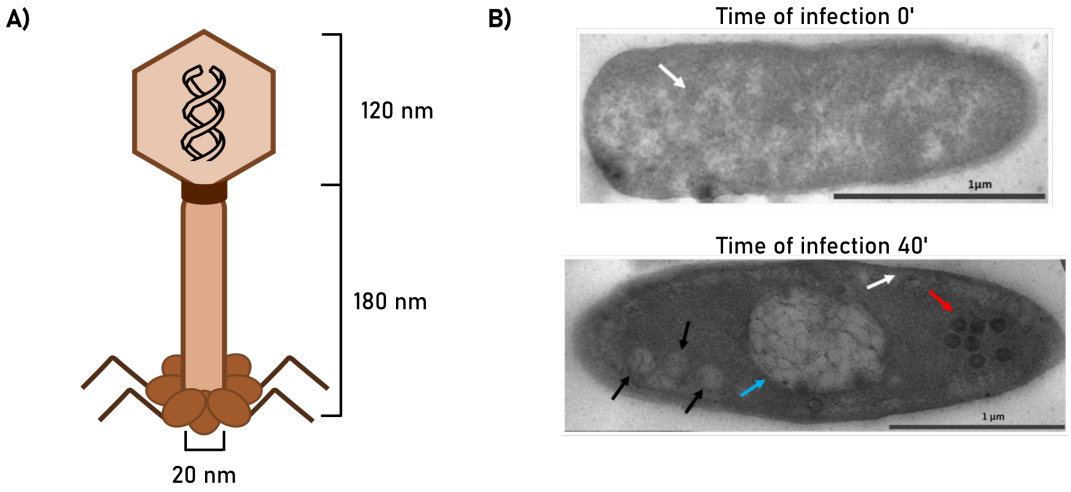


Figure 1.6: Bacteriophage ΦKZ morphology and pseudo-nucleus maturation in *P.aeruginosa* cells. A) Schematic diagram about the ΦKZ morphology. Measurements of the head and tail are representative, as they are variable between different studies. B) Electron microscopy images showing the time-course of the ΦKZ infection and pseudo-nucleus maturation in *P.aeruginosa* cells. Minutes after infection are indicated at the top of each image. White arrows point at bacterial nucleoid; black arrows at round compartments (RC); blue arrows at the pseudo-nucleus at different stages of maturation; red arrows at new bacteriophage capsids. Figure adapted from (Danilova et al., 2020) under a Creative Commons Attribution License.

The massive genome of bacteriophage ΦKZ was sequenced in the early 2000s (Mesyanzhinov et al., 2002). The ΦKZ genome has a low GC content in comparison to its host. It has 306 open reading frames (ORFs) and six tRNAs. Intriguingly, only 59 bacteriophage-encoded proteins show significant sequence homology to other proteins with known function and only 11 of them are homologous to other *double-stranded DNA* (dsDNA) bacteriophage proteins, indicating that ΦKZ is a distinct evolutionary strand of the *Myoviridae* family. The ΦKZ genome sequencing was the starting point for a large number of subsequent studies that aimed to characterise this unique bacteriophage (Miroshnikov et al., 2006; Yakunina et al., 2015; Orekhova et al., 2019; Chaikeeratisak et al., 2017a; Danilova et al., 2020; Mendoza et al., 2020; Aylett et al., 2013). One of these studies found another interesting feature in the bacteriophage ΦKZ genome: it included three virion proteins and two others that were not associated with the virion with high

homology to the two largest subunits of bacterial *DNA-directed RNA polymerase* (RNAP)s, β and β', suggesting that ΦKZ encodes two RNAPs. Later, ΦKZ development and transcription proved to be insensitive to rifampicin, confirming that the bacteriophage relies on its self-encoded enzymes for gene expression (Ceyssens et al., 2014). Recently it was reported that during infection, ΦKZ forms a proteinaceous "nucleus" within the host cytoplasm that protects its genome from the host immunity mechanisms, such as CRISPR-Cas and restriction enzyme, (Chaikeeratisak et al., 2017a,c; Danilova et al., 2020; Mendoza et al., 2020) and traffic through a bacteriophage-encoded tubulin-like cytoskeleton (Chaikeeratisak et al., 2017a,b; Aylett et al., 2013; Chaikeeratisak et al., 2019) (Figure 1.6).

Overall, bacteriophage ΦKZ encodes many proteins that the host would typically be expected to provide, making it a really interesting life-from to study. These sophisticated mechanisms imply a unique infection cycle with intriguing evolutionary parallels to their eukaryotic equivalents. A deep understanding of the bacteriophage ΦKZ life-cycle in *P.aeruginosa*, including a detailed characterisation of all the proteins involved in its regulation, would be crucial to fully understand the evolution of this interesting family of giant bacteriophages and would hopefully contribute to improve potential bacteriophage therapy approaches.

1.2.3 The transcriptional apparatus of bacteriophage ΦKZ

1.2.3.1 DNA-dependent RNA polymerase, the central enzyme of transcription

During the development of all living organisms, cells specialise into distinct cell types thanks to the expression of different genes. The first step of gene expression is transcription, which is the synthesis of RNA from a DNA template. Gene transcription is tightly regulated and is performed by a large molecular machine, the *DNA-directed RNA polymerase* (RNAP) (Cramer, 2002). Whereas bacteria and archaea usually use a single RNAP to transcribe all genes, eukaryote organisms use specialised RNAPs which transcribe different gene classes (Werner and Grohmann, 2011). Based on subunit composition and conserved amino acid sequence, RNAPs can be divided into two evolutionarily unrelated families. The first family comprises

single-subunit RNAP (ssRNAP)s, which are composed of a single polypeptide. This group of polymerases usually transcribe genes of mitochondria, chloroplasts and some bacteriophages (Cermakian et al., 1997). The second family is formed by *multi-subunit RNAP* (msRNAP)s which transcribe genes from all organisms, from bacteria to humans. Together with some RNA-dependent RNAPs, msRNAPs form a superfamily of "two-barrel" RNAPs (Iyer et al., 2003), with eubacterial genomes encoding the simplest form of msRNAPs. The conserved *double psi beta-barrel* (DPBB) domains are split between the two largest subunits (β and β' in bacteria), and together form the catalytic site (Figure 1.7). The DPBB of the largest subunit (β' in bacteria) contains the universally conserved motif DxDGD, with the three aspartates coordinating one Mg^{2+} ion essential for catalysis. The two largest RNAP subunits on their own are catalytically inactive and require additional universally conserved subunits. In bacteria, a dimer of α subunits and the ω subunit are required to assemble a functional complex (Zhang et al., 1999; Heyduk et al., 1996; Minakhin et al., 2001). The RNAP core ($\beta\beta'\alpha_2\omega$ in bacteria) is evolutionarily conserved amongst all three domains of life (Sweetser et al., 1987) and is responsible for DNA binding and RNA synthesis.

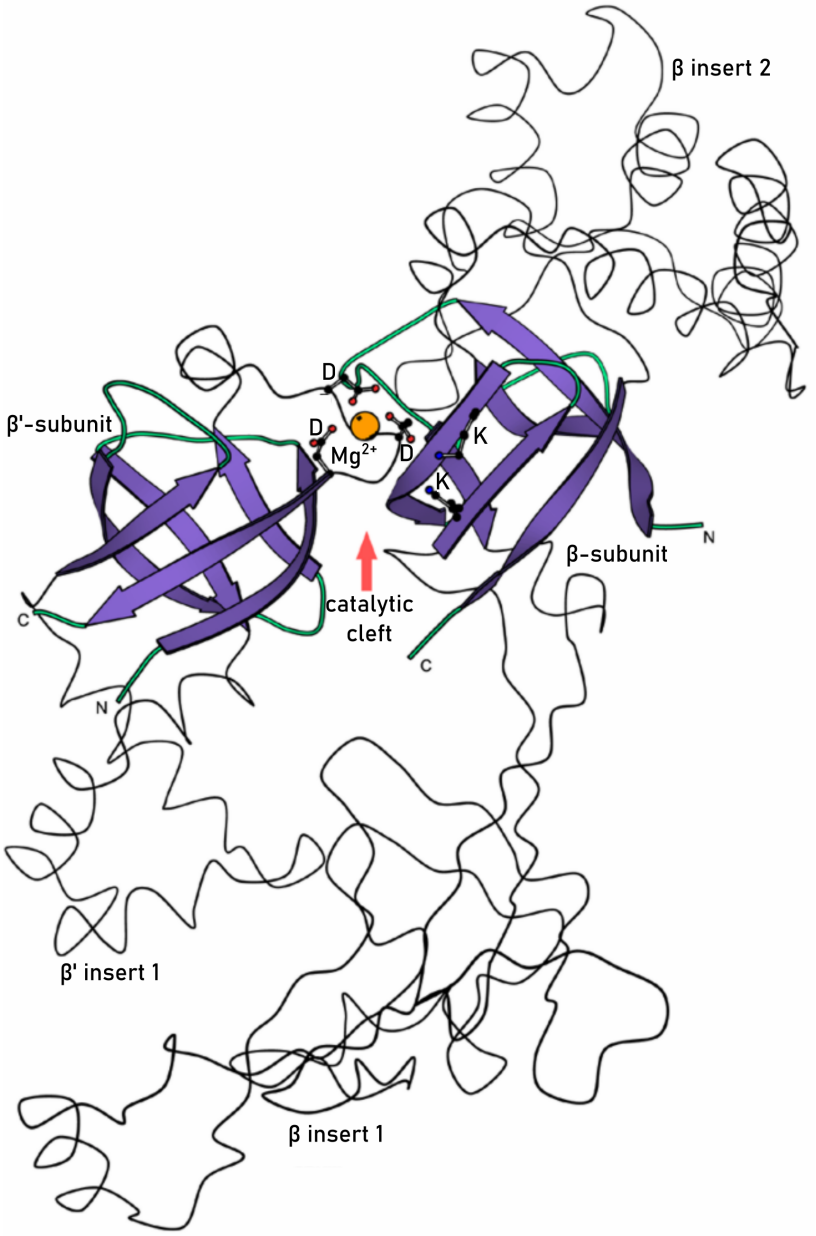


Figure 1.7: Structure of the catalytic centre of the "two-barrel" RNAPs family formed by interacting DPBBs from the β and β' subunits. The metal-coordinating DxDGD motif included in the β' subunit and the functionally important lysines of the β subunit projecting into the catalytic cleft are shown in ball and stick representation. The two DPBBs are juxtaposed in an asymmetric head to tail configuration. Figure adapted from (Iyer et al., 2003) under a Creative Commons Attribution License.

A transcription cycle can be divided into three stages: initiation, elongation and termination. The RNAP core ($\beta\beta'\alpha_2\omega$ in bacteria) is catalytically sufficient and can perform elongation and termination on its own, however, it requires accessory factors to initiate transcription from specific promoters (Saecker et al., 2011). On one hand, bacterial transcription initiation is modulated by one of the several σ factors, whose binding to the RNAP core allows the

enzyme to specifically recognise and melt DNA promoters, hence positioning the catalytic site at the transcription start point (Paget, 2015; Lonetto et al., 1992). On the other hand, in archaeal and euakryotes transcription initiation is more sophisticated and involves a wide range of transcriptional factors which are evolutionarily unrelated to bacterial σ factors (Werner, 2007; Vannini and Cramer, 2012). Usually, bacterial genomes encode a principal σ factor which transcribes constitutive genes needed for the maintenance of the basic cell functions, and a variable number of alternative σ factors involved in specific functions, such as stress responses or morphological changes (Gruber and Gross, 2003; Browning and Busby, 2004). In *E.coli* σ factors can be classified in two main families, the σ^{70} and the σ^{54}. σ^{70}-related factors are encoded in all bacterial genomes and contain four conserved regions (Lonetto et al., 1992; Gruber and Gross, 2003). The most N-terminal region, region 1, is thought to have an inhibitory effect on DNA binding. Region 2 can be divided into four subregions and is involved in the recognition of the -10 element of bacterial promoters and DNA melting for transcription bubble formation, while region 3 is responsible for the recognition of the -10 extended element. Finally, region 4 recognises the -35 promoter region and interacts with transcriptional activators (Paget, 2015).

Several structures of bacterial RNA polymerases have been determined since the crystal structure of *Thermus aquaticus* RNAP (PDB: 1HQM) was first reported in 1999 (Zhang et al., 1999), all of them contributing to a full understanding of the molecular basis of transcription. Bacterial RNAPs have a particular crab-claw shape with an internal channel running alongside the enzyme (Figure 1.8). The two largest subunits, β and β' form the catalytic centre with their respective DPBBs, and also serve as a scaffold for the DNA, the DNA-RNA hybrid and the RNA formed during transcription. One of the pincers of the crab-claw interacts with the downstream dsDNA, which runs alongside the main DNA channel until it reaches the catalytic cleft, while the RNA-DNA hybrid usually expands perpendicularly to the downstream dsDNA. The two largest subunits are highly conserved amongst bacteria, however they also contain large sequences of lineage-specific insertions (Lane and Darst, 2010b), which are usually located at the periphery of the complex, where density maps are usually of relatively low quality. Comparison with other cellular RNAPs, revealed a conserved architecture amongst msRNAPs with the active site buried at the centre of the enzyme and other conserved structural elements (Lane

and Darst, 2010a). The β' subunit comprises a long bridge helix that separates the main secondary channels and a trigger loop that alternates between a helical and extended conformation throughout nucleotide addition reaction. The N-terminus of the β' subunit forms the clamp, a flexible element that encloses the DNA and DNA-RNA hybrid within the catalytic centre of the enzyme. The lid is included in the clamp and separates the dsDNA strands. The $\beta 1$ and $\beta 2$ domains included in the β subunit, also known as protrusion and lobe, form a channel that holds the non-template *single-stranded DNA* (ssDNA) during transcription bubble formation. The flap, also included in the β subunit, guides the elongated RNA transcript through the RNA-exit channel.

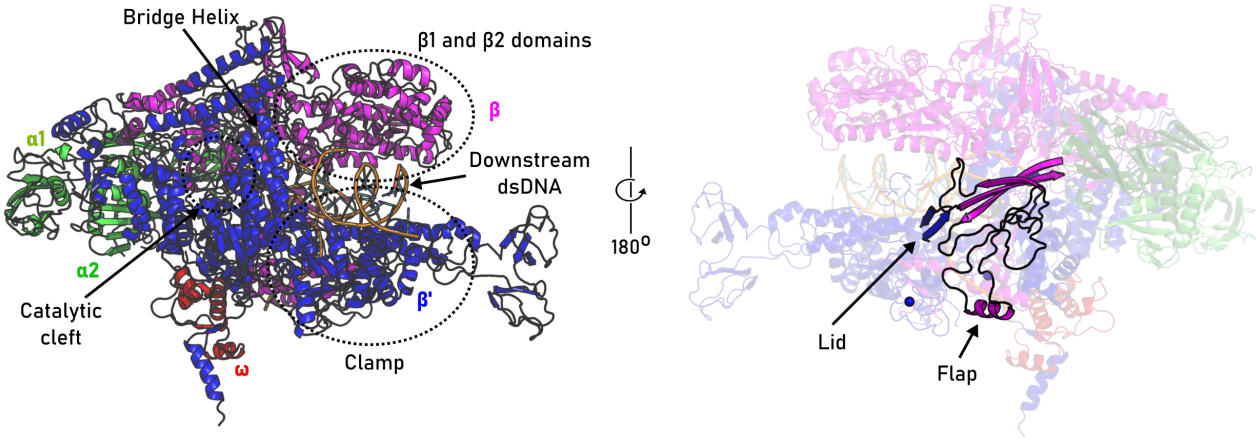


Figure 1.8: Architecture of bacterial msRNAPs. The structure of the *Thermus thermophilus* elongation complex (PDB 2O5J) (Vassylyev et al., 2007) in two different views. Relevant conserved structural elements encircled in black dotted lines and pointed out with black arrows.

1.2.3.2 Bacteriophage ΦKZ encodes its own multi-subunit RNAPs

The majority of bacteriophages have evolved efficient mechanisms to redirect the host transcription machinery to recognise viral promoters, hence ensuring sequential gene expression throughout the distinct developmental stages of their life cycle (Nechaev and Severinov, 2003). An increase in genome sequencing and bioinformatic studies over the last decades has unveiled the presence of distant homologues of RNAP subunits in some eukaryotic viruses, bacteriophages and prophages, which are known as "non-canonical" RNAPs (Sokolova et al., 2020, 2017; Yakunina et al., 2015; Ahn and Moss, 1992). Some bacteriophages, such as T7, N4 or Xp10, encode their own ssRNAPs responsible for transcribing specific subsets of genes during

different stages of infection (Lenneman and Rothman-Denes, 2015; Kroger and Schroeder, 1981; Falco et al., 1977, 1980; Semenova et al., 2005). These viruses use different strategies. T7-like phages initially recruit the host RNAP to transcribe early bacteriophage genes, which encode a ssRNAP that will transcribe middle and late bacteriophage genes (Steitz, 2004). Alternatively, N4-like bacteriophages encode two ssRNAPs which transcribe early and middle bacteriophage genes respectively, and recruit the host RNAP to transcribe bacteriophage late genes (Cho et al., 1995; Gleghorn et al., 2008). One rare exception is a multisubunit RNAP isolated from *Bacillus subtilis* PBS2 infected cells (Clark et al., 1974; Clark, 1978), however minimal information about the PBS2 genomic sequence has complicated the study of this intriguing enzyme and its relationship to other RNAPs. Similarly to bacteriophage PBS2, the giant members of the bacteriophage ΦKZ family were reported to encode two sets of proteins homologous to the two largest subunits of canonical msRNAPs (Ceyssens et al., 2014).

Bacterial RNAPs are inhibited by the antibiotic rifampicin, which binds the β subunit of the enzyme deep in the active site channel and prevents elongation of the RNA transcript after 2 to 3 nucleotides have been added (Campbell et al., 2001). The life cycle of ΦKZ in *Pseudomonas aeruginosa* is unaffected by the host RNAP inhibitor rifampicin, which indicates that ΦKZ is independent of the host transcription machinery and provides its own RNA polymerases throughout the infection cycle (Ceyssens et al., 2014). Whereas almost all known bacteriophage-encoded RNA polymerases are single subunits, ΦKZ is one of only a few viruses to encode multisubunit RNAPs (msRNAPs) (Clark et al., 1974; Yakunina et al., 2015; Sokolova et al., 2017). Bacteriophage ΦKZ encodes two sets of proteins homologous to subregions of the two largest subunits of bacterial RNAP, β and β' subunits and it does not seem to possess homologues for the α or ω subunits, which are usually required for the assembly of a catalytically active enzyme in bacterial msRNAPs (Heyduk et al., 1996; Minakhin et al., 2001).

The ΦKZ *virion RNAP* (vRNAP) is formed by at least four polypeptides (gp178, gp149, gp180, and gp80), and is is injected together with the bacteriophage genome to express ΦKZ early genes (Ceyssens et al., 2014) (Figure 1.9). The ΦKZ vRNAP specifically initiates transcription from early promoters characterised by an AT-rich consensus element and an upstream non-conserved motif. Early transcription occurs independently of bacteriophage protein production, which

indicates that all elements needed for early transcription are provided by the virion. Conversely, transcription of middle and late promoters needs bacteriophage protein synthesis. Amongst the transcribed early bacteriophage genes, there are those encoding a *non-virion RNAP* (nvRNAP) which transcribes all further genes needed for the assembly of new bacteriophages that will end up lysing the cell host (Figure 1.9). Many interesting questions remain unanswered regarding the biology of these non-canonical msRNAPs, including how the subunits assemble without canonical α and ω subunits, or how they recognise specific bacteriophage promoters. Moreover, further investigation of the function, structure and evolution of non-canonical msRNAPs may reveal new insights into the evolutionary origins of msRNAPs.

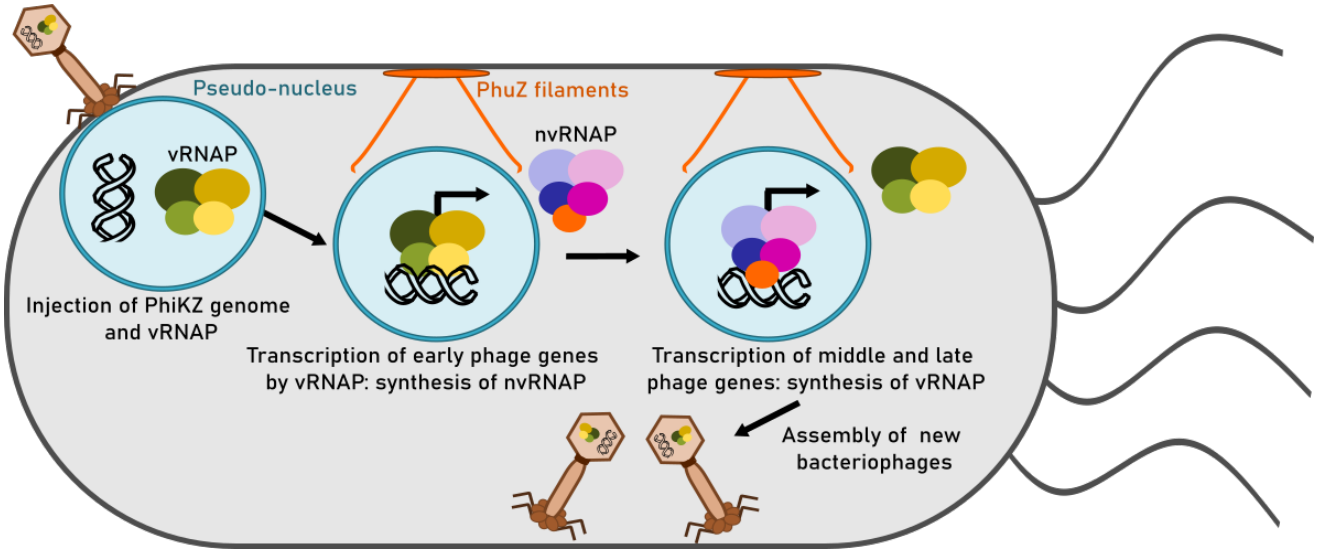


Figure 1.9: Bacteriophage ΦKZ transcription strategy. Each transcription stage is represented with an image and a description below. ΦKZ particles are represented in brown and the host cell (*P.aeruginosa*) in grey. The rest of the elements are labelled in the figure.

1.2.3.3 The ΦKZ non-virion RNA polymerase

In 2015, the ΦKZ nvRNAP was isolated from ΦKZ-infected *P.aeruginosa* cells (Yakunina et al., 2015). The same authors demonstrated that the enzyme is formed by five polypeptides: gp55, gp68, gp71-73, gp74, and gp123. Gp55 and gp74 were found to be homologous to the N- and C-terminus of the bacterial β' subunit, with gp55 including the first DPBB domain (Figure 1.10) (Yakunina et al., 2015). The gp71-73 subunit was found to be homologous to the C-terminal half of the bacterial β subunit including the second DPBB, while gp123 appeared to be a highly divergent homologue of the N-terminus of the bacterial β subunit. The fifth

subunit of the enzyme, gp68, had no similarity to any other protein family, although several homologues were found in other members of the ΦKZ family, suggesting that it is crucial to perform the nvRNAP function (Yakunina et al., 2015). Intriguingly, the ΦKZ non-virion RNAP is sufficient to recognise and transcribe from bacteriophage late promoters without any homologues to bacterial subunits α or ω, or any initiation factors, such as bacterial σ factors or archaeal or eukaryotic transcription factors, suggesting a different mechanism for assembly and promoter recognition. Furthermore, the ΦKZ nvRNAP recognises a late promoter consensus sequence that overlaps with the transcription start point (5'-TATG-3') (Figure 1.11) (Yakunina et al., 2015), in contrast to the -10 and -35 promoter consensus elements of bacterial promoters, suggesting a novel mechanism for promoter recognition. Conversely, ΦKZ middle promoters are not efficiently recognised by the nvRNAP *in vitro* (Yakunina et al., 2015). This observation suggests that middle promoters might be recognised by a modified version of either the vRNAP or nvRNAP or, alternatively, that a bacteriophage-encoded DNA-binding factor could specifically recruit one of the non-canonical polymerases to the middle promoters.

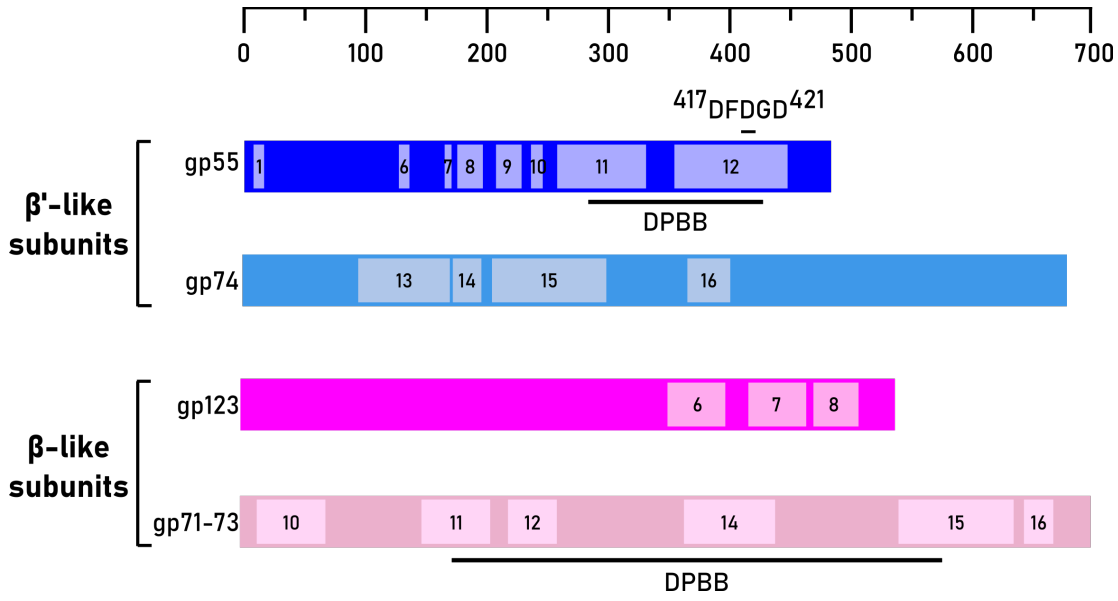


Figure 1.10: Schematic diagram of the conserved sequence and structural regions of the ΦKZ nvRNAP subunits. Each subunit is represented as a box with a different colour. Evolutionarily conserved sequence regions similar to bacterial msRNAPs are labeled according to the Lane and Darst nomenclature (Lane and Darst, 2010a,b). The DPBB domains that form the nvRNAP catalytic centre as well as the aspartate triad are marked with a black line underneath the corresponding subunit. Figure adapted from (Yakunina et al., 2015) under a Creative Commons license.

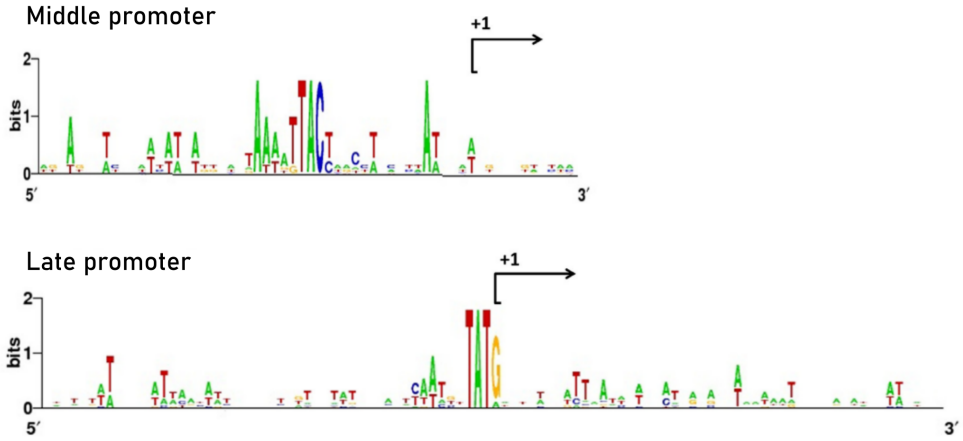


Figure 1.11: Middle and late promoters of the ΦKZ nvRNAP subunits. Sequence logos of the ΦKZ middle and late promoters. Figure adapted from (Yakunina et al., 2015) under a Creative Commons license.

To further investigate the role of the intriguing fifth subunit of the ΦKZ nvRNAP, Orekhova *et al.* created a co-expression system with two plasmids: one encoding the four β/β'-like subunits, and a second encoding only the fifth subunit, gp68 (Orekhova et al., 2019). With this system they could produce *four-subunit* (4s) recombinant nvRNAP, including the four proteins homologous to the two largest bacterial RNAP subunits, and *five-subunit* (5s) recombinant complexes including all the components of ΦKZ nvRNAP. *In vitro* transcription assays revealed that complexes lacking gp68 were unable to elongate existing RNA or bind to a DNA-RNA hybrid, demonstrating that gp68 is essential for even the most basic biological functions of the complex (Orekhova et al., 2019). Similarly, the nvRNAP of another member of the family of giant bacteriophages, *Bacillus subtilis* AR9, also encodes four β/β'-like subunits and requires a fifth subunit for promoter recognition (Lavysh et al., 2016; Sokolova et al., 2017). However, AR9 remains catalytically active in the absence of the fifth subunit, which suggests a less intimate integration into the complex. An unusual msRNAP was previously isolated from *B.subtilis* infected with PBS2, however the enzyme was isolated in two forms from infected cells: a five-subunit complex and a four-subunit complex, in contrast with ΦKZ nvRNAP, which was only purified as a 5s complex from ΦKZ-infected *P.aeruginosa* cells (Yakunina et al., 2015).

Overall, bacteriophage ΦKZ encodes its own msRNAPs which are able to sequentially transcribe all genes needed for the bacteriophage-life cycle, hence being completely independent from the host transcription machinery. The ΦKZ nvRNAP has been purified and biochemically charac-

terised, demonstrating that it is a five-subunit complex including four subunits homologous to the two largest bacterial subunits, and a fifth one with no homology to any other protein with known function, which plays an important role in promoter recognition and transcription initiation. Apart from encoding its own transcription machinery, ΦKZ forms a pseudo-nucleus upon infection which is positioned at the centre of the cell through tubulin-like filaments. Hence, the giant bacteriophage ΦKZ encodes multiple proteins with intriguing evolutionarily parallelisms to both prokaryotes and eukaryotes. A proper study of all the components involved in the ΦKZ life-cycle in *P.aeruginosa* would provide useful knowledge on its evolution and its relationships with its host.

1.3 Objectives

Most important cellular functions are regulated by dynamic multiprotein complexes that can alternate between different conformational states. These complexes are usually difficult to characterise due to their size and flexibility, however recent advances in the cryo-EM field, in particular direct electron detectors and improved software, have allowed the structural determination of large multiprotein complexes at high resolution by single-particle analysis. mTORC1 is a clear example of a large complex composed of several subunits which acts as a central regulator for metabolism, ensuring that cell growth only occurs under favourable conditions. The structure of mTORC1 was determined by cryo-EM (Aylett et al., 2016b; Yang et al., 2016, 2017) providing a complete structure of this kinase. The regulation of mTORC1 involves several complexes that need to be coordinated to integrate multiple signals and generate a single output: the activation or the inhibition of the kinase. The recruitment of mTORC1 to the lysosome and the complexes involved in its regulation have been relatively well characterised in recent years, and the discussion is now moving towards the kinetics and inter-communication between these regulatory complexes.

This PhD project was initially focused on the study of one of the protein complexes regulating Rag GTPase signaling within the mTORC1 pathway that was still poorly characterised; the FLCN complex. We aimed to use structural biology to provide a mechanistic model to explain at a molecular level how the Rag GTPase nucleotide state was modulated by the FLCN complex. To do so we had planned to use single-particle cryo-EM to determine the structure of the FLCN-FNIP2 complex and the molecular basis of its interaction with the Rag GTPase heterodimer. However, after we made substantial progress on the structural characterisation of the FLCN complex on its own, two structures of its complex with the Rag GTPases and Ragulator, responsible for anchoring the Rag GTPases at the lysosome, were published by other groups (Shen et al., 2019; Lawrence et al., 2019). Moreover, immediately after that, there was a college closure for 3 months due to the COVID-19 pandemic. After these three months, access to the lab was still limited, causing major disruptions to the progress of the project.

Considering that I had previously been involved in structural studies of bacteriophage MS2 capsids (de Martín Garrido et al., 2020) and my supervisor, Dr. Aylett, had previously studied components of the interesting bacteriophage ΦKZ, we established a collaboration with Dr. Yakunina's group based at Peter the Great St. Petersburg Polytechnic University. This group is working on the study of the bacteriophage ΦKZ, in particular its transcriptional machinery, and they have recently purified the ΦKZ nvRNAP (Yakunina et al., 2015; Orekhova et al., 2019). Bacteriophage ΦKZ is independent of the host transcriptional machinery and is one of a few bacteriophage exceptions to encode its own set of multi-subunit RNAPs, a virion-RNAP and a non-virion RNAP (nvRNAP). Our collaborators showed that the non-virion RNAP is formed by four subunits homologous to the bacterial β/β'-subunits and a fifth one with no homology to other proteins with known function, but essential for transcription initiation. However, how this fifth subunit is involved in transcription initiation and promoter recognition at a mechanistic level is still poorly understood. Therefore, the aim of this collaboration was to use cryo-EM to determine for the first time the structure of the ΦKZ nvRNAP and elucidate the function of the intriguing fifth subunit in transcription.

Chapter 2

Materials and Methods

2.1 General Procedures

2.1.1 PCR

Routine *polymerase chain reactions* (PCRs) were performed using the Phusion® High-Fidelity PCR Kit (New England Biolabs, UK). We used the conditions recommended by the manufacturer and adjusted the volume accordingly depending on the final reaction volume (Table 2.1).

Component	Volume per 50 µL reaction (µL)	Final concentration
Phusion DNA Polymerase	0.5	1.0 units
5X Phusion HF or GC Buffer	10	1X
10 µM Forward Primer	2.5	0.5 µM
10 µM Reverse Primer	2.5	0.5 µM
10 mM dNTPs	1	200 µM
Template DNA	variable	< 250 ng
Nuclease-free water	To 50 µL	

Table 2.1: Volumes and final concentrations used for standard PCR reactions. Volumes are indicated for a 50 µL reaction. Phusion DNA Polymerase, 5X Phusion HF or GC Buffer and dNTPs were provided by the Phusion® High-Fidelity PCR Kit (New England Biolabs, UK).

After adding the appropriate volume of each component in a tube, reactions were mixed and liquid collected at the bottom of the tube. Reactions were performed using a Biometra TRIO Series Thermal Cycler (Analytik Jena, Germany) using the parameters indicated in Table 2.2. Samples were left at 4 °C until collection from the Thermal Cycler and stored at -20 °C until use.

Step	Temperature (°C)	Time
Initial Denaturation	98	1 min
32 Cycles	98	15 s
	Annealing temperature (variable)	30 s
	72	15s/kb (variable)

Table 2.2: Conditions used to run standard PCR reactions. Table showing the cycle conditions used to run standard PCR reactions. The annealing temperature was adapted for each set of primers and the extension time to the length of the amplified fragment.

2.1.2 *E.coli* transformation

For cloning procedures, plasmids were transformed by heat shock into NEB 10-beta Competent *E.coli* cells (New England Biolabs, UK). After thawing a tube of cells on ice, we added 1-2 µL of plasmid into one aliquot and carefully tapped the tube to mix the cells with the DNA (vortexing or mixing with the pipette could damage the cells). We left the mixture on ice for 30 *minutes* (min) and then tubes were incubated at 42 °C for 45 *seconds* (s). After heat shock, we left the cells on ice for 5-10 min and plated them on a *Lysogeny broth* (LB) (Bertani, 2004) agar plate containing the required antibiotics (100 µg/mL ampicillin).

2.1.3 Golden Gate Reaction

To transfer the genes into their corresponding expression vectors we used the Golden Gate strategy (Engler et al., 2008). This methodology utilises from the restriction enzyme BsaI to specifically insert a DNA fragment flanked with specific restriction sites into a destination plasmid containing those cleavage sites. The advantage of using BsaI is that this enzyme cuts

outside the restriction enzyme sites. We prepared reactions at a final volume of 20 µL using the conditions indicated in Table 2.3.

Component	Quantity per 20 µL reaction (µL)
Destination plasmid	100-200 ng
Insert	2:1 molar ratio
DpnI*	1 µL
BsaI*	1 µL
T7 ligase*	1 µL
T7 ligase buffer*	2 µL
Nuclease-free water	Up to 20 µL

Table 2.3: Standard volumes used to perform Golden Gate reactions. Table showing the components used to run a golden gate reaction. Components marked with a * were obtained from New England Biolabs (UK).

After adding the appropriate volume for each component, liquid was collected at the bottom of the tube and the assembly reaction was run in a Biometra TRIO Series Thermal Cycler (Analytik Jena, Germany) for 30 cycles alternating 2 min and 30 s at 37°C and 2 min and 30 s at 16 °C. After the last cycle, reactions were incubated at 37°C for 10 min and left at 4°C until collection. Reactions were transformed into NEB 10-beta Competent *E.coli* cells as indicated in section 2.1.2.

2.1.4 HiFi assembly

We used the NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs, UK) to assemble multiple DNA fragments. This method is very efficient and allows assembly of several DNA fragments with varied overlapping ends (15-30 bp) regardless of their length. Generally, 0.1 pmol of each fragment were mixed. Quantities of each fragment were added, and then the same volume of the 2X NEBuilder® HiFi DNA Assembly Master Mix was added. If the number of fragments was higher than 5 the quantity of each fragment was reduced to <0.1 pmol to avoid having more than 0.5 pmols in total, as recommended by the manufacturer. After proper mixing of the DNA fragments with the HiFi Assembly Master Mix, the reaction was incubated for 1 h at 50 °C. For assemblies with more than five fragments, incubation time was usually

increased to 2h.

2.1.5 Agarose Gel Electrophoresis

We prepared 1 % agarose gels by mixing one tablet containing 0.5 g of agarose (Agarose Tablets, Meridian Bioscience, USA) with 50 mL of 1x TBE buffer (89 mM Tris, 89 mM boric acid and 3.2 mM EDTA) and left it to dissolve for 5 min at room temperature. After that, the mixture was heated for 1 min and let to cool down at room temperature for 10-15 minutes. Then, 5 µL of SYBR Safe (Invitrogen, Thermo Fisher Scientific, USA) were added and the liquid was poured into a horizontal electrophoresis system (Bio-Rad Laboratories, Inc., USA). The gel was left to cool down at room temperature for 30min-1h. Samples were prepared by mixing them with Gel Loading Dye, Purple (6X) (Bio-Rad Laboratories, Inc., USA). After loading the samples into the gel wells, gels were run in TBE buffer at 75-100V for 40-60 minutes (voltage and time were adjusted depending on the length of the DNA fragments analysed).

2.1.6 SDS-PAGE

Sodium dodecyl sulphate–polyacrylamide gel electrophoresis (SDS-PAGE) was performed using 4–20 % Mini-PROTEAN® TGX™ Precast Protein Gels (Bio-Rad Laboratories, Inc., USA). 15 µL of sample were diluted with 5 µL of 4x LDS Sample Buffer (Expedeon, now Abcam, UK) and heated at 90 °C for 2-3 min. Gels were run at 200 V for 36-40 min in 1x SDS buffer (25 mM Tris, 192 mM glycine, 0.1 % SDS). To visualise proteins, gels were stained with InstantBlue® Coomassie (Abcam, UK) for 2h-overnight and subsequently destained with water for 2h-overnight. Gels were imaged on a Gel Doc XR+ (Bio-Rad Laboratories, Inc., USA) using the preset protocol for Coomassie stain.

2.1.7 Native PAGE and Gel-shift

Native PAGE was performed using 4–20 % Mini-PROTEAN® TGX™ Precast Protein Gels (Bio-Rad Laboratories, Inc., USA). Samples were prepared by mixing 10 µL of each sample with 10 µL of 2X Loading Buffer without SDS (63.5 mM Tris-Cl pH=6.8 and 40% Glycerol). Loading Buffer was prepared without any tracking dye to avoid masking the signal of Cy5-oligonucleotides. Hence, one well was loaded with 4x LDS Sample Buffer (Expedeon, now Abcam, UK) to monitor the front of the gel. 20 µL of samples were loaded on the gel, that was run for 20-24 min at 200 V if it was run at room temperature (for DNA gels) or 40 min at 120 V if the gel was run at 4 °C (for gel-shifts to check DNA-protein binding).

2.1.8 Mass spectrometry

Protein bands of interest were manually excised from a Coomassie-stained SDS-PAGE gel. Individual gel slices were prepared for mass spectrometry through extensive washing with ultrapure water. Samples were analysed at the Mass spectrometry and Proteomics facility of the University of St Andrews.

2.1.9 General Protocols for the preparation of EM grids

2.1.9.1 Grid washing and plasma cleaning

Grids were washed with ultrapure water and ethyl acetate (Sigma-Aldrich, USA) to eliminate any kind of plastic contaminants arising from the Quantifoil grid production. After washing, grids were left to air-dry.

Usually, grids are plasma cleaned prior to sample application to remove any contaminants accured during production or washing, and to render them hydrophilic. Grids were plasma cleaned using a Basic Plasma Cleaner PDV-23G-2 (Harrick Plasma, USA) for 10-15 s in air using the high mode setting.

2.1.9.2 Preparation of Graphene Oxide covered grids

Preparation of cryo-EM samples on *Graphene Oxide* (GrOx) covered grids is routinely done in our lab using a well-established protocol that has been successful in obtaining high-resolution structures of different complexes (de Martín Garrido et al., 2021c,b; Ramlaul et al., 2021). This method uses GrOx flakes usually prepared using the Hummers method (Hummers and Offeman, 1958). A 1 % GrOx and 0.01 % *n-Dodecyl-beta-Maltoside* (DDM) solution was prepared by mixing 2.5 µL of Graphene Oxide 2 mg/mL (Sigma-Aldrich, USA), 2.5 µL of 1 % DDM (GLYCON Biochemicals GmbH, Germany) and 245 µL of ultrapure water (final volume 250 µL). Immediately before use, the mixture was vortexed vigorously for 1 min to ensure that the GrOx flakes were properly suspended.

To prepare GrOx-covered *transmission electron microscope* (TEM) grids, 5 µL of GrOx-DDM suspension and 5 µL of ultrapure water were dropped onto a piece of paraffin. Firstly, the grid was gently deposited on top of the GrOx suspension droplet and incubated for 1 min with the carbon-covered side in contact with the liquid. After incubation, the grid was lifted with negative-action tweezers and the droplet of water subsequently recovered by gently touching the liquid gently with the copper side of the grid. At this point, while holding the grid with the tweezers, there should be a droplet of GrOx-DDM suspension on the carbon-covered side of the grid and a droplet of water on the copper side. Then, the grid was carefully blotted with the water droplet-side down against a piece of filter paper. The principle here is that blotting of the water will settle the GrOx suspension throughout the grid by capillary action.

2.2 The FLCN complex and its interacting partners

2.2.1 Materials

2.2.1.1 Plasmids

All *Spodoptera litura* genes were individually cloned in a pG plasmid (Figure 2.1) making use of the Golden Gate method (Engler et al., 2009)(see section 2.1.3). The pG plasmid is a derivative of the expression vector used by the Prof. Wigley group (Rees et al., 2017). It contains an Ampicillin resistance gene for selection of transformed colonies and Tn7 transposition sites for subsequent site-specific integration into the bacmid in DH10MultiBac cells (Berger et al., 2004). After that, genes were assembled into multiple-gene pG vectors using the NEBuilder HiFi Assembly kit (see section 2.1.4).

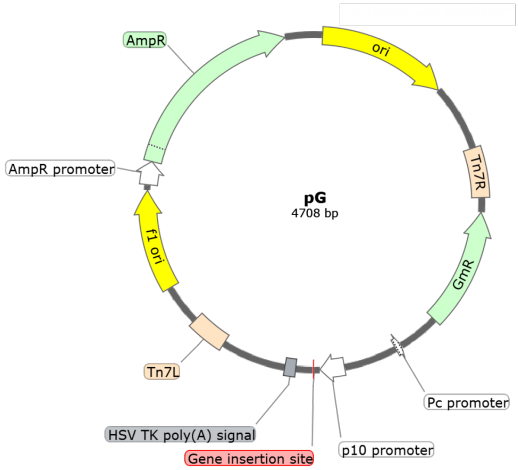


Figure 2.1: The pG plasmid map. Representative diagram of the pG plasmid map with the main elements of the plasmid indicated. Figure made with SnapGene.

2.2.1.2 Genes and proteins

S.litura gene sequences were identified by multiple alignment using *Basic Local Alignment Search Tool* (BLAST). Sequences were codon-optimized for *Spodoptera frugiperda* expression, checked for restriction enzymes sites, and ordered through Integrated DNA technologies (USA). All constructs included an SBP tag (Keefe et al., 2001) for affinity purification in one of the

proteins. All protein constructs used are listed in (Table 2.4) and were sequenced verified using Eurofins TubeSeq services.

Vector	Genes	Tagged Gene
pG	*Sl* FLCN-FNIP2	FNIP2-SBP
pG	*Sl* RagA-RagC	RagC-SBP

Table 2.4: List of protein constructs used for *Sl* FLCN-FNIP2 and RagA-RagC purification. Table showing the list of constructs used in Chapter 3. The tagged protein in each construct is indicated.

2.2.2 Protein Expression

2.2.2.1 Virus generation and amplification

We expressed both *Sl* FLCN-FNIP2 and *Sl* RagA-RagC in insect cells (Sf9) using the MultiBac baculoviral expression system (Berger et al., 2004). Genes encoding subunits of each protein complex were cloned together into a pG plasmid (Figure 2.1) containing Tn7 transposition sites, which allowed subsequent integration of the expression cassette into the bacmid in DH10MultiBac cells (Berger et al., 2004). After transposition, positive colonies were selected and the bacmid DNA was purified by ethanol precipitation using the buffers included in the QIAGEN Plasmid Plus Kit (QIAGEN, Netherlands).

We usually screened 2-4 bacmids from different positive colonies for each construct. Transfection of purified bacmids was achieved by mixing the DNA with Cellfectin^{TM} II Reagent (Thermo Fisher Scientific, USA). 2 mL of Sf9 cells grown in Insect-XPRESS^{TM} Protein-free Insect Cell Medium (Lonza, Switzerland) at 0.5 x 10^6 cells/mL were grown in a 6-well plate and then transfected with the bacmid-Cellfectin mix. Cells were left to incubate for 72 h at 27°C. Viruses were amplified in Sf9 cells, by infecting 50 mL of Sf9 cells at 1 x 10^6 cells/mL with the 2 mL transfected cells, and left incubating at 27 °C and agitated at 130 rpm. After 72h, cells were harvested by centrifugation at 1000 rcf for 10 min and viruses were isolated collecting the supernatant, which was designated as generation of viruses 1 (P1). A second round of amplification was performed by infecting 50 mL of Sf9 cells at 1 x 10^6 cells/mL with 1 mL of

P1 (infection rate 1:50) and leaving them incubating for 72 h at 27 °C and agitated 130 rpm. The second generation of viruses (P2) were finally isolated by centrifugation at 1000 rcf for 10 min. Viruses were kept at 4 °C until use with 1:50 dilution of *Fetal Bovine Serum* (FBS) (Sigma-Aldrich, USA).

2.2.2.2 Establishment of protein expression conditions

Expression of *Sl* FLCN-FNIP2 and *Sl* RagA-RagC was firstly tested with harvested cells from the P2 virus amplification. After virus isolation by centrifugation, the pellet was resuspended in 5 mL of the corresponding Resuspension Buffer (see Table 2.5 and 2.6), supplemented with Halt^{TM} Protease Inhibitors Cocktail 100X (Thermo Fisher Scientific, USA). Cells were lysed by sonication in a beaker placed on ice, for 2 min with a pulse of 1 s on and 4 s off, at 40% amplitude. Cell debris was separated from the soluble fraction by centrifugation for 25 min at 18 000 rpm, 4 °C and the supernatant was collected for subsequent analysis on SDS-PAGE.

After testing that both *Sl* FLCN-FNIP2 and *Sl* RagA-RagC were soluble and successfully expressed in the P2 generation of viruses, protein expression was scaled up by infecting larger volumes of Sf9 cells under the same conditions. For *Sl* FLCN-FNIP2 complexes, 400 mL of Sf9 cells were infected with 8 mL of previously tested P2 virus (keeping an 1:50 infection ratio) and incubated at 27 °C and 130 rpm for 72h. For *Sl* RagA-RagC complexes, 200 mL of Sf9 cells were infected with 4 mL of previously tested P2 virus (1:50 infection ratio) and incubated at 27 °C and agitated at 130 rpm for 48h.

2.2.3 Purification of the *Sl* FLCN-FNIP2 complex

After 72h, 400 mL of Sf9 cells infected with P2 *Sl* FLCN-FNIP2 virus were harvested by centrifugation for 10 min at 1000 rcf. Cells were resuspended in 20 mL of Resuspension buffer (Table 2.5). Lysis was performed by sonication in a beaker placed on ice for 2 min 30 s, with a pulse of 1 s an 4 s off, and at a 40% amplitude. Cell debris was separated from the soluble fraction by centrifugation for 45 min at 18 000 rpm, 4 °C, and the supernatant fraction was

collected for subsequent purification.

All the affinity purification steps were performed at 4 °C. 20 mL of supernatant was loaded onto a 5 mL StrepTrap™ HP (Merck KGaA, Germany) pre-equilibrated with 5 *column volumes* (CV) (50 mL) of Resuspension Buffer (Table 2.5). Proteins that bound to the column non-specifically were washed out with 5 CV of Resuspension Buffer (Table 2.5). To try to minimise aggregation, SBP-tagged proteins were eluted with the minimum possible volume. The column was filled with 4 mL of Elution buffer (Table 2.5) and left incubating for 10 min. After incubation, elution was performed using a syringe in 2.5 mL fractions of Elution Buffer (Table 2.5). After purification, the column was washed with 5 CV of 100 mM Tris-Cl pH=8.0 and stored in 20 % ethanol. Elution fractions were left on ice until injection into the gel-filtration column.

Samples from affinity purification were analysed by SDS-PAGE and the elution fraction displaying the highest amount of protein in the gel (usually Elution fraction 1) was subsequently injected in a Superdex 200 10/300 GL column (GE Healthcare, now Cytiva, USA) pre-equilibrated with SEC Buffer (Table 2.5). The protein was monitored through changes in absorbance in 230 nm and 280 nm and the retention volume was compared to Gel-filtration standards (Bio-Rad Laboratories, Inc., USA) composed of: thyroglobulin (670 000 Da), γ-globulin (158 000 Da), ovalbumin (44 000 Da), myoglobin (17 000 Da), and vitamin B12 (1 350 Da). Gel-filtration was performed on an ÄKTA purifier 10 (GE healthcare, now Cytiva, USA) for 24 mL at 0.25 mL/min and collecting 0.25 mL fractions. Representative fractions were analysed by SDS-PAGE. Protein concentration was determined using a DS-11+ Spectrophotometer (DeNovix, USA). Samples were immediately used for cryo-EM sample preparation (0.1-0.3 mg/mL). Samples used for GTPase assays were stored at 4 °C for a maximum of 24 h until use.

Buffer Name	Components
Resuspension Buffer	25 mM K-HEPES pH=7.6, 250 mM KCl, 0.5 mM EDTA, 1 mM TCEP, 10 % Glycerol
Elution Buffer	25 mM K-HEPES pH=7.6, 250 mM KCl, 0.5 mM EDTA, 1 mM TCEP, 10% Glycerol, 5 mM Desthiobiotin
SEC Buffer	25 mM K-HEPES pH=7.6, 250 mM KCl, 0.5 mM EDTA, 1mM TCEP

Table 2.5: List of buffers used for *Sl* FLCN-FNIP2 purification.

2.2.4 Purification of the RagA-RagC complex

After 48 h, 200 mL of Sf9 cells infected with P2 *Sl* RagA-RagC virus were harvested by centrifugation for 10 min at 1000 rcf. Lysis and affinity purification was performed as explained in Section 2.2.3 using the buffers in Table 2.6.

All the eluted fractions from the affinity purification that contained *Sl* RagA-RagC complexes according to SDS-PAGE, were concentrated with Amicon Ultra centrifugal filter units Ultra-15, MWCO 30 kDa (Millipore, USA) up to 2 mL by sequential centrifugation rounds of 12 min at 4000 rpm, 4 °C. 2 mL of concentrated sample were injected into a Superose 12 100/300 GL column (GE Healthcare, now Cytiva, USA) pre-equilibrated in SEC buffer (Table 2.6) using an ÄKTA purifier 10 (GE healthcare, now Cytiva, USA). Gel-filtraion was run for 24 mL at 0.25 mL/min and collecting 0.25 mL fractions. Representative fractions were analysed by SDS-PAGE and samples containing *Sl* RagA-RagC complexes were concentrated up to 200 µL. Protein concentration was determined using a DS-11+ Spectrophotometer (DeNovix, USA). 10 µL aliquots at 37 mg/mL were flash frozen and stored at -80 °C until use.

Buffer Name	Components
Resuspension Buffer	25 mM K-HEPES pH = 7.6, 175 mM KCl, 0.5 mM EDTA 1 mM TCEP 10% Glycerol
Elution Buffer	25 mM K-HEPEs pH=7.6, 175 mM KCl, 0.5 mM EDTA, 1 mM TCEP, 10% Glycerol, 5 mM Desthiobiotin
SEC Buffer	25 mM K-HEPES pH=7.6, 175 mM KCl, 0.5 mM EDTA, 1mM TCEP

Table 2.6: List of buffers used for *Sl* RagA-RagC purification.

2.2.5 FLCN-FNIP2-RagA-RagC complex formation

2.2.5.1 Size-exclusion chromatography

To prepare FLCN-FNIP2 complexes with RagA-RagC, the limiting factor was usually the FLCN-FNIP2 concentration obtained after size-exclusion chromatography, which was very vari-

able (0.1-2 mg/mL), as the sample was prone to aggregation, concentration was avoided during FLCN-FNIP2, hence limiting the achievable final concentration.

To prepare FLCN-FNIP2-RagA-RagC complexes in the presence of GTPγS, we initially diluted *Sl* RagA-RagC complexes to a final concentration of 30 µM with Resuspension Buffer supplemented with 1 mM $MgCl_2$ and 0.5 mM GTPγS. Then we mixed equal volumes of the *Sl* RagA-RagC mixture with *Sl* FLCN-FNIP2 complexes at 15 µM (2:1 ratio) and incubated these at 4 °C for 15 min. After incubation, complexes were injected onto a pre-equilibrated Superdex 200 10/300 GL (GE Healthcare, now Cytiva, USA) with SEC Buffer (Table 2.6) supplemented with 1 mM $MgCl_2$. 0.25 mL fractions were analysed by SDS-PAGE.

To prepare FLCN-FNIP2-RagA-RagC complexes in the presence of GppNHp and GDP, we initially mixed *Sl* FLCN-FNIP2 with *Sl* RagA-RagC at 1:2 ratio (2.3 µM and 4.6 µM respective final concentrations) in SEC Buffer (Table 2.6) supplemented with 100 µM GppNHp, 100 µM of GDP and 1 mM $MgCl_2$. The mixture was left incubating for 5 h at 4 °C. After incubation, complexes were injected onto a pre-equilibrated Superdex 200 10/300 GL (GE Healthcare, now Cytiva, USA) with SEC Buffer (Table 2.6) supplemented with 1 mM $MgCl_2$. 0.25 mL fractions were analysed by SDS-PAGE.

2.2.5.2 Density gradients: GraFix and Prefixation

For asymmetric and dynamic macromolecular complexes, structural heterogeneity can be a limiting factor to achieve high-resolution reconstruction by cryo-EM. Commonly used separation techniques, such as chromatography or dialysis, may result in increased complex disruption because they involve exposure to various surfaces. The combination of density gradients with chemical fixation, also know as GraFix (Kastner et al., 2008; Stark, 2010) provides a highly reproducible method which successfully stabilises macromolecular complexes and separates them from aggregates and other contaminants. In our laboratory we use several variations of GraFix.

Prior to prefixation, we performed a cross-linking screening using serial dilutions of *glutaraldehyde* (GTA) and analysing complex formation on SDS-PAGE. Before the addition of the fixation

agent, *Sl* FLCN-FNIP2 complexes were mixed with *Sl* RagA-RagC at a final concentration of 0.6 µM and 1.2 µM respectively in SEC buffer supplemented with 100 µM GTPγS and 1mM $MgCl_2$. Samples were mixed with ten serial 2-fold dilutions of cross-linking reagent, from 0.5 % to 0.002 %. After cross-linking, samples were mixed with SDS-loading buffer and subsequently analysed by SDS-PAGE.

For prefixation we used a modification of PreFix (de Martín Garrido et al., 2021a). 10-30% glycerol gradients were prepared with resuspension buffer (Table 2.5) supplemented with the appropriate percentage of glycerol. The top buffer containing 10% glycerol was supplemented with 0.05% GTA. *Sl* FLCN-FNIP2 complexes and *Sl* RagA-RagC complexes were mixed at a final concentration of 0.8 µM and 1.6 µM respectively with resuspension buffer (Table 2.5) supplemented with 1 mM $MgCl_2$ and 100 µM of GTPγS, and were left incubating for 10 min on ice. To apply the sample to the gradients, some liquid was removed from the top (500 µL). Gradients were centrifuged at 30 000 rpm for 19 h at 4 °C. After centrifugation, gradients were fractionated into 0.25 mL fractions with a peristaltic pump connected to an AKTA purifier 10 (GE healthcare, now Cytiva Life Sciences, Global Life Sciences Solutions, USA). Fractions were then analysed by SDS-PAGE.

2.2.6 GTPase assays

2.2.6.1 Malachite green Phosphate assay kit

GTPase assays were initially performed using the Malachite green Phosphate assay kit (Sigma Aldrich, UK). This kit uses a malachite green reagent that forms a stable dark green color with free phosphate released after GTP hydrolysis, resulting in a colorimetric product that can be detected at 620 nm (600-660nm). Therefore the kit provides a direct procedure for measuring the NTPase activity in a microplate by measuring the absorbance at 620 nm, which is proportional to the enzyme activity.

All samples and standards of the assay were run in triplicate in a final volume of 40 µL on a 96-well plate. The assay provides a Phosphate Standard solution to prepare Phosphate Standards

to associate the detected absorbance with the quantity of phosphate. Phosphate Standards were prepared as recommended by the provider: 0, 12.5, 25, 31.25, 37.5 and 50 µM. After plotting the absorbance for each known phosphate concentration ([Pi]), the experimental data was fitted by linear regression (Figure 2.2).

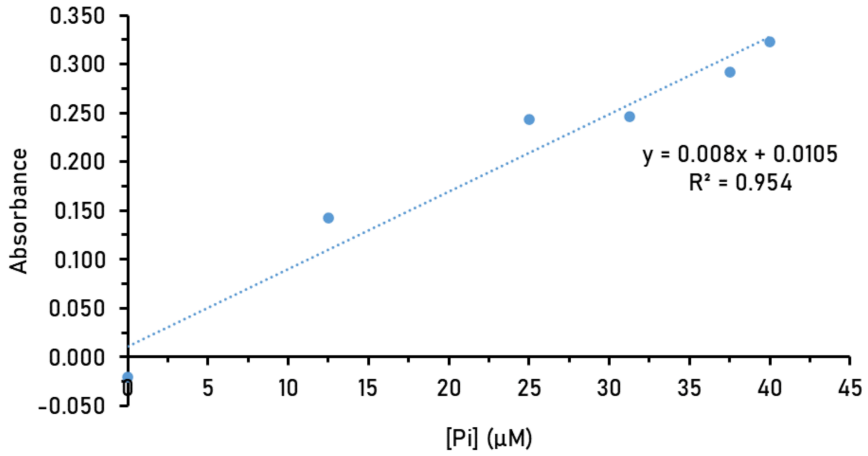


Figure 2.2: Malachite Green GTPase assay Phosphate Standards curve. After plotting the absorbance (y axis) for each known phosphate concentration (x axis, [Pi] in µM) experimental data was fit into a linear equation by linear regression. The equation is displayed on the graph.

2.2.7 GTPase-Glo™ Assay

The GTPase-Glo™ Assay (Promega, USA) has a different principle to that of the previous assay (Malachite green Phosphate assay): it detects the remaining GTP after a GTP hydrolysis reaction. The GTP is initially converted to ATP using the GTPase-Glo™ reagent, and the ATP is finally detected using a luciferase and luciferin substrate to produce bioluminescence. Based on that, the GTPase activity is inversely correlated to the amount of light produced.

Each reaction was run in triplicate in a final volume of 10 µL on a 96-well plate. A positive control to measure the initial quantity of GTP in the reactions was prepared by diluting GTP at a final concentration of 5 µM in GTPase/GAP buffer supplemented with 1 mM DTT. Negative controls were prepared by adding 10 µL of GTPase/GAP buffer and 10 µL of SEC buffer (used to purify the proteins Figure 2.5) on the plate wells. A 2x RagA-RagC GTPase solution (1 µM, as recommended by the manufacturer) was prepared in GTPase/GAP Buffer containing 1 mM DTT. Then, serial dilutions of 2x FLCN-FNIP2 were prepared in GTPase/GAP Buffer

containing 1 mM DTT containing 10 µM GTP (2.4, 1.2, 0.6, 0.3, 0.15, 0.8 µM). Reactions were started by adding 5 µL of 2x RagA-RagC GTPase to 5 µL of each 2x FLCN-FNIP concentration. Reactions were left for 3h at room temperature.

During incubation we prepared the Reconstituted GTPase-Glo™ Reagent by mixing 0.8 µL of GTPase-Glo™ Reagent 500x with 0.2 µL of ADP 10 mM in 399 µL of GTPase/GAP buffer, following the instructions of the manufacturer (Promega, USA). After 3h, 10 µL of the reconstituted GTPase-Glo ™ Reagent were added to each well and the plate was incubated for a further 30 min at room temperature. After incubation, 20 µL of Detection Reagent were added and luminescence was measured in a CLARIOstar microplate reader (BMG LABTECH, Germany) in *relative light units* (RLU). For data analysis, background luminescence corresponding to the average of the wells containing GTPase/GAP buffer wells was substracted from samples measurements, and then triplicate values for each condition were averaged between them. Error bars represent the standard deviation between triplicates. One-way ANOVA tests performed with GraphPad Prism were used for statistical analysis, with $P < 0.05$ taken to be significant.

2.2.8 Single-Particle Electron Microscopy

2.2.8.1 Negatively-stained specimens

Negatively-stained grids were prepared using a variation of the carbon floating method (Harris, 1982) using a support floatation block (de Martín Garrido et al., 2021a,c). Mica sheets with a deposited layer of amorphous carbon were floated in a well containing *Sl* FLCN-FNIP2 sample (0.1 mg/mL), before floating in another well containing 2 % (w/v) uranyl acetate. The floating carbon was then recovered onto a washed and plasma-cleaned Quantifoil R 1.2/1.3 300 copper mesh grid (Quantifoil, Germany) and left to air-dry. Grids were imaged on a Tecnai T12 BioTwin (Thermo Fisher Scientific, USA) electron microscope at a magnification of 52 000 fold operated at 120 kV.

2.2.8.2 Vitreous specimens

Quantifoil R 1.2/1.3 on 300 copper mesh grids (Quantifoil, Germany) were washed and plasma-cleaned for 40 s on low mode using a Basic Plasma Cleaner PDC-23G-2 (Harrick Plasma, USA). Samples were then applied to the grids and blotted for 0 s with a blot force of -25 using a Vitrobot Mark IV (Thermo Fisher Scientific, USA) at 4 °C and 100 % humidity and plunged into liquid ethane. The blotting pads were prepared using Whatman filter paper number 41, 55 mm (GE Healthcare Life Sciences, now Cytiva, USA).

2.2.8.3 Data Collection and Image Processing

Cryo-grids of *Sl* FLCN-FNIP2 complex were imaged on a Titan Krios G3i (Thermo Fisher Scientific, Waltham, MA, USA) equipped with a K3 direct electron detector (GATAN, San Diego, USA) at the London Consortium for Electron Microscopy microscope located at the Francis Crick Institute institute (London, UK). Micrographs were collected at a magnification of 81 000 fold, an acceleration voltage of 300 kV, and a total dose of 40 $e^-/Å^2$, at a defocus of -1 µm and using a *Volta phase plate* (VPP) (Table 2.7).

Pixel size (Å/pix)	1.10
Defocus (µm)	-1
Magnification (X)	81 000
Exposure time (s)	3.2
Total dose ($e^-/Å^2/s$)	40
Frame dose ($e/Å^2$)	1.2
C2 aperture (µm)	70
Gatan imaging energy filter (GIF) slit width (eV)	20

Table 2.7: Summary of data collection parameters for the *Sl* FLCN-FNIP2 high-resolution data using a VPP. Data was collected on a Titan Krios G3i (Thermo Fisher Scientific, USA) equipped with a K3 direct electron detector (GATAN, San Diego, USA) at the London Consortium for Electron Microscopy microscope located at the Francis Crick Institute (London, UK).

A total of 3 146 movies were aligned and weighted by dose using MotionCor2 (Zheng et al., 2017)

to obtain a final image. Poor quality micrographs were discarded visually based on irregularities of the Thon rings estimated by CTFFIND4 (Rohou and Grigorieff, 2015). From a total of 2 409 processed images, 1 889 319 particles were selected using BATCHBOXER (Tang et al., 2007). Particles were classified into two-dimensional class averages over several rounds of refinement using RELION3 (Zivanov et al., 2018), with only two views of the complex converging into high-resolution classes. A total of 181 177 particles corresponding to those classes were further classified in three-dimensional averages using a reference map filtered to 10 Å created with the human FLCN-FNIP2 structure as a model (Shen et al., 2019; Lawrence et al., 2019). These were further classified without alignment into five classes in three-dimension. Three class averages including 180 891 particles were retained for further refinement with 152 767 yielding a final reconstruction of the *Sl* FLCN-FNIP2 complex reaching 5.7 Å according to an independent half-set FSC of 0.143.

2.2.9 Structural model generation

Structural models were initially generated with Swiss-Model (Waterhouse et al., 2018) using the recent human FLCN-FNIP2 structure as a model (Shen et al., 2019; Lawrence et al., 2019). Initially, the *Sl* protein sequences of FLCN and FNIP2 were aligned with the respective human sequences using CLUSTAL Omega (Sievers et al., 2011). Then the FLCN-FNIP2 PDB files (PDB 6NZD, 6ULG) were provided to Swiss-Model which generated structural models for both *Sl* FLCN and *Sl* FNIP2 independently. Large unstructured regions of *Sl* FNIP2 were deleted prior to fitting into the 5.7 Å resolution map. The final models were fitted in the map and included the following sections: 1-430 for *Sl* FLCN (all sequence) and 27-138, 297-501, 894-1047 for *Sl* FNIP2.

2.3 The ΦKZ non-virion polymerase

2.3.1 ΦKZ non-virion RNAP expression and purification

ΦKZ nvRNAP complexes were purified by Dr. Yakunina's group at Peter the Great St. Petersburg Polytechnic University (Russia). Reproduced here for the readers benefit. We performed the gel shift for the P119L RNA/DNA template and all the cryo-EM.

2.3.1.1 Plasmids

Our collaborators designed a co-expression system to express recombinant ΦKZ nvRNAP which is functionally equivalent to the native polymerase and can be obtained at larger amounts with a less laborious protocol (Orekhova et al., 2019).

Briefly, the ΦKZ nvRNAP individual genes, gp55, gp74, gp71-73 and gp123 were cloned into a set of pET-based plasmids and later assembled in a single pnvCo-Ex plasmid (Orekhova et al., 2019). The pGp68 plasmid was also produced earlier as detailed in Orekhova et al. (2019). This system allows expression of 4s nvRNAP, which only includes the four β/β' -like subunits, and co-expression of the full complex, 5s ΦKZ nvRNAP.

2.3.1.2 Protein expression and purification

BL21(DE3) *E.coli* cells were transformed by one or two plasmids with compatible origin to express 4s and 5s nvRNAP respectively. Expression was induced by the addition of 1 mM IPTG on OD_{600} = 0.5-0.7 and performed at 22 °C for 3 h.

Recombinant ΦKZ nvRNAP complexes were purified as previously described (Orekhova et al., 2019). For subunit complex validation, 1 g wet mass of each corresponding cell pellet was disrupted by sonication in 10 mL of buffer A containing 5 mM Imidazole (Table 2.8) followed by centrifugation at 11 000 rcf for 30 min at 4 °C. The soluble fraction was loaded onto a HisTrap

HP 1 mL (GE Healthcare Life Sciences, now Cytiva, USA) previously equilibrated, and then washed with buffer A with 5 mM Imidazole (Table 2.8). Recombinant complexes were then eluted with buffer A containing 250 mM Imidazole (Table 2.8). Size-exclusion chromatography was performed on a Superdex 200 Increase 10/300 GL (GE Healthcare Life Sciences, now Cytiva, USA) in TGED buffer with 200 mM NaCl (Table 2.8).

Buffer Name	Components
Buffer A	40 mM Tris-Cl pH 8.0, 10% glycerol, 500 mM NaCl1 mM DTT
TGED buffer	20 mM Tris-Cl pH 8.0, 5% glycerol, 0.5 mM EDTA, 1 mM DTT
Sample buffer	15 mM Tris-Cl pH 8.0, 150 mM NaCl
Transcription buffer	40 mM Tris-Cl pH 8.0, 10 mM MgCl2, 5 mM DTT

Table 2.8: List of buffers used for the ΦKZ nvRNAP purification and biochemical assays.

For purification of gp68 without a tag, 1 g wet mass of cell pellet was resuspended in 10 mL TGED buffer with 50 mM NaCl and disrupted by sonication. Clarified lysate was passed through a HiTrap Q XL column (GE Healthcare Life Sciences, now Cytiva, USA) equilibrated in the same buffer. The resulting sample was loaded onto a HiTrap Heparin HP (GE Healthcare Life Sciences, now Cytiva, USA) followed by washing using TGED buffer with 100 mM NaCl. The protein was eluted using TGED buffer with 200 mM NaCl. The eluted fractions were concentrated (Amicon Ultra-4 Centrifugal Filter Unit with Ultracel-30 membrane, EMD Millipore, Merck, USA) and purified by size exclusion chromatography using a Superdex 200 Increase 10/300 (GE Healthcare Life Sciences, now Cytiva, USA) in TGED buffer with 200 mM NaCl.

2.3.2 DNA-template preparation

The following experiments were performed by Dr. Yakunina's group unless stated.

The RNA-DNA hybrid and all types of DNA templates used for biochemical assays were prepared by our collaborators in Dr. Yakunina's group as previously described (Orekhova et al., 2019). All the oligonucleotides used are listed in Table 2.9.

The DNA-template type	Oligonucleotide name	Sequence
DNA-RNA hybrid	Cy3-RNA	5'-Cy3-GUAGCGGA-3'
	Template DNA	5'-GGTCCTGTCTGAAATTGTTATCCGCTAC-3'
	Non-template DNA	5'-ACAATTTCAGACAGGACC-3'
Long DNA (from -64 to +148)	KZP119F	5'-GAGATGTACAGTGTATCATTTAGATAGC-3'
	KZ119prex	5'-TCGGAAACTATATAGTTGGTGTCTAAC-3'
Short DNA (from -12 to +15 on template strand)	KZP119_12t	5'-TTGGGAACTGTCTACATAGCAATATAG-3'
	FJ.KZP119(+3)nt	5'-AGACAGTTCCCAA-3'

5'-ACAATTTCAGACAGGACC-3'
5'-GUAGCGGA-3'
3'-CATCGCCTATTAAAGTCTGTCCTGG-5'

+1
5'GAGATGTACAGTGTATCATTTAGATAGCTAGTAATTTTAGTGAATGTATTTGCTATATTGCTATGTAGACAGTTCCCAAAAGCC
TAAAGTTACAATATAGGTACTTTCACAATGACCCAAGTTCTGGTTATGCCAAAAGCTGTTTTTGAATCATATAAAACAGCAGGTGT
ACAATTTAACAAATTGTTAGACACCAACTATATAGTTTCCGA 3'

+1
5'-AGACAGTTCCCAA-3'
3'-GATATAACGATACATCTGTCAAGGGTT-5'

Table 2.9: List of oligonucleotides used for biochemical assays with the ΦKZ nvRNAP purification and biochemical assays. Data provided by our collaborators in Dr. Yakunina's group.

The P119L-4nt DNA template and P119L RNA-DNA template used for cryo-EM were prepared by myself at Imperial College London. Oligonucleotides were ordered through Integrated DNA technologies (IDT). DNA and RNA/DNA templates were prepared by mixing an equal volume of each oligonucleotide at a final concnetration of 1 µM, incubated at 95 °C for 5 min and cooled down to 30 °C by increments of -1.1 °C/min. After that, annealed oligonucleotides were left on ice for 10 min and kept at -20 °C until use. To check that oligonucleotides were properly annealed they were analysed on native PAGE and stained with SYBR Safe (Invitrogen, Thermo Fisher Scientific, USA). All the oligonucleotides used for cryo-EM purposes are listed in Table 2.10.

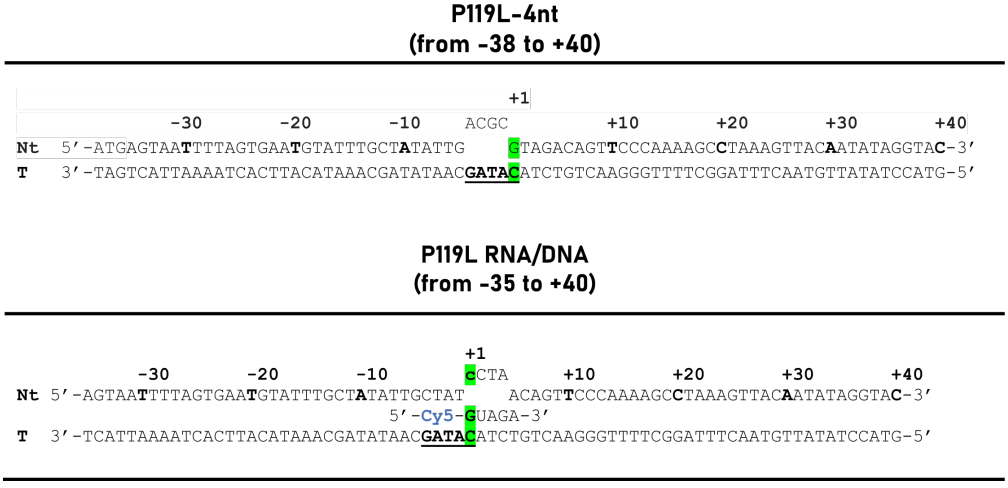


Table 2.10: List of oligonucleotides based on the ΦKZ P119L promoter used for cryo-EM. All the oligonucleotides were ordered from IDT and were annealed as indicated in section 2.3.2.

2.3.3 Mass spectrometry

The following experiments were performed by Dr. Yakunina's group unless stated.

Protein bands of interest were manually excised from the Coomassie-stained SDS or native TBE-PAGE. Individual slices from SDS or native PAGE were prepared for mass-spectrometry through *in situ* trypsin digestion at 37 °C for 4h as previously described (Yakunina et al., 2015). In the case of TBE-PAGE slice gels, no washing step was performed.

2.3.4 Electrophoretic mobility shift assay and photo-crosslinking

The following experiments were performed by Dr. Yakunina's group unless stated.

The *Electrophoretic mobility shift assay* (EMSA) was performed as previously described (Orekhova et al., 2019). For photocrosslinking reactions, 10 μL containing 5 pmol of the template DNA and 7.5 pmol of ΦKZ nvRNAP were incubated in transcription buffer (see Table 2.8) for 15 min at 37 °C. To induce the formation of covalent cross-linking, the mixture was irradiated for 15 seconds with *ultraviolet* (UV) radiation, using a RAPOP-100 femtosecond laser system and an ATsG800-17 third harmonic generator ("Avesta-Project"). The radiation parameters (the 3rd harmonic) were; radiation pulse energy of 10 μJ, pulse duration of 80 fs, repetition frequency of 2 kHz, and beam diameter of 2 mm. The resulting average power was 20 mW, and the

radiation power density in the pulse was 4 $GWcm^{-2}$. To break down unsuccessfully crosslinked complexes, heparin was added to the reaction mixture to a concentration of 100 g/L and the reaction mixture was incubated at 37 °C for 10 minutes. Samples were resolved by EMSA in DNA loading buffer (8.3% glycerol) or SDS-DNA loading buffer (8.3% glycerol, 0.2% SDS, 30 mM EDTA) before loading onto the 8% TBE-polyacrylamide gel.

2.3.5 Native PAGE and *in vitro* transcription

The following experiments were performed by Dr. Yakunina's group unless stated.

Native PAGE and *in vitro* transcription reactions were performed by our collaborators as previously described (Orekhova et al., 2019).

Native PAGE performed at Imperial College to confirm ΦKZ nvRNAP binding to the P119L RNA/DNA template were performed as indicated in section 2.1.7. To visualise DNA, gels were stained with SYBR Safe (Invitrogen, Thermo Fisher Scientific, USA) and imaged on a Gel Doc XR+ (Bio-Rad Laboratories, Inc., USA). To visualise the Cy5-RNA, gels were imaged on a ChemiDoc XRS+ (Bio-Rad Laboratories, Inc., USA) using the Cy5 channel.

2.3.6 Limited proteolysis

The following experiments were performed by Dr. Yakunina's group unless stated.

10 pmol of ΦKZ nvRNAP and 20 pmol of its complexes with DNA-templates were digested in 100 mM Tris-Cl pH 8.5 supplemented with 0.09 mg/mL trypsin (Sigma-Aldrich, USA) for 30 min at 22 °C. The proteolysis reactions were quenched through the addition of SDS-PAGE sample buffer, followed by immediate incubation of each sample at 100 °C for 5 minutes. The resulting samples were analyzed by SDS-PAGE (8 and 10 % polyacrylamide gels).

2.3.7 Single-Particle Electron Microscopy for the ΦKZ nvRNAP complex

2.3.7.1 Cryo-EM sample preparation

ΦKZ nvRNAP complexes were exchanged into sample buffer, to a final concentration of 0.1-0.3 mg/mL. Samples of ΦKZ nvRNAP were adsorbed to a thin film of graphene oxide deposited upon the surface of holey carbon copper grids (R2/1, 300 mesh, Quantifoil, Germany). Grids were blotted for 1-2 s, left incubating with sample for 4 s and plunge-frozen with a blot force of -6.

2.3.7.2 Calculation of reference density

Using an FEI F20 electron microscope (Thermo Fisher Scientific, USA), a total of 257 micrographs were collected from a nvRNAP cryo-grid on an early Falcon direct electron detector. Micrographs were collected at 81 000 fold magnification, an acceleration voltage of 200 kV, and a total dose of 50 $e^-/Å^2$, over a defocus range of -2 to -5 µm (Table 2.11). After aligning the frames and dose-weighting them using MotionCor2 (Zheng et al., 2017) and rejecting poor-quality micrographs based on irregularities in the Thon rings estimated by CTFFIND4 (Rohou and Grigorieff, 2015), a total of 45 980 particles were selected semi-automatically using BATCHBOXER (Tang et al., 2007). Particles were then classified into 2D class averages using RELION 3.0 (Zivanov et al., 2018) and well-defined classes including 18 508 particles were selected for initial three-dimensional reconstruction. Initial models were generated using the stochastic gradient descent approach in RELION 3.0 (Zivanov et al., 2018), yielding an initial reconstruction to 14 Å resolution by an independent half-set FSC of 0.143. The 14 Å resolution map was then filtered to 50 Å and used as reference for automatic refinement. Reconstructions from the initial models were consistent with class-averages, therefore they were used for further refinement of high-resolution data.

	ΦKZ nvRNAP Reference density
Magnification (×)	81 000
Voltage (kV)	200
Electron exposure ($e^-/Å^2$)	50
Defocus range (μm)	-2 to -5
Pixel size (Å/pix)	1.65
Symmetry imposed	C1
Initial number of particles	45 980
Final number of particles	18 508
Map resolution (Å) at FSC = 0.143	14

Table 2.11: Summary of data collection parameters for the ΦKZ nvRNAP data collection for calculation of reference density. Data was acquired at the EM facility of the Centre for Structural Biology located at Imperial College London using an FEI F20 electron microscope (Thermo Fisher Scientific, USA).

2.3.7.3 Data Collection and Image Processing

ΦKZ nvRNAP cryo-grids were imaged on a Titan Krios G3i (Thermo Fisher Scientific, USA) equipped with a K3 direct electron detector (GATAN, USA) at the London Consortium for Electron Microscopy microscope located at the Francis Crick Institute institute (London, UK). The microscope was operated at 300 kV, at a magnification of 48 000 fold and over an applied defocus range of -0.75 to -3.25 μm, and a total exposure of 51.2 $e^-/Å^2$. A total of 10 741 frames were recorded automatically using EPU with a pixel size of 0.85 Å (Table 2.12).

A total of 10 741 movies were aligned, summed and weighted by dose according to the method of Grant and Grigorieff using MotionCor2 (Zheng et al., 2017) to obtain a final image. Poor-quality micrographs were rejected based on decreased estimated maximum resolution on CTF estimation using CTFFIND4 and visually based on irregularity of the observed Thon rings. Particles were selected using BATCHBOXER (Tang et al., 2007), and refinement thereafter performed using RELION 3.0 (Zivanov et al., 2018). 2D reference-free alignment was performed on 3 179 799 initial particles to exclude those that did not yield high-resolution class averages. Of these, 1 043 026 particles were retained for further refinement, and after a further screening based on CTF parameters for higher resolution, 855 016 particles that generated high-resolution class averages of micrographs collected at defocus values over 1 μm contributed to an initial

refinement to high resolution using the auto-refinement procedure in RELION 3.0 (Zivanov et al., 2018), reaching just under 3 Å according to an independent half-set FSC of 0.143. These were then classified without alignment into nine classes in three-dimensions and the class with the largest ordered fraction of the molecule was retained for final high-resolution refinement. This final gold-standard refinement of ΦKZ nvRNAP from 50 085 particles reached 3.3 Å according to an independent half-set FSC of 0.143.

	ΦKZ nvRNAP EMD-12886	ΦKZ nvRNAP Clamp EMD-12885
Magnification (×)	48,000	48,000
Voltage (kV)	300	300
Electron exposure (e^-/Å^2)	51	51
Defocus range (μm)	-0.75 to -3.25	-0.75 to -3.25
Pixel size (Å/pix)	0.85	0.85
Symmetry imposed	C1	C1
Initial number of particles	3 179 799	3 179 799
Final number of particles	855 016	50 085
Map resolution (Å) at FSC = 0.143	3	3.3

Table 2.12: Summary of high-resolution data collection parameters for the ΦKZ nvRNAP. Data was acquired on a Titan Krios G3i (Thermo Fisher Scientific, USA) equipped with a K3 direct electron detector (GATAN, San Diego, USA) at the London Consortium for Electron Microscopy microscope located at the Francis Crick Institute institute (London, UK).

2.3.7.4 Modelling and refinement

Sequence analysis using BLASTp implied that the ΦKZ subunits, overall, were more similar to the RNAP structure of *M.tuberculosis* amongst other structures available in the PDB (PDB 6EDT) (Boyaci et al., 2019). Therefore, initial models for the most homologous regions of the ΦKZ nvRNAP were prepared using SWISS-MODEL (Waterhouse et al., 2018) based on *Mycobacterium tuberculosis* RNAP structure (PDB 6EDT) (Boyaci et al., 2019). Despite *Mycobacterium tuberculosis* RNAP being the closest structural homologue available overall, homology was variable amongst the four β/β'-like subunits.

Structural models for the full-length of gp71-73, the full-length of gp74 and the C-terminus of gp55 (290-419) were well-defined and could be fitted into the density guided by secondary structural elements. After fitting models into the density, molecular models were then rebuilt

with COOT (Emsley et al., 2010; Emsley and Cowtan, 2004) and extended outwards from well-defined reference points to the remaining regions of the four homologous proteins. A structural model was built for most of gp71-73, gp74, excluding its central region, and the C-terminus of gp55 (Table 2.13) using the high-resolution 3 Å map. The 3.3 Å map was used to build the N-terminus of gp55, corresponding to the clamp, and assist with the building of other regions (Table 2.13). Gp123 structural models were variable and contained large sections without defined secondary structure, which was expected given its weak homology to the N-terminus of the bacterial β subunit (Yakunina et al., 2015). After building gp71-73, gp74 and gp55, guided by secondary structure predictions (Drozdetskiy et al., 2015), a structural model for the N-terminus and C-terminus of gp123 was built in COOT (Emsley et al., 2010; Emsley and Cowtan, 2004) (Table 2.13).

		Nº of residues			
Subunit	Chain ID	Model	Sequence	% Modelled Sequence	Modelled sections
gp55	A	358	489	73	9-44, 150-213, 216-233, 247-486
gp74	D	510	677	75	1-384, 552-677
gp71-73	C	693	700	99	2-592, 597-698
gp123	E	370	543	68	1-165, 322-471, 483-543
gp68	B	203	521	39	304-413, 429-521
Total	-	2134	2930	73	

Table 2.13: Summary of the coverage of the ΦKZ nvRNAP structural model. The sections of sequence that were built for each ΦKZ nvRNAP subunit are indicated in numbers.

The fifth subunit, ΦKZ-gp68 had no homology to any other protein with known function (Yakunina et al., 2015), therefore it was not possible to generate any reliable structural model. The C-terminus of gp68 was built into the remaining density using COOT (Emsley et al., 2010; Emsley and Cowtan, 2004), initially as a poly-alanine model, before the sequence register was eventually established based upon the identification of clear patterns of large side-chains and secondary structure elements predicted by the JPred server (Drozdetskiy et al., 2015) (Table 2.13).

The atomic model was refined with PHENIX real-space refine (Afonine et al., 2012). A summary of the refinement statistics can be found in Table 2.14 including the B-factor calculated with B-average (CCP4). The cryo-EM density maps for the ΦKZ nvRNAP complex are publicly available in the EM Databank under accession codes EMD-12885 and EMD-12886, and the corresponding molecular models have been deposited in the protein data bank under accession codes PDB 7OGP and 7OGR. Electron micrographs allowing reproduction of the structural results have been deposited as EMPIAR-10709. Homology searches and Cα comparisons were carried out using the DALI-lite server (Holm and Sander, 1995), while surface area calculations were performed using the PISA server.

	PDB ID 7OGR	PDB ID 7OGP
Mean model-map correlation / B-factor	0.8032 / 40.261	0.8360 / 109.761
Molprobity overall score / clash score	1.5 / 3.80	1.44 / 3.33
RMS deviations in bond lengths / bond angles	0.003 Å / 0.720°	0.003 Å / 0.689°
Ramachandran favoured / allowed / disallowed (%)	95.31 / 4.69 / 0.00	95.37 / 4.63 / 0.00
Rotamers favoured / allowed / disallowed (%)	95.55 / 4.38 / 0.07	95.39 / 4.34 / 0.27

Table 2.14: Model refinement statistics for the ΦKZ nvRNAP structure.

2.3.8 Single particle Cryo-EM for ΦKZ nvRNAP complexes with the P119L promoter

2.3.8.1 Cryo-EM sample preparation

To prepare ΦKZ nvRNAP complexes bound to DNA, nvRNAP complexes were mixed at 0.1 mg/mL final concentration with the P119L-4nt DNA template (Table 2.10) at a 1:3 ratio and left incubating for 10 min at room temperature. After that, nvRNAP-DNA complexes were separated from free nvRNAP and excess nucleic acids by size-exclusion over a Superose 6 10/300 GL (GE Healthcare Life Sciences, now Cytiva, USA) run at 0.25 ml/min. Fractions of 0.25 mL were collected and analysed by SDS-PAGE.

Cryo-EM samples for the ΦKZ nvRNAP complexes bound to the P119L RNA/DNA template were prepared immediately after the transcription reaction. We mixed ΦKZ nvRNAP at 0.1 mg/mL final concentration and the RNA/DNA scaffold P119L (Table 2.10) at a 1:1 ratio in transcription buffer (Table 2.8) in the presence of 1 mM ATP, CTP and GTP. Reaction were incubated for 30 min at 37 °C. After incubation, ΦKZ nvRNAP complexes bound the P119L RNA/DNA template were adsorbed to a thin film of graphene oxide deposited upon the surface of holey carbon copper grids (Quantifoil^{TM} R 2/1 on 300 copper mesh)(Quantifoil, Germany). After application of sample to the grids, samples were plunge-frozen in liquid ethane using a Vitrobot Mark IV (Thermo Fisher Scientific, USA) at 4 °C and 100 % humidity, -6 blot force, 4 s waiting time and 0.5-1 s blotting. The blotting pads were prepared with Whatman number 41 filter paper 55 mm discs (GE Healthcare Life Sciences, now Cytiva, USA).

2.3.8.2 Screening of vitrification conditions

ΦKZ nvRNAP-P119L RNA/DNA template cryo-grids were imaged on a Glacios cryo-TEM (Thermo Fisher Scientific, USA) equipped with a Falcon 4 direct electron detector located at the EM facility at the Centre for Structural Biology at Imperial College London. The microscope was operated at 200 kV, at a magnification of 150 000 fold and over an applied defocus range of -1.2 to -2.5 µm, and a total exposure of 52 $e^-/Å^2$ (Table 2.15). A total of 1 600 movies were recorded automatically using EPU with a pixel size of 0.94 Å/pixel (Table 2.15).

Movies were aligned, summed and weighted by dose according to the method of Grant and Grigorieff using MotionCor2 (Zheng et al., 2017) to obtain a final image. Poor-quality micrographs were rejected based on low estimated resolution on CTF estimation, and irregularities of the observed Thon rings. Particles were selected using BATCHBOXER (Tang et al., 2007), and refinement performed using RELION 3.0 (Zivanov et al., 2018). 2D reference-free alignment was performed on 535 289 initial particles to exclude those that did not yield high-resolution class averages. Of these, 156 304 particles yielding high-resolution 2D class averages were retained for further refinement. Those particles contributed to an initial refinement to high resolution using the auto-refinement procedure in RELION 3.0 (Zivanov et al., 2018) and reaching 4.7 Å

according to an independent half-set FSC of 0.143.

	ΦKZ nvRNAP-DNA Glacios
Magnification (×)	150 000
Voltage (kV)	200
Electron exposure ($e^-/Å^2$)	52
Defocus range (μm)	-1.2 to -2.5
Pixel size (Å/pix)	0.94
Symmetry imposed	C1
Initial number of particles	535 289
Final number of particles	156 304
Map resolution (Å) at FSC = 0.143	4.7

Table 2.15: Summary of collection parameters for the ΦKZ nvRNAP complexes bound to P119L RNA/DNA (initial model). Data was acquired on a Glacios cryo-TEM (Thermo Fisher Scientific, USA) equipped with a Falcon 4 direct electron detector located at the EM facility at the Centre for Structural Biology at Imperial College London.

2.3.8.3 High-resolution Data Collection and Image Processing

ΦKZ nvRNAP complexes bound to the P119L RNA/DNA template were imaged on a Titan Krios G3i equipped with a Falcon 4i direct electron detector and a Selectris Filter (Thermo Fisher Scientific, USA) at the Electron Bio-Imaging Centre (eBIC) at Diamond Light Source. The microscope was operated at 300 kV, at a magnification of 150 000 fold and over an applied defocus range of -1.4 to -2.9 µm, and a total exposure of 51 $e^-/Å^2$. A total of 3 092 frames were recorded automatically using EPU with a pixel size of 0.92 Å/pixel (Table 2.16).

Frames were were aligned and dose-weighted with MotionCor2 (Zheng et al., 2017) and CTF parameters were estimated with CTFFIND4 (Rohou and Grigorieff, 2015). Micrographs displaying irregularities in their Thon rings and those that did not contain high-resolution information according to the power spectrum fitting were discarded, retaining 2 545 micrographs for further refinement. 1 198 587 particles were selected semi-automatically with BATCHBOXER (Tang et al., 2007) and were classified in 2D class averages with RELION4 (Kimanius et al., 2021) leaving CTF information up to the first peak intact. Particles yielding high-resolution

class averages (265 839) were retained for further refinement. This set of particles was further classified in six 3D classes without alignment using a high tau fudge value (tau2_fudge = 16). One of the classes containing 83 389 particles had better defined density for the DNA and the peripheral regions of the complex. CTF and aberration refinement (including beam-tilt, anisotropic magnification and per-particle CTF estimation) and Bayesian polishing using RELION 4 (Kimanius et al., 2021) were performed on this set of particles, which finally contributed to an 3.3 Å initial reconstruction, according to an independent half-set FSC of 0.143, using the auto-refinement procedure in RELION4. Particles contributing to this high-resolution reconstruction were further refined to improve density in the catalytic core, DNA and gp68 NTD regions. 3D classification with fine angular search (sigma_ang = 2) yielded a class with well-defined DNA at the downstream region and the catalytic core, which included 39 919 particles. Those particles contributed to a high-resolution reconstruction using the auto-refinement procedure in RELION4 which reached 3.3 Å resolution according to an independent half-set FSC of 0.143.

	ΦKZ nvRNAP-DNA **Krios**
Magnification (×)	150 000
Voltage (kV)	300
Electron exposure ($e^-/Å^2$)	51
Defocus range (μm)	-1.4 to -2.9
Pixel size (Å/pix)	0.92
Symmetry imposed	C1
Initial number of particles	1 198 587
Final number of particles	39 919
Map resolution (Å) at FSC = 0.143	3.3

Table 2.16: Summary of high-resolution data collection parameters for the ΦKZ nvRNAP bound to the RNA/DNA scaffold P119L template. Data was acquired on a Titan Krios G3i equipped with a Falcon 4i direct electron detector and a Selectris Filter (Thermo Fisher Scientific, USA) at the Electron Bio-Imaging Centre (eBIC) at Diamond Light Source.

2.3.8.4 Modelling and refinement

Structural models for the uncompleted regions of the ΦKZ nvRNAP structure (7OGP) (de Martín Garrido et al., 2021b) were generated using AlphaFold2 (Jumper et al., 2021). These models cor-

responded to middle domain of gp74 (367-545), middle domain of gp123 (143-337), gp55-clamp (1-291) and N-terminus of gp68 (1-318). Molecular models were then rebuilt with COOT (Emsley and Cowtan, 2004; Emsley et al., 2010) and expanded from the well-defined regions, as most of the secondary structure elements were already in agreement with the map (Table 2.17). The density for the gp68 N-terminal region was not resolved at high-resolution in the 3.3 Å reconstruction, but the Alphafold2 structural model could be docked into the density.

At the catalytic core, the 6 bp RNA hybridised along the template ssDNA strand was built in COOT (Emsley and Cowtan, 2004; Emsley et al., 2010) according to the RNA scaffold sequence (5'-GUAGA-3'). The downstream dsDNA was well-defined in our reconstruction and, based on the RNA sequence, sequence for the template dsDNA strand was assigned at the catalytic core, and expanded to the downstream dsDNA up to position +23 relative to the transcription start site. The ssDNA strand forming the transcription bubble was visible up to position -5 relative to the transcription start site, whereas the non-template ssDNA strand was not visible. The connection of the template strand forming the transcription bubble with the upstream dsDNA was not well-resolved in our reconstruction. The upstream dsDNA region was lower resolution than the downstream dsDNA, however at low contours the phosphate backbone was visible and a 17 bp dsDNA structural model with unassigned sequence was built in COOT (Emsley and Cowtan, 2004; Emsley et al., 2010).

		Nº of residues				
Subunit	Chain ID	Model	Sequence	% Modelled Sequence	% Modelled sequence (β/β')	Modelled sections
gp55	A	486	489	99	99	1-486
gp74	D	655	677	97	97	1-469, 474-546, 565-677
gp71-73	C	697	700	100	100	2-698
gp123	E	510	543	94	94	2-192, 196-232, 247-310, 327-543
gp68	B	229	521	44		274-287, 297-414, 425-521
Total	-	2577	2930	88	97	

Table 2.17: Summary of the coverage of the ΦKZ nvRNAP complex bound to P119L RNA/DNA structure Sections of sequence that were built for each ΦKZ nvRNAP subunit are indicated with numbers.

The atomic model was refined with PHENIX real-space refine (Afonine et al., 2012). A summary of the refinement statistics can be found in Table 2.18 including the B-factor calculated with B-average (CCP4). Homology searches and Cα comparisons were carried out using the DALI-lite server (Holm and Sander, 1995), while surface area calculations were performed using the PISA server.

Mean model-map correlation / B-factor	**0.8218 / 43.046**
Molprobity overall score / clash score	1.57 / 4.51
RMS deviations in bond lengths / bond angles	0.003 Å / 0.694°
Ramachandran favoured / allowed / disallowed (%)	95.06 / 4.90 / 0.04
Rotamers favoured / allowed / disallowed (%)	95.05 / 4.64 / 0.31

Table 2.18: Model refinement statistics for the structure of the ΦKZ nvRNAP bound to the P119L RNA/DNA promoter.

Chapter 3

The Folliculin complex and its interacting partners

FLCN is a tumor suppressor associated with Birt-Hogg-Dubé (BHD) syndrome, a hereditary syndrome which, ultimately, causes kidney cancer. Since its discovery, it has been established that FLCN forms stable complexes with two folliculin interacting proteins (FNIPS). Both FLCN and FNIPs are thought to comprise longin and DENN domains, however little information is available about the interaction within them at a structural level. Longin and DENN domains are typically involved in the regulation of small GTPases. Accordingly, the FLCN complex is a positive modulator of mTORC1 through its GAP activity towards the RagC/D GTPase. Intriguingly, other studies have shown that FLCN requires RagA/B-GDP to be recruited to the lysosome. The double nucleotide-bound setting, with the Rag GTPase heterodimer being able to simultaneously bind to both GTP and GDP, is a unique feature of the Rags in comparison to other GTPases, which suggests that the mechanism by which FLCN perform its GAP activity towards the RagC/D GTPase is likely to be novel. In this chapter we aimed to provide structural insights about folliculin (FLCN) and its interacting partners. We used cryo-EM to investigate the structure of the FLCN-FNIP2 complex on its own, as well as its interaction with the Rag GTPase heterodimer which would hopefully provide further details about their roles in the mTORC1 signalling pathway.

3.1 Structural studies of the *Sl* FLCN-FNIP2 complex

3.1.1 Expression and purification of the *Sl* FLCN-FNIP2 complex

The FLCN-FNIP2 complex was expressed using the MultiBac baculoviral expression method (Berger et al., 2004) at the well-established Baculovirus expression facility at Imperial College London. This expression system had previously been used to express dynamic and labile complexes, such as mTORC1 or mTORC2 (Aylett et al., 2016a; Stuttfeld et al., 2018). Previous attempts in our lab to express the human *(Homo sapiens)* FLCN-FNIP2 and the yeast orthologues Lst7-Lst4 *(Saccharomyces cerevisiae)* using the baculoviral expression system were unsuccessful. Hence, FLCN-FNIP2 complexes from different insect species were tested in the baculoviral expression system, with *Spodoptera litura* complexes giving the highest expression yield (data obtained prior to the start of this project by Kailash Ramlaul). Both *S.litura* FLCN and FNIP2 are similar to their respective homologues in humans, with more than 30% of the sequence being identical and conserving most of the secondary structure elements (sequence similarity 32.87 % for FLCN, and 35.7% for FNIP2)(Figure 3.1 and 3.2).

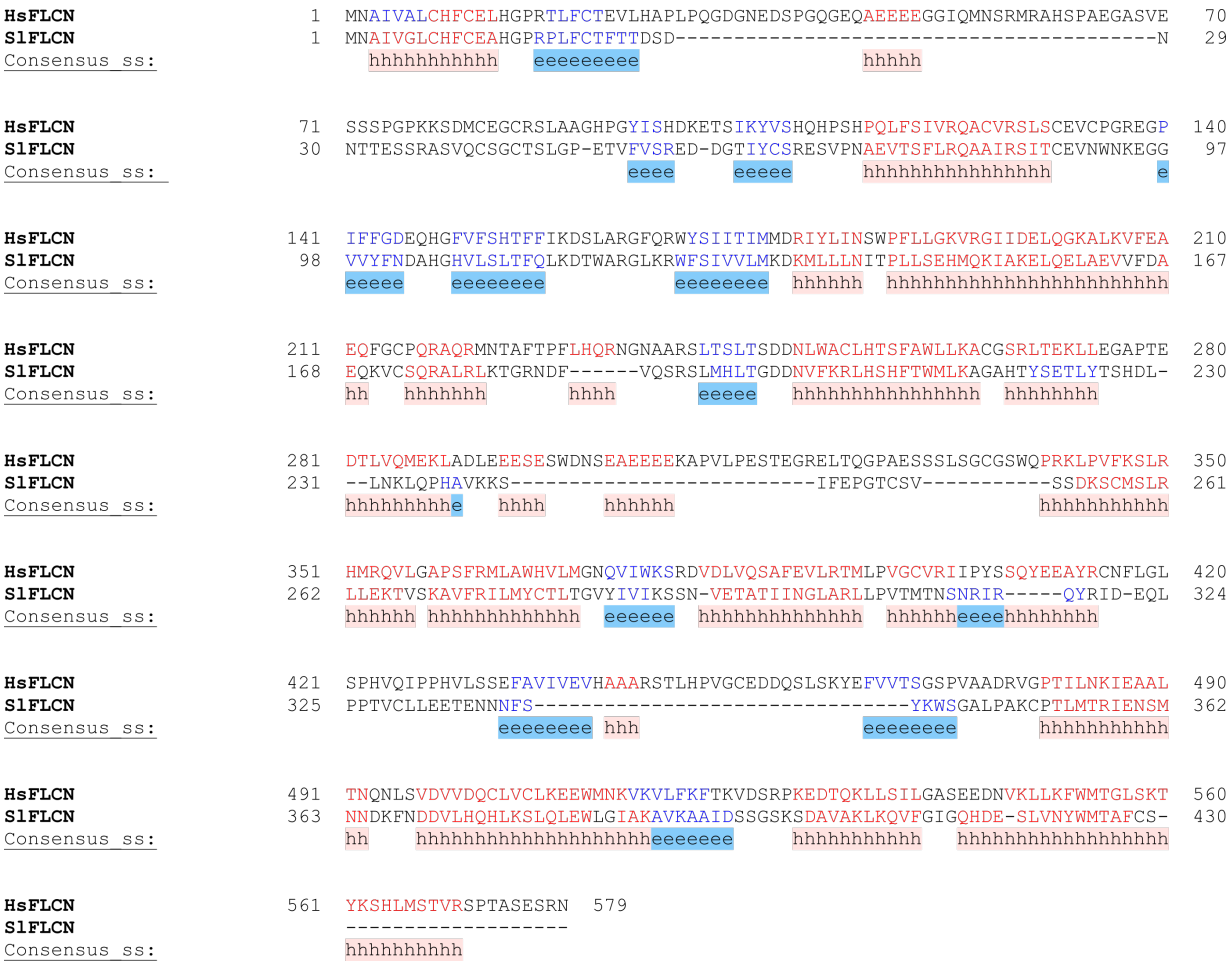


Figure 3.1: Sequence alignment of *S.litura* and human FLCN The results were computed with PROMALS3D. Residues in blue represent β-sheets whereas residues in red represent α helices. The last line represents the consensus secondary structure following the same colour code.

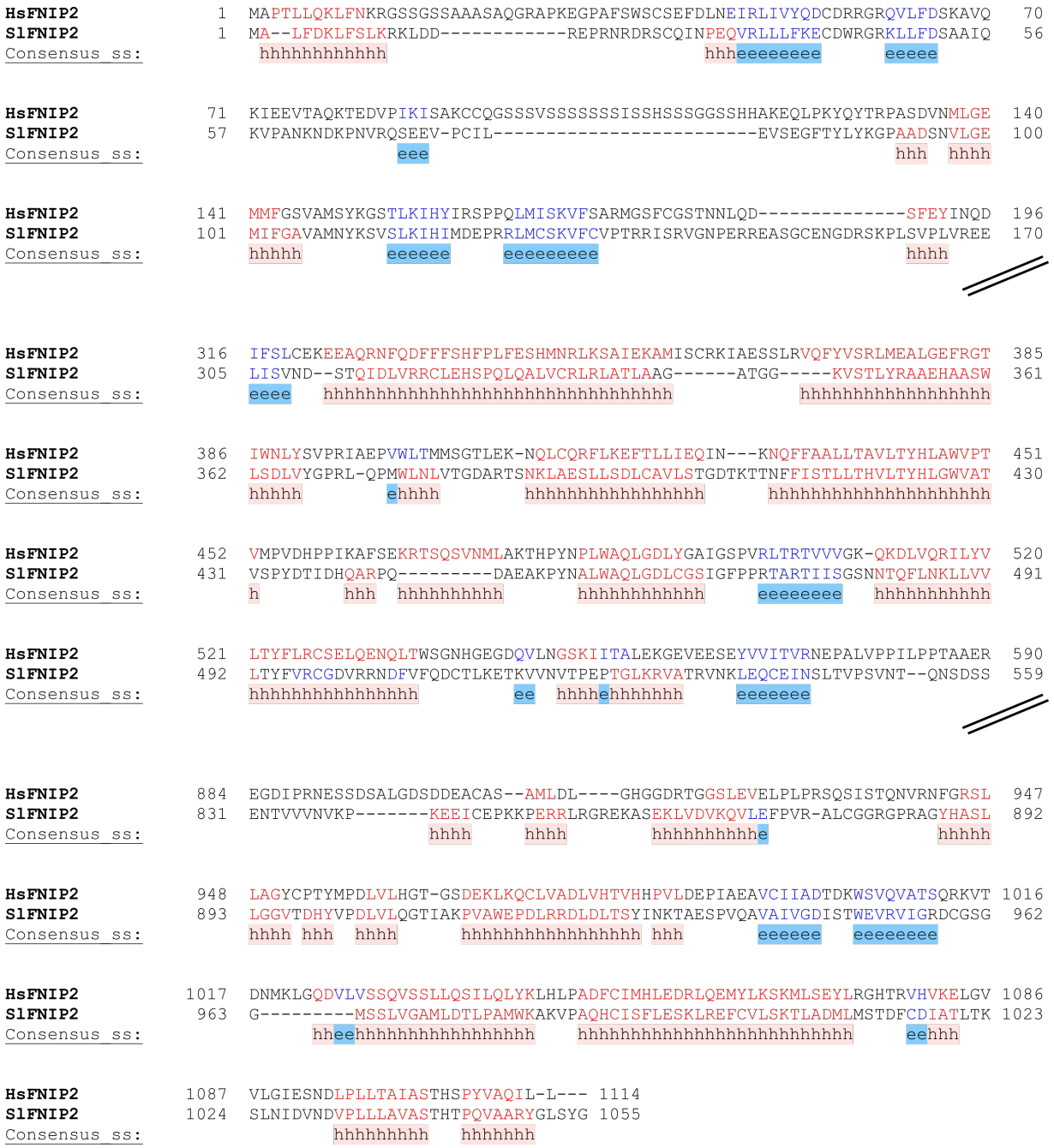


Figure 3.2: Sequence alignment of *S.litura* and human FNIP2 The results were computed with PROMALS3D. Residues in blue represent β-sheets whereas residues in red represent α helices. The last line represents the consensus secondary structure following the same colour code. Black lines represent regions where the sequence is omitted as there is a large unstructured region.

We therefore cloned both *Sl* FLCN and FNIP2 genes into a pG vector (see Chapter 2) (Rees et al., 2017) including Tn7 transposition sites to allow specific integration of the expression cassette into the bacmid. We then transected Sf9 cells with the generated bacmids and collected the assembled viruses to infect a large amount of cells for protein production. The largest protein of the complex, FNIP2, included an SBP tag (Keefe et al., 2001) at the C-terminus, while the FLCN protein was left without a tag. After harvesting the cells, the complex was purified by means of affinity chromatography using a Streptrap resin, and subsequent *Size-exclusion chromatography* (SEC) to separate the complex from aggregates and other contaminants (see Chapter 2). Analysis by SDS-PAGE revealed the presence of two proteins with similar molecular weights to FLCN and FNIP2 (48 kDa and 117 KDa respectively) that co-eluted together after affinity purification (Figure 3.3). The presence of both proteins was later validated by *mass-spectrometry* (MS) (Figure 3.3), thus confirming that the FLCN and FNIP2 proteins form a stable complex in insect cells.

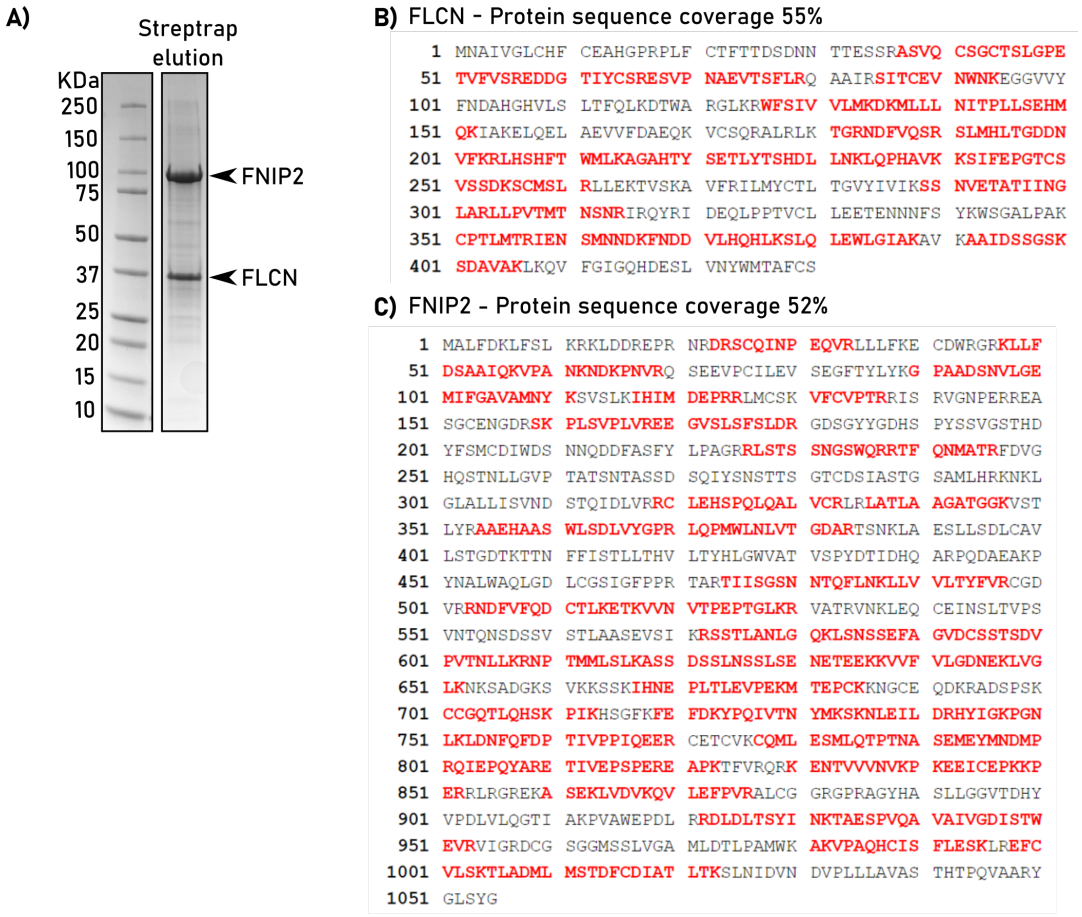


Figure 3.3: Affinity purification and mass spectometry of the FLCN-FNIP2 complex. A) SDS-PAGE showing the eluted fraction obtained after affinity purification using a Streptrap resin. B) *Sl* FLCN sequence with peptides identified by MS highlighted in red. Peptides were analysed against the Mascot internal database. C) *Sl* FNIP2 sequence with peptides identified by MS highlighted in red. Peptides were analysed against the Mascot internal database.

Initially, we were concentrating eluted fractions after the affinity purification, however we observed that samples tended to aggregate. We therefore minimised the elution volume during affinity chromatography to 2 mL and immediately loaded it onto the gel-filtration column. The analysed sample yielded a broad peak at around 12 mL of elution volume (Figure 3.4). SDS-PAGE and MS later confirmed that this peak corresponded to the FLCN-FNIP2 complex. Although in SEC macromolecules are eluted according to their *hydrodynamic radius* (R_h), which accounts for their shape and size, we compared the elution volume of the *Sl* FLCN-FNIP2 complex with that of protein standards with known molecular weight to have an estimation of its oligomeric state. The *Sl*FLCN-FNIP2 complex has a theoretical molecular weight of 165 KDa and its gel-filtration profile showed that it had a similar retention volume to protein standards of 158 kDa (Figure 3.4), which indicated that the complex was probably in a monomeric form

in solution. To confirm the oligomeric state of the complex in solution, we attempted to perform SEC-MALS, however we could only observe an aggregated peak at void volume, which indicated that complexes were probably aggregating at room temperature.

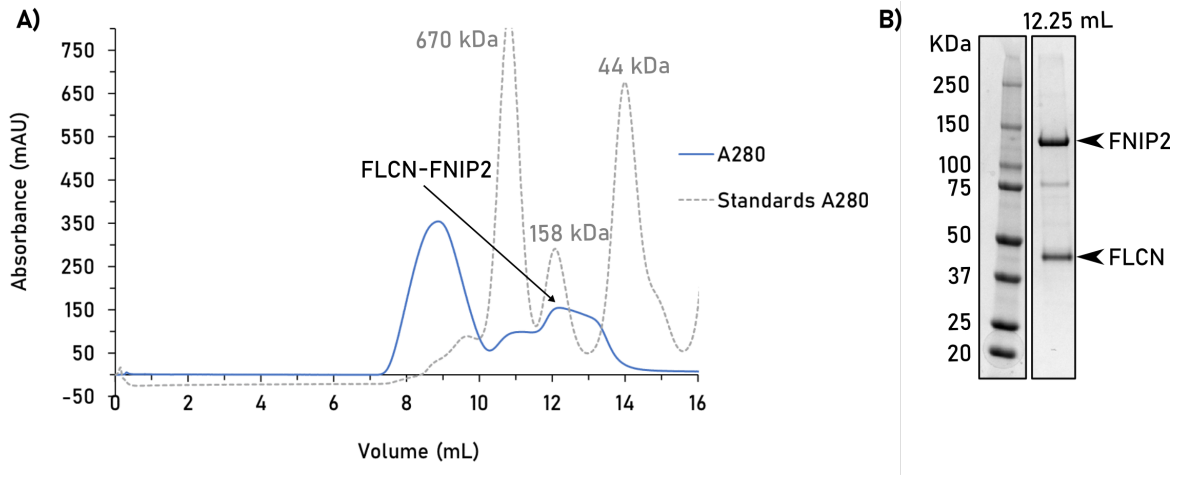


Figure 3.4: Size-exclusion chromatography of the FLCN-FNIP2 complex. A) SEC profile of *Sl* FLCN-FNIP2 complex (blue) and protein standards (dotted grey line). B) Coomassie-stained SDS-PAGE of eluted SEC fraction containing the assembled FLCN-FNIP2. Molecular weight standards are displayed on the left for comparison.

3.1.2 Screening of vitrification conditions

Before preparing frozen samples, we initially investigated the overall architecture of the complex by negative staining. We imaged negatively-stained specimens on an FEI Tecnai 12 Twin equipped with an LaB_6 filament and a CCD camera (F216, TVIPS). Observed micrographs were well-populated (≈120 particles/micrograph) and exhibited clear single particles (Figure 3.5). Although the size of the particles appeared to be quite small by visual inspection, we collected enough images to generate 2D class averages using RELION 3 (Zivanov et al., 2018). The classes were smaller than expected for a complex of this size (165 kDa) measuring 90 Å in length. Presumably, due to the large disordered regions comprised within both proteins only the core of the complex was converging into 2D class averages. Amongst the classes, we could observe a consistent elongated shaped with two differentiated domains (Figure 3.5).

Due to the small size of the particles, we considered that we would need a high-resolution data collection using a VPP to generate a three-dimensional reconstruction of the FLCN-FNIP2 complex. The VPP is a technology that has previously been successful in resolving

the structures of sub-200 kDa protein complexes at atomic resolution (Khoshouei et al., 2017; Liang et al., 2017). We optimised the vitrification conditions in-house using an FEI Tecnai F20 (Thermo Fisher Scientific, USA) microscope equipped with a Falcon II at the Imperial College Centre for Structural Biology. We prepared cryo-EM grids using the eluted sample from the gel-filtration at a final concentration of 0.1 mg/mL, avoiding any further concentration steps that caused sample aggregation. To obtain thin layers of vitreous ice we treated the grids mildly with the plasma cleaner on low mode on a Harrick Basic Plasma cleaner (see Chapter 2) to render them sufficiently hydrophilic for the sample to spread properly across the grid, but to avoid high blotting times which could result in dry sections throughout the grid. We obtained thin layers of vitreous ice using plasma-cleaned grids for 40 s on low mode, blot force of -25 and blot time of 0s (Figure 3.6) (see Chapter 2). Due to the small appearance of the particles we could not generate high-resolution 2D classes or observe many features of the particles with the F20, but we could observe single particles homogeneously distributed throughout the grid.

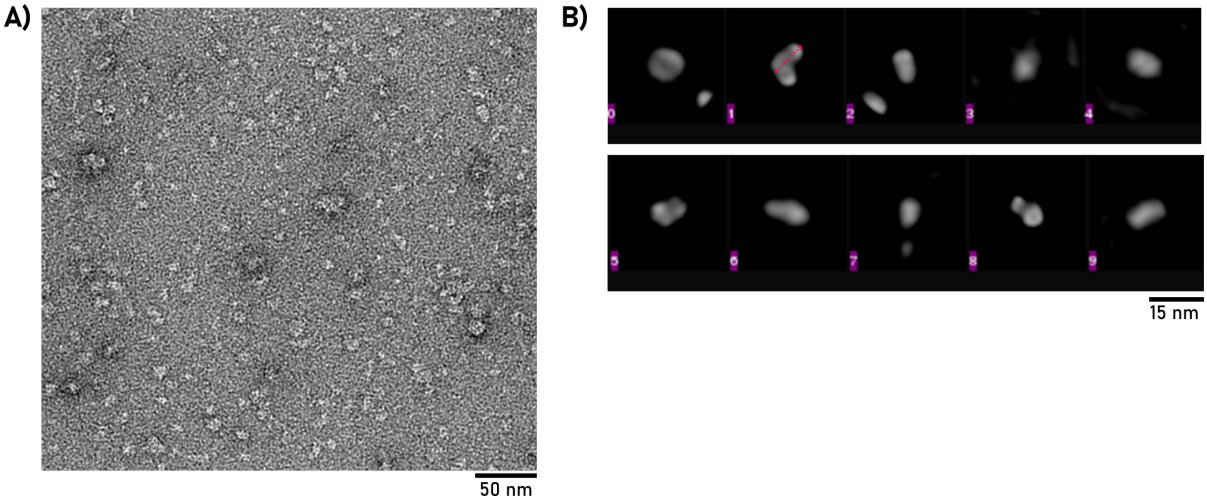


Figure 3.5: Negatively-stained grids of the FLCN-FNIP2 complex. A) Representative micrograph of a negatively-stained grid of the FLCN-FNIP2. Image taken with a Tecnai T12 Twin microscope equipped with a CCD camera. B) Representative 2D class averages obtained with RELION. The pink line indicates a measurement of 90 Å.

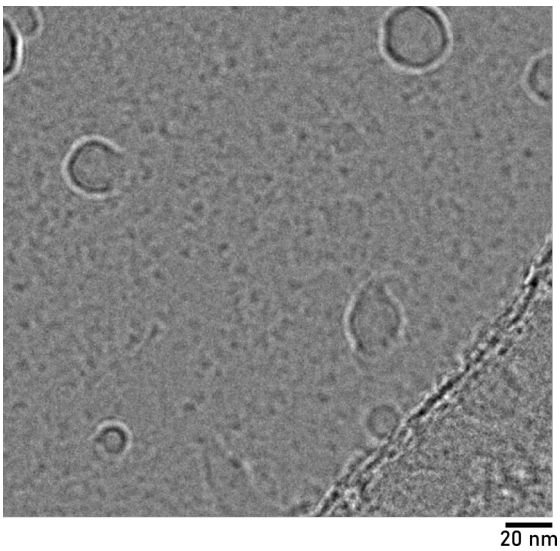


Figure 3.6: Cryo-EM screening of FLCN-FNIP2 samples. Representative micrograph of a cryo-EM grid of the FLCN-FNIP2 complex. The small size of the particles prevented the generation of any meaningful 2D classes.

3.1.3 High-resolution data collection using a Volta phase plate and image processing

Having established the vitrification conditions to visualize FLCN-FNIP2 complex by TEM, we collected a high-resolution dataset on a Titan Krios equipped with a Gatan K3 camera at LonCEM (The Francis Crick Institute) using a Volta phase plate. Introducing a phase plate at the back focal plane results in a phase shift of the unscattered beam from the scattered beam, which are combined again where the image is formed, generating a phase delay. Optimally in TEM, such phase shift converts the CTF from a sine-shape function to a cosine-shape function, hence preserving the signal at low spatial frequencies and enhancing contrast (Wang and Fan, 2019). The use of a phase plate allowed us to collect movies very close to focus (-1 µm). Overall, the particle density was lower than would have been ideal as we were observing between 10-20 particles/micrograph. Despite the low particle density, single particles were clearly visible on the micrographs and had an apparent consistent size and "banana-like" shape (Figure 3.7).

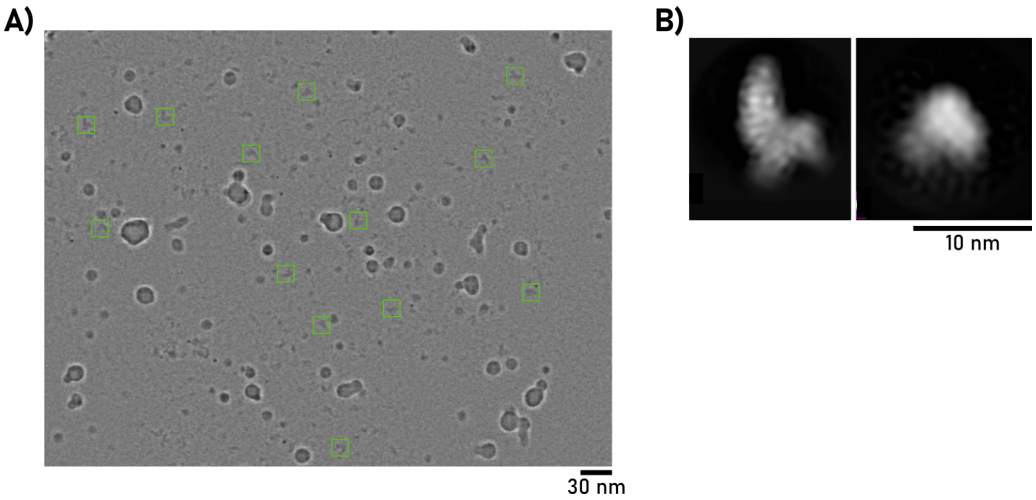


Figure 3.7: High-resolution data collection using a Volta phase plate of the FLCN-FNIP2 complex. A) Representative micrograph of the data collection. Particles are highlighted with green boxes. B) Representative 2D class averages obtained with RELION.

By the time we were processing the high-resolution data for the *Sl* FLCN-FNIP2 complex, two structures of the human FLCN-FNIP2 complex had been released (Shen et al., 2019; Lawrence et al., 2019). We therefore used the human structure to generate references to run our 3D classification. We initially aligned and dose-weighted the frames with MotionCor2 (Zheng et al., 2017) and then we estimated the CTF parameters and the phases with CTFFIND4 (Rohou and Grigorieff, 2015). Then, we semi-automatically selected nearly 2 million particles that were further classified into 2D class averages with RELION 3 (Zivanov et al., 2018). Particles converged into two 2D classes corresponding to two different views of the complex (Figure 3.7, an indication that the sample was likely to be preferentially oriented on the grid probably due to its elongated shape. We further sorted out the set of particles that converged into high-resolution 2D class averages by means of 3D classification without alignment and subsequent 3D classification with a fine local angular search (sigma_ang 2). After classification, 152 767 particles yielded a 3D reconstruction at 5.7 Å resolution according to an independent half-set FSC of 0.143 (Figure 3.8, 3.9).

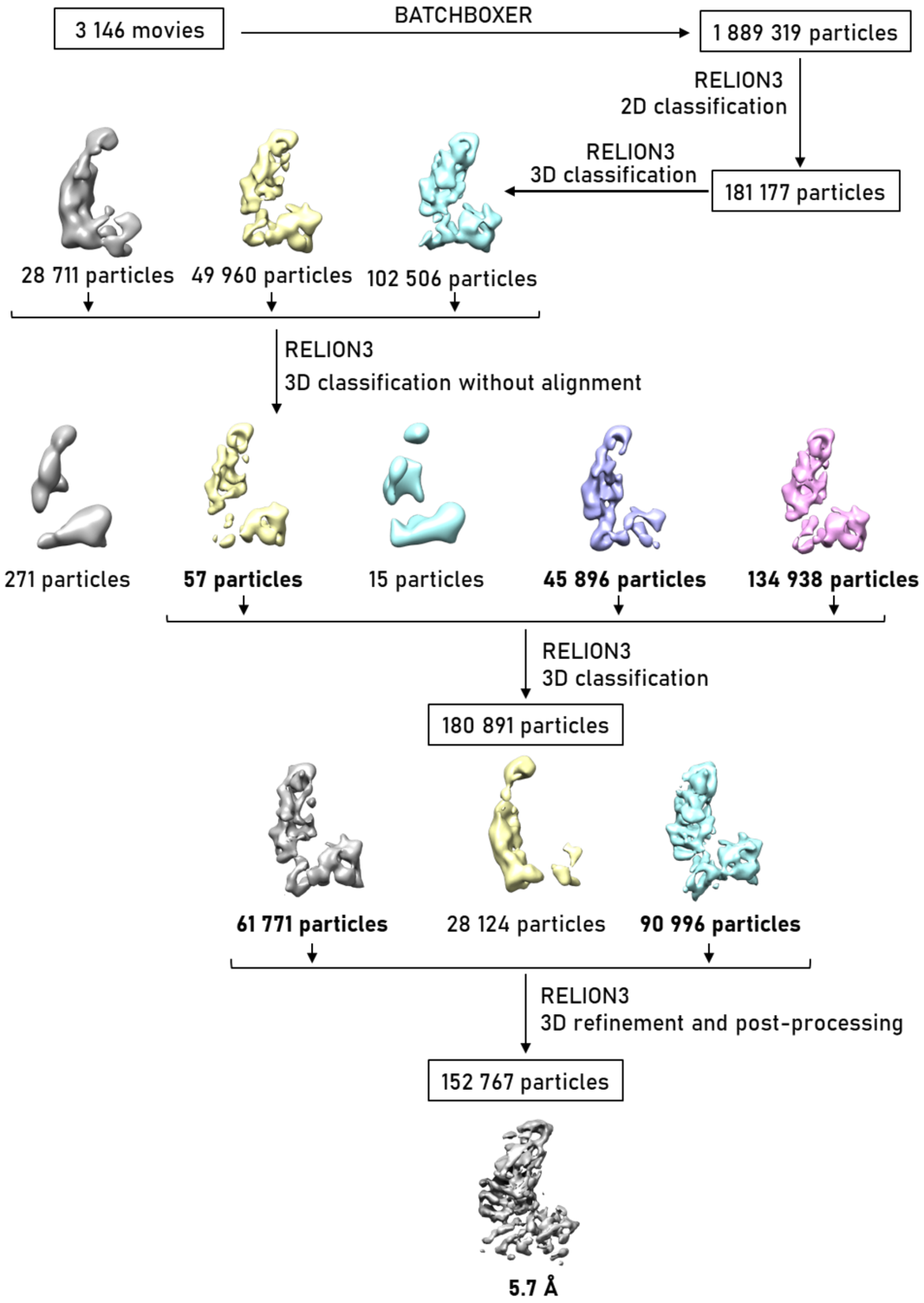


Figure 3.8: Data processing scheme for the cryo-EM high-resolution data collection of the FLCN-FNIP2 complex. Schematic diagram depicting the data processing procedure. Particle numbers retained at each stage of the process are indicated below each 3D reconstruction.

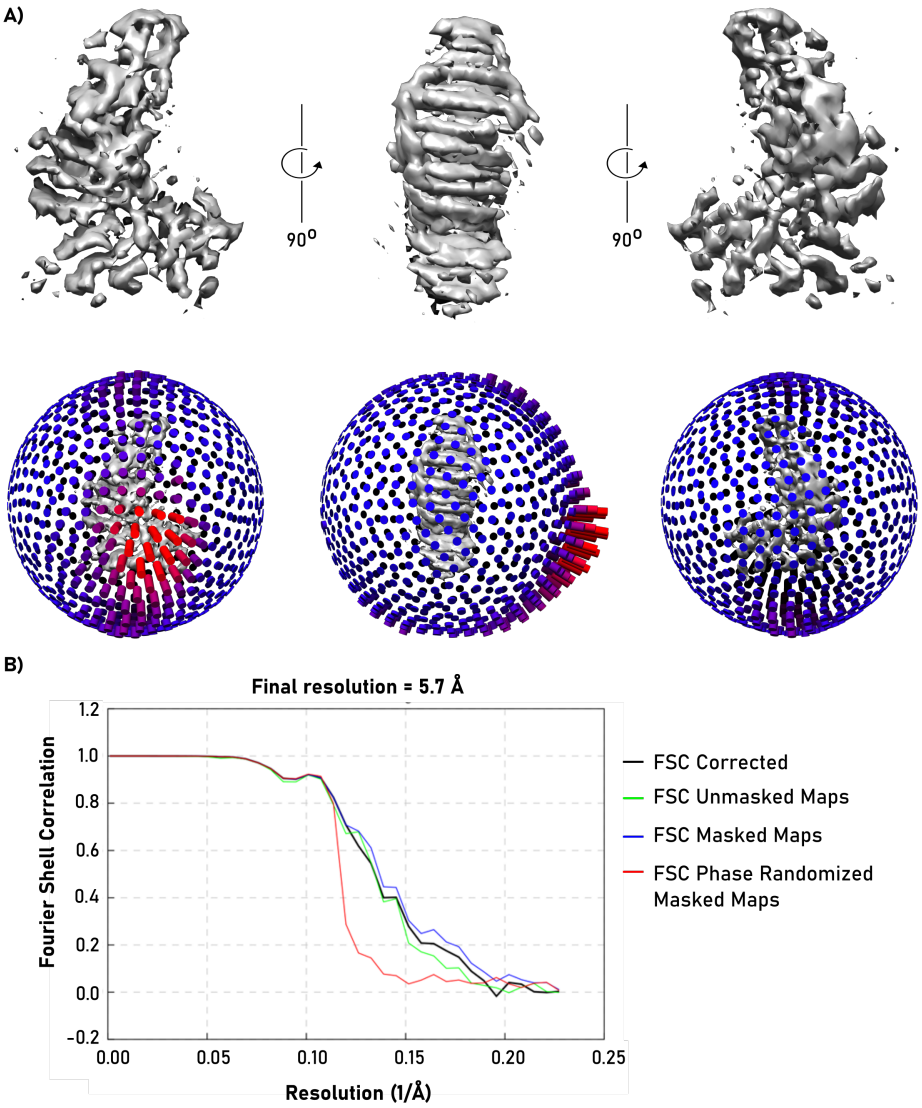


Figure 3.9: Cryo-EM reconstruction of the *Sl* FLCN-FNIP2 complex. A) At the top, different views of the *Sl* FLCN-FNIP2 5.7 Å map. At the bottom, the corresponding angular distribution for each view. Most of the views correspond to the front view (left) which indicates that the complex is strongly preferentially oriented in the grids. B) Global FSC curves for the final map.

Overall, the *Sl*FLCN-FNIP2 complex had an elongated C-shape architecture reminiscent to the human complex (Shen et al., 2019; Lawrence et al., 2019) (Figure 3.10), as we had previously observed in the micrographs (Figure 3.7). The angular distribution of the particles contributing to the model was uneven, with most of the particles displaying a front view covering the most elongated axis of the particle and bottom views (relative to Figure 3.9). The number of particles displaying side views (relative to Figure 3.9) was low, explaining the under-representation of these views in the 2D classes. The preferential orientation of particles at the air-water interface restricted the angular information available for the three-dimensional reconstruction and, ultimately, limited the achievable resolution. After applying a B-factor with RELION post-processing, the effect of preferential orientations was actually more pronounced, and there were no high resolution features at all in the areas with small number of projections. Despite the complex being highly preferentially oriented on the grid, our 3D model had sufficient detail to fit *Sl* FLCN and FNIP2 structural models generated using the human structures as template (Shen et al., 2019; Lawrence et al., 2019).

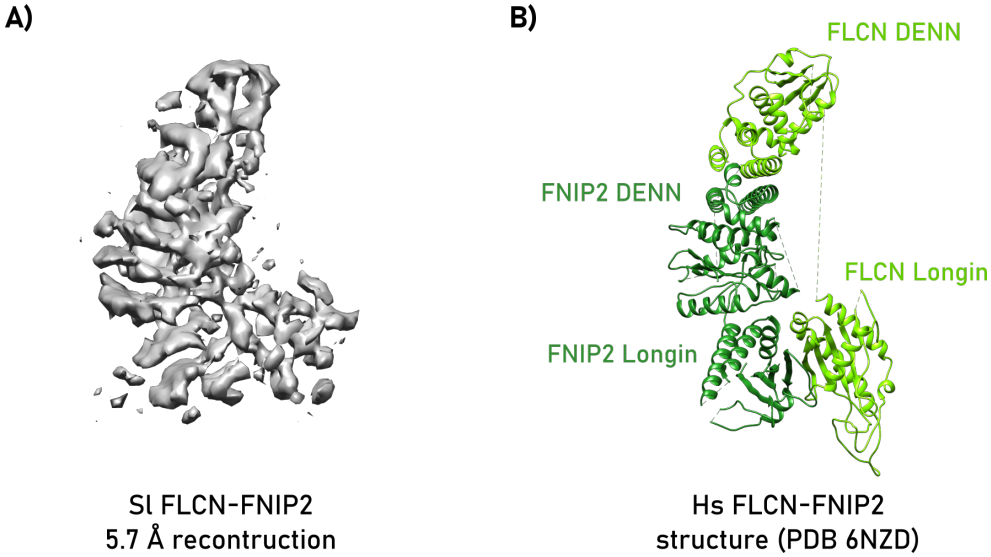


Figure 3.10: Structural comparison of the *Sl* FLCN-FNIP2 3D reconstruction with the human FLNC-FNIP2 structure. A) The 5.7 Å reconstruction of the *Sl* FLCN-FNIP2 complex. B) Structure of the human FLCN-FNIP2 complex determined by (Lawrence et al., 2019) (PDB 6NZD).

3.1.4 Architecture of the *Sl* FLCN-FNIP2 complex

The 5.7 Å three dimensional reconstruction of the *Sl* FLCN-FNIP2 had an elongated C-shape of 94 Å in length comprising an elongated tip and a wider base of 60 Å. Due to the particles being preferentially oriented on the grid (at the air-water interface), the map had several regions that were not resolved to high resolution. At the time the results of this chapter were obtained, AlphaFold2 had yet to be released (Jumper et al., 2021), so we initially generated structural models with Swiss-Model (Waterhouse et al., 2018) using the human structures as templates (Shen et al., 2019; Lawrence et al., 2019). We deleted the large unstructured regions of FNIP2 and fitted both, FLCN and FNIP2 structural models into the 5.7 Å reconstruction.

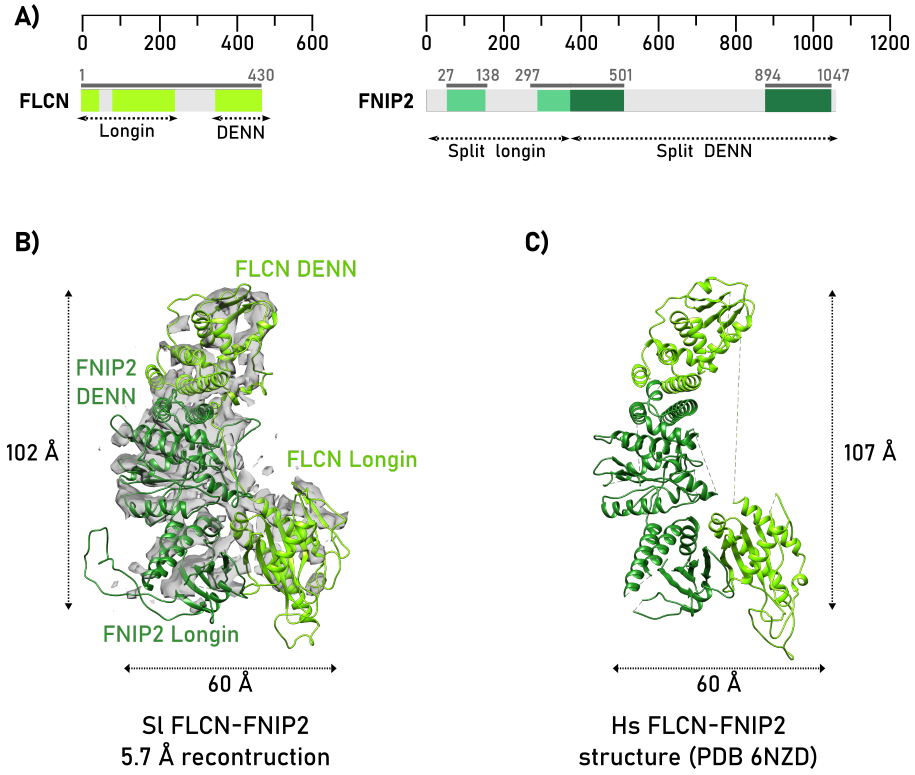


Figure 3.11: Architecture of the *Sl* FLCN-FNIP2 complex. A) Schematic representation of the domain organisation of *Sl* FLCN and *Sl* FNIP2. Regions covered by the Swiss-Model structural models are represented with a grey line. B) The Swiss-model structural models for FLCN (light green) and FNIP2 (dark green) fitted in the 5.7 Å map. The length of the reconstruction is indicated on the left and it is the distance between the C_α of Asn96 of the FNIP2 longin, and the C_α of Glu335 of the FLCN DENN model. The width of the base of the complex is indicated at the bottom and it is the distance between the C_α of Thr344 of FNIP2 longin domain and the C_α of Val201 of the FLCN longin domain. C) The structure of the human FLCN-FNIP2 complex (PDB 6NZD) (Lawrence et al., 2019). The length is the distance between the C_α of FNIP2 Asn136 and FLCN Pro428. The width is the distance between the C_α of FNIP2 Tyr371 and FLCN Asn249.

Both models fitted quite accurately into the reconstruction (real-space correlation coefficient at 5.7 Å calculated with Chimera of 0.73 for the *Sl* FLCN model and 0.66 for the *Sl* FNIP2 model) and showed that the overall architecture of the human FLCN-FNIP2 complex was preserved in *Spodoptera litura* (Figure 3.11). No extra regions were observed in the *Sl* FLCN-FNIP2 in comparison to that of the human structures (Shen et al., 2019; Lawrence et al., 2019). According to the structural models based on the human structures, the domain organisation of the human FLCN FNIP2 complex (Shen et al., 2019; Lawrence et al., 2019) was also conserved in *S.litura*. The FLCN longin and DENN domains were physically separated, whereas both FNIP domains were located at the centre of the complex interacting within them. The human structures of the FLCN-FNIP2 complex were part of a complex with the Rag GTPase heterodimer and Ragulator (Shen et al., 2019; Lawrence et al., 2019), whereas our 3D reconstruction was obtained from isolated *Sl* FLCN-FNIP2 complexes. Given that both structures had the same overall architecture and probably the same domain organisation according to structural models, we could confirm that the FLCN-FNIP2 complex retains its conformation in the absence of the Rag GTPase heterodimer.

3.2 Preliminary structural studies of the *Sl*FLCN-FNIP2-RagA-RagC complex

3.2.1 Expression and purification of the *Sl* RagA-RagC complex

The FLCN-FNIP2 complex is known to interact with the Rag GTPase heterodimer in mammalian cells to regulate the mTORC1 signaling pathway, and the same behaviour has been found in yeast (Péli-Gulli et al., 2015; Petit et al., 2013b; Tsun et al., 2013). We therefore attempted to form a FLCN-FNIP2-RagA-RagC complex and perform single-particle cryo-EM to provide the molecular mechanism for the role of FLCN in the regulation of mTORC1 activity. We initially cloned the *Sl* RagA and RagC sequences into a suitable vector to express them using the same baculoviral expression system used for the FLCN complex (Figure 2). The largest protein of the complex, RagC, included an SBP tag (Keefe et al., 2001) at the C-terminus, whereas RagA was left untagged. We generated bacmids containing the RagA-RagC genes, and transfected Sf9 cells to generate viruses. We amplified the virus obtained and used the final generation (P2) to infect a larger volume of Sf9 cells. After harvesting the cells, the RagA-RagC complex was purified by means of affinity purification using a Streptrap column and subsequent gel-filtration (see Chapter 2).

After affinity purification, two proteins of similar molecular weight to RagA (36 kDa) and RagC (44 kDa) eluted together. Subsequent mass spectrometry and gel-filtration confirmed that they formed a complex with an elution volume of 14 mL over a Superose 12 column. Overlapping the RagA-RagC elution profile with that of protein standards confirmed that the eluted peak was located as expected for a complex with the theoretical molecular weight of the Rag GTPase heterodimer (80 kDa) (Figure 3.12), although this comparison is not accurate as macromolecules are separated according to estimated shape an size (R_h) in SEC, it is commonly used as an estimation. Overall, both SEC and SDS-PAGE results corroborated the prior results that *Sl* RagA and RagC form a stable complex in insect cells.

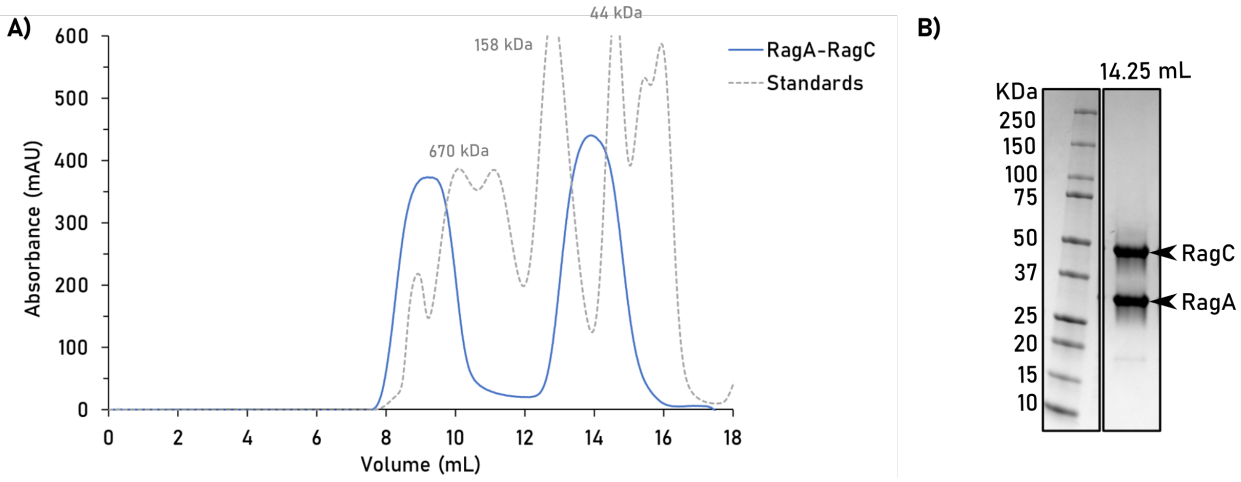


Figure 3.12: Purification of the RagA-RagC heterodimer. A) Size-exclusion chromatography profile of the RagA-RagC heterodimer after affinity purification (blue) and protein standards (dotted grey line). B) SDS-PAGE of the central fraction of RagA-RagC elution peak.

The FLCN complex acts as a GAP for the RagC/D GTPase, promoting the GDP-bound RagC/D state and, therefore, acting as a positive regulator of mTORC1. To verify that the insect complexes expressed using the baculoviral system were functional we attempted to assess the FLCN-FNIP2 GAP activity using a GTPase assay. Firstly, we used the Malachite Green Phosphate Assay kit (Sigma-Aldrich, USA), which generates a colorimetric product proportional to the GTPase activity of the present enzyme in the reaction. These results showed that the RagA-RagC GTPase heterodimer had a small amount of intrinsic GTPase activity ($[P_i]$=25 $\pm$ 1.8 μM), and that in the presence of FLCN-FNIP2 the GTPase activity was considerably increased ($[P_i]$=46 $\pm$ 2.9 μM). Unexpectedly, the FLCN-FNIP2 complex also displayed a considerable amount of GTPase activity ($[P_i]$=30 $\pm$ 1.4 μM) (Figure 3.13), which made these results difficult to interpret.

As we could not get reliable results with the Malachite Green assay, we then tried the GTPase-Glo^{TM} Assay (Promega), which detects the amount of GTP remaining after GTP hydrolysis (see Chapter 2) (Mondal et al., 2015). The GTPase activity is therefore inversely correlated to the amount of light produced. To assess the FLCN GAP activity towards RagA-RagC, we titrated serial dilutions of the GAP complex (FLCN-FNIP2) over a fixed concentration of the Rag GTPase heterodimer. Firstly, we measured the RLU without any protein which indicated the initial quantity of GTP (Figure 3.13). Our results showed that the RagA-RagC heterodimer was able to hydrolyse a small amount of GTP due to its intrinsic GTPase activity

(Figure 3.13). Consistently with the previous assay, the FLCN-FNIP2 complex also showed a substantial level of unexpected GTPase activity. We presumed that the activity could be caused by some contaminant that was pulled down with the FLCN complex and possessed GTPase activity, as there were still some extra bands in the SDS-PAGE gel after gel-filtration (Figure 3.4). The titration of FLCN-FNIP2 over a fixed concentration of RagA-RagC GTPases still promoted the GTPase activity in a dose-dependent manner, and this increase was still higher in comparison to the unexpected FLCN-FNIP2 basal activity. However, these results were not sufficient to distinguish whether or not this increase in GTPase hydrolysis was due to the specific FLCN GAP activity or the intrinsic GTPase activity observed in our FLCN-FNIP2 sample.

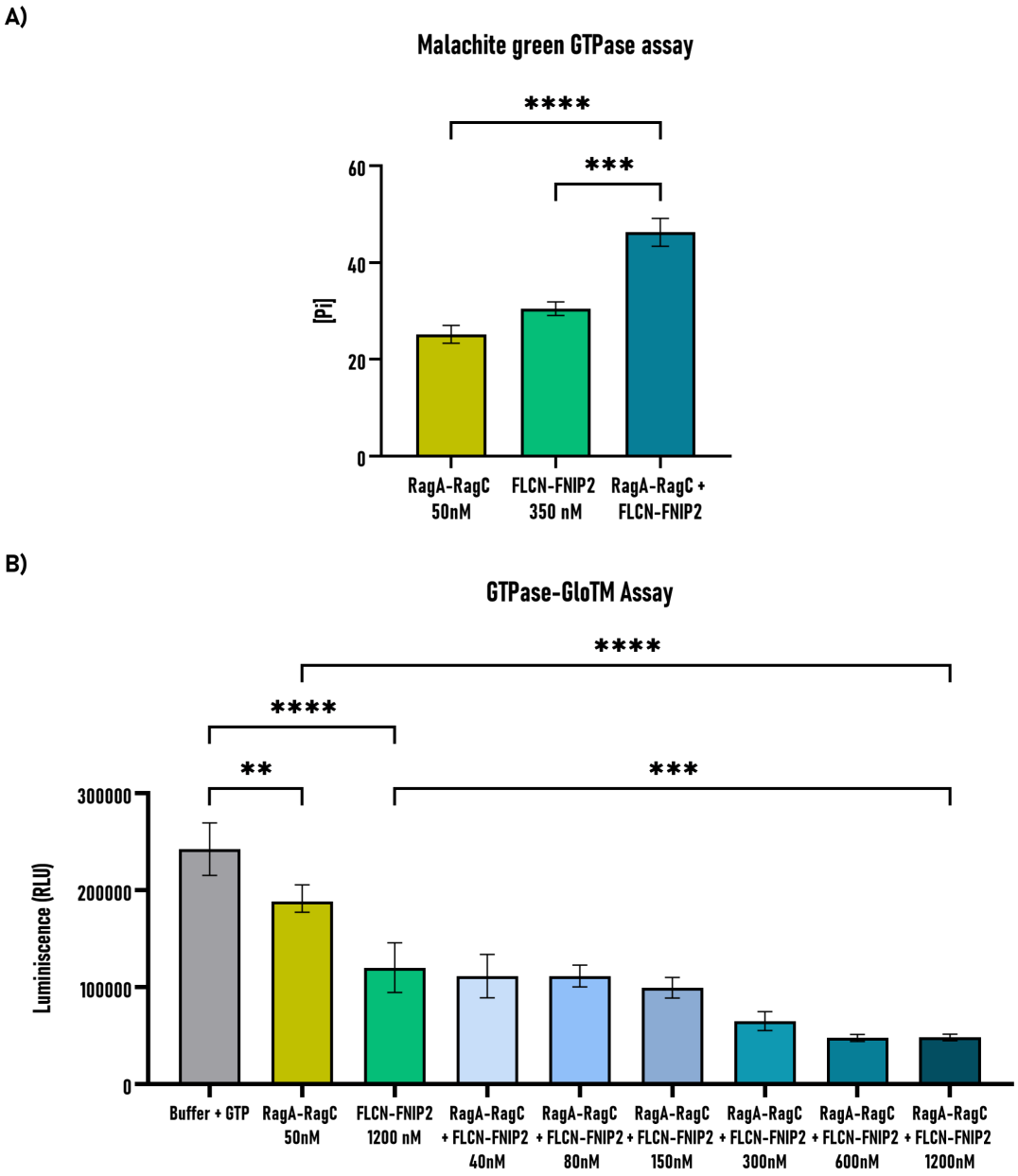


Figure 3.13: FLCN-FNIP2 GAP activity towards the RagA-RagC GTPase heterodimer. A) Malachite green GTPase assay. The [Pi] concentration represents the free phosphate liberated after GTP hydrolysis in three different conditions: RagA-RagC 50nM, FLCN-FNIP2 350 nM and combination of both complexes. B) GTPase-Glo^{TM} assay using serial dilutions of FLCN-FNIP2 (40, 80, 150, 300, 600 and 1200 nM) over a fixed concentration of RagA-RagC (50 nM). Luminiscence (RLU) represents the quantity of GTP left on the sample after the reaction. In both charts error bars on top of each column represent the standard deviation between triplicates. Both data sets were analysed with an ordinary one-way ANOVA with $P < 0.05$ taken to be significant.

3.2.2 Optimisation of *Sl* FLCN-FNIP2-RagA-RagC complex formation

The FLCN-FNIP2 complex is recruited to the lysosome interacting with GDP-bound RagA/B, where it performs its GAP activity towards GTP-bound RagC/D, to promote the mTORC1

activating state, RagC/D-GDP (Meng and Ferguson, 2018). To provide a detailed mechanism at a molecular level of how FLCN performs its GAP activity towards the Rag GTPase heterodimer, we firstly attempted to form a stable complex of FLCN-FNIP2 bound to RagA-RagC using the GTP analogue, GTPγS. This analogue replaces one of the oxygens of the γ-phosphate of GTP with a sulfur, thus creating a nucleotide that can only be slowly hydrolyzed, which facilitates the study of the active form of GTPases (Spoerner et al., 2007).

We therefore mixed the purified FLCN-FNIP2 with an excess of RagA-RagC in the presence of GTPγS. In the SEC profile, we observed a peak that was shifted in comparison to the FLCN-FNIP2 peak (elution volume of 10.5 mL and 12.25 mL respectively), and a later peak including the RagA-RagC excess (12.75 mL elution volume) (Figure 3.14). SDS-PAGE analysis confirmed that the 10.5 mL peak had the four components of the complex: FLCN, FNIP2, RagA and RagC. The FLCN-FNIP2-RagA-RagC would have a theoretical molecular weight of 245 kDa assuming a 1:1:1:1 stoichiometry. Although macromolecules are eluted according to their R_h in SEC, comparison of the complex SEC profile with that of protein standards showed that the putative complex had an elution volume lower than that expected for a complex of 245 kDa. This observation suggested three different scenarios: the complex could have a different stoichiometry than 1:1:1:1 between subunits, it was mostly aggregated, or the shape of the protein complex was very different to that of the protein standards (mostly globular), hence affecting its elution volume.

To analyse the quality of the complex, we prepared negatively-stained grids concentrating the fractions corresponding to the first eluted peak in the gel-filtration profile to a final concentration of 0.3 mg/mL and imaged them on a Tecnai T12 Twin microscope. We observed that most of the sample was forming large aggregates (Figure 3.15) and, although there were some areas that were free of aggregation that contained some single particles, the density within those areas was exceptionally low and the particles appeared smaller than would be expected (Figure 3.15), a sign of complex break down. The high level of aggregation observed in the micrographs was in line with the SEC profile, as the "complex peak" was too close to the void volume for a complex with a theoretical molecular weight of 245 kDa (Figure 3.14). Moreover, concentration of the sample after size-exclusion chromatography probably promoted aggregation, as we

previously observed for *Sl* FLCN-FNIP2 complexes (see section 3.1.1).

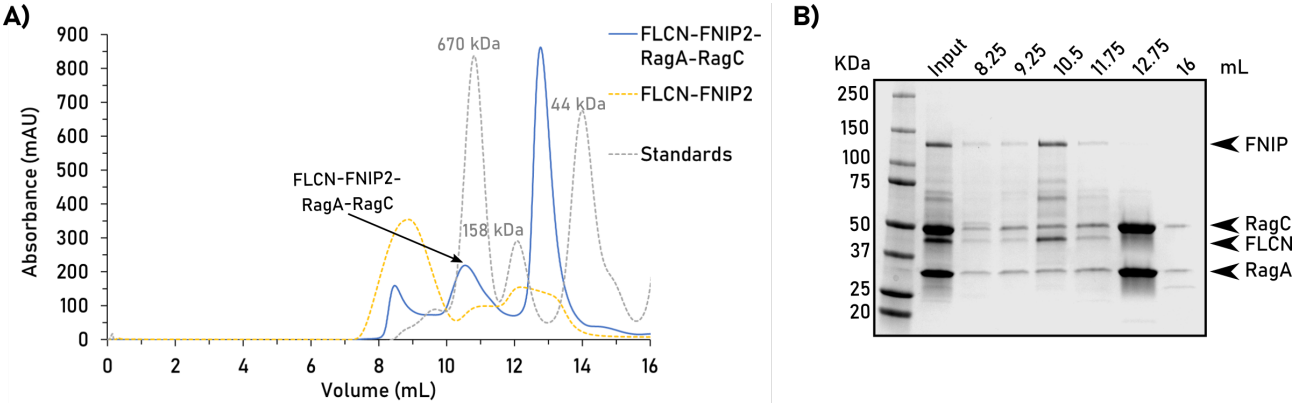


Figure 3.14: FLCN-FNIP2-RagA-RagC complex formation in the presence of GTPγS A) Size-exclusion chromatography profile of the FLCN-FNIP2-RagA-RagC complex in the presence of GTPγS (blue) overlapping with the FLCN-FNIP2 profile (orange dotted line) and protein standards trace (dotted grey line). B) SDS-PAGE of the representative fractions of the size-exclusion chromatography profile. The elution volume of each fraction is indicated at the top of the gel in mL.

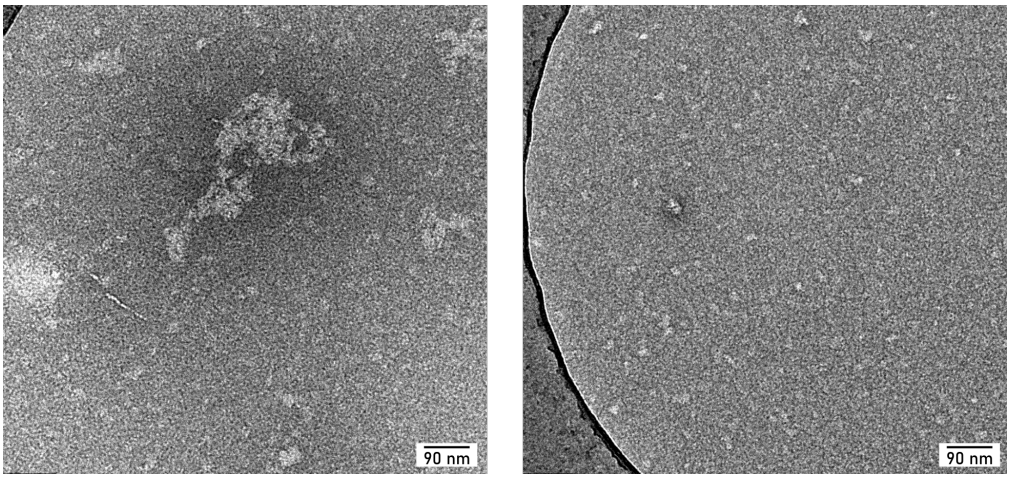


Figure 3.15: Negative staining screening of the *sl* FLCN-FNIP2-RagA-RagC complex formation in the presence of GTPγS. On the left, representative micrograph with aggregation. On the right, representative micrograph of an area free of aggregation. Scale bars are located at the bottom-right of each image.

The previous results showed an indication of FLCN complex binding to the Rag GTPase heterodimer in the presence of GTPγS according to the gel-filtration profile, however these complexes were prone to form large aggregates and to dissociate when applied to negatively-stained grids (Figure 3.15). For this reason, we decided to try a more aggressive method that would stabilise the complex: cross-linking. The addition of a fixation agent directly into the sample solution can sometimes result in the appearance of artefacts and aggregation due to non-specific interactions. Chemical fixation over a density gradient, also known as GraFix, is usually an efficient method that combines sedimentation of complexes in a density gradient with mild

chemical fixation (Kastner et al., 2008), successfully diluting aggregates and other contaminants. Prior to the density gradient, we performed a cross-linking screen to determine the optimal concentration of chemical fixation reagent that was needed to form complexes without obtaining a great deal of aggregation. We tested a wide range of glutaraldehyde percentages (0.004% to 0.5%), with concentrations between 0.032%-0.062% resulting in a proper compromise between complex formation and aggregation according to SDS-PAGE analysis (Figure 3.16)

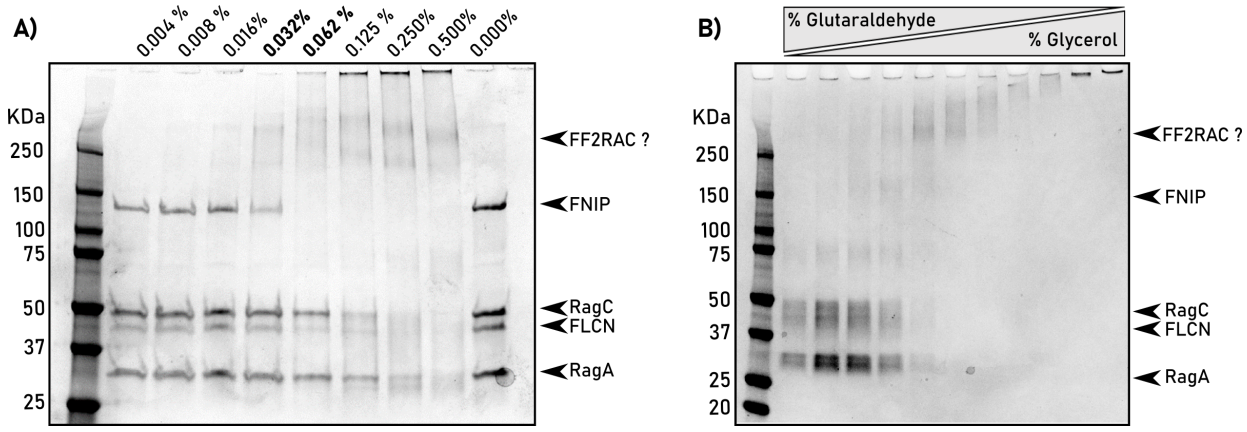


Figure 3.16: Cross-linking screening and density gradient fixation of the FLCN-FNIP2-RagA-RagC complex in the presence of GTPγS. A) SDS-PAGE of crosslinked complexes with serial dilutions of glutaraldehyde. Percentage of cross-linker is indicated at the top of the gel. B) SDS-PAGE after fractionation of density gradient.

To perform the density gradient cross-linking, we used a modified version of GraFix. Instead of adding the cross-linker to the denser solution to create a fixation reagent gradient in the same direction as the density gradient, the fixation agent was added to the less dense buffer, so that complexes were firstly fixated and then separated over the density gradient, hence minimising dissociation. After centrifugation, gradients were fractionated and analysed by SDS-PAGE. The FLCN-FNIP2-RagA-RagC complex has a theoretical molecular weight of 245 kDa assuming a 1:1:1:1 stoichiometry. In the gel we could observe that, while the FLCN-FNIP2 and RagA-RagC bands were disappearing proportionally to the quantity of fixation reagent added, a higher molecular weight band around 250 kDa appeared, which could presumably be the complex (Figure 3.16). We therefore prepared negatively-stained grids with fractions containing the 250 kDa band and imaged them using a Tecnai T12 Twin microscope. We could observe an even distribution of particles that appeared much larger than the FLCN-FNIP2 complex on

its own (Figure 3.5 and 3.17), we therefore proceeded to collected images to generate 2D class averages. Indeed, the 2D classes appeared larger than the FLCN-FNIP2 averages with some classes being 370 Å in length, however they were not consistent in size, nor in shape (Figure 3.17), which indicated that proteins were probably being randomly cross-linked, instead of being fixed through a specific interaction, and further optimisation was needed to achieve formation of a stable complex.

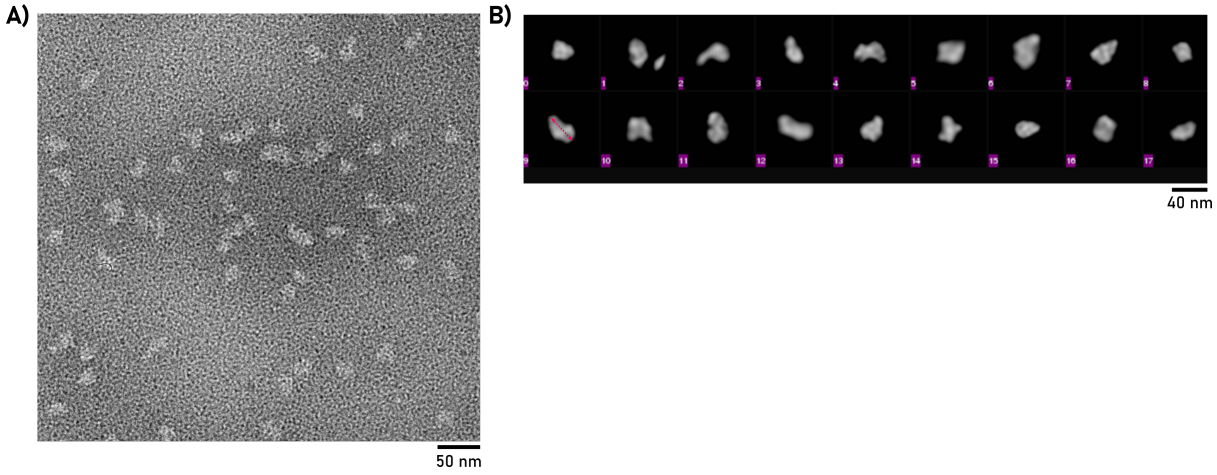


Figure 3.17: Screening of FLCN-FNIP2-RagA-RagC negatively-stained grids after density gradient fixation. A) Representative micrograph of crosslinked FLCN-FNIP2-RagA-RagC after density gradient. B) Representative 2D class averages. Pink arrow indicates 37a Å.

In parallel, we also investigated the effect of using other GTP analogues which provide different intermediate states, because the binding of FLCN-FNIP2 to the Rag GTPase heterodimer is dependent on the GTPase nucleotide state (Meng and Ferguson, 2018; Tsun et al., 2013; Petit et al., 2013a). We initially tried GDP-BeF_3, an analogue that mimics the GDP–Pi state, however, sample was aggregating immediately after adding the GDP-BeF_3 as we could observe a white precipitate in the solution. We then tried to form a complex in the presence of both GDP and GppNHp, a non-hydrolyzable GTP analogue that binds tightly to GTPases in the presence of Mg^{2+} and forms stable Ras complexes (Michael Spoerner et al., 2005). Apart from the GTP analogue, we also added GDP to promote the proper conformation of GDP-bound RagA, which is thought to be crucial for the FLCN-FNIP2 GAP activity towards RagC-GTP (Meng and Ferguson, 2018). We mixed the FLCN-FNIP2 complex (2.3 µM) with 2x excess of RagA-RagC (4.6 µM) in buffer supplemented with GDP and GppNHp. Complexes were left incubating on ice for 5h and then analysed by gel-filtration and SDS-PAGE. In the gel-filtration

profile, we could observe a peak just before the FLCN-FNIP2, which had the four proteins of the complex in the gel (Figure 3.18). We therefore prepared negatively stained grids and imaged them using a Tecnai T12 microscope (Thermo Fisher Scientific, USA). Grids displayed an even distribution of particles across the holes, however the particles appeared smaller than expected (80-90 Å long, similar to the FLCN-FNIP2 size) and they did not have a consistent size and shape (Figure 3.18). The low intensity of the peak and the large amount of unbound FLCN-FNIP2 and RagA-RagC indicated that complex was probably not stable enough to be held together on the grid.

The FLCN complex is thought to bind to the Rag GTPase heterodimer interacting with GDP-bound RagA and then performing its GAP activity towards GTP-RagC (Meng and Ferguson, 2018). To try to stabilise the complex, we intended to design a point mutation in RagA to lock its GDP-bound state. Unfortunately, while we were performing the cloning for the new construct, as well as optimising the conditions for the formation of a complex with FLCN-FNIP2 and the Rag GTPase heterodimer, two cryo-EM structures of this complex bound to Ragulator, responsible for anchoring the Rag GTPases to the lysosome, were published (Lawrence et al., 2019; Shen et al., 2019). In the following discussion, we will make a detailed comparison between the published structures and our results, as well as discussing their biological significance.

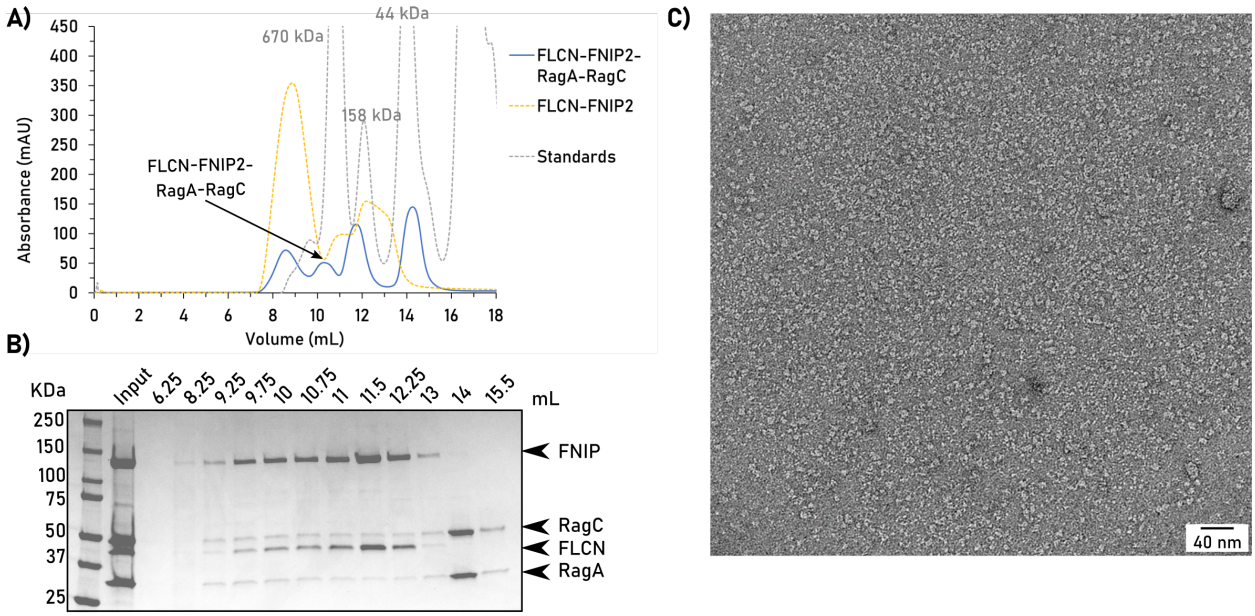


Figure 3.18: FLCN-FNIP2-RagA-RagC complex formation in the presence of GppNHp and GDP. A) Size-exclusion chromatography profile of the FLCN-FNIP2-RagA-RagC complex with GppNHp and GDP (blue) overlapped with the profiles of FLCN-FNIP2 (dotted orange line) and protein standards (dotted grey line). B) SDS-PAGE of eluted fractions from size-exclusion chromatography. The corresponding elution volume is indicated at the top each lane. C) Representative micrograph of a negatively-stained grid of FLCN-FNIP2. Image taken with a Tecnai T12 Twin microscope equipped with a CCD camera.

3.3 Discussion

mTORC1 is a central regulator of metabolism and cell growth that ensures that cells only grow under favorable conditions (Rabanal-Ruiz and Korolchuk, 2018). Regulation of mTORC1 is highly dependent on the nucleotide state of the Rag GTPase heterodimer, which in turn is regulated by several multiprotein complexes with GAP and GEF activity (Ramlaul and Aylett, 2018). FLCN-FNIP is one of the modulators of the Rag GTPase heterodimer nucleotide state because it possesses GAP activity towards RagC/D, thus promoting the activation of mTORC1. Furthermore, FLCN is involved in other cellular functions, such as membrane trafficking and lysosomal positioning, through the regulation of Rab GTPases (de Martín Garrido and Aylett, 2020). Due to its important function, mutations affecting the entire FLCN gene were found to be responsible for BHD syndrome (Nickerson et al., 2002), which ultimately, can result in the development of kidney cancer. Hence, a proper understanding of the FLCN complex at a molecular level can provide useful knowledge for the understanding of the development of tumours associated with BHD.

In this chapter, we aimed to use cryo-EM to determine a structure of the FLCN complex bound to its partner FNIP2, as well as the Rag-GTPase heterodimer to provide a detailed mechanism of the FLCN regulation in the mTORC1 signaling pathway. Firstly, we successfully expressed and purified the *Spodoptera litura* FLCN-FNIP2 complex and obtained an intermediate-resolution reconstruction of the *Sl* FLCN-FNIP2 complex. In parallel, two high-resolution structures of the human FLCN-FNIP2 complex bound to RagA-RagC-Ragulator were published (Shen et al., 2019; Lawrence et al., 2019) (Figure 3.19). These structures allowed us to generate structural models for the *Sl* FLCN and FNIP2 proteins and fit them into our reconstruction. Comparison between the *S.litura* structural model and the human structure, showed that they share the same architecture and domain organisation, and that both are missing large unstructured regions found in FNIP2, which are probably flexible elements involved in regulatory events events through phosphorylation (Hasumi et al., 2008a; Takagi et al., 2008). The human FLCN-FNIP2 structure show that both FLCN and FNIP2 contain the predicted longin and DENN domains

(Nookala et al., 2012; Zhang et al., 2012; Pacitto et al., 2015), which are typically encountered in proteins interacting with small GTPases (Yoshimura et al., 2010). The human structures also show that the FLCN and FNIP2 interact through two distinct interfaces (Shen et al., 2019; Lawrence et al., 2019). At the top of the complex, the two DENN domains interact through a long α-helix located towards the C-terminus of each domain. On the other side of the complex, the FLCN and FNIP2 N-terminal longin domains heterodimerize forming a ten β-sheet platform, in a similar manner to that found in the GATOR1 subunits NPRL2 and NPRL3, a GAP for Rag A/B (Shen et al., 2018). The heterodimerisation through the longin domains forms a platform that is accessible for the interaction with the Rag GTPase heterodimer, a typical behaviour for the longin domains that have previously been found to interact with small GTPases (Levine et al., 2013; Kinch and Grishin, 2006).

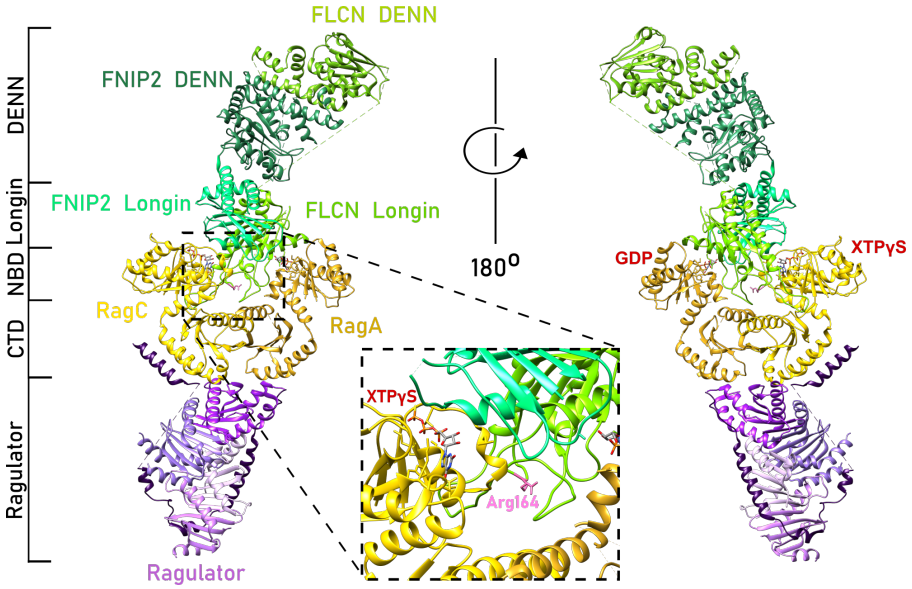


Figure 3.19: Organisation of the FLCN-FNIP2 complex at the lysosomal membrane. Full Structure of the FLCN-FNIP-Rag GTPase-Ragulator complex (PDB: 6NZD) (Lawrence et al., 2019) from two different views depicted in a cartoon representation. The inset shows how the residue Arg164 of the FLCN and the nucleotide bound to RagC are pointing different directions, thus making less plausible that this is the conformation of the complex with GAP activity. Figure adapted from (de Martín Garrido and Aylett, 2020) under the terms of a CC BY license.

In the second part of this chapter, we aimed to use cryo-EM to investigate at a molecular level how FLCN-FNIP2 binds the Rag GTPase heterodimer and exerts its GAP activity towards Rag C/D. We successfully expressed and purified the *Sl* RagA-RagC heterodimer but, unfortunately, while we were optimising the formation of an *Sl* FLCN-FNIP2-RagA-RagC complex, the two cryo-EM structures of the human complex bound to the lysosomal Ragulator were released at

a resolution of 3.6 Å (Lawrence et al., 2019) and 3.3 Å (Shen et al., 2019), respectively. In the following section I will use the human structure to understand the issues we were having while trying to form a stable complex between the *Sl* FLCN-FNIP and the Rag GTPase heterodimer. The FLCN-FNIP2-RagA-RagC-Ragulator complex displays an elongated architecture with the Rag GTPases located at the centre (Figure 3.19). The RagA-RagC GTPases contact Ragulator with their respective *C-terminal domain* (CTD)s, as previously reported in other structures (de Araujo et al., 2017; Yonehara et al., 2017), while the NBDs contact both FLCN and FNIP2 (Figure 3.19).

Interestingly, although FLCN is a GAP for RagC, both structures show that FLCN and FNIP2 contact both Rag GTPases, suggesting a unique mechanism for catalysis. The high-resolution of both models and comparison with GATOR1, a GAP for RagA/B (Shen et al., 2018), led to the identification of FLCN-Arg164 as the catalytic residue responsible for the GAP activity towards RagC (Shen et al., 2019; Lawrence et al., 2019), however, this residue is pointing away from the nucleotide bound to RagC in both structures (Figure 3.19). The fact that the catalytic residue is far from the nucleotide when FLCN-FNIP2 is bound to the lysosome, explains why the FLCN complex is inactive at the lysosomal membrane when amino acid levels are low. The authors suggest that a conformational change must occur so that FLCN can exert its GAP activity towards RagC/D, and that their structure is probably representing an intermediate towards GTP hydrolysis (Shen et al., 2019). Our *Sl* FLCN-FNIP2 reconstruction shows that the overall architecture of the complex is preserved when is not bound to the Rag GTPase, which suggests that the conformational change is likely to be local, around the catalytical arginine.

Furthermore, the architecture of the human FLCN-FNIP2-RagA-RagC-Ragulator complex provides a molecular mechanism to explain why the nucleotide state of the Rag GTPases controls FLCN recruitment to the lysosome under starvation conditions, and its further dissociation when nutrient levels are restored: when nutrient levels are high, the Rag GTPase heterodimer bound to the lysosomal membrane through Ragulator, is in its active state, RagA-GTP RagC-GDP and can bind the mTORC1 component, Raptor. Upon starvation, the inactive Rag GTPase heterodimer, RagA-GDP RagC-GTP, adopts a “wide-open” conformation, enlarging

the cleft between its NBDs, and leaving enough space to bind to the FLCN-FNIP complex (Figure 3.20). Because the FLCN complex uses the same interface as mTORC1 to bind the Rag GTPases, it needs to dissociate from the lysosome when nutrient levels are recovered, hence leaving the RagA-RagC NBDs cleft available for mTORC1 binding (Figure 3.20).

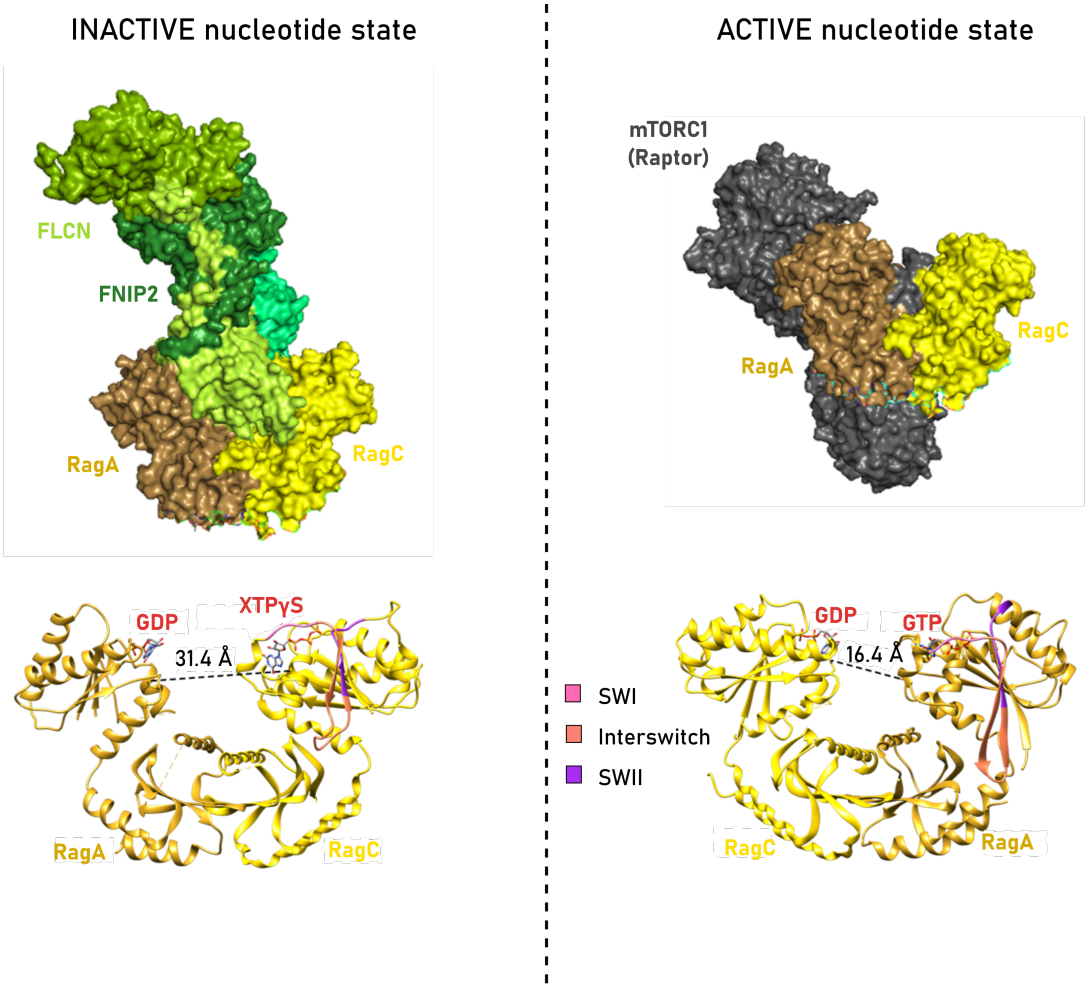


Figure 3.20: The Rag GTPase heterodimer nucleotide states. On the left, the RagA-RagC adopts a "wide-open" conformation when they are in their inactive nucleotide setting RagA-GDP and RagC-GTP which is able to accommodate the FLCN-FNIP2 complex between the NBDs (PDB 6NZD). On the right, when the Rag GTPase heterodimer is in its active nucleotide state (GTP-RagA-RagC-GDP) it adopts an open conformation able to bind to mTORC1 (PDB 6SB2) (Anandapadamanaban et al., 2019).

Both cryo-EM structures show that the FLCN-FNIP2 complex requires the Rag GTPase heterodimer to be in a "wide-open" conformation to fit within its NBDs. This conformation of the Rag GTPases is only achieved through a particular nucleotide state (GDP-RagA and RagC-GTP), which indicates that the binding of FLCN-FNIP2 is extremely dependent on the Rag GTPase nucleotide setting (Figure 3.20). In both human FLCN-FNIP2 structures bound to the Rag GTPases, their conformation was locked introducing mutations that promote their

appropriate nucleotide state to bind the FLCN-FNIP2 complex (Shen et al., 2019; Lawrence et al., 2019). In our attempts to form a complex, we were using wild-type RagA and RagC. Indeed, we were observing a small quantity of FLCN complex bound to RagA-RagC in the presence of GDP and GppNHp (Figure 3.18), but it was not stable enough to hold together under negatively-stained conditions. Considering the strict requirement for GDP-bound RagA and GTP-bound RagC for FLCN-FNIP2 to bind the Rag GTPase heterodimer, and the fact that the Rag GTPases have a basal GTPase activity, we presume that there was probably a fraction of the Rag GTPases in the right nucleotide setting which was able to bind FLCN-FNIP2, but such fraction was probably only a small proportion and not stable enough to hold together in the grids. With the available structures (Shen et al., 2019; Lawrence et al., 2019), we also checked whether or not any of our tags were disrupting any key interactions for complex formation. Both FLCN and FNIP2 interact with the Rag GTPase heterodimer through their N-terminal longin domains, and our constructs had the tag on the C-terminus of FNIP2. The tag is also non-disrupting for the FLCN-FNIP2 interaction as we obtained an intermediate-resolution reconstruction that displayed the same conformation as the human high-resolution structure. On the other hand, our RagA-RagC construct contained a tag in the C-terminus of RagC. The Rag GTPase heterodimer interacts with the FLCN-FNIP2 complex through the cleft formed between their NBDs, while the CTDs heterodimerize. We could express and purify a RagA-RagC stable complex, which indicates that the tag was not affecting the heterodimerisation of the Rag GTPases, and given than the interaction through FLCN occurs via their NBDs, it is very unlikely that the CTD tag was interfering in complex formation. Another noticeable feature of the published structures is that they both include the Ragulator complex (Shen et al., 2019; Lawrence et al., 2019). This complex is responsible for anchoring the Rag GTPases to the lysosome and to some extent, this interaction can also affect the proper Rag GTPases conformation.

The nucleotide setting of the Rag GTPase heterodimer is key for the FLCN-FNIP2 binding. RagC binds GTP 4-fold tighter than RagA (RagC K_d=11 nM, and RagA K_d=47 nM) and both bind GDP with less affinity (RagC K_d=41 nM, and RagA K_d=76 nM) (Shen et al., 2017). Moreover, GTP dissociates from both Rag GTPases much more slowly than GDP (Shen et al.,

2017). However, the unique heterodimeric architecture of the Rag GTPases allows them to have intersubunit communication so that binding of GTP to one subunit inhibits association and promotes the hydrolysis of GTP on the other subunit (Shen et al., 2017). This mechanism ensures that the Rag GTPase heterodimer is maintained in its active or inactive state. Therefore, we would expect that without introducing any mutations in the Rag GTPase heterodimer, we would have a major population of either GTP-RagA-RagC-GDP (active) or GDP-RagA-RagC-GTP (inactive) and that half of our FLCN-FNIP2 complexes should be able to bind to the Rag GTPase inactive conformation. According to our SEC profiles and our grids we had a very small pool of FLCN complexes binding to the Rag GTPase heterodimer, suggesting that the complex was being disrupted even with the appropriate Rag GTPase nucleotide state. The introduction of mutations that lock the appropriate nucleotide state of the Rag GTPases and the presence of Ragulator was probably key to obtain stable FLCN-FNIP2 complexes with the Rag GTPases. If the human structures would not have been published, we would have generated Rag GTPase mutations that would lock their nucleotide state, as some studies were already pointing to it being key for FLCN binding (Meng and Ferguson, 2018), and ordering of the Ragulator genes had already mentioned in our discussions. Furthermore, if Alphafold had been available to us we would have been able to generate a structural model of the FLCN-FNIP2 complex that we could have used to optimise our FLCN-FNIP2 and Rag GTPase expression constructs, for example, deleting the large unstructured regions of FNIP2, which are not part of the core of the FLCN-FNIP2 complex. Without accurate structural models it was not clear which regions should be deleted without compromising complex formation.

The published structures of the human FLCN-FNIP2 complex bound to RagA-RagC-Ragulator shed light on how FLCN regulates the Rag GTPase heterodimer nutrient state at the lysosome under starvation conditions, however, there are still many questions that remain unanswered. FLCN mutations affecting the entire gene are responsible for BHD syndrome, including mutations in the C-terminal DENN domain of FLCN. Based on the structure, mutations in this domain would not disrupt the interaction with the Rag GTPases, which indicates that this domain must be involved in other important functions. As we previously highlighted in the introduction, FLCN DENN domains have been reported to interact with Rab GTPases to

modulate membrane trafficking (Nookala et al., 2012), and FLCN has also been involved in lysosomal positioning through its interaction with other members of the Rab family (Starling et al., 2016) and with autophagy (Goodwin et al., 2021). Moreover, FLCN-FNIP can also interact with AMPK, another regulator of nutrient signalling. Hence, FLCN is a multifunctional protein involved in several cellular processes which require a wide range of different interacting partners. How this versatile protein coordinates all these cellular functions still remains unclear.

Chapter 4

Structure of the ΦKZ non-virion RNA polymerase

Bacteriophage ΦKZ is one of a few exceptional bacterial viruses to encode two multisubunit RNAPs which transcribe all genes needed throughout the different stages of the infection cycle. The study of these non-canonical msRNAPs can provide useful information with respect to the foundations of transcription, as well as the intriguing evolution of bacteriophages. Previous biochemical studies have demonstrated that the ΦKZ non-virion RNAP is a 5-subunit complex able to initiate transcription from specific promoters. However, little is know about the function of each subunit during the transcription reaction. Intriguingly, the fifth subunit, gp68, has no homology to any other protein with known function but has been reported to be crucial for the assembly of a functional enzyme, promoter recognition and transcription initiation. In this chapter, we aimed to determine the structure of the ΦKZ non-virion RNA polymerase in order to provide first insights into its architecture, the organisation of the four split β/β' -like subunits, and the role of gp68. To determine the structure of this five-subunit complex (329 KDa) we chose cryo-EM, which is particularly powerful for large multiprotein complexes and requires minimal quantities of sample.

4.1 Preparation of ΦKZ non-virion RNA polymerase complexes

ΦKZ nvRNAP complexes were prepared by Dr. Yakunina's group at Peter the Great St. Petersburg Polytechnic University (St. Petersburg, Russia) (de Martín Garrido et al., 2021b).

Complexes were expressed in *Escherichia coli* using an established co-expression system with one plasmid containing the four β/β'-like proteins and a second one containing the fifth subunit, gp68 (Orekhova et al., 2019). This system allowed the generation of both 4-subunit (4s) complex containing the four nvRNAP subunits with homology to the two largest bacterial subunits, and 5-subunit (5s) nvRNAP complexes including all the subunits of the enzyme. To determine if the purified complexes were active, our collaborators performed an *in vitro* transcription assay with an RNA/DNA scaffold. 5s complexes containing all the subunits of the enzyme were able to elongate the existing RNA, however 4 subunit complexes were unable to start transcription, suggesting a crucial role of gp68 in transcription initiation (Orekhova et al., 2019).

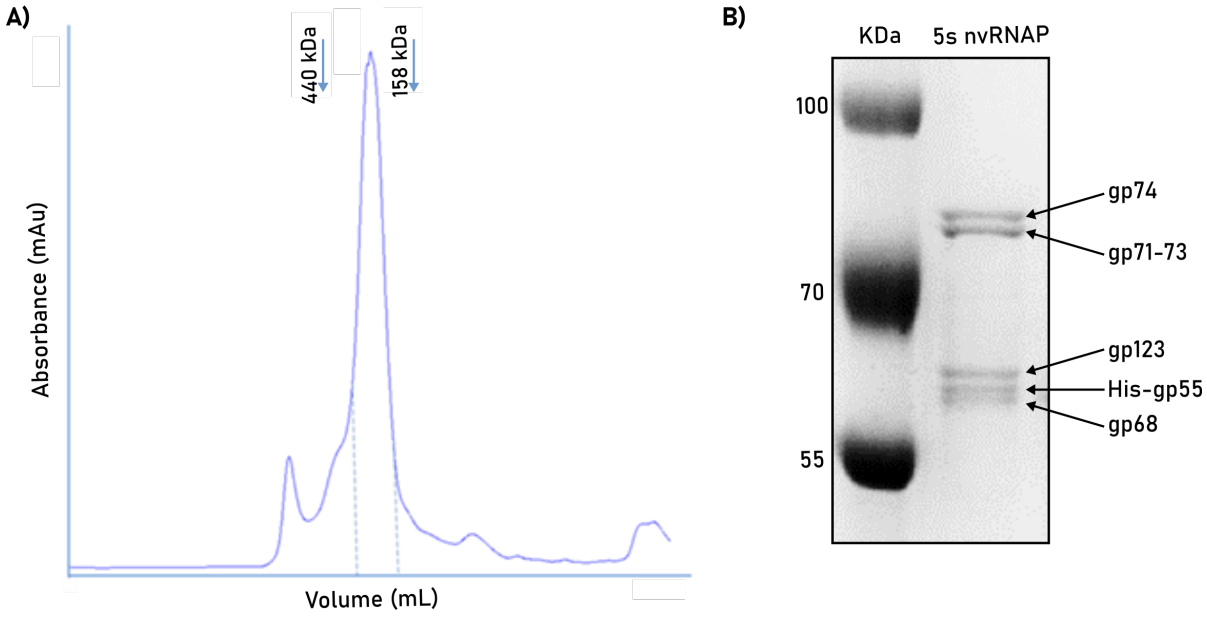


Figure 4.1: Size-exclusion chromatography of the ΦKZ nvRNAP. A) SEC profile of the ΦKZ nvRNAP (blue). B) Coomassie-stained SDS-PAGE of eluted nvRNAP fraction containing the five subunits of the complex. Data generated by Dr. Yakunina's group. Figure adapted from (de Martín Garrido et al., 2021b) under a Creative Commons Attribution License.

As 5s complexes proved to be functional, we used them for further structural analysis using cryo-EM (de Martín Garrido et al., 2021b). Complexes were expressed in *E.coli* using the same protocol described in Orekhova et al. (2019); de Martín Garrido et al. (2021b). Briefly, after harvesting and lysing the cells, ΦKZ nvRNAP complexes were purified by nickel affinity chromatography and subsequent gel-filtration (Chapter 2). The presence of all the subunits forming the complex was validated by SDS-PAGE, which showed unique bands for each subunit and no contaminants: gp74 (77 kDa), gp71-73 (74 kDa), gp123 (63 kDa), His_6-gp55 (56 kDa), and gp68 (59 kDa) (Figure 4.1 A, B). After that, bands were analysed by mass-spectrometry, confirming the presence of the five subunits that form the nvRNAP complex (Orekhova et al., 2019; de Martín Garrido et al., 2021b). To investigate the role of gp68, this subunit was also cloned into a single plasmid and expressed independently using *E.coli* cells. The purified free gp68 was only used in biochemical experiments (de Martín Garrido et al., 2021b) performed by our collaborators.

4.2 Optimisation of vitrification conditions for the ΦKZ nvRNAP

Initial screening of frozen samples was performed in-house at the Electron Microscopy facility of the Centre for Structural Biology (Imperial College London, UK), using an FEI Tecnai F20 electron microscope equipped with a Falcon II direct electron detector. Purified nvRNAP complexes were frozen and shipped at 4 °C by our collaborators in Dr. Yakunina's group. Upon arrival, samples were immediately aliquoted, flash frozen in liquid nitrogen and stored at -80 °C until use. To prepare cryo-EM samples we thawed one of the frozen aliquots and diluted it to a final concentration of 0.3 mg/mL. We adsorbed purified ΦKZ nvRNAP complexes to a thin film of graphene oxide deposited onto holey carbon copper grids. We chose this method because it had already been widely optimised in our lab, and it had been successful in the preparation of frozen samples of multiprotein complexes de Martín Garrido et al. (2021c); Ramlaul et al. (2021).

We found that nvRNAP complexes were easily adsorbed to the graphene oxide film and we could prepare well-populated cryo-EM grids (≈70-90 particles/micrograph) with a low concentration of sample (0.1-0.3 mg/mL) (Figure 4.2 A), which was very convenient as we were working with shipped sample from our collaborators in Russia. After screening different vitrification conditions, we determined that we were consistently achieving thin layers of vitreous ice with a blot force of -6, waiting times of 4-10 s to allow complexes to adhere to the graphene oxide film, and 1-2 s blotting times (Figure 4.2 A)(see Chapter 2). Micrographs of ΦKZ nvRNAP contained single particles with consistent size (≈ 100 Å), and several orientations of these could already be appreciated by eye (Figure 4.2 A). We therefore set up a collection using the FEI Tecnai F20 electron microscope available in the college facility (see Chapter 2).

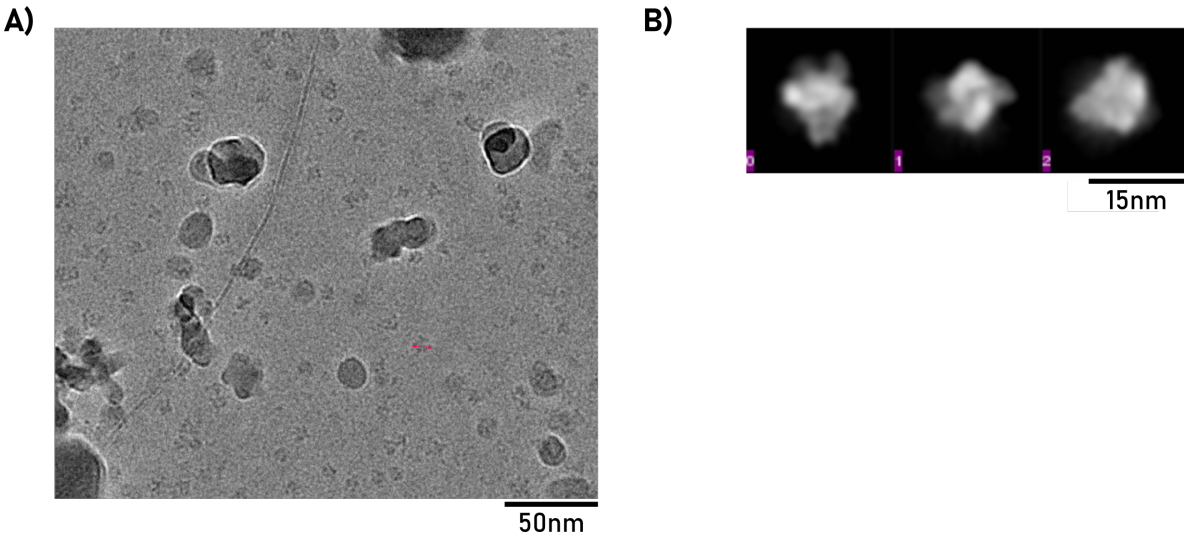


Figure 4.2: Screening of vitrification conditions for the ΦKZ nvRNAP. A) Representative micrograph of frozen nvRNAP complexes adsorbed to a graphene oxide film deposited onto a copper grid. Pink line represents the length of a particle measuring 100 Å. B) 2D class averages representing different views of the complex.

We performed single-particle analysis to obtain an initial 3D reconstruction that would provide information about the overall shape of nvRNAP complexes. After initial pre-processing of the collected movies with MotionCor2 (Zheng et al., 2017) and CTFFIND4 (Rohou and Grigorieff, 2015), 18 508 particles that classified into well-defined 2D class averages (Figure 4.2 B) were selected to generate an initial 3D reconstruction. Initial models were generated using the *stochastic gradient descent* (SGD) approach in RELION 3 (Zivanov et al., 2018) and yielded an initial three-dimensional reconstruction of the complex at 14 Å resolution according to an independent half-set FSC of 0.143 (Figure 4.3). This low resolution reconstruction was then filtered to 50 Å and used as an initial reference for further refinement. The initial reconstruction showed a reminiscent shape to that of bacterial msRNAPs, which was not surprising considering the reported sequence homology (Yakunina et al., 2015), and it also displayed a clear internal channel typical of these enzymes.

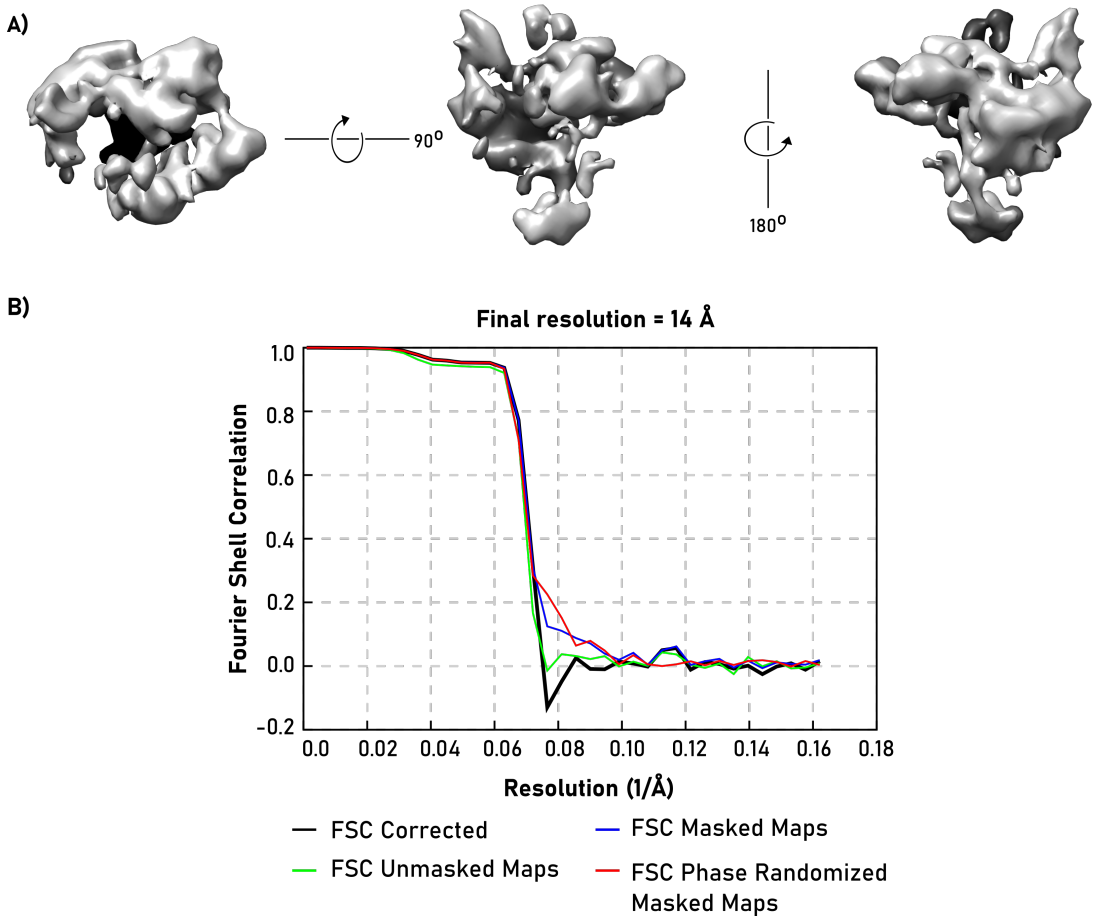


Figure 4.3: Initial reference-free three-dimensional reconstruction of the ΦKZ nvRNAP. A) Reference-free 3D reconstruction of the complex at 14 Å resolution showing different views of the structure. B) Global FSC curve for the 14 Å reconstruction.

4.3 High-resolution data collection, image processing and model building for the ΦKZ nvRNAP

After determining the optimal vitrification conditions for ΦKZ nvRNAP complexes, we collected a larger data set on a Titan Krios G3i equipped with a K3 direct electron detector at the London Consortium for Electron Microscopy (The Francis Crick Institute). Micrographs were well-populated (≈ 150 particles/micrograph) and single particles were clearly visible, which had the same shape and dimensions as previously observed while screening in-house (Figure 4.4 A). We therefore collected a set of 10 741 movies (see details in Chapter 2). Frames were aligned and dose-weighted using MotionCor2 (Zheng et al., 2017) and CTF estimation subsequently performed with CTFFIND4 (Rohou and Grigorieff, 2015). Based on the CTF estimation, micrographs with low estimated resolution and irregularities in the observed Thon Rings were discarded. After particle selection and extraction, particles were refined in 2D reference-free alignment to exclude ice contaminants and other types of particles. Over a million particles (1 043 026) classified into high-resolution 2D class averages, which included different views of the particle (Figure 4.4).

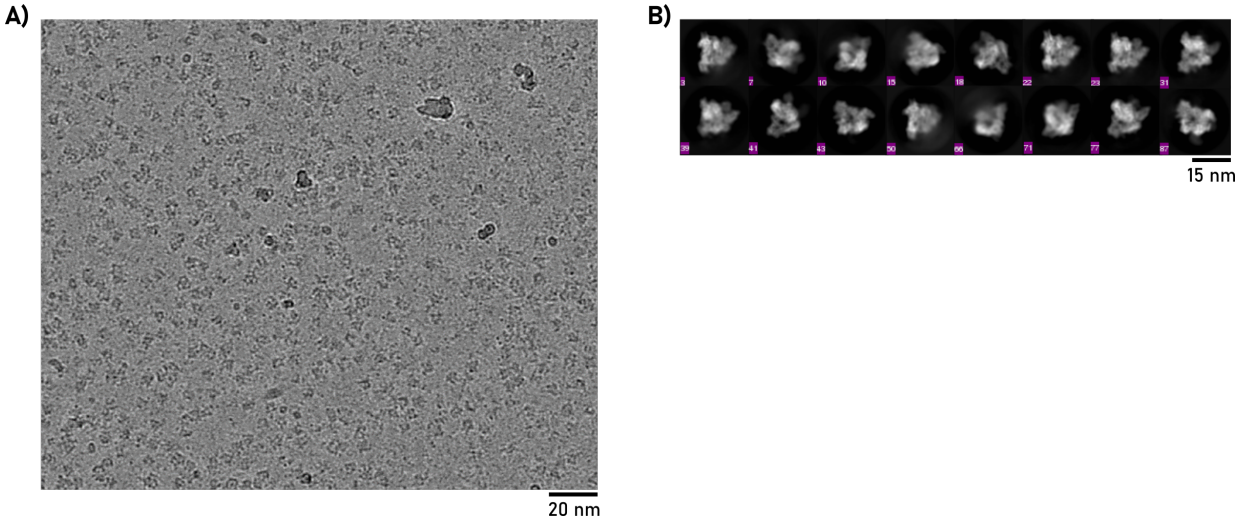


Figure 4.4: High-resolution data collection of ΦKZ nvRNAP complexes. A) Representative micrograph of the ΦKZ nvRNAP complex adsorbed to a graphene oxide film deposited onto a copper grid. B) Representative 2D class averages displaying different views of the complex.

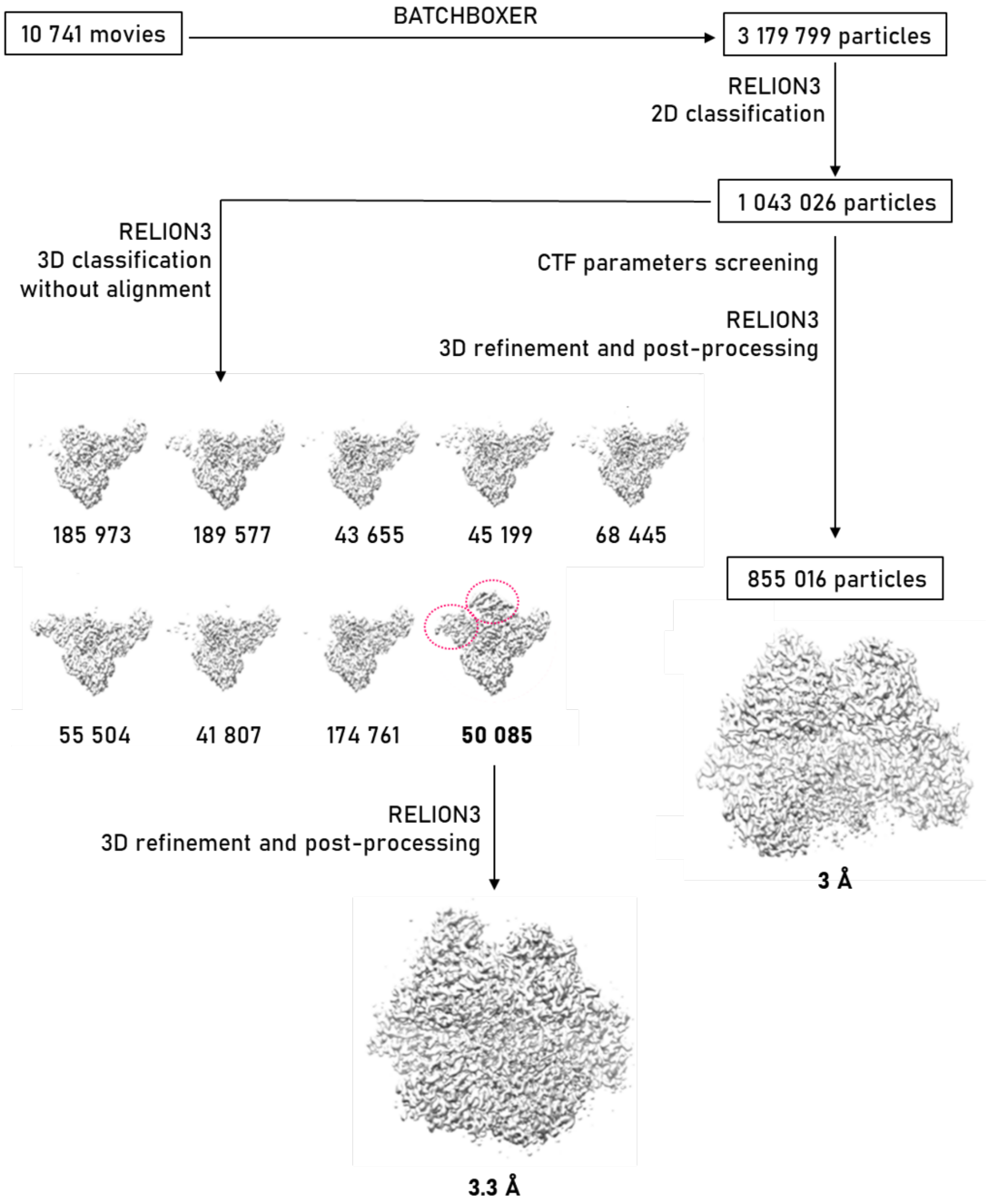


Figure 4.5: Data processing scheme for the cryo-EM high-resolution data collection for the ΦKZ nvRNAP complex. Flow-chart depicting the data processing procedure. Particle numbers retained at each stage of the process are indicated below each 3D reconstruction. Well-resolved peripheral regions are highlighted in pink circles in the class retained for further refinement to obtain the 3.3 Å reconstruction. Figure adapted from de Martín Garrido et al. (2021b) under a Creative Commons Attribution License.

Due to the large amount of particles we performed further screening based on the CTF parameters to retain particles that were likely to contain high-resolution information. We therefore selected all particles that generated high-resolution 2D class averages of micrographs collected at defocus values over 1 µm. Those particles yielded an initial reconstruction at 3.0 Å resolution

according to an independent half-set FSC of 0.143 (Figure 4.6 A-D). This initial reconstruction included the better ordered regions of the core of the enzyme at high-resolution, with most areas being resolved at higher resolution than 3.5 Å (Figure 4.6 B). However, some peripheral areas still showed weak density (Figure 4.6 A), particularly the area corresponding to the clamp, a mobile element that can adopt different conformations which are key to perform the enzyme's function (Chakraborty et al., 2012). Extensive sorting by means of 3D classification without alignment and focused refinement, finally yielded a 3.3 Å resolution with the peripheral areas of the enzyme better resolved (Figure 4.7). The 3.3 Å map was used to assist the building of a molecular model for the peripheral regions of the nvRNAP, particularly, the clamp (Figure 4.5).

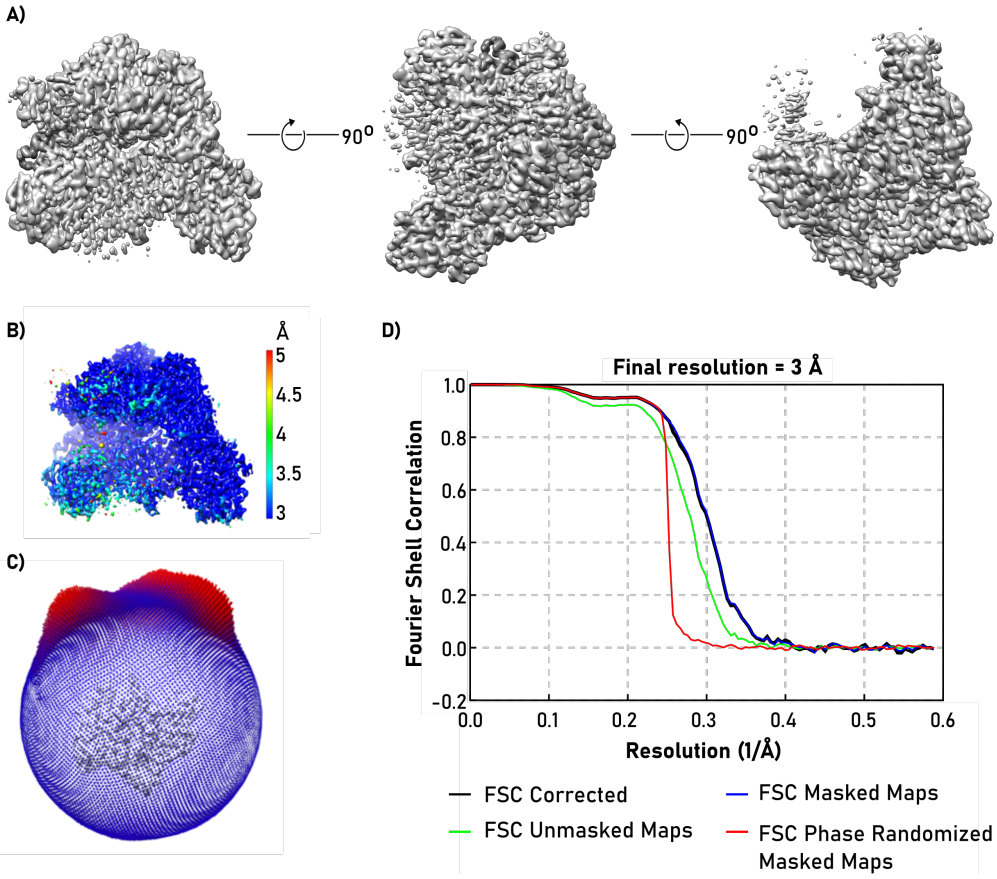


Figure 4.6: High-resolution 3D reconstruction of the ΦKZ nvRNAP core. A) 3 Å reconstruction of the ΦKZ nvRNAP in different views. The map was generated including particles that classified into high-resolution class averages from micrographs collected over 1 µm defocus. B) Local resolution map of the ΦKZ nvRNAP 3 Å reconstruction calculated using BLOCRES. C) Angular distribution plot of the ΦKZ nvRNAP 3 Å reconstruction. D) Global FSC curve for the ΦKZ nvRNAP 3 Å reconstruction.

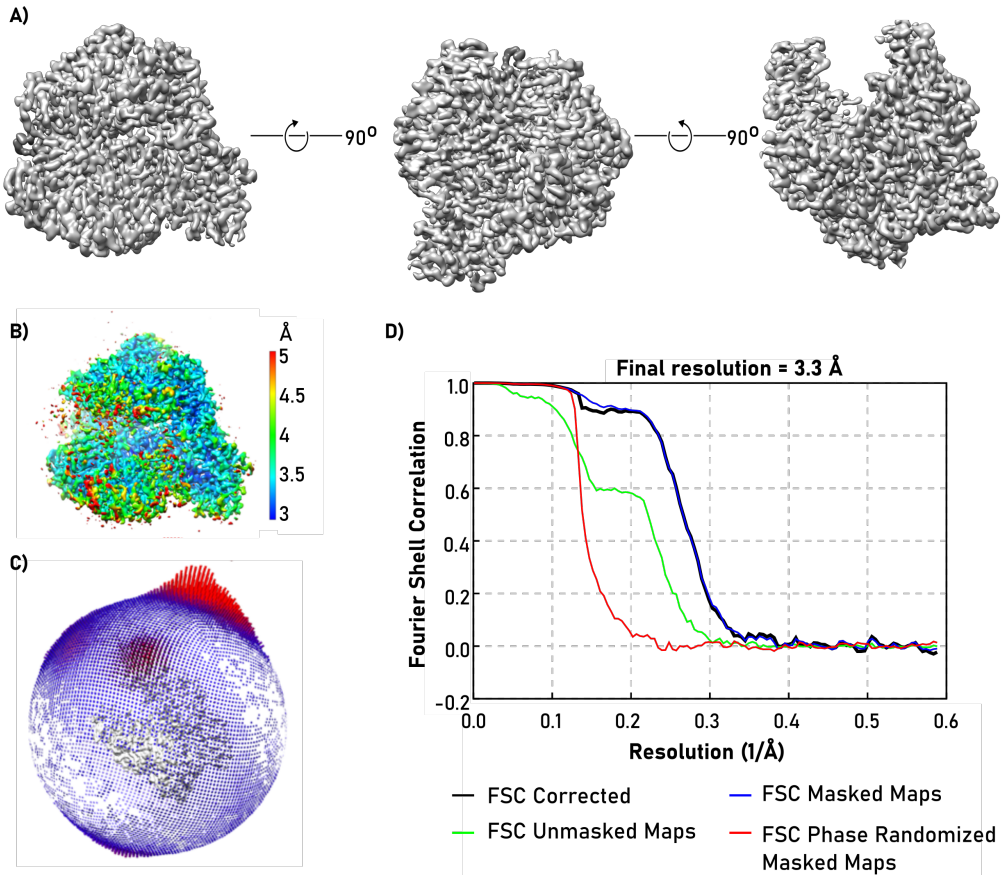


Figure 4.7: High-resolution 3D reconstruction of the ΦKZ nvRNAP including the clamp. A) 3.3 Å reconstruction of the ΦKZ nvRNAP including the clamp depicted in different views. The map was generated after extensive sorting of the particles with 3D classification without alignment and focus refinement. B) Local resolution map of the ΦKZ nvRNAP 3.3 Å reconstruction calculated using BLOCRES. C) Angular distribution plot of the ΦKZ nvRNAP 3.3 Å reconstruction. D) Global FSC curve for the ΦKZ nvRNAP 3.3 Å reconstruction.

Overall, the nvRNAP reconstruction had a shape reminiscent of the typical "crab-claw" appearance of bacterial msRNAPs with an internal channel running along the side of the enzyme, and two pincers at each side of the channel (Sutherland and Murakami, 2018). This resemblance was not unexpected given the previously reported sequence similarity between four of the nvRNAP subunits and the two largest bacterial subunits β and β' (Yakunina et al., 2015). Sequence analysis by BLASTp identified that the ΦKZ subunits were mostly similar to the RNAP structures of *Mycobacterium tuberculosis* (PDB 6EDT) (Boyaci et al., 2019). We therefore used this structure as a template to generate homology models for the four β/β'-like subunits of the nvRNAP using SWISS-MODEL (Waterhouse et al., 2018) (Alphafold2 was not available at that time). Models for the gp71-73, the gp74 and the C-terminus of gp55 (290-419) were well-defined, whereas models for gp123 were very variable and had large sections without defined

secondary structure, which was not unexpected given its weak homology to the C-terminus of the bacterial β subunit. As gp68 had no homology to any other protein with known function (AR9 structure yet to be released) and Alphafold2 was not available at that time, we could not generate any reliable structural model for this subunit. Using the well-defined models of gp71-73, gp74 and gp55, and the highest-resolution map (3 Å resolution) (Table 4.8), positions for the most conserved structural elements of the four β/β'-like subunits were established and expanded from them to provide an atomic model in Coot (Emsley and Cowtan, 2004; Emsley et al., 2010) (Figure 4.8) covering the majority of gp71-73, most of gp74 and gp123, excluding their middle domains, and the C-terminus of gp55 (Figure 4.9) (see Chapter 2).

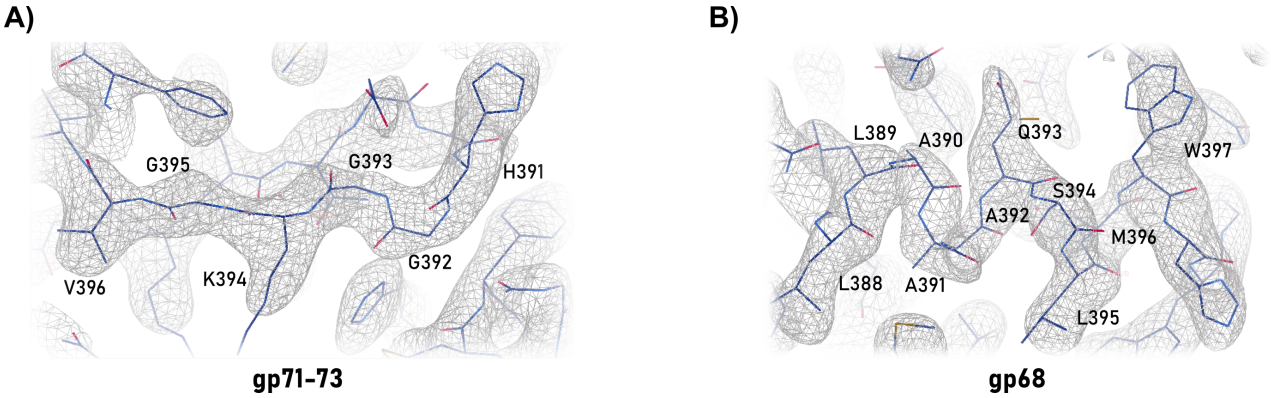


Figure 4.8: Representative high-resolution density for the ΦKZ nvRNAP 3 Å map. gp71-73 (A) and gp68 (B) molecular models fitted into the 3 Å map. Molecular models are displayed in blue, and the map is displayed as a grey mesh.

The N-terminus of gp55, including the clamp, was completed using the 3.3 Å reconstruction, which had better defined density for some peripheral regions (Figure 4.9). After building an atomic model for the the four subunits with homology to bacterial msRNAPs, which formed the core of the enzyme, we noticed that there was an extra domain displaying a mostly helical topology located at the bottom of the map, between gp55 and gp71-73 (Figure 4.9). Guided by secondary-structure predictions (Drozdetskiy et al., 2015), we could assign the additional domain to the C-terminus of the nvRNAP fifth subunit, gp68 (covering residues 304-413 and 429-521) (Figure 4.9). Overall, combining both the 3 Å and the more completed Å 3.3 reconstructions, we built an atomic model for 73% of the complex, providing a structure of the ΦKZ nvRNAP for the first time and identifying the location of its intriguing fifth subunit, gp68 (Figure 4.9 and 4.10).

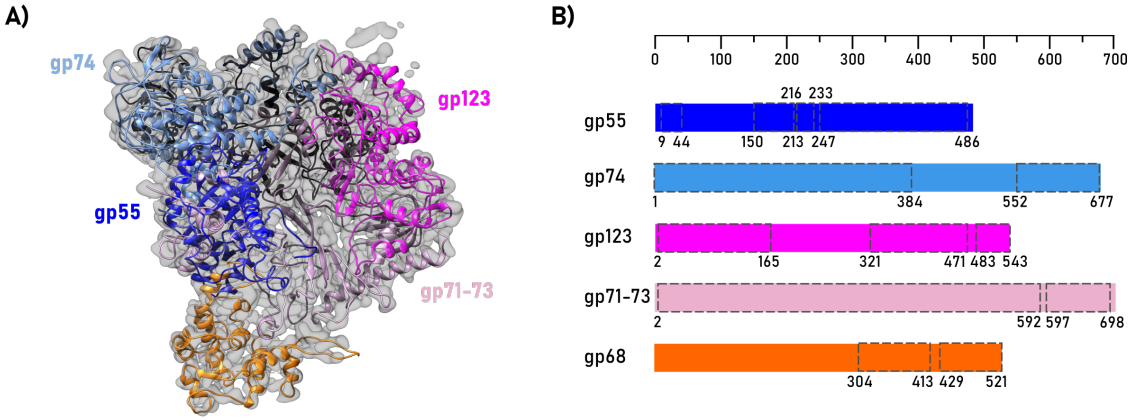


Figure 4.9: Building a molecular model for the ΦKZ nvRNAP. A) 3.3 Å map with the molecular model of the five-subunit ΦKZ nvRNAP depicted in cartoon representation fitted into the density. B) Schematic diagram of the built sections for each ΦKZ nvRNAP subunit.

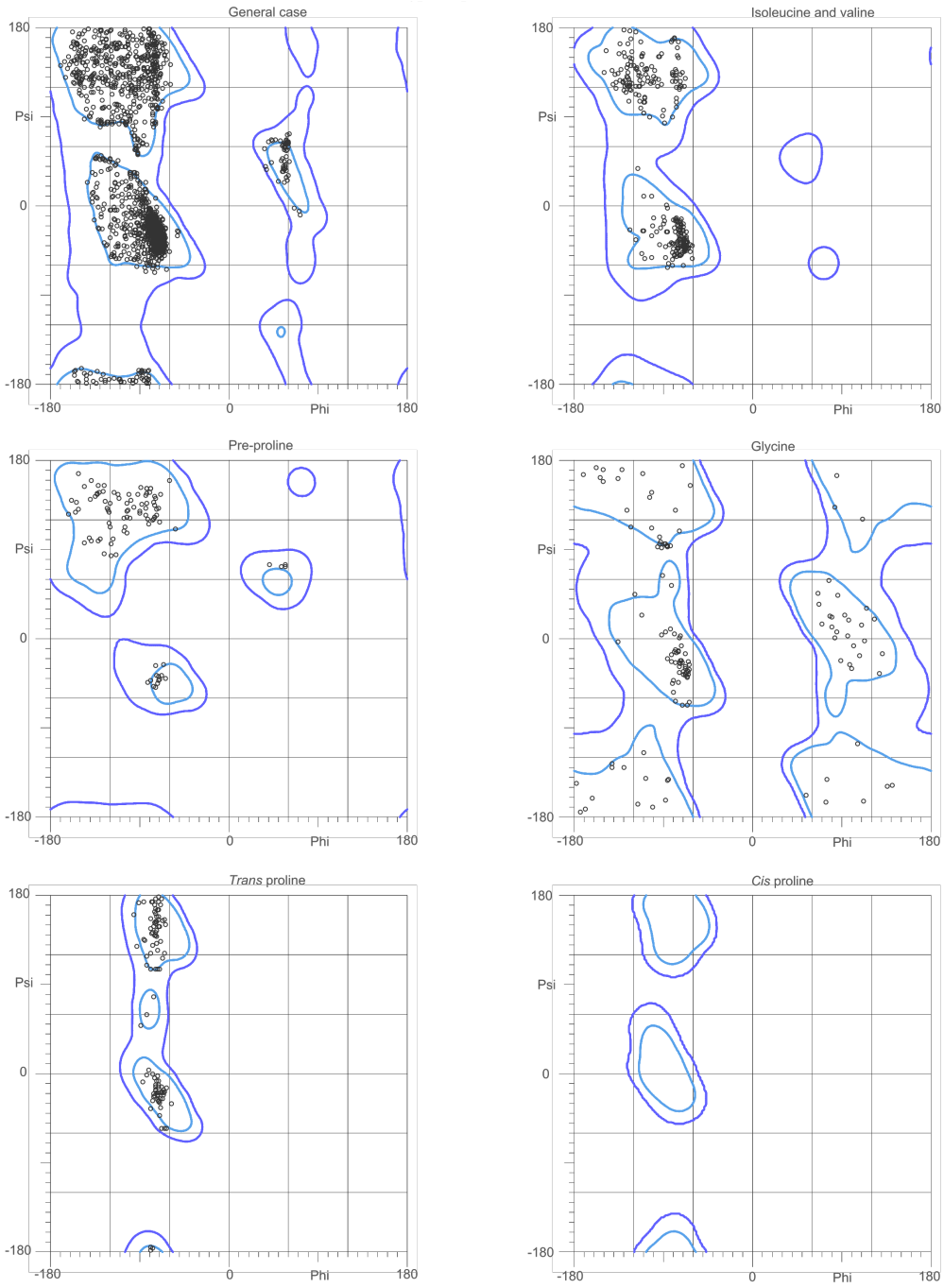


Figure 4.10: Ramachandran plot of the ΦKZ nvRNAP structure including the clamp. Ramachandran plots were computed with Molprobity (Lovell et al., 2003).

4.4 Architecture of the ΦKZ nvRNA polymerase

The five-subunit ΦKZ nvRNAP complex preserved the β/β' architecture of eubacterial msRNAPs (Sutherland and Murakami, 2018; Zhang et al., 1999). One side of the enzyme was formed by the β'-like subunits, gp55 and gp74, whereas the β-like subunits, gp71-73 and gp123, were located on the opposite side, with the internal DNA channel running between them (Figure 4.11). The most peripheral regions of the complex that would encircle the channel at the top of the enzyme were weakly resolved in our 3D reconstruction. However, through comparison to other eubacterial msRNAP structures, we could confidently assign them to the middle domains of gp74 and gp123 (Figure 4.12). In bacterial msRNAPs the location of these two domains corresponds to one of the pincers that interacts with the downstream dsDNA, which would explain their flexibility in the absence of nucleic acids. Sequence analysis of these domains did not identify any of the universal conserved regions of the β and β' subunits (Lane and Darst, 2010b; Yakunina et al., 2015), which suggested that they had probably evolved from lineage-specific insertions of different bacteria species.

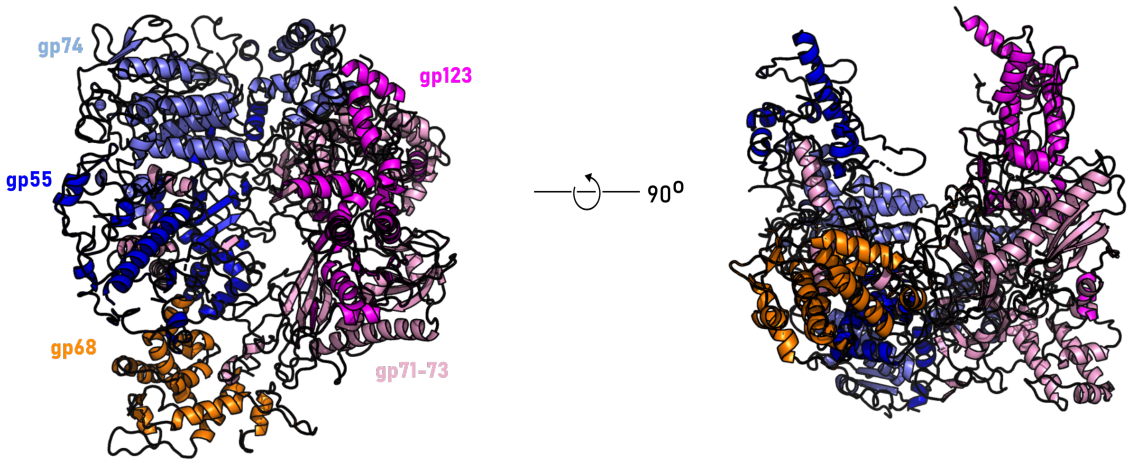


Figure 4.11: Architecture of the ΦKZ nvRNAP. Two different views of the ΦKZ nvRNAP structure represented in cartoon. Each subunit is coloured as indicated in the figure. Figure adapted from de Martín Garrido et al. (2021b) under a Creative Commons Attribution License.

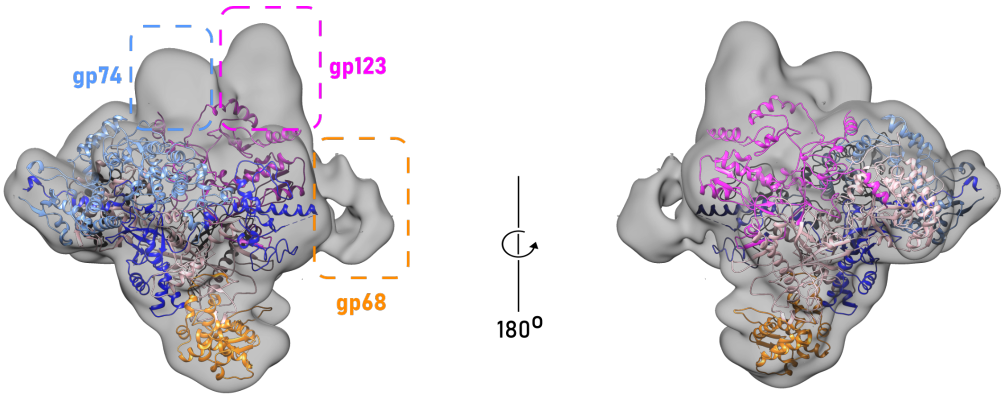


Figure 4.12: Low resolution features of the ΦKZ nvRNAP reconstruction. Three-dimensional map with a Gaussian filter applied coloured in light grey, with fitted nvRNAP structure represented as a cartoon. Extra density is highlighted in dotted-line rectangles coloured according to the subunit they are expected to correspond to. Figure adapted from de Martín Garrido et al. (2021b) under a Creative Commons Attribution License.

The structure revealed that the fifth subunit of the ΦKZ nvRNAP, gp68, was integrated into the complex making extensive contacts with gp71-73 (1 848 Å^2 of interface buried surface area) and, to a lesser extent, with gp55 (446.10 Å^2 of interface buried surface area) (Figure 4.11, Table 4.1). We found that the interaction within them was mostly governed by hydrogen bonds and charge-charge interactions (Figure 4.13), with 4% and 15.5% of the solvent-accessible area of gp68 contributing to the interaction with gp55 and gp71-73, respectively (Table 4.1). Our collaborators (Dr.Yakunina's group) also performed co-purification experiments in which they co-expressed pairs of subunits with one affinity-tag and evaluated recovery of both proteins. Consistent with our structure, their results showed that gp68 co-purified with both gp71-73 and gp55 de Martín Garrido et al. (2021b). The N-terminus of gp68 was not visible in our model. However, we could observe some weakly-resolved density at the front of the enzyme encircling the channel that we hypothesized might correspond to the N-terminal domain of the fifth subunit, gp68 (Figure 4.12).

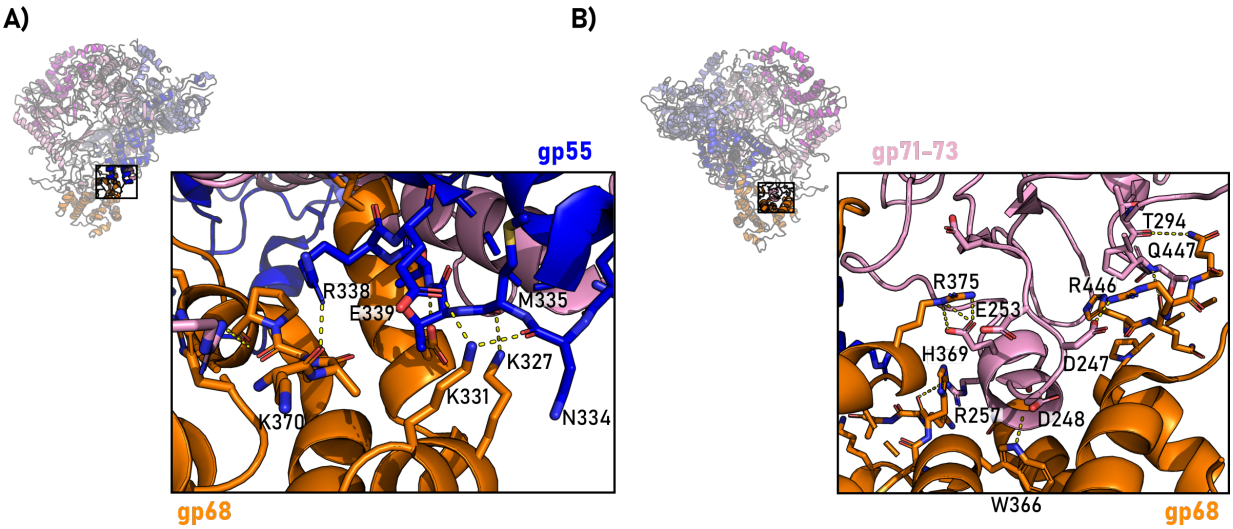


Figure 4.13: Integration of gp68 in the ΦKZ nvRNAP complex. A) Inset displaying a region of the gp68-gp55 interface. B) Inset displaying a region of the gp71-73-gp68 interaction. In both panels, hydrogen bonds are depicted as a dotted yellow line and residues involved in the interaction are labelled accordingly. Figure adapted from de Martín Garrido et al. (2021b) under a Creative Commons Attribution License.

	Gp68 interactions		
Interacting partner	**Total solvent-accessible surface area (Å²)**	**Interface buried surface area (Å²)**	**% of buried-surface area**
Gp71-73	11 928.40	1 848.00	15.5
Gp55	11 928.40	463.10	3.9

Table 4.1: Buried surface area of the gp68 interaction with gp55 and gp71-73. Values extracted from PDBePISA server (https://www.ebi.ac.uk/pdbe/pisa/). Table adapted from de Martín Garrido et al. (2021b) under a Creative Commons Attribution License.

4.5 Structural comparison with bacterial multisubunit RNAPs

The ΦKZ nvRNAP is a pared down version of bacterial msRNAPs. The five-subunit complex with split β and β' -like subunits proved to be sufficient to initiate transcription without accessory subunits α and ω, which are present in bacterial, archaeal and eukaryotic msRNAPs (Yakunina et al., 2015). Accordingly, the overall appearance of the nvRNAP was smaller than bacterial msRNAPs (Figure 4.14). Comparison of each ΦKZ nvRNAP β' -like subunit with structures available in the PDB (Holm and Sander, 1995) showed that the closest structures to each subunit corresponded to different bacterial species (Table 4.2).

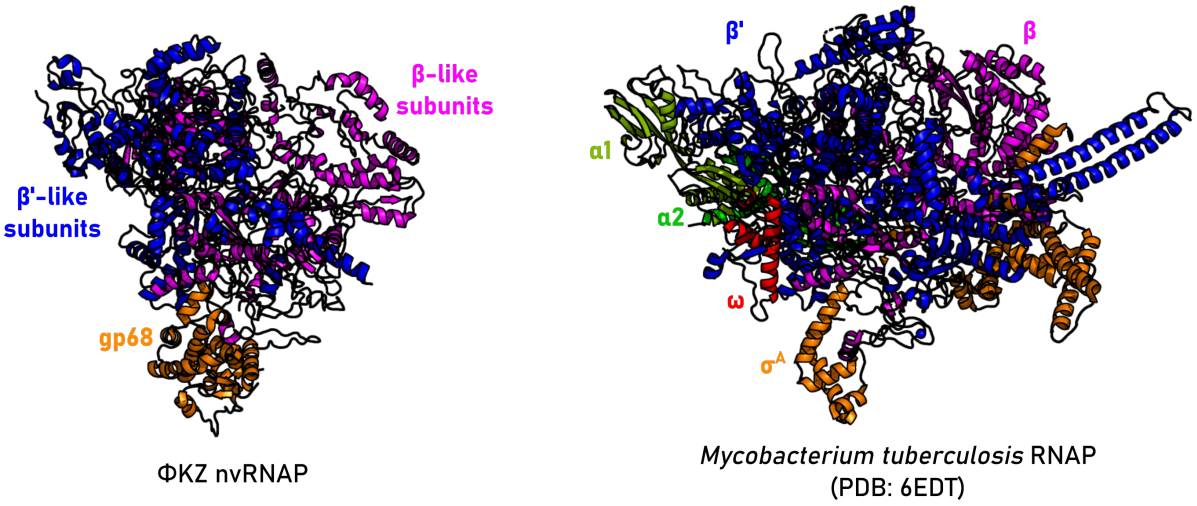


Figure 4.14: Structural comparison of the ΦKZ nvRNAP with bacterial msRNAPs. A) Structure of the ΦKZ nvRNAP depicted in cartoon representation with the β-like subunits coloured in magenta, the β'-like subunits coloured in blue, and gp68 in orange. B) *M.tuberculosis* RNAP structure (PDB 6EDT). Subunits are coloured as indicated in the figure.

	Bacterial RNAP structures		
ΦKZ nvRNAP subunit	PDB code and chain	Species	RMSD
gp55	3DXJ-D	*Thermus thermophilus*	3.6
	6DCF-D	*Mycobacterium smegmatis*	3.7
gp74	3DXJ-D	*Thermus thermophilus*	3.9
	5BYH-D	*Escherichia coli*	4.1
gp123	5UAC-C	*Escherichia coli*	4.0
	6JCY-C	*Mycobacterium tuberculosis*	3.9
gp71-73	5UAC-C	*Escherichia coli*	5.8
	6JCY-C	*Mycobacterium tuberculosis*	5.1

Table 4.2: Structural similarity of the ΦKZ nvRNAP subunits and bacterial msRNAPs structures. Table showing the top DALI restults for each nvRNAP subunit. The RMSD was computed by DALI (Holm and Sander, 1995).

In line with with the observations made by previous sequence analysis (Yakunina et al., 2015), we found that the key active structural elements and the catalytic residues of the ΦKZ nvRNAP were well-conserved. At the catalytic cleft, the two DPBBs were contributed by gp71-73, containing two conserved lysines (K386 and K394) believed to interact with the RNA transcript, and by gp55, which included the universally conserved DxDGD (^{417}DFDGD^{421}) near its C-terminus, responsible for the coordination of a catalytic Mg^{2+} conserved in all nucleotidyltransferases (Zhang et al., 1999; Iyer et al., 2003; Yokoyama et al., 2002).

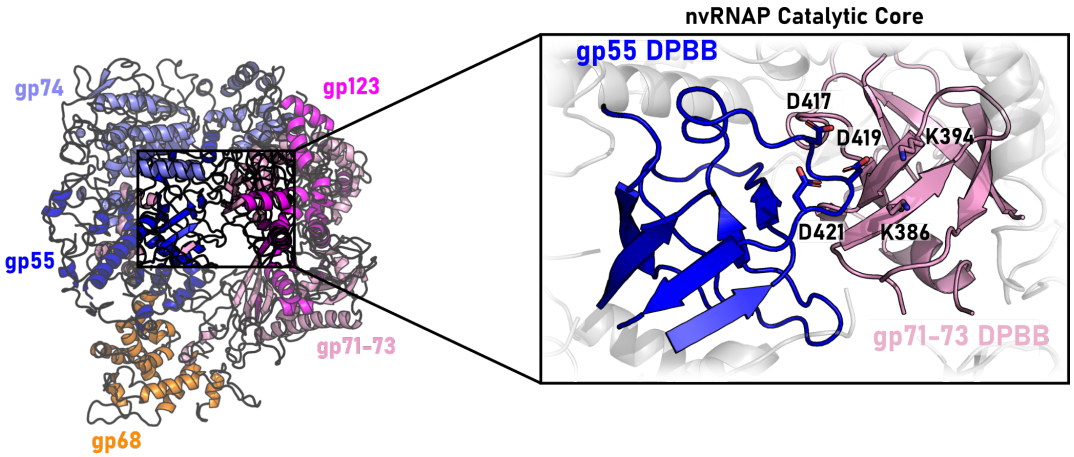


Figure 4.15: The catalytic core of the ΦKZ nvRNAP. Cartoon representation of the ΦKZ nvRNAP structure with each subunit coloured following the same scheme as in previous figures. Inset of the catalytic core showing the two DPBBs in gp55 and gp71-73. Side chains of universally conserved residues are represented as sticks and indicated with labels.

The structural elements that govern the transcription reaction were also conserved in the ΦKZ nvRNAP structure. The bridge helix and trigger helix/loop, which switch between alternate conformations during nucleotide addition (Yokoyama et al., 2002; Wang et al., 2006), were found in the β-like subunit gp74 (bridge helix: 243-279, trigger helix/loop: 366-393). Similarly to the situation in bacterial msRNAPs, the bridge helix crossed from the β' side of the enzyme to the β side, however, we found that it was partially disordered at the central region, losing the typical helical conformation. As for the the trigger loop, only the beginning of the helix was clearly visible in our reconstruction (366-384), probably due to the absence of nucleic acids (Figure 4.16). In gp74, we also found one of the Zn^{2+} binding domains (C288, C341, C350, C353) observed in bacterial msRNAPs (Figure 4.16) (Yokoyama et al., 2002).

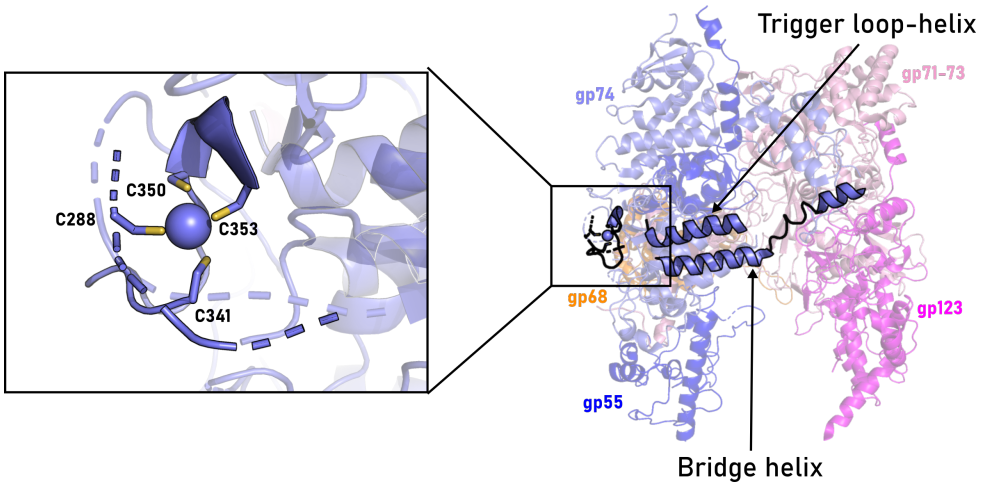


Figure 4.16: Architectural elements of the ΦKZ nvRNAP gp74 subunit. nvRNAP structure with bridge helix and trigger loop-helix indicated in arrows. Color scheme is the same as in previous figures. On the left, inset showing the Zn^{2+} binding domain with each Cys labelled accordingly. Figure adapted from (de Martín Garrido et al., 2021b) under a Creative Commons Attribution License.

The clamp, a flexible element whose position is key to regulate entry of single-stranded DNA to the active site, was also conserved in the ΦKZ nvRNAP, and was located within the C-terminus of gp71-73 (643-692) and the N-terminus of gp55 (9-286) (Figure 4.17). We could partially observe the lid (211-223) and the two clamp helices (225-273) (Figure 4.7), only in the subset of particles that yielded the most completed structure (3.3 Å map). However, other parts of the clamp were still weakly resolved in our reconstruction, probably due to the ability of this element to adopt different conformations throughout the reaction cycle (Chakraborty et al., 2012).

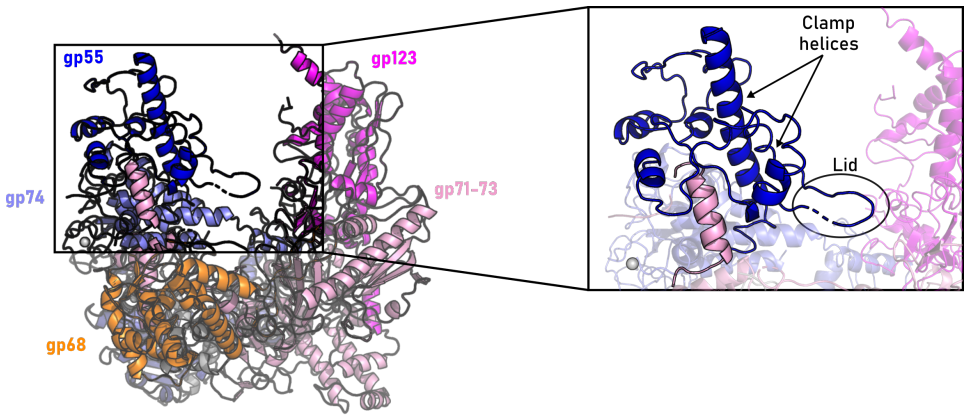


Figure 4.17: Architecture of the ΦKZ nvRNAP clamp. nvRNAP structure with clamp highlighted. Color scheme is the same as in previous figures. On the right, inset showing the clamp comprised within gp55 and the C-terminus of gp71-73. The clamp helices are pointed with arrows and the lid is indicated by a circle. Figure adapted from (de Martín Garrido et al., 2021b) under a Creative Commons Attribution License.

The flap was also visible in our structure. The core of the flap (198-206 and 368-375) retained the typical β-sandwich configuration, however the flap-tip-helix was not conserved in our structure. Instead, gp71-73 contained a small loop with a single helix turn (249-254) that interacted with gp68 (Figure 4.17). The flap-tip-helix is not conserved in archaeal or eukaryotic msRNAPs because it plays bacterial-specific roles during different stages of transcription (Lane and Darst, 2010a), as well as regulatory functions, which would explain why this element became obsolete in the ΦKZ nvRNAP.

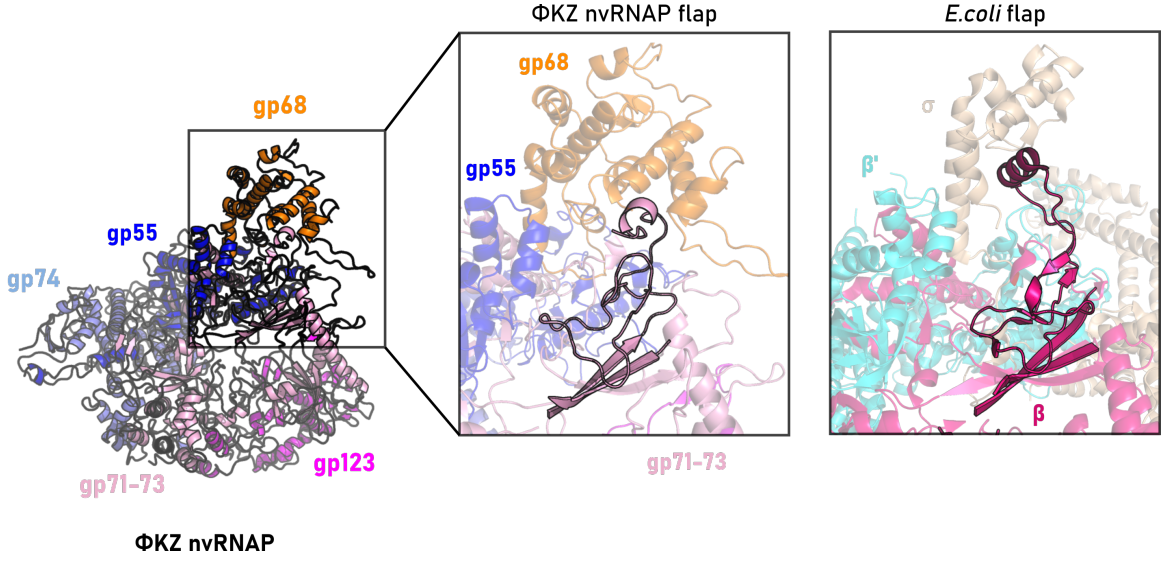


Figure 4.18: Comparison between the ΦKZ nvRNAP and the bacterial flap. The ΦKZ nvRNAP structure with the flap highlighted with a square. Color scheme is the same as in previous figures. On the right, inset showing the flap found in gp71-73 and the *E.coli* flap (PDB 5UAC). Figure adapted from (de Martín Garrido et al., 2021b) under a Creative Commons Attribution License.

4.6 Gp68 is involved in the assembly of a functional nvRNAP

The five-subunit ΦKZ nvRNAP is sufficient to initiate transcription from late phage promoters, without needing accessory subunits α and ω to assemble a catalytically active enzyme (Yakunina et al., 2015; Orekhova et al., 2019). We previously observed that the four β/β'-like subunits contained conserved structural elements and catalytic residues from bacterial subunits, we therefore wanted to further investigate if gp68 was involved in complex assembly. For this purpose, our collaborators (Dr.Yakunina's group) expressed and purified gp68 independently in a single plasmid instead of coexpressing it with the rest of subunits (de Martín Garrido et al., 2021b). Using the independently purified gp68, they assembled the full nvRNAP complex *in vitro* by adding gp68 to 4s, which included the four β/β'-like subunits. The *in vitro*-assembled complexes contained all five subunits forming the ΦKZ nvRNAP and bound DNA templates containing the specific promoter consensus sequence, similar to 5s-complexes assembled *in vivo* (Figure 4.19 A). However, the *in vitro* assembled complexes were unable to initiate transcription from a specific phage late promoter (Figure 4.19 B). These results suggested that gp68 probably needs to be included in the complex at early stages of complex assembly to form a catallytically active polymerase.

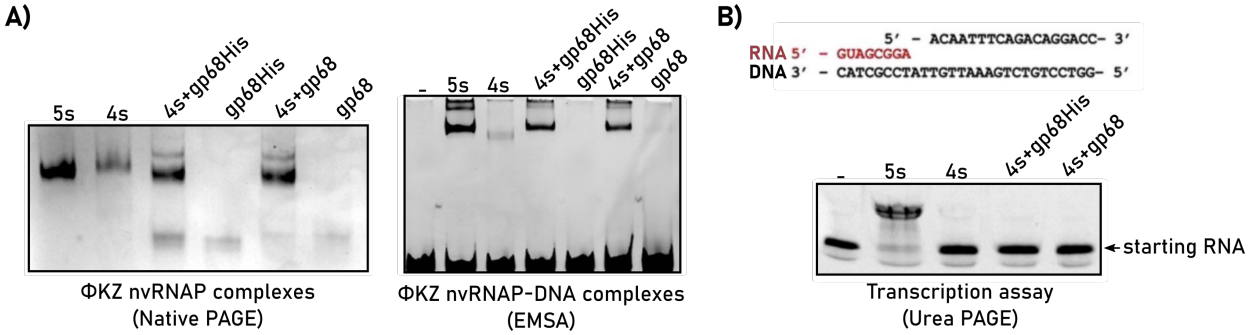


Figure 4.19: Gp68 is necessary for the assembly of an active nvRNAP. A) Native PAGE of 4, and 5 subunit complexes prepared through the addition of gp68 at different points, and EMSA with a 212 bp specific dsDNA-template containing the KZP119L promoter sequence (long DNA, from -64 to +148, (Table 2.9), and assemblies of recombinant nvRNAP complexes and independently purified gp68-His and gp68 samples. B) Urea-PAGE of the transcription assay from the RNA-DNA hybrid shown above. Activity is indicated by decreased mobility (lane 5s). Data provided by Dr.Yakunina's group. Figure adapted from (de Martín Garrido et al., 2021b) under a Creative Commons Attribution License.

In eubateria, the dimeric α subunits serve as an organising platform for the assembly of the catalytically active core. Both β and β' interact with the α-dimer, with universally conserved regions βa14 and βa15 being essential (Lane and Darst, 2010b) for this interaction. Sequence analysis showed that these regions are well-conserved in the ΦKZ nvRNAP subunit gp71-73 (Yakunina et al., 2015). Consistently, we observed that the structure of βa14 (368-442) and βa15 (547-656) was also preserved in the nvRNAP, except for a small insertion loop in gp71-73 (551-556) close to the α binding interface (Figure 4.20). Furthermore, we observed that gp68 interacted with the bottom-helix (643-656) of the universally conserved region βa15 of gp71-73 (Figure 4.20), hence supporting a role of gp68 in the assembly of a catallytically active centre, as previously suggested by biochemical assays (Figure 4.19). This interaction, however, was in a different region than the bacterial β-α interface.

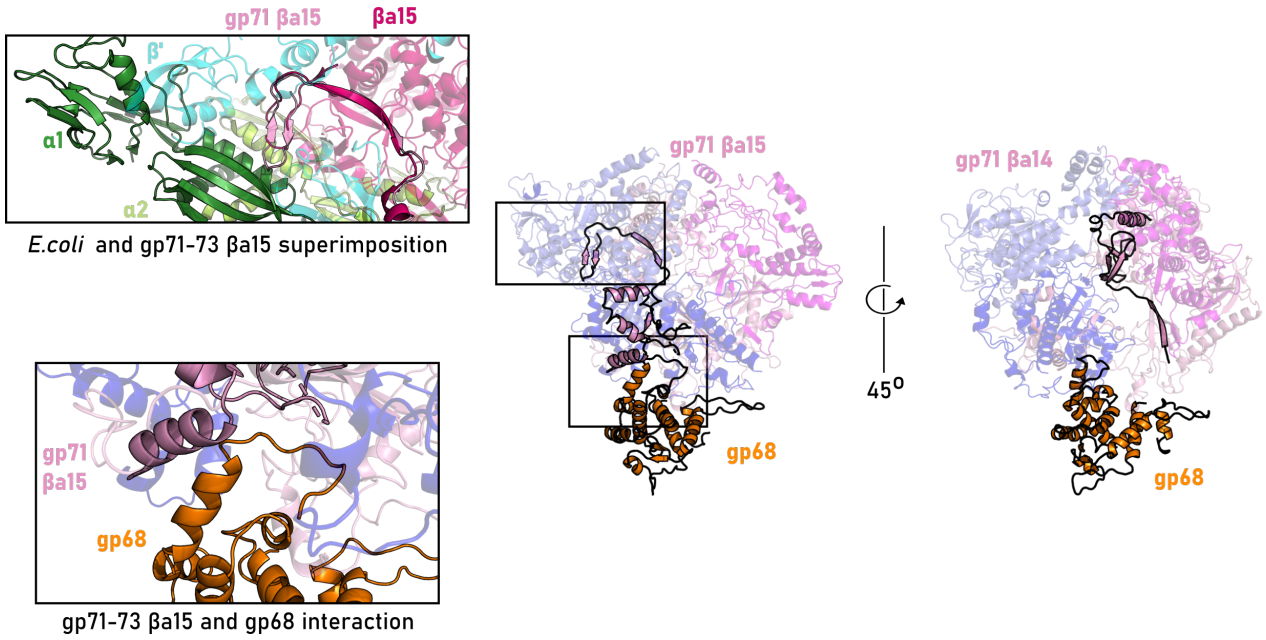


Figure 4.20: Interaction of gp68 with the universally conserved motif βa15. On the right, two different views of the nvRNAP with gp68 in orange and regions βa14 and βa15 of gp71-73 in pink. (Top-left) Inset showing the *E.coli* βa15 interaction with one of the α subunits. gp71-73 βa15 (pink) is superimposed on the *E.Coli* structure (legend in the figure) to show that it has a longer loop including residues 551-556. (Bottom-left) Inset showing the interaction of gp68 with region βa15 of gp71-73. Figure adapted from (de Martín Garrido et al., 2021b) under a Creative Commons Attribution License.

4.7 Gp68 is structurally and functionally similar to bacterial σ factors

Considering that the five-subunit polymerase can initiate transcription from specific promoters without requiring any extra subunits (Yakunina et al., 2015; Orekhova et al., 2019), we wanted to investigate the potential involvement of the the fifth subunit, gp68, in DNA binding. To identify which subunits were involved in the interaction with DNA, our collaborators in Dr. Yakunina's group (Peter the Great St. Petersburg Polytechnic University St. Petersburg, Russia) performed an EMSA combined with photo-crosslinking (de Martín Garrido et al., 2021b). Mass spectrometry identified both gp123 and gp68 within the band corresponding to the fully denatured complex containing the DNA binding subunits (Figure 4.21 A), which confirmed our hypothesis that gp68 was involved in DNA binding. After that, they performed limited proteolysis with two different DNA templates: a fork junction short DNA (from -12 to +15 nucleotide on non-template strand), and a long DNA (from –64 to +148 bp)(Table 2.9). In the absence of DNA, an N-terminal fragment of gp68 was released (1-382) after trypsin treatment. When the short DNA was present, there was a weak protective effect of the full gp68 and in the presence of long DNA such protective effect against gp68 proteolysis (Figure 4.21 B) was even stronger (de Martín Garrido et al., 2021b). The protection of gp68 in the presence of DNA confirmed the involvement of gp68 in DNA binding. Furthermore, if the DNA was long enough, the binding resulted in a more compact complex with the full-length of gp68 being involved in DNA binding.

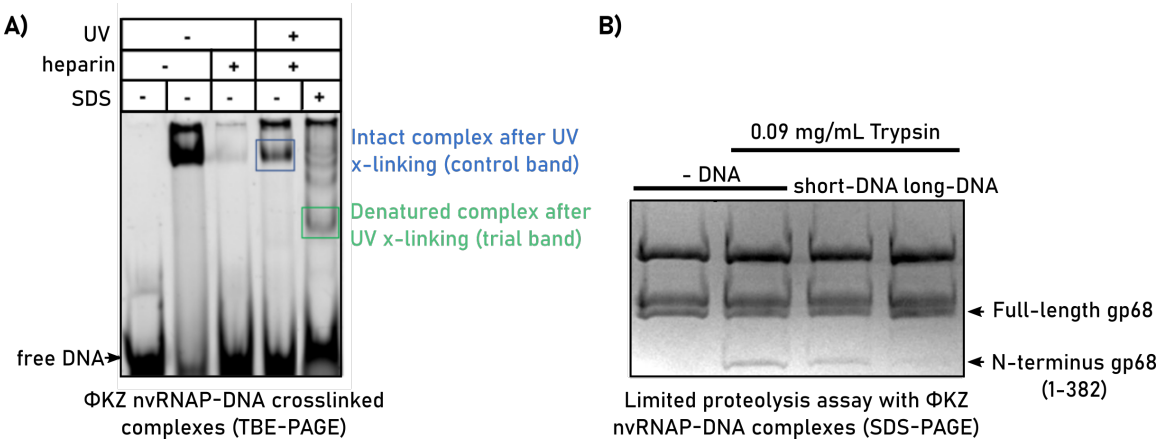


Figure 4.21: Gp68 of the ΦKZ nvRNAP is involved in DNA binding. A) TBE-PAGE of UV cross-linked DNA bound complexes. Bands extracted for subsequent mass-spectrometry analysis are highlighted in squared boxes. B) Protection of gp68 from limited proteolysis by DNA templates. Both short and long templates are based on the KZP119L promoter sequence (Table 2.9); the short template extends from -12 to +15 nucleotides from transcription start (marked +1), whereas the long template indicates a full 212 bp PCR-fragment. The full gp68 and the N-terminal fragment of gp68 extending to 382 residues are marked by arrows. Data provided by Dr.Yakuninas group (at Peter the Great St. Petersburg Polytechnic University (St. Petersburg, Russia) Figure adapted from (de Martín Garrido et al., 2021b) under a Creative Commons Attribution License.

Superimposition of the nvRNAP structure with that of bacterial msRNAPs, using *M.tuberculosis* RNAP as an example, showed that the C-terminus of gp68 was at the same location within the complex as in bacterial σ factors, covering region 4 (Figure 4.22). After this observation, we performed a comparison of the gp68 structure with other structures available in the PDB using the DALI server (Holm and Sander, 1995). The comparison showed that the C-terminus of gp68 which is in the same position as the σ region 4, had a weakly similar fold to a more N-terminal part of other bacterial σ factors (region 2), making it difficult to interpret (C_α RMSD 3.964 when residues 36-91 of σ^C were superimposed on residues 324-356 and 379-401 of gp68, as suggested by DALI (Holm and Sander, 1995))(Figure 4.23). The closest similar structure to the C-terminus of gp68 was a fragment of the *M.tuberculosis* σ^C, a member of the *extracytoplasmic function* (ECF) σ factors (Ebastien Rodrigue et al., 2006).

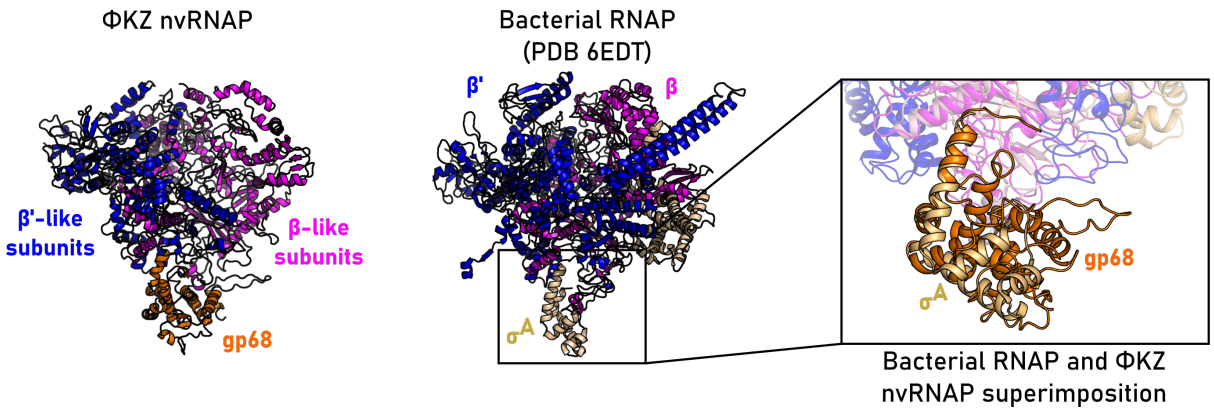


Figure 4.22: Gp68 is positioned identically to eubacterial σ-factor On the left, comparisons of the *M. tuberculosis* RNAP (PDB 6EDT) and ΦKZ nvRNAP and, on the right, superimposition indicating colocation of gp68 and bacterial σA factor. Figure adapted from (de Martín Garrido et al., 2021b) under a Creative Commons Attribution License.

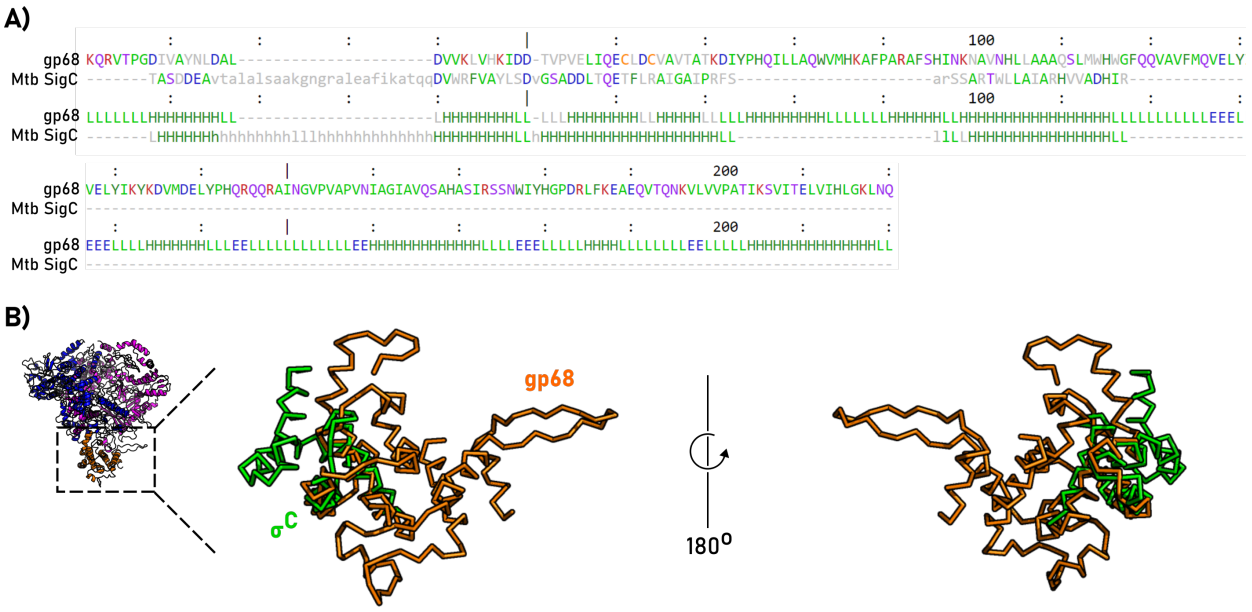


Figure 4.23: Gp68 is structurally similar to eubacterial σ-factor. A) Sequence alignment based on structural similarity made with DALI. The top part shows the amino acid sequences of gp68 and Mtb σ ^{c}. Below, secondary assignments are represented with H (helix), E (strand) and L (coil). The most frequent amino acid type is coloured in each column B) Superimposition of Mtb σ^{c} (PDB: 2o7g) on gp68 depicted in ribbon representation. Figure adapted from (de Martín Garrido et al., 2021b) under a Creative Commons Attribution License.

Comparing the gp68 structure with that region of σ that was in the same location (region 4), we noticed that the gp68 C-terminus was considerably more expanded, making a larger interaction with the rest of the complex (2 311.10 versus 1 495.40 Å^2), and therefore anchoring gp68 steadily in the complex (Table 4.3). Moreover, we noted that purification of the ΦKZ nvRNAP from *P.aeruginosa*-infected cells failed to obtain a complex without gp68, and the use of several chromatographic steps, including buffers with high NaCl concentration (up to 600

mM) did not disrupt the complex (Yakunina et al., 2015; Orekhova et al., 2019). Observing the electrostatics of the gp68 region visible in our model (C-terminus), we could not observe any region with a positively charged pocket that could potentially interact with DNA (Figure 4.24), apart from some patches located at the peripheral region.

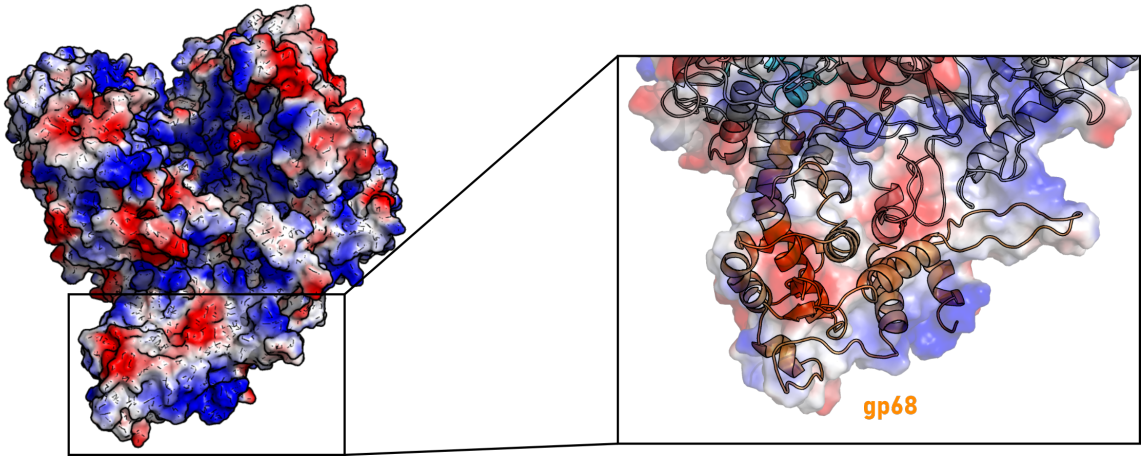


Figure 4.24: Electrostatic potential surface of the ΦKZ nvRNAP. Electrostatic potential surface of the ΦKZ nvRNAP molecular model with blue regions representing positive charges, red representing negative and white neutral. Inset, cartoon representation of gp68 and electrostatic potential surface of the same region. Figure adapted from (de Martín Garrido et al., 2021b) under a Creative Commons Attribution License.

			Gp68 or σ interactions	
	Total solvent-accessible surface area (Å²)	Total buried-surface area (Å²)	Interacting partner	Interface buried surface area (Å²)
nvRNAP	228 903.00	18 284.90	gp71-73	1 848.00
			gp55	463.10
			Total	**2 311.10**
***M.tuberculosis* (6EDT)**	128 108.00	9 775.30	β	1 130.90
			β'	364.50
			Total	**1 495.40**

Table 4.3: Comparison of the buried surface area of nvRNAP gp68 and σ factor A from *M. tuberculosis*. Values extracted from PDBePISA server (https://www.ebi.ac.uk/pdbe/pisa/). Table adapted from (de Martín Garrido et al., 2021b) under a Creative Commons Attribution License.

4.8 Discussion

Bacteriophage ΦKZ is one of a few exceptional bacterial viruses that encode a pair of non-canonical multi-subunit msRNAPs that are able to transcribe all genes needed to complete the bacteriophage life-cycle (Ceyssens et al., 2014). In 2015, purification of the ΦKZ nvRNAP revealed that it is formed by four subunits homologous to the two largest bacterial subunits and a fifth one shared amongst ΦKZ-like phages but with no homology to any other protein with known functionality that is essential for promoter recognition and transcription initiation (Yakunina et al., 2015; Orekhova et al., 2019). The low sequence similarity of the fifth subunit and its crucial role for the enzyme function suggested that the transcriptional mechanism was likely to be novel and posed new questions about the evolution of these non-canonical msRNAPs.

In this chapter we have determined for the first time the three-dimensional structure of the ΦKZ nvRNAP, revealing that this enzyme preserves the conserved β/β'-architecture of eubacterial RNAPs, but includes a fifth subunit, gp68, embedded in the complex. Structural comparison with bacterial msRNAPs showed that most of the structural elements involved in the reaction cycle, as well as the catalytic centre are conserved, suggesting that the mechanisms of polymerisation and catalysis are likely to be similar to those in bacterial msRNAPs. ΦKZ, however, has developed a minimal but sophisticated polymerase with split subunits that does not rely on the canonical subunits α and ω for the assembly of a catalytically active core. The structure also revealed that the fifth subunit, gp68, is embedded in the complex making extensive contacts with gp71-73 (homologous to the C-terminus of the β subunit), and to a lesser extent to gp55 (homologous to the N-terminus of the β' subunit), identifying for the first time the position of this subunit within the complex and confirming that it is an integrated member of the ΦKZ nvRNAP.

Comparison with bacterial msRNAPs showed that gp68 is positioned identically to, and exhibits structural homology to eubacterial σ factors. Moreover, biochemical assays suggested that gp68 is involved in DNA binding, which is congruent with a σ-like activity. Any of these observations

made on their own would not be sufficient to draw any conclusion, however, if we observe all of them in combination we can suggest that gp68 has probably evolved from an ancestral σ factor but has diverged considerably. The gp68 region visible in our structure overlaps with the location of region 4 of σ^{70}-like factors of group 1 (Paget, 2015) and, conversely, the same region has a weakly similar fold to region 2 of another bacterial σ factor belonging to the ECF group, making it difficult to interpret and indicating that the structural homology is low. The ECF group comprises the most abundant and heterogeneous σ factors, which are important for bacterial stress adaptation and are associated with virulence. Hence, similarity of the C-terminus of gp68 with a member of the most diverse σ group supports our hypothesis of gp68 being highly divergent from bacterial σ factors, yet conserving their function. Region 4 of bacterial σ factors interacts with the -35 element of bacterial promoters (Paget, 2015). ΦKZ nvRNAP specifically recognizes late promoters with a consensus sequence 5'-TATG-3' overlapping with the transcription initiation site, and there is no conserved sequence in the -35 region of these promoters. Likewise, the assembly and stabilisation of the active site must occur through a different mechanism, and the fifth subunit gp68 probably plays an important role in both processes. Given the extreme sequence and structural divergence between gp68 and bacterial σ factors, and the fact that gp68 is integrated in the complex, promoter recognition and transcription initiation are also expected to happen through a different mechanism.

The structure of the ΦKZ nvRNAP shows that the fifth subunit, gp68, is embedded in the complex and single-particle cryo-EM analysis failed to identify any nvRNAP class completely missing the gp68 subunit after sorting from over 3 million particles. Purification of native nvRNAP from ΦKZ-infected *P.aeruginosa* cells showed that the five subunits were present in equimolar amounts, which would be unexpected if gp68 could be separated from the complex (Yakunina et al., 2015). Furthermore, during purification of both native nvRNAP (Yakunina et al., 2015) and recombinant 5s nvRNAP (Orekhova et al., 2019) several chromatographic steps were used, which included use of high-salt buffers (up to 600 mM NaCl) which did not disrupt the 5s nvRNAP complex. Comparison of the C-terminal fragment of gp68 with the C-terminal fragment of bacterial σ factors (region 4), showed that gp68 was substantially expanded, establishing a larger interaction with the rest of the complex. Furthermore, we could

not observe any putative DNA-binding pocket in the C-terminal region of gp68 modelled in our structure. We therefore suggest that the C-terminus of gp68 may have a mainly integrative and stabilising role, whilst the N-terminus, flexible in our structure, would be more intimately involved in DNA binding. All these observations support a conclusion that gp68 is steadily embedded in the ΦKZ nvRNAP complex. Integration of the σ-like subunit in the ΦKZ nvRNAP complex suggests that the exchange of σ factor for alternate promoter selection probably became unnecessary for the bacteriophage life cycle. Consistently, the gp71-73 β-like subunit of ΦKZ has evolved to reduce the flap-tip helix to a small loop with a single helix turn, which makes extensive interactions with gp68. The flap-tip-helix is associated with the proper positioning of the σ to recognise the -35 promoter element (Lane and Darst, 2010a). In the nvRNAP structure, the orientation of the σ factor probably became redundant, as the putative σ-like subunit, gp68, is already held in place by being embedded in the polymerase complex.

The ΦKZ nvRNAP is structurally similar to several bacterial msRNAPs available in the PDB, however different RNAPs seem closer depending on the metric used and the subunit that is being compared. Moreover, the β/β'-like subunits of the ΦKZ nvRNAP have been split into halves. These observations are an example of the great evolutionary distance between the ΦKZ nvRNAP and canonical msRNAPs. We speculate that the initial transfer of the rpoB and rpoC bacterial genes (encoding β and β' respectively) would suppose an evolutionary advantage for ΦKZ, and that the β/β'-like split subunits might have appeared due to intron mobility events resulting involving genome rearrangements. Possessing its own transcriptional system with split subunits would suppose increased tolerance of host variation, which will ultimately increase the bacteriophage fitness. Furthermore, we would speculate that the split subunits might have evolved to meet the import requirements to the ΦKZ "pseudo-nucleus" (Chaikeeratisak et al., 2017a,b). Although the rules that govern protein internalisation into the proteinacious "pseudo-nucleus" are not fully understood, it has been demonstrated that smaller proteins are more easily imported (Mendoza et al., 2020).

The five-subunit ΦKZ nvRNAP is catalytically active on its own and does not require accessory subunits α and ω. Purification of the ΦKZ nvRNAP from ΦKZ infected cells resulted in a five-subunit complex and purification of the ΦKZ nvRNAP with proteins from the bacterial host

has been previously unsuccessful (Ceyssens et al., 2014; Yakunina et al., 2015), which would exclude the possibility of the ΦKZ nvRNAP requiring host-encoded α or ω subunits. Genes encoding these two subunits (rpoA and rpoZ) are located in a different location in the host genomic DNA than that of genes encoding the two largest subunits β and β', rpoB and rpoC respectively, which are located in the same operon (Barry et al., 1979). Considering that the ultimate reason for gene transfer would be evolutionary fitness, the lack of α and ω subunits in the ΦKZ genome implies that they are not required to produce a functional msRNAP and their transfer would have implied extra DNA to be replicated by bacteriophage, which represents a fitness disadvantage because the life cycle becomes longer.

Our *de novo* structural model for the C-terminus of gp68 is in the same position within the complex as in bacterial σ factors and matches σ structures in DALI (Holm and Sander, 1995). Moreover, biochemical studies have shown that this subunit is involved in DNA binding and is essential for promoter recognition and transcription initiation (Orekhova et al., 2019). Therefore, we have evidence that gp68 is probably a distant homologe of bacterial σ factors, but that it has diverged to a greater degree in comparison to other four β/β'-like subunits to be permanently anchored to the nvRNAP complex. We would speculate that the acquisition of a σ factor by the bacteriophage would have happened before the operon containing rpoB and rpoC genes, providing an evolutionary advantage that led to its retention, while rpoB and rpoC genes would have been acquired later by the bacteriophage to favour the compatibility of the existing phage σ-like subunit with bacteriophage promoters, which would explain the greater divergence of gp68 in comparison to the other nvRNAP subunits.

Apart from bacteriophage ΦKZ there are a few other exceptions that encode msRNAPs, which is the case of *Bacillus subtilis* bacteriophage PBS2 (Clark et al., 1974; Clark, 1978) and AR9 (Lavysh et al., 2016; Sokolova et al., 2017). Due to the lack of information about its genome sequence, little is known about PBS2. However, it has been found that *Bacillus subtilis* bacteriophage AR9 also encodes four proteins homologous to eubacterial β and β' subunits, and a fifth subunit, gp226, crucial for promoter recognition, suggesting a similar mechanism to ΦKZ nvRNAP. The AR9 nvRNAP complex, however, remains catalytically active in the absence of the fifth subunit (Sokolova et al., 2017), indicating reduced dependence on this subunit to

assemble a catalytically active enzyme, in comparison to what we report for the ΦKZ nvRNAP.

Recently, a structure of the AR9 nvRNAP bound to its promoter was published, identifying the fifth subunit, gp226, as a promoter-specificity subunit (Fraser et al., 2021). Given the great similarity between the AR9 nvRNAP polymerase structure and the ΦKZ nvRNAP structure reported in this chapter, we surmise that the location of the gp68 N-terminus is probably similar within the rest of the complex and that gp68 is likely to perform a similar function as the AR9 promoter-recognition subunit, consistent with our hypothesis that gp68 has probably evolved from an ancestral σ factor. The consensus sequences of specific promoters recognised by these two non-canonical polymerases are completely different. While the AR9 nvRNAP recognises a -11 element in the template strand containing uracyls, the ΦKZ nvRNAP recognises a conserved sequence overlapping with the transcription initiation site, which suggests that promoter recognition and bubble formation could occur by divergent mechanisms. In AR9, the requirement for uracils in the promoter consensus sequence would allow the bacteriophage to avoid unnecessary transcription from the host genome, whereas in ΦKZ, the formation of the "pseudo-nucleus" physically separates the viral genome from the host genetic material. Furthermore, despite the fact that AR9 infects a Gram-positive host and ΦKZ infects a Gram-negative one they encode similar transcriptional machinery to be independent of their respective host, implying intriguing evolutionary relationships. Further structural studies of the ΦKZ nvRNAP bound to specific promoters will be required to properly understand how transcription is achieved in these interesting giant bacteriophages.

Chapter 5

Structure of the ΦKZ non-virion RNAP bound to its promoter P119L

In the previous chapter, we described for the first time the architecture of the ΦKZ non-virion RNA polymerase, a non-canonical msRNAP comprising four β/β'-like subunits and a fifth subunit with no homology to any other protein with known function which is essential for promoter recognition and transcription initiation. We also provided structural comparisons to eubacterial msRNAPs and suggested that the fifth subunit, gp68, had probably evolved from a common ancestor with bacterial σ factors. However, the reported structure was lacking important parts of the enzyme, such as the N-terminus of gp68 or the middle domains of gp123 and gp74. In this chapter, we aimed to determine the structure of the ΦKZ non-virion RNA polymerase bound to its specific promoter ΦKZ P119L, to provide more details about the role of the intriguing fifth subunit and the transcription mechanism of this non-canonical nvRNAP.

5.1 Preparation of ΦKZ nvRNAP complexes with the late promoter P119L promoter

The ΦKZ nvRNAP specifically recognizes and initiates transcription from late bacteriophage promoters, which are characterised by a consensus 5'-TATG-3' motif overlapping with the transcription start site in the last guanosine (+1) (Yakunina et al., 2015). Substitutions to any of the positions of the consensus region lead to a severe decrease in transcription, and the downstream sequence has also been reported to be essential to form a functional promoter specifically recognised by the nvRNAP (Yakunina et al., 2015).

To prepare stable nvRNAP-DNA complexes, we designed a DNA template based on the specific ΦKZ late promoter P119L containing a pre-opened transcription bubble to facilitate polymerase binding. The DNA template covered positions -38 to +40 relative to the transcription start site (+1) and included a pre-opened 4-nucleotide artificial bubble created through the introduction of missmatches between -1 and -4 in the non-template strand, just before the transcription starting site (P119L-4nt)(Figure 5.1). Our collaborators tested whether or not the ΦKZ nvRNAP could bind the P119L-4nt DNA template by gel-shift (Figure 5.1). Their results confirmed that the nvRNAP could indeed bind to the newly designed DNA template, we therefore prepared nvRNAP-DNA samples with the P119L-4nt template to perform single-particle analysis using cryo-EM.

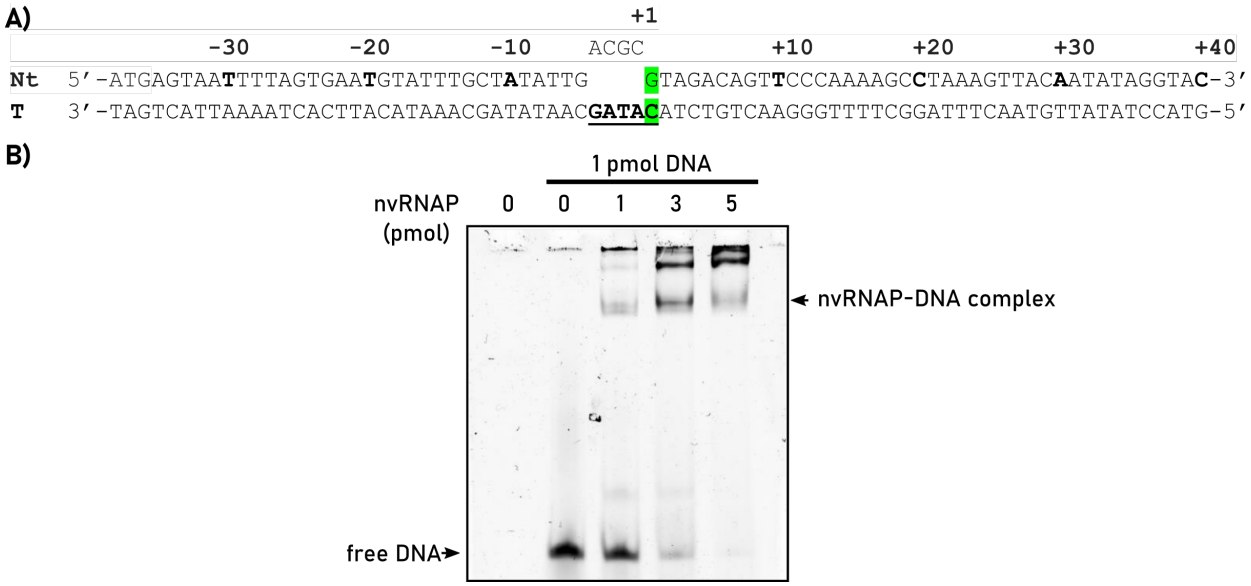


Figure 5.1: Analysis of the ΦKZ nvRNAP binding to the P119L-4nt DNA template. A) The form and sequence of the P119L-4nt DNA template. Transcription start site is coloured in green and the consensus region is highlighted in bold and underlined font. Nt indicated the non-template strand and T indicates the template strand. B) Analysis of the binding of the ΦKZ nvRNAP to the designed DNA template by gel shift (6% TBE-gel stained with SYBR-Gold). ΦKZ nvRNAP and DNA quantities are indicated on top of each lane. Data provided by our collaborators in Dr.Yakunina's group (Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia)

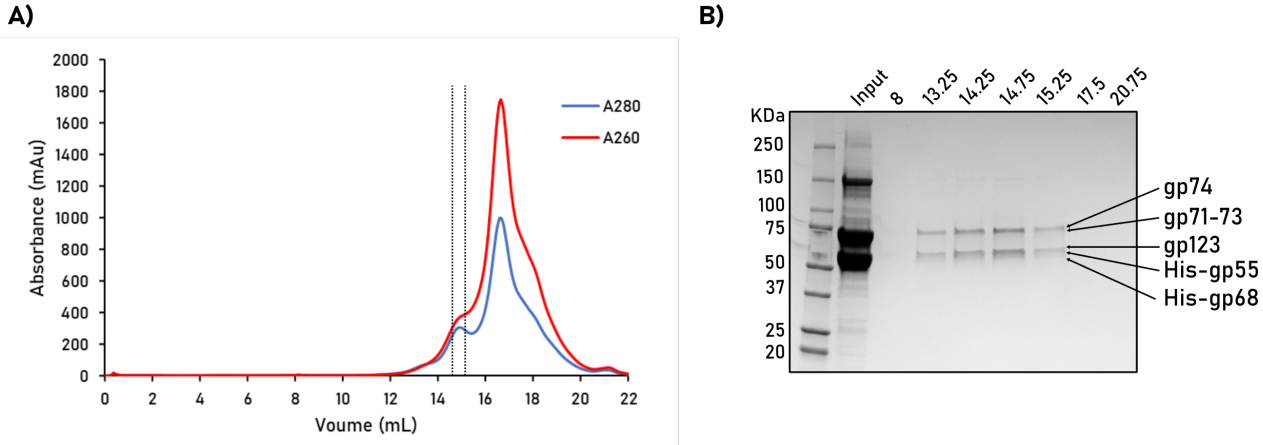


Figure 5.2: Gel filtration profile of ΦKZ nvRNAP complexes bound to the P119L-4nt DNA template. A) Gel-filtration profile of ΦKZ nvRNAP complexes in the presence of the P119L-4nt DNA template. The dotted lines indicate the peak selected to prepare cryo-EM grids. B) SDS-PAGE analysis of the gel filtration fractions of ΦKZ nvRNAP complexes in the presence of P119L-4nt DNA template.

We mixed ΦKZ nvRNAP complexes with the P119L-4nt template at a 1:3 ratio and analysed the resulting complexes by gel-filtration (Figure 5.2). We could observe two peaks corresponding to the ΦKZ nvRNAP according to SDS-PAGE analysis (Figure 5.2). Both peaks had higher A260 than A280, indicating the presence of nucleic acids. We selected the central fraction of the first

peak (higher retention volume), and prepared cryo-EM grids using the same conditions that we had previously used to prepare ΦKZ nvRNAP frozen samples: graphene-oxide coated grids, -6 blot force, waiting time of 10s, and blotting time 0.5-1s (see Chapter 2). We could observe thin layers of ice and single particles distributed within the holes. Particles appeared very similar to the ΦKZ nvRNAP holoenzyme in shape and size, however it was not possible to appreciate any nucleic acids bound by visual inspection. Hence, we collected 5 311 movies in a Titan Krios G3i equipped with a K3 direct electron detector at the London Consortium for Electron Microscopy (The Francis Crick Institute) (Figure 5.3). We performed initial single-particle analysis to identify if there was any DNA bound to the enzyme (Figure 5.4). After aligning frames with MotionCor2 (Zheng et al., 2017) and estimating CTF parameters with CTFFIND4 (Rohou and Grigorieff, 2015), we extracted nearly 3 million particles. We performed initial 2D classification to clean up the dataset and particles yielding high resolution class averages were further classified in three dimensions. None of the classes appeared to have DNA bound to them despite extensive sorting (Figure 5.4). We hypothesized that the complexes were probably disrupted upon freezing, as ΦKZ nvRNAP complexes with late promoters have reported to be highly unstable (Yakunina et al., 2015).

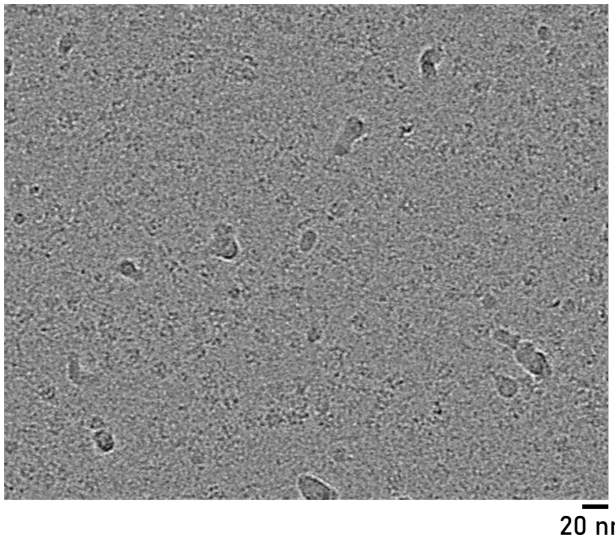


Figure 5.3: Representative micrograph of ΦKZ nvRNAP complexes bound to the P119L-4nt DNA template. Representative micrograph of the high-resolution data collection of ΦKZ nvRNAP complexes in the presence P119L-4nt DNA template.

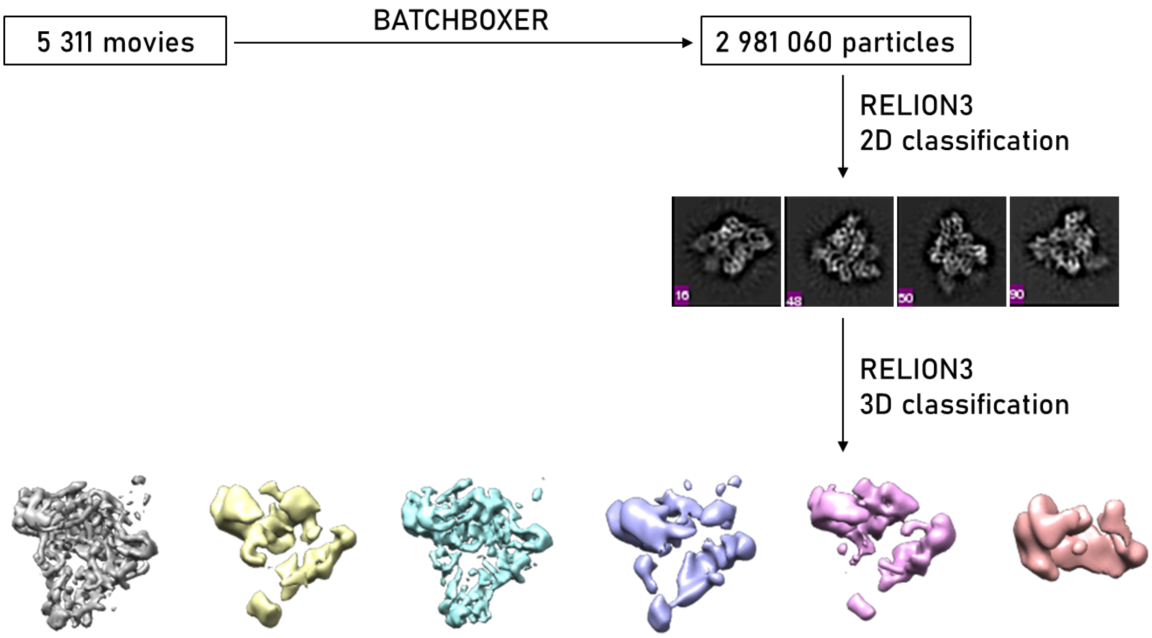


Figure 5.4: Data processing scheme for the ΦKZ nvRNAP complexes and P119L-4nt DNA template. Flow-chart representing the data processing procedure followed to sort out particles with and without DNA, however we could not identify any particles bound to the P119L-4nt template. Particle numbers retained at each stage are indicated.

We could not observe ΦKZ nvRNAP-DNA complexes using the P119L-4nt template in the results of our single particle analysis (Figure 5.4). To further stabilise the complex, we aimed to obtain a transcribing complex of the ΦKZ nvRNAP, a strategy that has been successful in obtaining cryo-EM structures for other polymerases (Glyde et al., 2018; Liu et al., 2017). For this purpose, we designed a DNA template covering positions -35 to +40 of the specific ΦKZ late promoter P119L, containing a 4-nucleotide artificial bubble and a 5 nucleotide RNA scaffold at the transcription start site. We moved the pre-opened bubble to minimise mutations within the consensus region that could compromise polymerase binding, and we placed it at the transcription start point (introducing mismatches between +1 and +4 in the non-template strand) to ease accommodation of the RNA template (Figure 5.5 A). To start transcription, we mixed ΦKZ nvRNAP complexes at a final concentration of 0.1 mg/mL with the RNA/DNA scaffold at a 1:1 ratio and added ATP, CTP and GTP. We did not add any UTP so that the reaction could not proceed beyond the first A encountered by the polymerase in the template strand and we could control the length of the transcript. Under these conditions, the nvRNAP should transcribe +3 nucleotides (up to +8 position). We let the reaction run for 30 minutes and, after that, we confirmed the presence of a RNA/DNA-nvRNAP complex by observing a gel

shift in the DNA band in the presence of polymerase (Figure 5.5 B). Once we had stable ΦKZ nvRNAP-RNA-DNA complexes, we prepared samples for cryo-EM using the same conditions and the same freezing parameters that we used for the preparation of nvRNAP complexes without DNA (see Chapter 4 and 2).

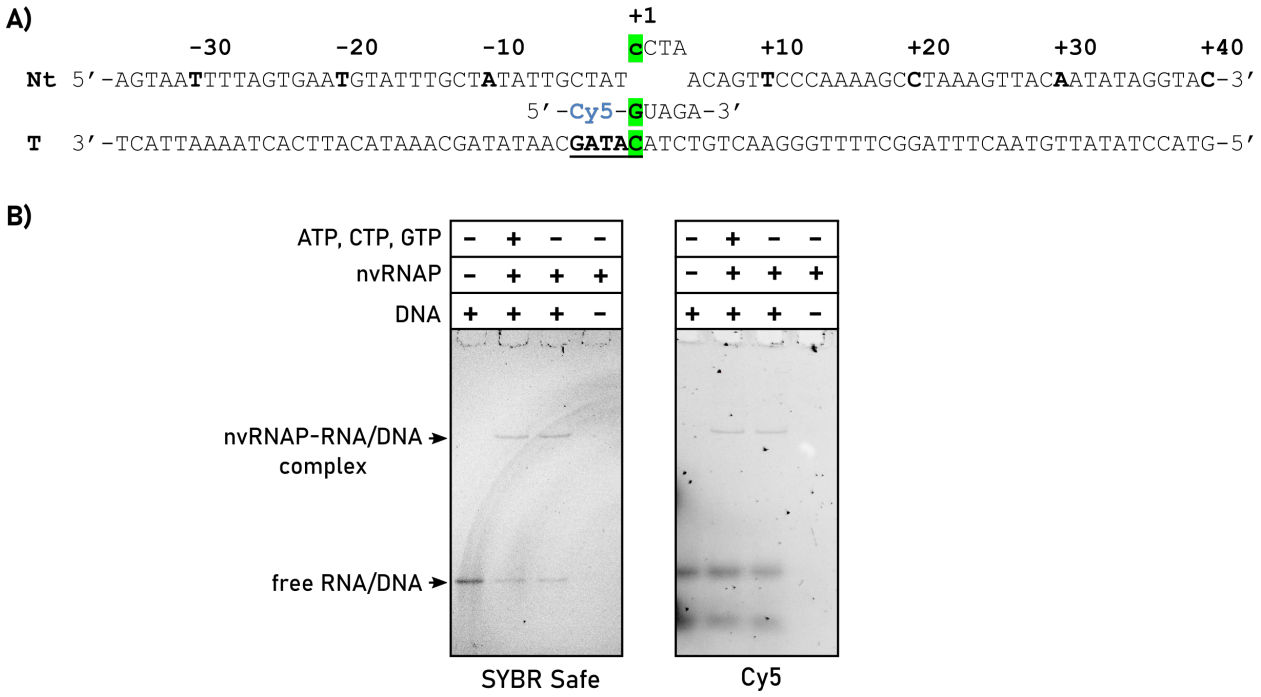


Figure 5.5: Analysis of the ΦKZ nvRNAP binding to the P119L RNA/DNA template. A) The form and sequence of the P119L RNA/DNA tempalate. The transcription start site is coloured in green and the consensus region is highlighted in bold and underlined font. B) Analysis of the binding of the nvRNAP to the designed RNA/DNA template by gel shift. 4–20 % Mini-PROTEAN TGX Precast Protein Gel (Bio-Rad Laboratories, USA) stained with SYBR-Safe (left) and imaged in the Cy5 channel (right). Samples run in each lane of the gel are indicated at the top of each image.

5.2 Preliminary structural analysis of the ΦKZ nvRNAP complex bound to the P119L promoter

We performed screening of the ΦKZ nvRNAP-DNA cryo-EM grids, at the EM facility at the Centre for Structural Biology at Imperial College London using a Glacios cryo-TEM (Thermo Fisher Scientific, USA) equipped with a Falcon 4 direct electron detector. The Glacios microscope included an auto-loader that allowed us to screen up to twelve grids at a time, which sped up the optimisation of the cryo-EM conditions. We found thin layers of vitreous ice in graphene oxide-covered grids blotted for 0.5s (Figure 5.6 A). Micrographs were well-populated with particles distributed across the grid holes (≈150 particles/micrograph) that displayed a similar shape and size to those seen in previous nvRNAP collections (see Chapter 4), (Figure 5.6 A). Despite using a more powerful microscope, it was still hard to observe any nucleic acids by visual inspection, so we collected 1 600 movies with the Glacios cryo-TEM (see Chapter 2 for collection parameters).

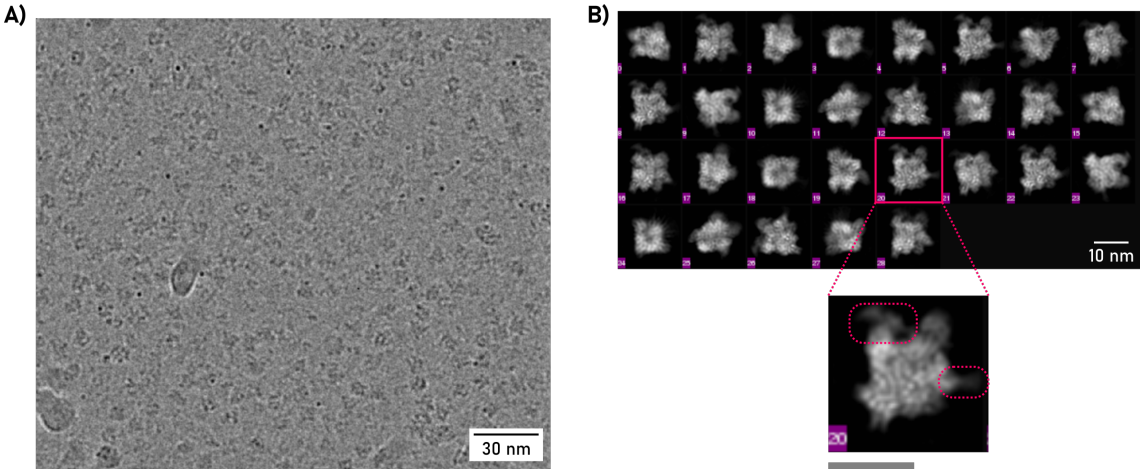


Figure 5.6: Initial structural analysis of ΦKZ nvRNAP complexes bound to the P119L RNA/DNA template. A) Representative micrograph of ΦKZ nvRNAP complexes bound to the P119L RNA/DNA template adsorbed to a graphene oxide film deposited onto a copper grid. B) Representative 2D class averages representing different views of the ΦKZ nvRNAP complex bound to the P119L RNA/DNA template. Inset, showing a magnified view of a high-resolution 2D class average with DNA highlighted in pink squares.

Frames were aligned and dose-weighted with MotionCor2 (Zheng et al., 2017) and CTF parameters were estimated using CTFFIND4 (Rohou and Grigorieff, 2015). Initially, particles were semi-automatically selected with BATCHBOXER (Tang et al., 2007) and then classified

in 2D class averages with RELION 3 (Zivanov et al., 2018). 2D classification was performed leaving the frequencies up to the first peak of the CTF intact to generate references to re-select particles more accurately. After re-boxing, a total of 528 792 particles were further classified in two dimensions, with 156 304 yielding high-resolution class averages (Figure 5.6 B). Amongst these classes, we could observe different views of the nvRNAP and we could also appreciate DNA entering and exiting the polymerase in some classes (Figure 5.6 B, inset). Particles yielding high-resolution 2D classes were retained for further refinement (156 304), yielding a 4.7 Å reconstruction according to an independent half-set FSC of 0.143 (Figure 5.7).

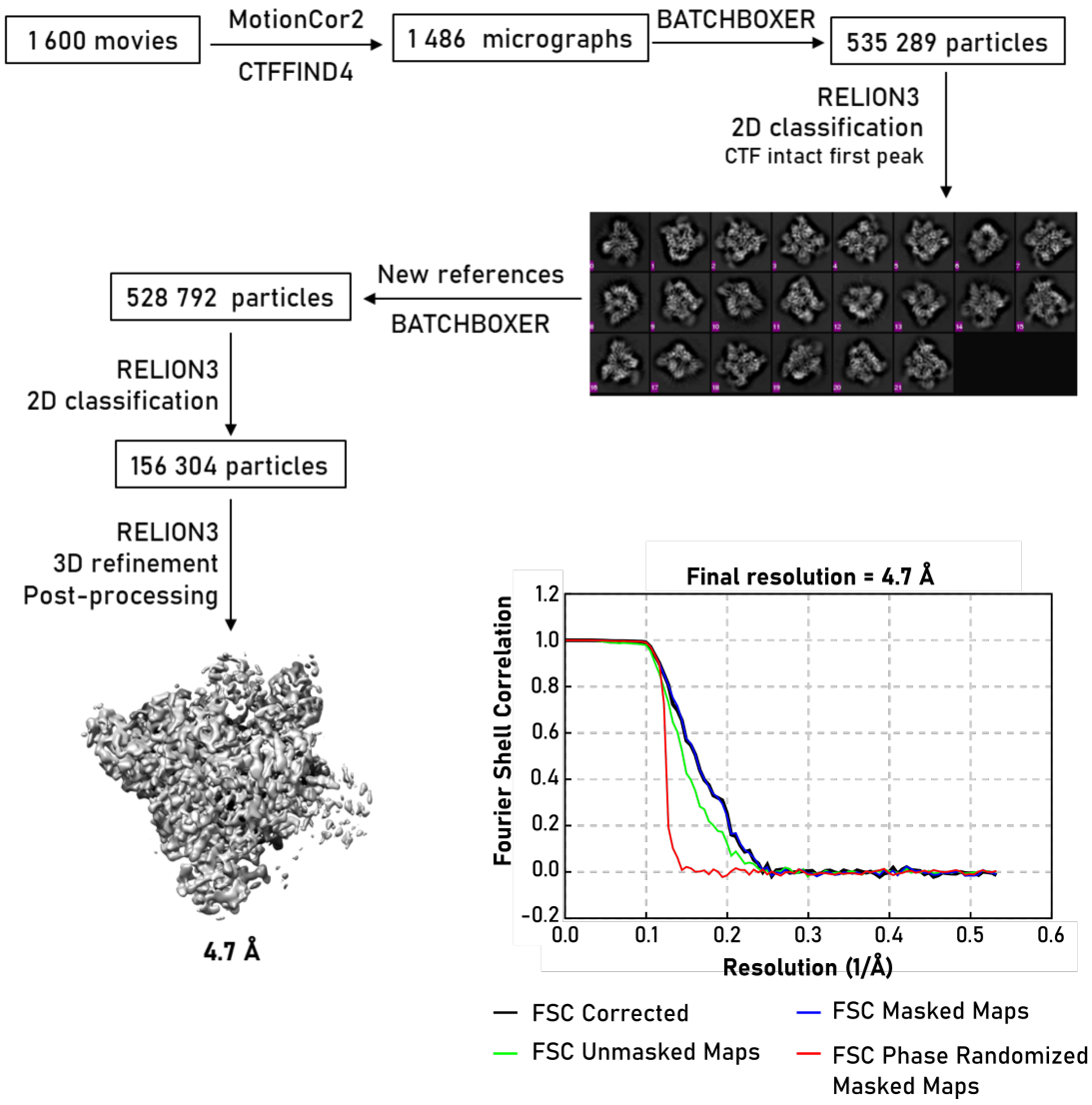


Figure 5.7: Data processing scheme for initial cryo-EM data collection of ΦKZ nvRNAP complexes bound to the P119L RNA/DNA template. Flow-chart depicting the data processing procedure. Particle numbers retained at each stage of the processing are indicated below each 3D reconstruction. On the bottom-right, the global FSC curve for the 4.7 Å reconstruction.

We fitted the nvRNAP structure that we obtained in the previous chapter (de Martín Garrido et al., 2021b)(PDB 7OGP) into the new 4.7 Å 3D reconstruction to make a direct comparison between the two conditions, and we could clearly observe new regions of density in the new map. Comparison with an *E.coli* RNAP transcribing complex (PDB 6GH5), confirmed that the new 3D reconstruction had DNA entering and exiting the ΦKZ nvRNAP channel in a similar manner to that in bacterial msRNAPs, as previously seen in some 2D class averages. The density for the downstream DNA was well-defined, whereas the density for the upstream DNA was weakly resolved in the 4.7 Å initial reconstruction, an indication that this region of DNA had a certain degree of flexibility in its position within the complex. We could also observe clear density with defined secondary structure elements in the regions that we previously allocated to the middle domains of gp74 and gp123, which were not resolved in the previous ΦKZ nvRNAP reconstruction. Furthermore, the new map had some extra density at the front of the enzyme in contact with the DNA in the region that we previously suggested that could correspond to the N-terminus of gp68.

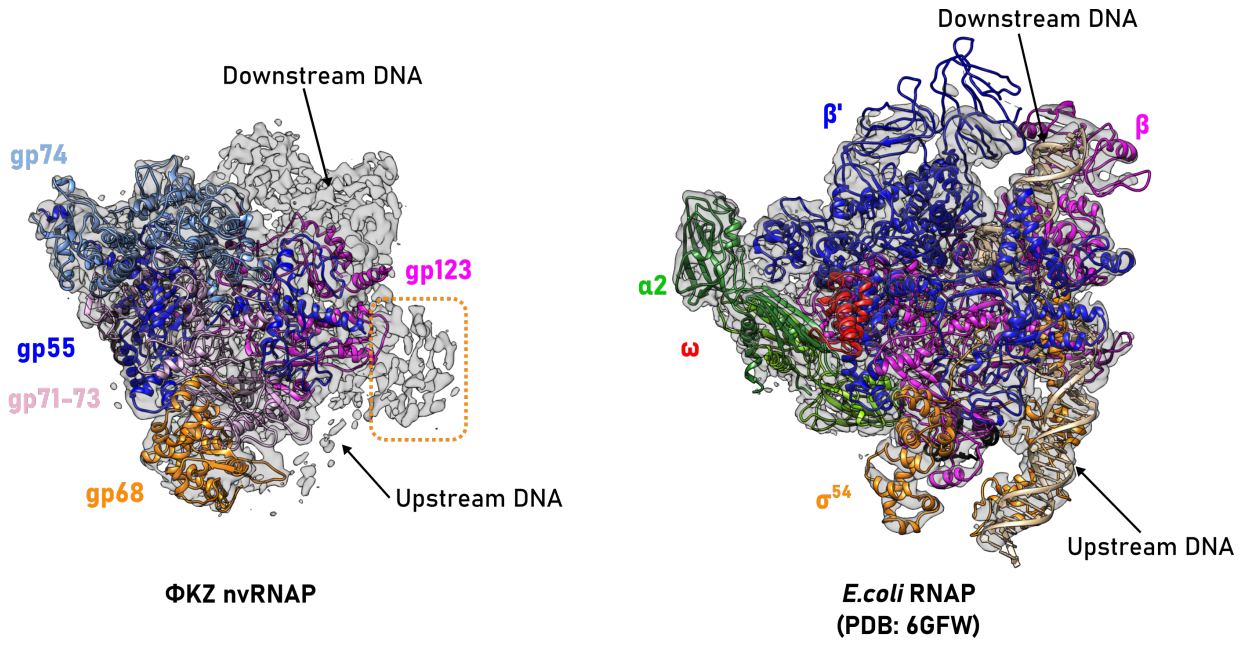


Figure 5.8: Initial comparison of the ΦKZ nvRNAP-DNA reconstruction with an *E.coli* transcribing complex. On the left, nvRNAP structure (see Chapter 4, PDB 7OGP) fitted to the 4.7 Å map. On the right, *E.coli* RNAP structure (PDB 6GFW) fitted to the corresponding EM map (EMDB-4397). Each subunit is coloured accordingly. The density which presumably corresponds to the DNA is indicated with black arrows and the region of the map that could correspond to the N-terminus of gp68 is highlighted in a dotted orange square.

5.3 High-resolution data collection and model building

We recovered the grid used for the Glacios collection and set up a high resolution data collection at eBIC (Diamond Light Source) using a Titan Krios G3i equipped with a Falcon 4i Direct Electron Detector and a Selectris filter (Thermo Fisher Scientific, USA). We selected new grid squares that were not imaged in the previous collection and micrographs appeared to be equally populated with nvRNAP particles (≈150 particles/micrograph) (Figure 5.9). A total of 3 092 movies were aligned and dose-weighted with MotionCor2 and RELION4 (Zheng et al., 2017; Kimanius et al., 2021) and CTF parameters were estimated with CTFFIND4 and RELION4 (Rohou and Grigorieff, 2015; Kimanius et al., 2021), with 60% of the micrographs (1 827) containing Thon rings extending to higher resolutions than 4 Å (Figure 5.9).

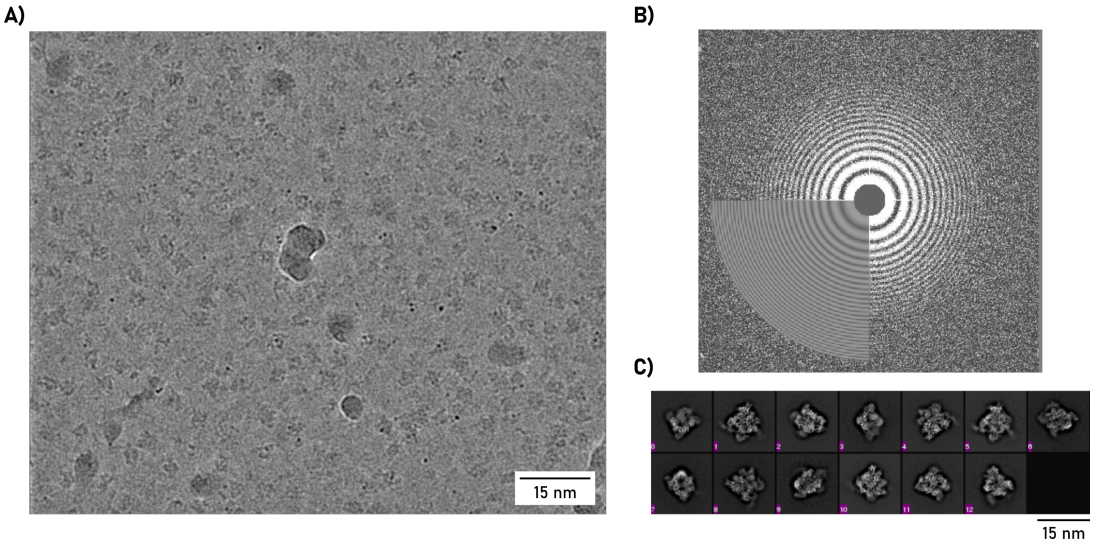


Figure 5.9: High-resolution data collection of the ΦKZ nvRNAP in complex with an RNA/DNA P119L template. A) Representative micrograph of nvRNAP complex bound to P119L RNA/DNA scaffold adsorbed to a graphene oxide film deposited onto a copper grid. B) Thon rings simulated with the estimated CTF parameters (bottom left) and experimental power spectrum fitted. C) Representative 2D class averages representing different views of the ΦKZ nvRNAP in complex with the P119L RNA/DNA template.

A total of 2 981 060 particles were semi-automatically selected with BATCHBOXER and further classified in 2D class averages with RELION4 (Kimanius et al., 2021). After that, the subset of particles yielding high-resolution class averages was further classified in 3D without alignment using a high tau fudge value (tau2_fudge 16) to try to separate particles with and without DNA. We selected 83 389 particles that yielded a high resolution class with the downstream

DNA region and the catalytic core well-defined. After CTF and aberration refinement (beam tilt, anisotropic magnification and per-particle CTF estimation), and Bayesian polishing, this set of particles yielded a 3.3 Å resolution reconstruction according to an independent half-set FSC of 0.143, that was used as reference for further refinement. After 3D classification with fine angular search (sigma_ang 2), 39 919 particles yielded a 3.3 Å resolution reconstruction according to an independent half-set FSC of 0.143 (Figure 5.10), which had better definition of the catalytic core and some peripheral regions.

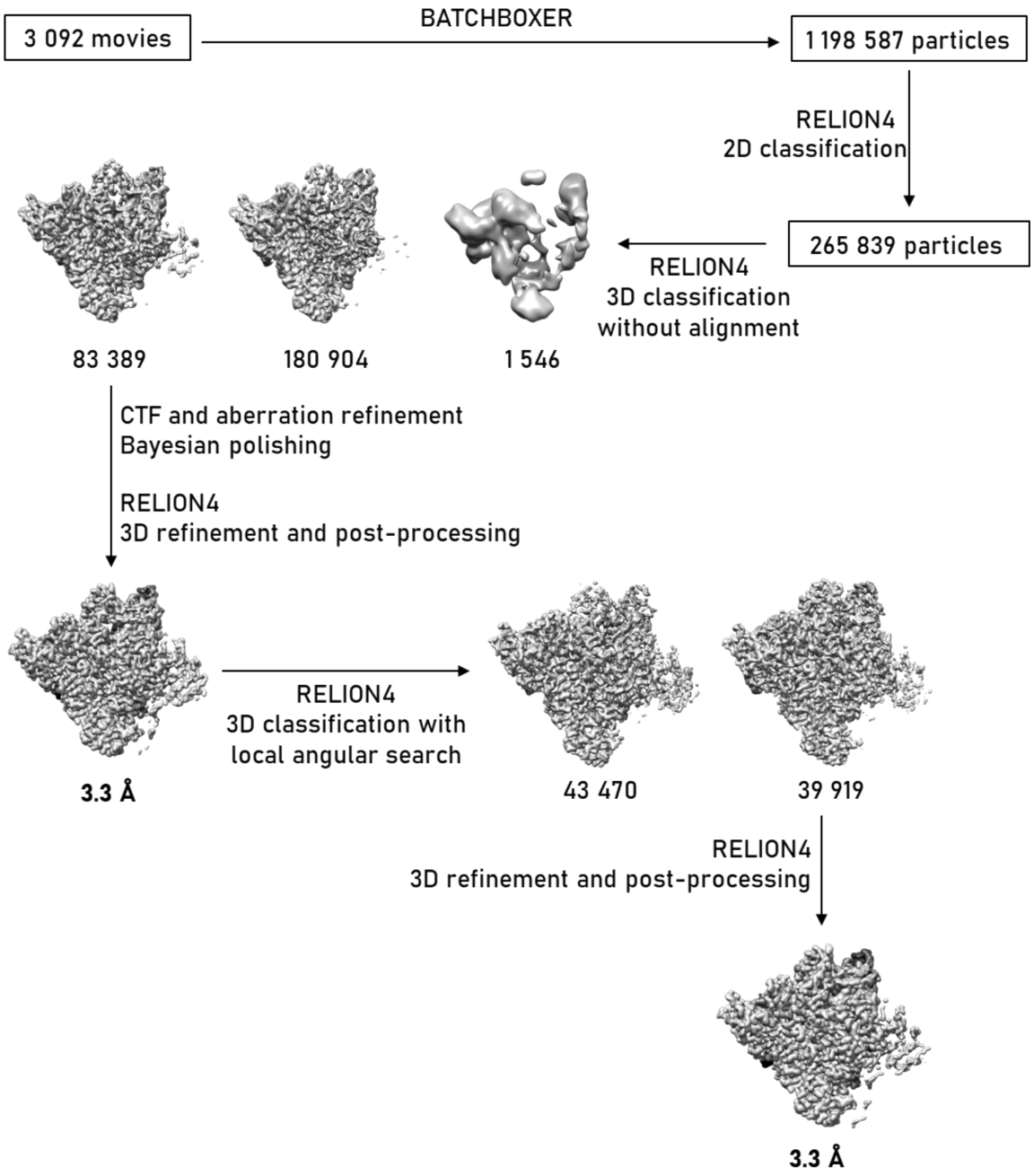


Figure 5.10: Data processing scheme for the high-resolution cryo-EM data collection of the ΦKZ nvRNAP complex bound to the P119L RNA/DNA template. Flow-chart depicting the data processing procedure. Particle numbers retained at each stage of the process are indicated below each 3D reconstruction.

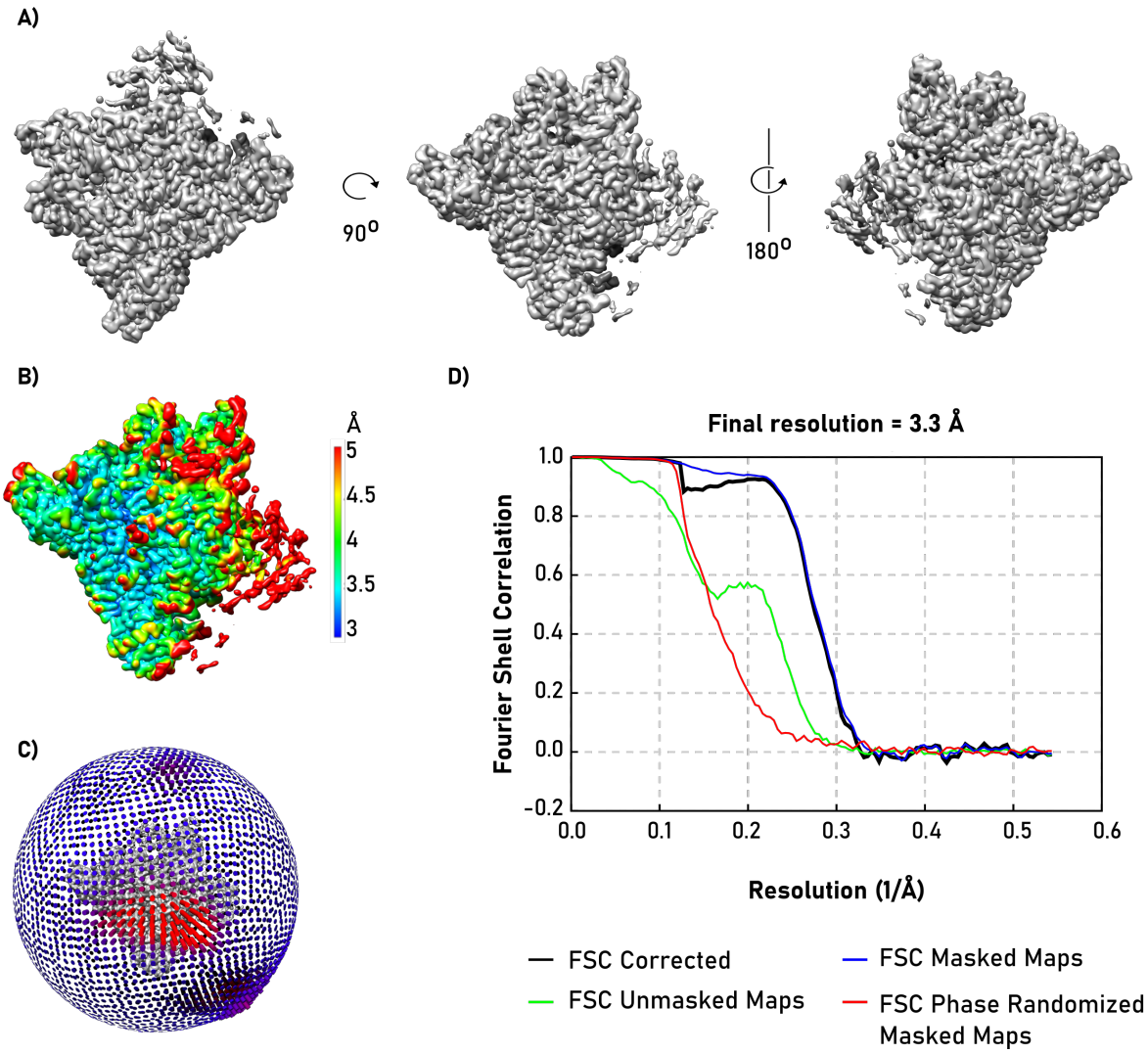


Figure 5.11: High-resolution 3D reconstruction of the ΦKZ nvRNAP bound to the P119L RNA/DNA template. A) 3.3 Å reconstruction of the ΦKZ nvRNAP bound to the P119L RNA/DNA template in different views. B) Local resolution map of the 3.3 Å reconstruction calculated using BLOCRES. C) Angular distribution plot of the 3.3 Å reconstruction. D) Global FSC curve for the 3.3 Å reconstruction.

The map provided better resolution for some sections that were weakly-resolved in the previous ΦKZ nvRNAP structure, hence providing better coverage of the structural model (Figure 5.11). Particularly, regions of density forming the two pincers encircling the downstream DNA that we had previously assigned to the middle domains of gp123 and gp74, and the clamp were better resolved, probably because of the presence of nucleotide. We therefore generated models of these regions with Alphafold2 (Jumper et al., 2021) (gp74 middle domain: 367-545, gp123 middle domain: 143-337) and fitted them in the 3.3 Å map (Figure 5.14). The Alphafold2 model for the gp123 middle domain fitted very accurately in the density, with a real-space correlation coefficient of 0.7 at 3.3 Å calculated with Chimera, we therefore used it as a guide to build a molecular model for this domain. The gp74 structural model fitting was not that accurate (real-space correlation coefficient of 0.4 on fitting), however we were still able to use it to establish sequence register and build a molecular model of this domain. The ΦKZ nvRNAP-DNA map at 3.3 Å resolution allowed us to built 97 % of the sequence of the four β/β'-like subunits (gp71-73, gp123, gp55 and gp74) (Figure 5.12 and 5.13).

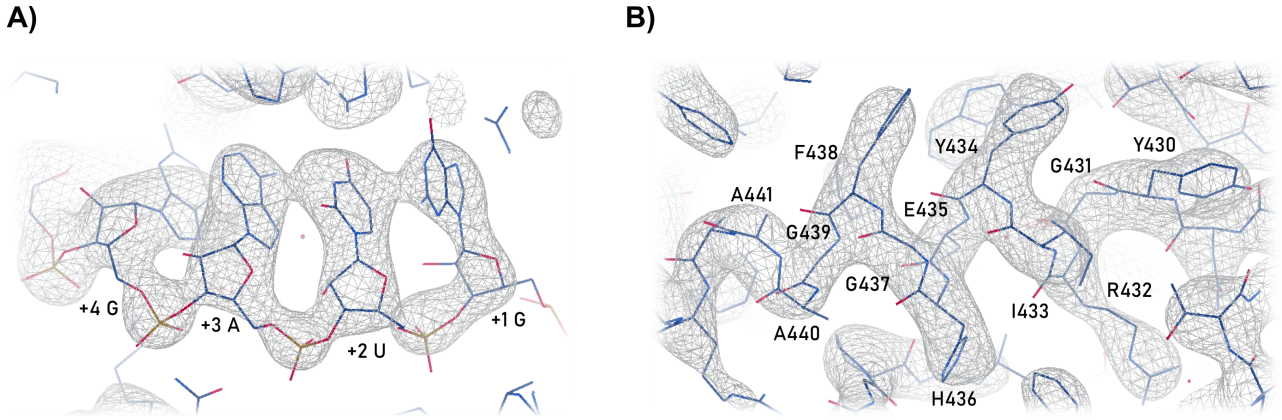


Figure 5.12: Representative high-resolution density of the ΦKZ nvRNAP-P119L RNA/DNA template 3.3 Å map. RNA (A) and gp71-73 (B) molecular models fitted into the 3.3 Å map. Molecular models are displayed in blue, and the map is displayed as a grey mesh.

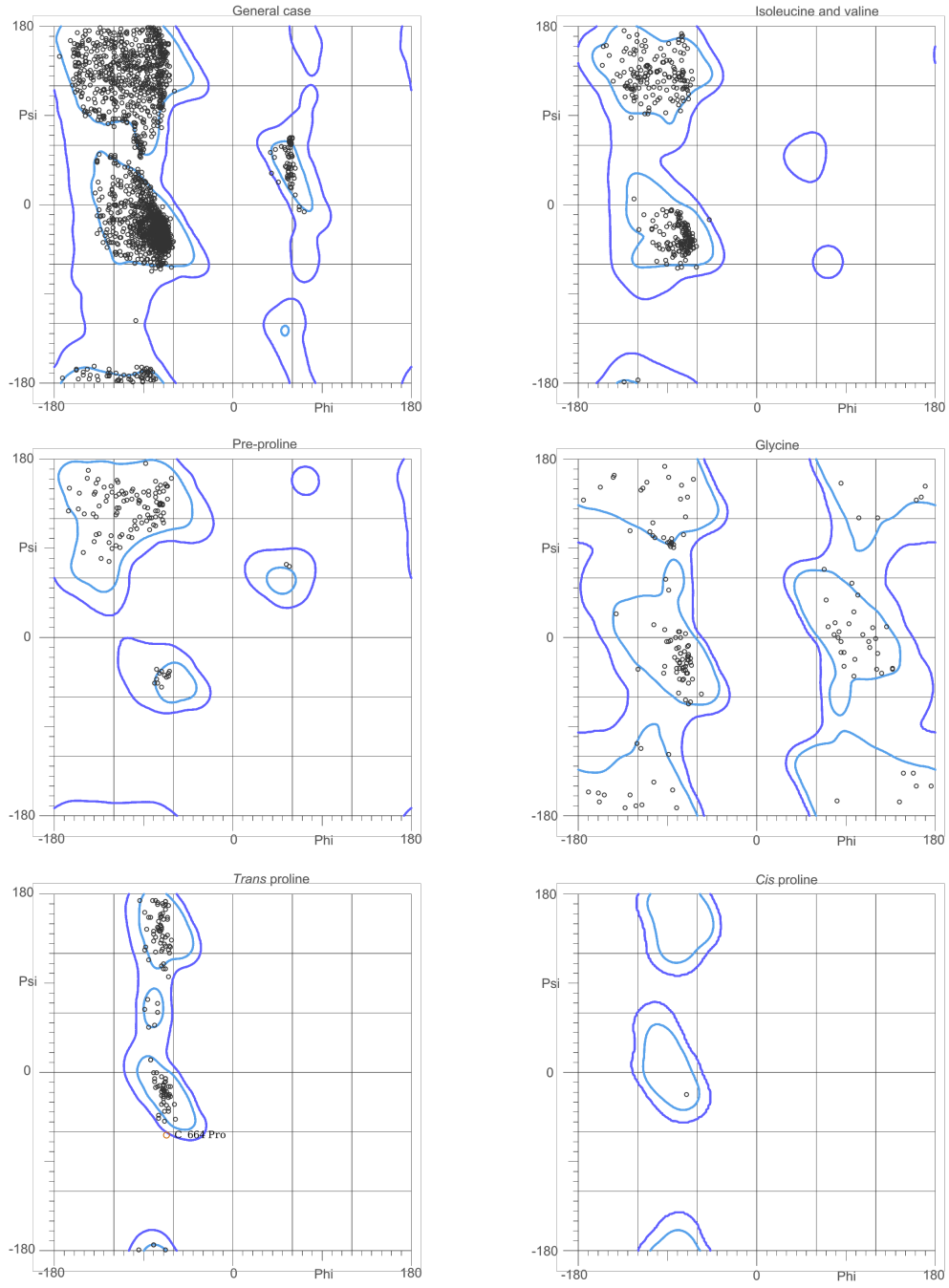


Figure 5.13: Ramachandran plot of the ΦKZ nvRNAP structure bound to the P119L RNA/DNA template. Ramachandran plots were computed with Molprobity (Lovell et al., 2003).

After building models for the four subunits with homology to the two largest bacterial msRNAPs subunits, the new 3D reconstruction had a lower resolution extra density region located at the front of the DNA channel. We could clearly appreciate α helices in that region, indicating that the density corresponded to a new protein domain. The only sequence of protein that could correspond to this novel region was the N-terminus of gp68. Indeed, a structural model of this region generated with Alphafold 2 (Jumper et al., 2021) (residues 1-318) fitted accurately

into the density with a real-space correlation coefficient of 0.6 at 3.3 Å resolution according to Chimera, and most of the α helices already coinciding with the density, thus identifying the position of the N-terminus of gp68 in the nvRNAP structure (Figure 5.14).

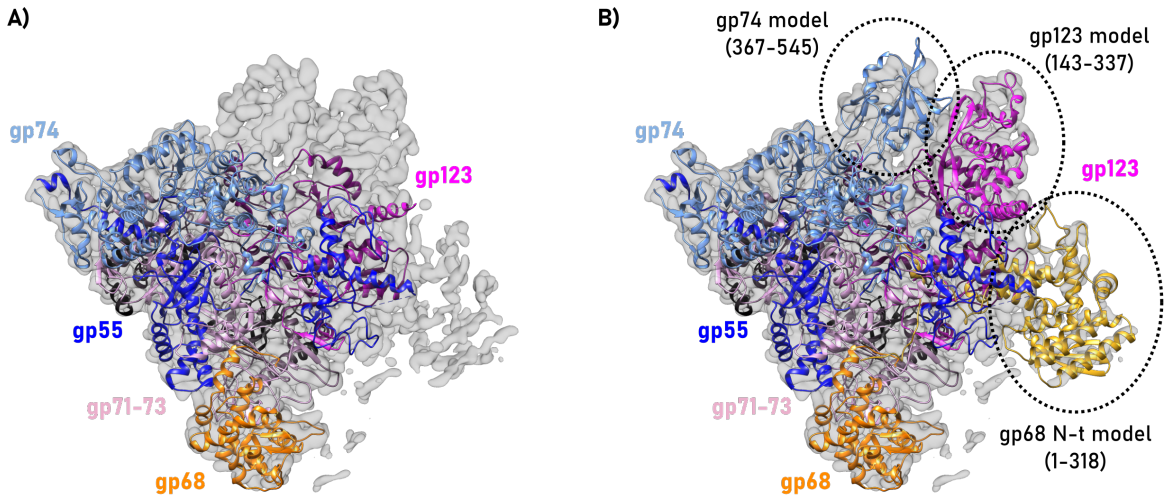


Figure 5.14: New features of the high-resolution 3D reconstruction of the ΦKZ nvRNAP bound to the P119L RNA/DNA template. A) 3.3 Å resolution reconstruction of the ΦKZ nvRNAP complex bound to the P119L RNA/DNA template with the previous nvRNAP structure without DNA fitted (PDB 70GP) (de Martín Garrido et al., 2021b). B) 3.3 Å resolution reconstruction of the ΦKZ nvRNAP complex bound to the P119L RNA/DNA template with the previous nvRNAP structure without DNA fitted (PDB 70GP) (de Martín Garrido et al., 2021b) and models for gp74 middle domain (396-533), gp123 middle domain (146-336), and gp68 N-terminus and linker (1-273) fitted.

Below the N-terminus of gp68 we could see some weakly resolved density that clearly corresponded to dsDNA, as the phosphate backbone was visible. Comparison with bacterial msRNAPs transcribing complexes allowed us to assign this density to upstream DNA (Figure 5.15). In our map, the resolution of this section was not high enough to assign sequence and it was poorly connected with the core of the enzyme, despite extensive sorting. However, at low contour (0.0046), we could trace the phosphate backbone and fit a 17 bp dsDNA model with unknown sequence. In the Glacios collection, we had the same occupancy issue with the 4.7 Å resolution reconstruction (Figure 5.8), we therefore presumed that this DNA region was flexible. The downstream DNA and the core were better resolved than the upstream DNA. At the core of the enzyme in close proximity to the two DPBB domains, we could observe a 6-nucleotide RNA transcript (Figure 5.15). We placed the bases of our initial RNA scaffold with one additional transcribed base (5'-GUAGAC-3') into the density and they fit accurately. Based on the position of the RNA, we could assign sequence for the DNA and expanded it

towards the downstream region, however the density was too weak to expand the sequence towards the upstream region, and the connection through the ssDNA forming the transcription bubble was not resolved.

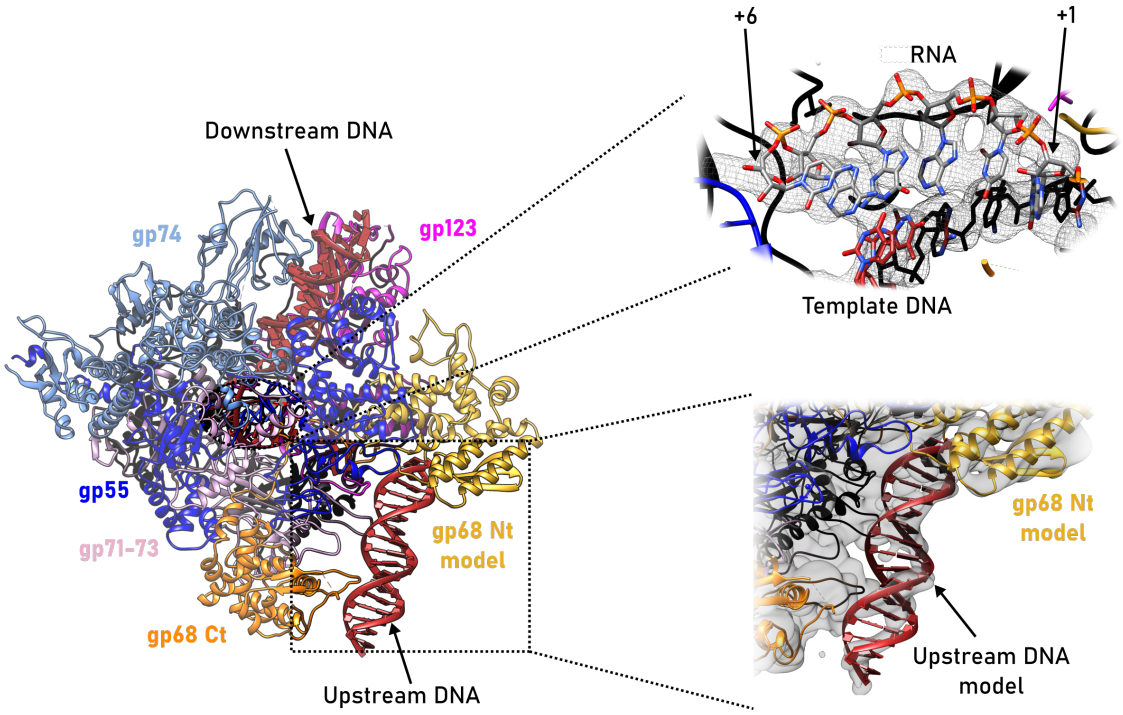


Figure 5.15: The upstream DNA and the RNA of the the ΦKZ nvRNAP bound to the P119L RNA/DNA template. Structure of the 3.3 Å ΦKZ nvRNAP complex bound to the P119L RNA/DNA template including a docked model of the upstream DNA and the N-terminus of gp68. Each subunit is coloured as indicated. (Top-right) Inset of the catalytic centre of the ΦKZ nvRNAP. RNA is represented in grey and template DNA in firebrick red, whereas density is represented in a grey mesh. (Bottom-right) Density represented in grey at low contour (0.0046) with a model with unknown sequence of dsDNA fitted.

5.4 Architecture of the ΦKZ nvRNAP bound to the KZP119L promoter

The new 3.3 Å resolution reconstruction allowed us to provide an atomic model for most of the four β/β'-like subunits (gp71-73, gp123, gp55 and gp74) covering 97 % of their sequence, and to dock an accurate structural model for the N-terminus of gp68 (Figure 5.16). Moreover, the structure included a 6-bp RNA transcript along the template strand at the catalytic core, as well as the downstream dsDNA accommodated in the polymerase channel. As we mentioned above, the upstream dsDNA was not resolved to high enough resolution to assign sequence, however we could dock a dsDNA model tracing the phosphate backbone. As we saw in the ΦKZ nvRNAP holoenzyme structure without DNA, the overall shape resembles that of bacterial msRNAPs and most of the structural elements involved in the transcription reaction were observed to be conserved (de Martín Garrido et al., 2021b). Accordingly, the DNA entering and exiting the nvRNAP channel was positioned in a similar manner to that in bacterial msRNAPs (Figure 5.16).

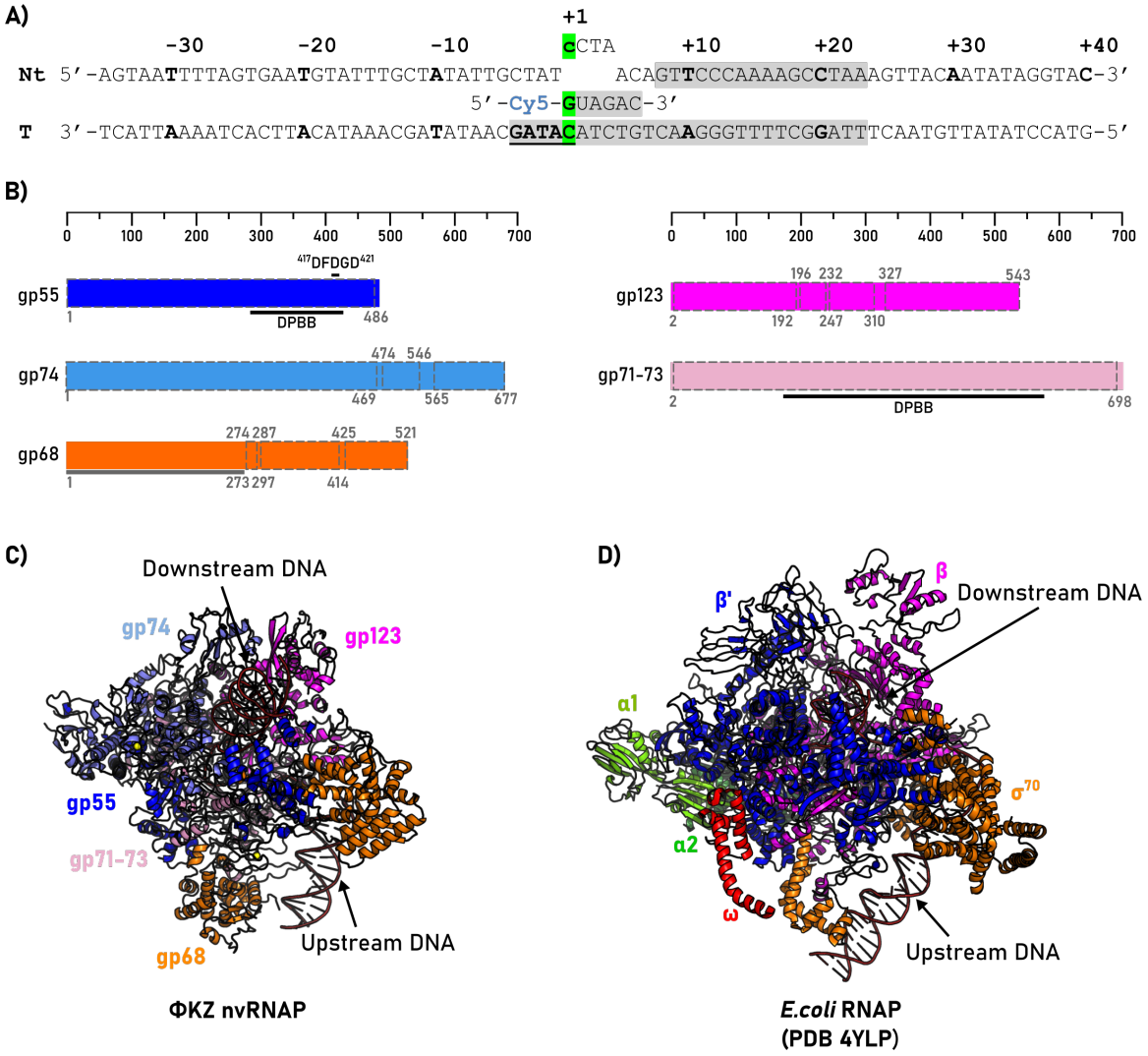


Figure 5.16: Structure of the ΦKZ nvRNAP complex bound to the P119L RNA/DNA template and structural comparison to an *E.coli* transcribing complex. A) Form and sequence of the P119L RNA/DNA template used to prepare nvRNAP complexes bound to DNA. The consensus sequence is represented in bold and underlined font. The transcription initiation site is coloured in green. Residues that were visible in our structure are represented in grey. Nt indicates non-template strand and T indicates template strand. B) Schematic representation of the coverage of our structural model. Each nvRNAP subunit is represented in a different color. Regions built in the structure are represented with dotted grey boxes. Regions with an accurate structural model are represented with grey lines, and important features with black lines. C) Structure of the ΦKZ nvRNAP bound to the P119L RNA/DNA template including a docked model of the upstream DNA and the N-terminus of gp68. Each subunit is coloured as indicated. D) Structure of an *E.coli* RNAP transcription initiation complex (PDB 4YLP)(Yuhong Zuo et al., 2015).

Comparing the structure of the ΦKZ nvRNAP complex with and without DNA (PDB 7OGP, Chapter 4), we found three regions that were better resolved in the presence of nucleic acids: the clamp (included in gp55), the middle region of gp74, and the middle domain of gp123 (Figure 5.14). Furthermore, other structural elements of the the holoenzyme which are key for transcription had different conformations in the ΦKZ nvRNAP structure bound to the

P119L RNA/DNA template. The trigger loop was fully visible in the ΦKZ nvRNAP structure bound to the RNA/DNA template, displaying an helix/loop conformation, and the bridge helix was in a fully helical conformation, in comparison to the nvRNAP structure with DNA (PDB 7OGP)(see Chapter 4)(de Martín Garrido et al., 2021b).

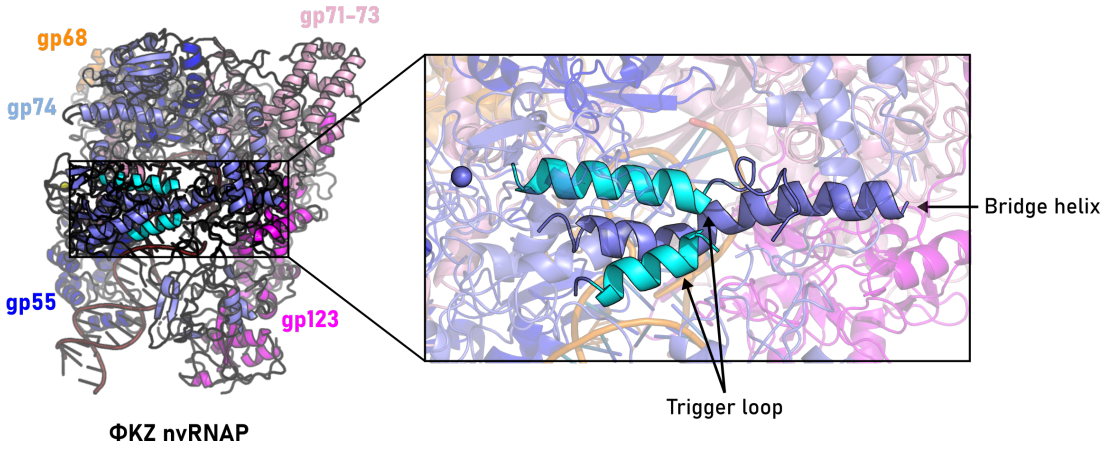


Figure 5.17: The ΦKZ nvRNAP trigger loop. (Left) The ΦKZ nvRNAP complex bound to the P119L promoter with the trigger loop region highlighted. (Right) Inset showing the ΦKZ nvRNAP trigger loop and bridge helix.

The clamp is a mobile element that oscillates between different conformations throughout the transcription reaction, adopting a closed conformation that stabilises the DNA during the RNAP promoter open complex (Chakraborty et al., 2012). In the ΦKZ nvRNAP structure bound to RNA/DNA template, we could observe that the clamp was fully ordered with the two clamp helices and the lid adopting similar conformations to those in bacterial msRNAP transcribing complexes (Figure 5.18). The lid interacted with the flap to enclose the RNA exit channel in the same way as in bacterial msRNAPs. Notably, we found that the rudder, responsible for stabilising ssDNA within the transcription bubble (Kuznedelov et al., 2002), was absent in the ΦKZ nvRNAP (Figure 5.18). At the N-terminus of gp55, close to the clamp, we also built the second Zn^{2+} binding domain conserved in all multisubunit msRNAPs, which is involved in promoter recognition (Lane and Darst, 2010a,b).

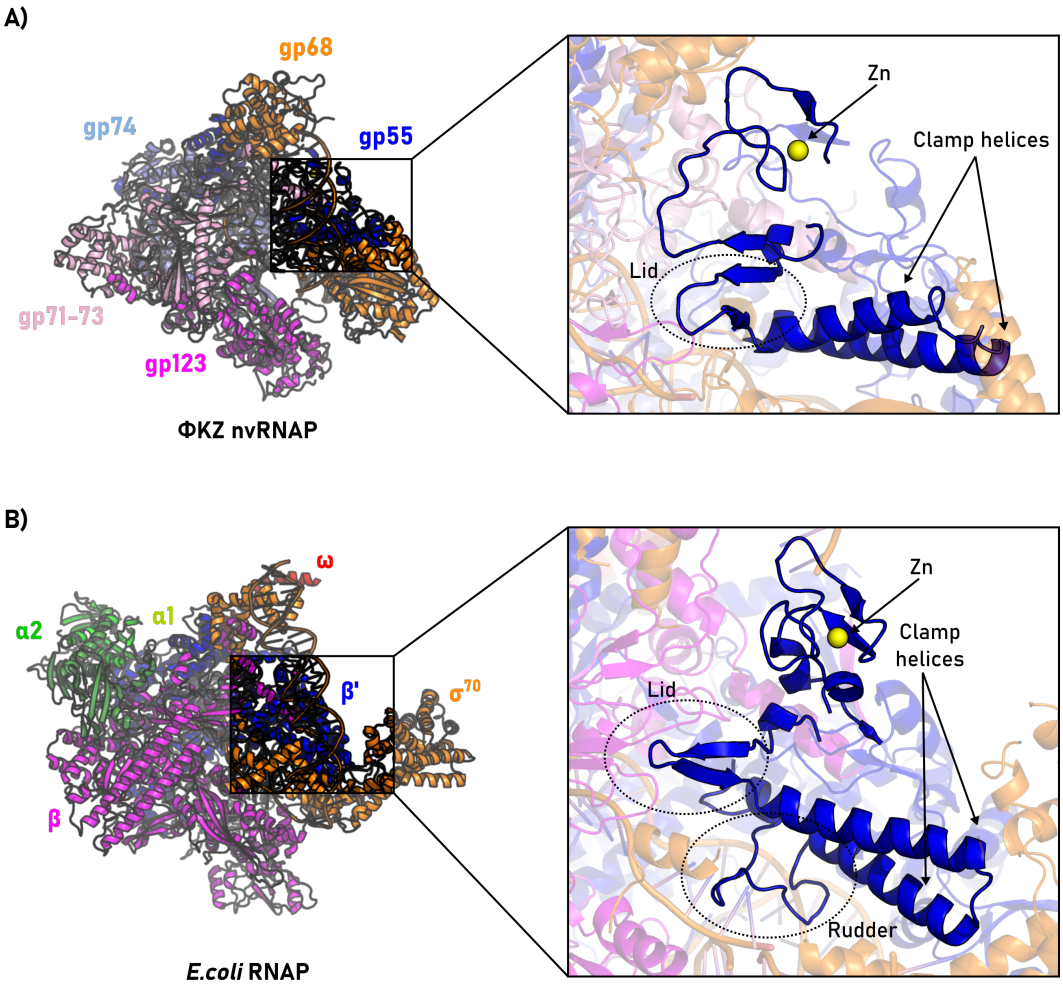


Figure 5.18: Structural comparison of the ΦKZ nvRNAP and *E.coli* RNAP structural elements included in the clamp. A) Structure of the ΦKZ nvRNAP complex bound to the P119L RNA/DNA template. Inset shows the structural elements of the clamp and the Zn^{2+} binding domain comprised at the N-terminus of gp55. B) *E.coli* RNAP transcribing complex (PDB 4YLP) (Yuhong Zuo et al., 2015). Inset showing the structural elements of the clamp. Note, that the rudder is absent in the ΦKZ nvRNAP structure.

The middle domains of gp123 and gp74 were also better ordered in the structure of the ΦKZ nvRNAP complex bound to the P119L RNA/DNA template. We found that these domains were located at the top of the enzyme forming the polymerase channel in the downstream DNA section. Indeed, we observed that the gp123 middle domain (146-336) formed the tip of one of the pincers that holds the downstream dsDNA (Figure 5.19). This domain corresponded to the bacterial β2 domain (also known as lobe), which covers the downstream dsDNA (Lane and Darst, 2010a). In the structure of the ΦKZ nvRNAP bound to the P119L RNA/DNA template, we observed that the gp123 middle domain interacted with the downstream DNA through a disordered loop containing positively charged residues (^{192}TKRDLGN^{198}) preserving the same function observed in bacterial msRNAPs (Figure 5.19). We performed a structural

comparison of the gp123 β2 domain using the DALI server and found that the closest structural homologue was the β subunit of *M.tuberculosis* (RMSD 2.9 according to DALI), although gp123 had additional insertions (Figure 5.20).

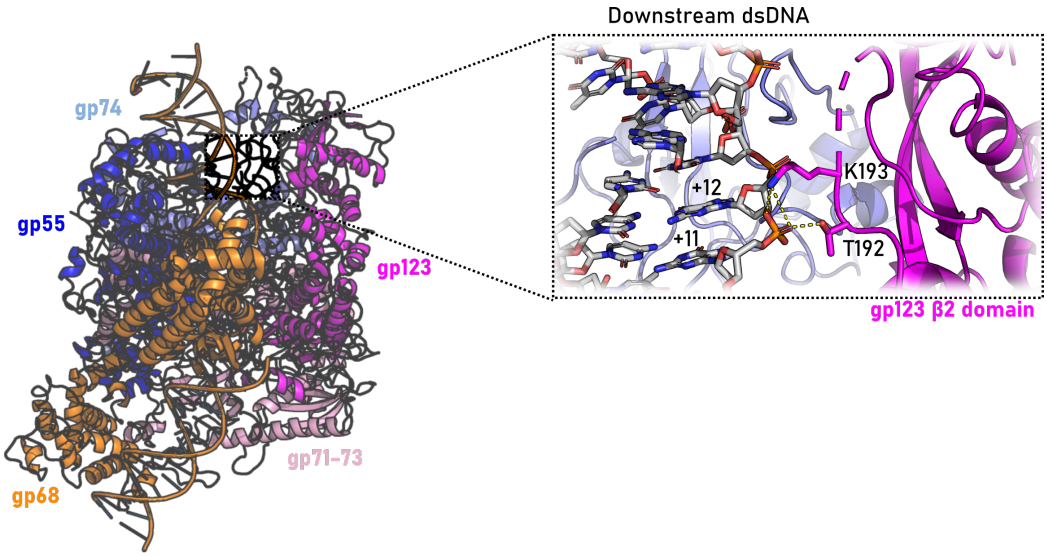


Figure 5.19: The ΦKZ gp123 middle domain interacts with the downstream DNA. ΦKZ nvRNAP structure bound to the P119L RNA/DNA template. On the right, inset corresponding to the squared region of the ΦKZ nvRNAP structure. The inset shows the interaction of gp123 with the upstream dsDNA through a half-disordered loop. Residues T192 and K193 (visible in our structure) interact with the DNA phosphate backbone at residues +11 and +12 relative to the transcription start point.

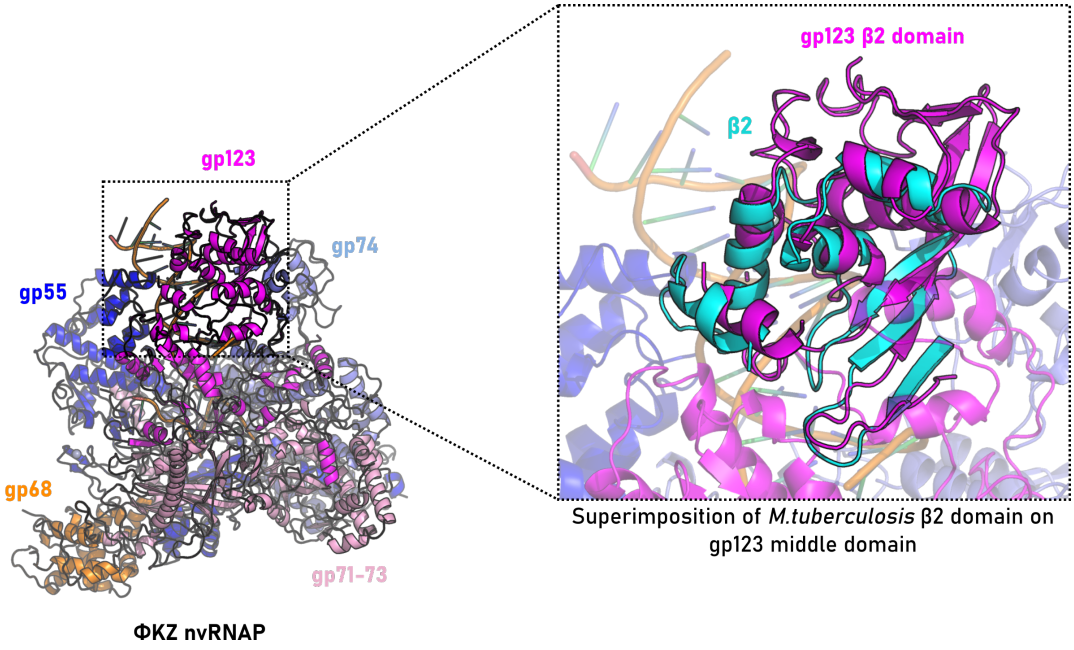


Figure 5.20: The ΦKZ nvRNAP and *M.tuberculosis* RNAP β2 domain comparison. ΦKZ nvRNAP structure bound to the P119L RNA/DNA template. On the right, the inset corresponds to the squared region of the ΦKZ nvRNAP structure and it shows superimposition of the *M.tuberculosis* β2 domain (PDB 6JCY) on the gp123 middle domain using the structural alignment provided by DALI.

The other domain that was newly-resolved in the ΦKZ nvRNAP structure bound to the P119L RNA/DNA template was found in gp74. This domain was an insertion in the trigger loop (396-533). The *E.coli* RNAP also has a lineage-specific insertion in the trigger loop, comprising residues 942-1129 of the β' subunit, which is involved in different stages of transcription (Sutherland and Murakami, 2018). We observed that the ΦKZ nvRNAP trigger loop insertion in gp74 had a different topology to that of the equivalent *E.coli* insertion (Figure 5.21).

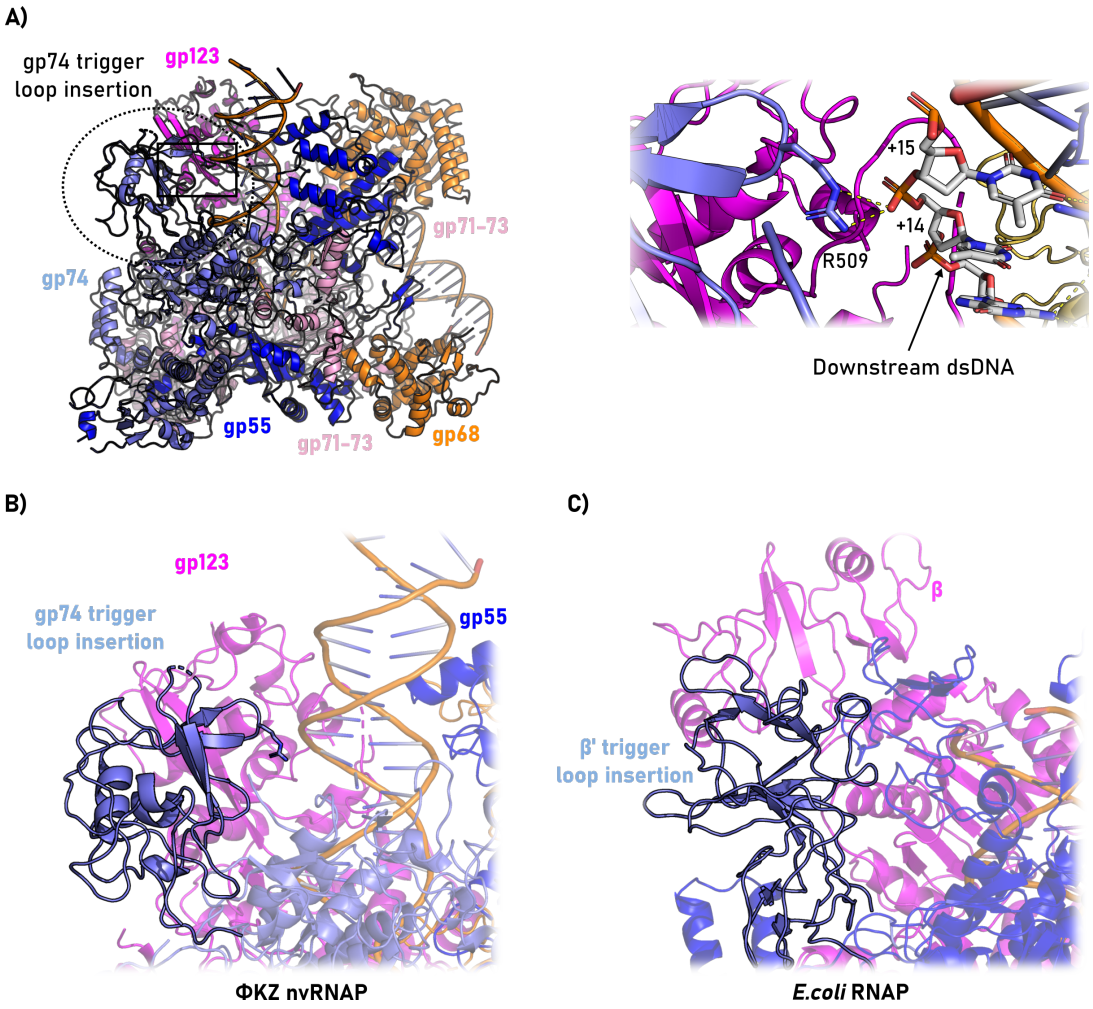


Figure 5.21: Comparison of the ΦKZ nvRNAP and the *E.coli* RNAP trigger loop insertions. A) ΦKZ nvRNAP structure bound to the P119L RNA/DNA template showing the location of the gp74 trigger loop insertion. Inset shows the interaction between the Arg508 of gp74 and phosphate backbone of +14 base relative to the transcription start site. B) ΦKZ nvRNAP bound to the P119L RNA/DNA template showing the gp74 trigger loop insertion C) *E.coli* RNAP transcribing complex (PDB 4YLP) showing the position of the β' insertion loop (942 to 1129).

This insertion is absent in the *M.tuberculosis* RNAP structure and the DALI server failed to identify any bacterial msRNAP subunits available in the PDB with structural homology to the gp74 trigger loop insertion. The trigger loop insertion was located next to the gp123 middle

domain, enclosing the downstream dsDNA and interacting with its phosphate backbone through positively charged residues (Arg508) (Figure 5.21). After the second helix of the trigger loop, bacterial msRNAPs have another structural domain known as the jaw, which is not conserved in archaeal and eukaryote msRNAPs.

5.5 Organisation of the ΦKZ nvRNAP catalytic core

At the catalytic center of ΦKZ nvRNAP structure bound to the P119L RNA/DNA template, we observed clear density for a 6-bp RNA transcript hybridised along the single-stranded DNA template (Figure 5.15). However, the non-template ssDNA forming the transcription bubble was not visible in our map. In bacterial msRNAPs, the *fork loop 2* (FL2) and *switch 3* (SW3) of the β subunit, and the lid and rudder included in the β' subunit form a network of interactions with the DNA and RNA strands to stabilise the transcription bubble. Most of the structural elements at the catalytic centre of the ΦKZ nvRNAP involved in transcription bubble formation were similar to those in bacterial msRNAPs (Figure 5.22 A), except for the rudder of the β' subunit, which was absent in the homologous nvRNAP subunit gp55 (Figure 5.18).

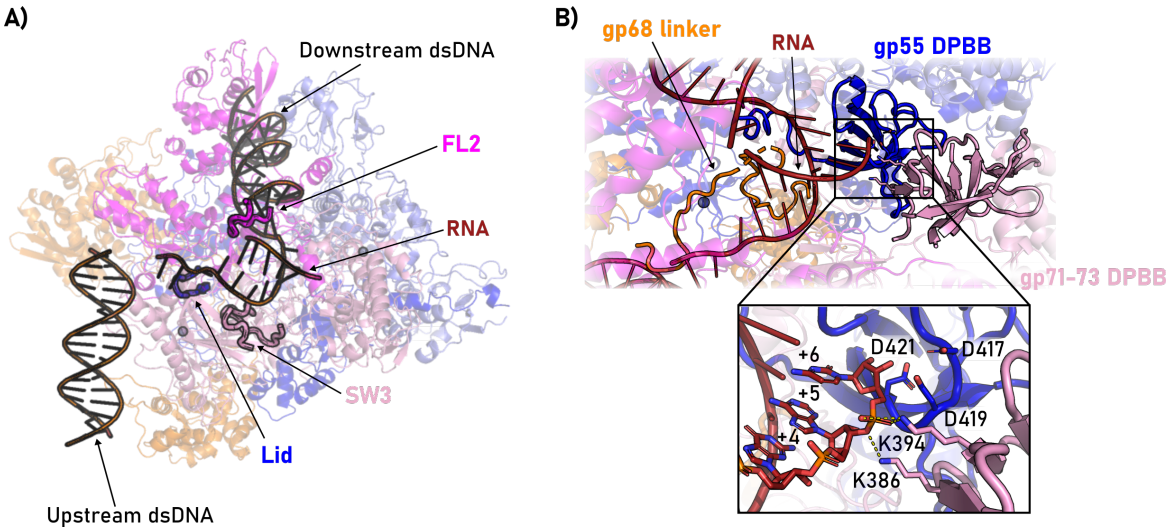


Figure 5.22: The catalytic center of the ΦKZ nvRNAP complex bound to the P119L RNA/DNA promoter. A) Structural elements involved in the transcription bubble formation in the ΦKZ nvRNAP. Each subunit is depicted with a different colour: gp123 in magenta, gp71-73 in pink, gp55 in blue, gp74 in light blue and gp68 in orange. B) The catalytic core of the ΦKZ nvRNAP. The two DPBBs are contributed by gp55 and gp71-73 and the gp68 linker invades the catalytic cleft. Inset shows the aspartate triad (^{417}DFDGD^{421}) contacting the last residue transcribed (+6) and the two universally conserved lysines (K386 and K394) contacting the RNA transcript in the upstream nucleotides.

As we observed in the ΦKZ nvRNAP structure without DNA (chapter 5), the catalytic cleft was formed by the two conserved *double-ψ-β-barrel* (DPBB) domains found in gp55 and gp71-73. The universally conserved motif ^{417}DFDGD^{421} was included in gp55 forming a similar loop to that found in *E.coli* transcribing complexes. The three aspartates contained in this motif are meant to coordinate the first catalytic Mg^{2+} needed for the nucleotide addition reaction. The three aspartates are located as if the Mg^{2+} would be contacting the recently added nucleotide (3'), however the ion was not visible at the resolution of our map. The other DPBB found in gp71-73 contained two positively charged lysines (K386, K394) that interacted with the nascent RNA. At the catalytic cleft, we could also observe the gp68 linker forming a loop in a similar manner to the bacterial $\sigma_{3.2}$ loop (Figure 5.22 B), which will be later discussed in this chapter.

5.6 The role of Gp68

Previous biochemical studies demonstrated that the fifth subunit of the ΦKZ nvRNAP, gp68, is required to assemble a functional enzyme and is essential for promoter recognition and transcription initiation. Furthermore, structural analysis of the nvRNAP holoenzyme suggested that gp68 had probably evolved from an ancestral bacterial σ factor (de Martín Garrido et al., 2021b; Yakunina et al., 2015; Orekhova et al., 2019). Resolving the structure of the ΦKZ nvRNAP complex bound to a specific P119L RNA/DNA template allowed us to partially build the gp68 linker and to dock a structural model for the gp68 N-terminus. We observed that the gp68 subunit was formed by two domains, the CTD (314-521) and the *N-terminal domain* (NTD)(1-272) which were physically separated by a linker (273-313). The overall structure of gp68 and its position within the complex resembled that of bacterial σ factors (Figure 5.23)

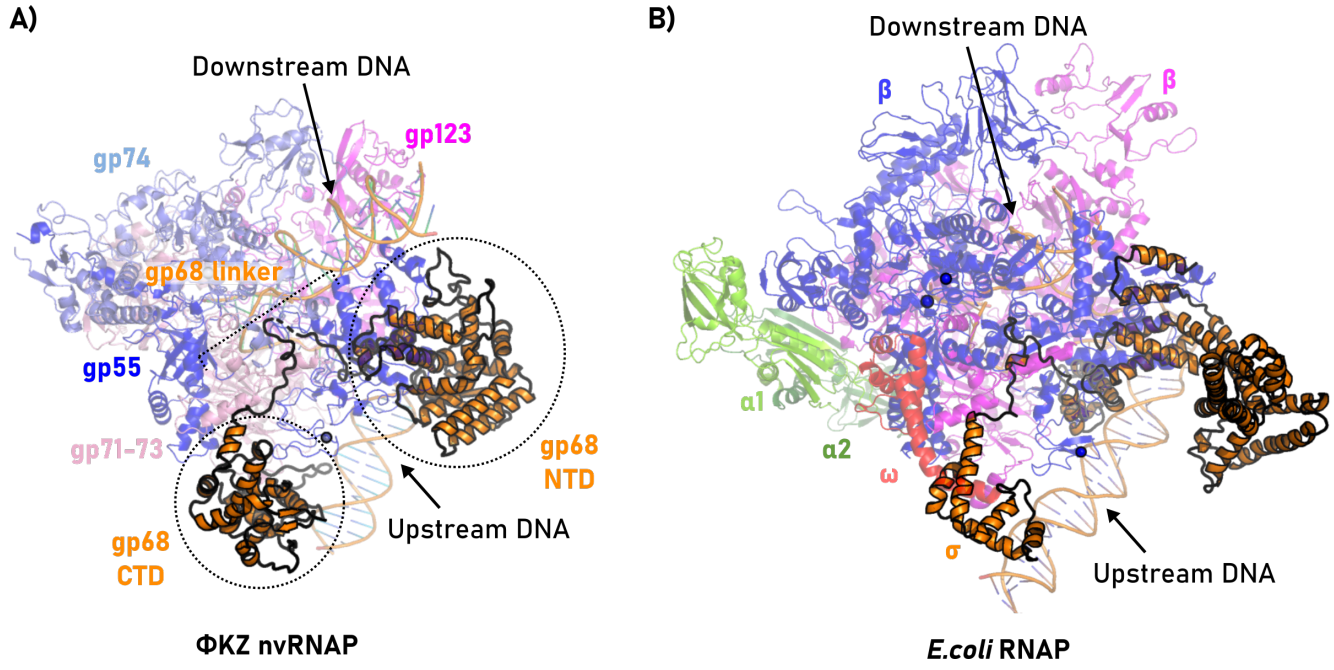


Figure 5.23: The gp68 architecture and position within the complex is similar to bacterial σ factors. A) Structure of the ΦKZ nvRNAP complex bound to the P119L RNA/DNA template with gp68 highlighted. B) *E.coli* RNAP transcription initiation complex with the σ factor highlighted (PDB 4YLP).

As we previously observed in the ΦKZ nvRNAP structure without DNA, the gp68 CTD domain corresponded to the bacterial σ region 4, however, we observed that the gp68 CTD was more expanded and established a larger interaction with the rest of the complex making extensive contacts with gp71-73 and gp55 (de Martín Garrido et al., 2021b)(Chapter 4). Region 4 of the bacterial σ factor is a mostly helical domain that interacts with the conserved -35 element of bacterial promoters through an *helix-turn-helix* (HTH) motif (Murakami et al., 2002; Campbell et al., 2002). According to our structure, the gp68 CTD was in a position suitable to interact with the upstream DNA through two different interfaces. The first contact involved a disordered loop (415-427) that was positioned in a similar position within the complex than the helix-turn-helix motif of σ (Figure 5.24). The second putative gp68-DNA contact involved an extended loop (445-460), which would contact the DNA in a more downstream region through its tip. The electrostatic potential and the amino acid composition of both the disordered loop and the extended loop (rich in positively charged and hydrophobic amino acids), supported a role of these two gp68 CTD regions in DNA binding. Due to the limited resolution of our map in some regions, we could not assign sequence to the upstream DNA, hence assignment of specific promoter regions involved in interactions with the gp68 CTD was not achievable.

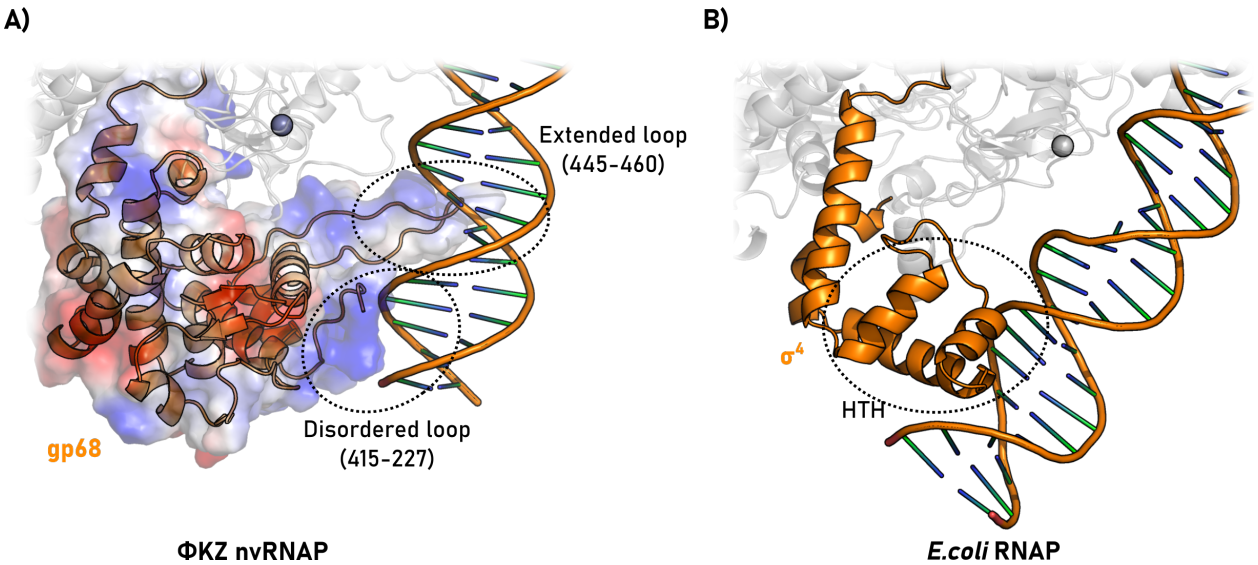


Figure 5.24: Structural comparison of the gp68 CTD with bacterial σ^4 region. A) The ΦKZ nvRNAP gp68 CTD structure. The electrostatic potential is depicted as a transparent surface with blue regions representing positively charged regions, white neutral, and red negatively charged regions. The black dotted line encircles the two putative DNA binding interfaces of the gp68 CTD. B) *E.coli* σ region 4 (PDB 4YLP). The helix-turn-helix motif is encircled with a black dotted line.

At the active center of the nvRNAP, we observed that the gp68 linker invaded the catalytic cleft in a similar manner to the $\sigma^{3.2}$ loop. The tip of the gp68 linker was disordered in our structure, however it ran towards the 5' end of the RNA transcript, in a similar manner to the bacterial σ finger (Figure 5.25). At the end of the gp68 linker, before starting the gp68 NTD, we found a loop (274-280) that interacted with the consensus sequence of the P119L promoter (Figure 5.25). We found that Thr278 of gp68 specifically interacted with nucleotide -4 in the template strand (relative to the transcription start site). Interestingly, this loop is at the same position as a HTH motif found in σ region 3.1 (Figure 5.25), which binds to an extension of the -10 element.

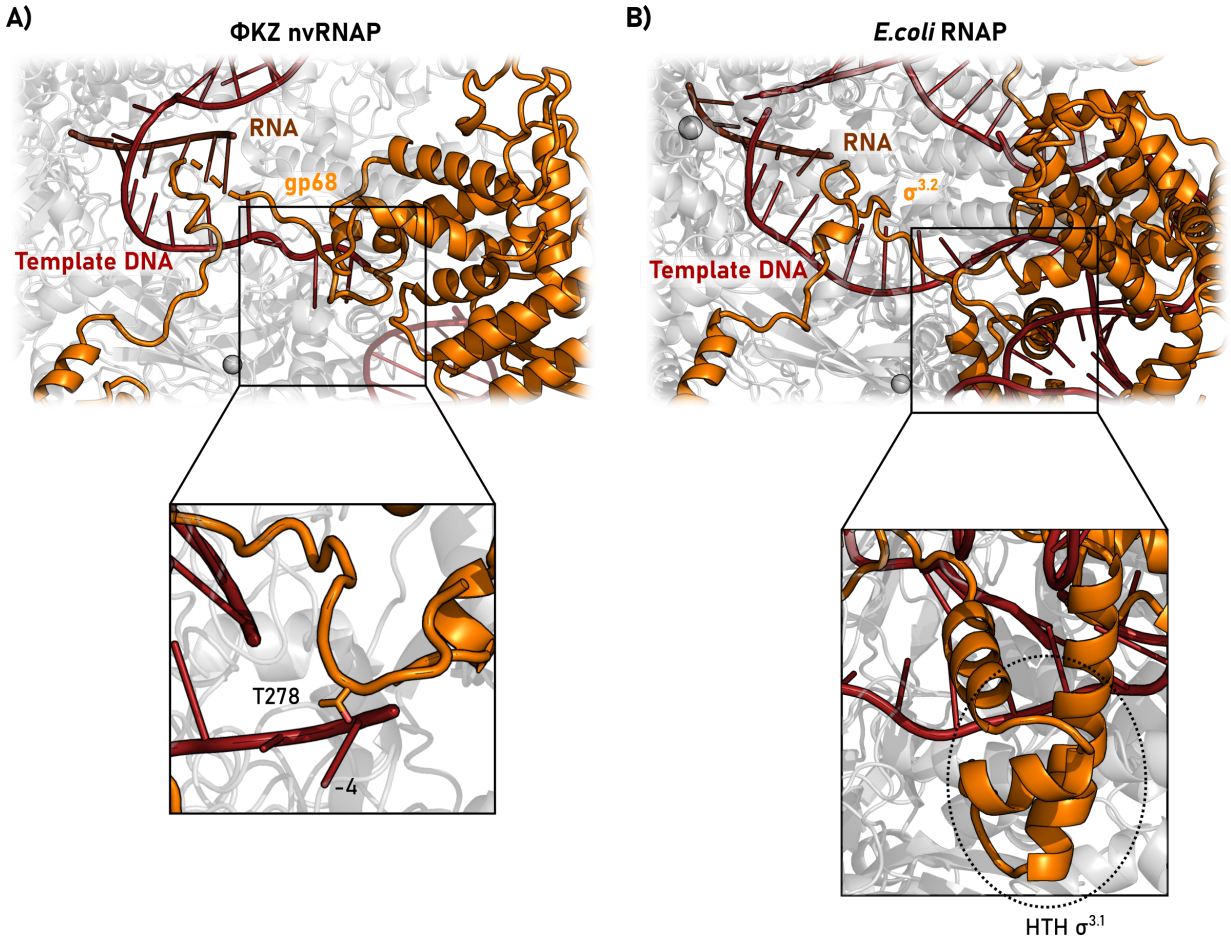


Figure 5.25: Structural comparison of the gp68 linker with region 3 of bacterial σ factors. A) The structure of the ΦKZ nvRNAP gp68 linker in orange, the RNA in brown and the DNA in firebrick red. Inset showing the loop region. B) *E.coli* σ factor region 3 (PDB 4YLP) in orange, RNA in brown and DNA in firebrick red. Inset showing the HTH motif comprised in the σ region 3.1.

The N-terminus of gp68 was not resolved at high-resolution in the ΦKZ nvRNAP-P119L RNA/DNA template structure, however we could dock an accurate structural model generated with Alphafold2 (Jumper et al., 2021). According the to the model, the gp68 NTD was located in a similar position to region 2 of bacterial σ factors. Moreover, the N-terminus of gp68 would interact with the two α helices of the clamp, which are a binding surface for σ factors in eubacterial msRNAPs (Lane and Darst, 2010a). Although the topology of the gp68 NTD is quite divergent from that of bacterial σ factors, the electrostatic potential surface of that domain supported a role in DNA binding (Figure 5.26).

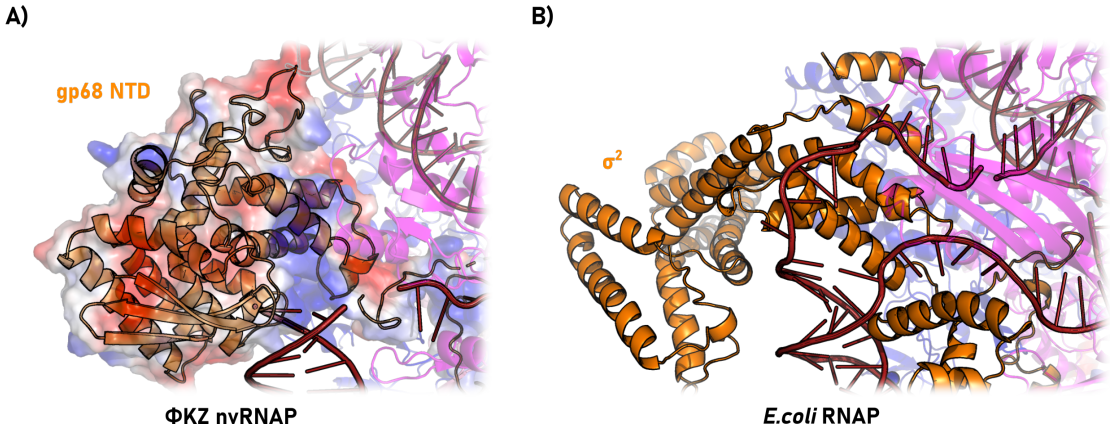


Figure 5.26: Structural comparison of the gp68 NTD with bacterial σ region 2. A) Structural model of the gp68 NTD of the ΦKZ nvRNAP. Electrostatic potential is represented as a transparent surface with blue regions representing positive charges, red representing negative and white neutral. gp68 is depicted in orange and DNA in firebrick red. B) Region 2 of the *E.coli* σ^{70} (PDB 4YLP). σ is depicted in orange and DNA in firebrick red.

5.7 Structural comparison of the ΦKZ and AR9 nvRNAPs

By the time we had obtained the structure of the ΦKZ nvRNAP complex bound to the P119L RNA/DNA template, a pre-print of the AR9 nvRNAP structure bound to its promoter had been released (Fraser et al., 2021). Similarly to bacteriophage ΦKZ, AR9 encodes two multi-subunit RNAPs that are sufficient to transcribe all genes needed for the assembly of new viruses (Lavysh et al., 2016; Sokolova et al., 2017). The AR9 nvRNAP is formed by four subunits with homology to the two largest eubacterial RNAP subunits, β and β', and a fifth subunit, gp226, with no homology to any other protein with known function which appears to be crucial for transcription (Sokolova et al., 2017; Lavysh et al., 2016). Considering the common features of the nvRNAPs encoded by these two distantly related bacteriophages, in the following section we performed an extensive comparison of the structural information available for these two non-canonical msRNAPs.

Before making any comparisons, it is important to note that our ΦKZ nvRNAP-DNA structure is bound to an RNA/DNA template based on its specific promoter P119L containing a 4-nucleotide transcription bubble at the transcription start point, whereas the AR9 nvRNAP structure was bound to a 3'-overhang dsDNA template mimicking one half of the transcription bubble, without any RNA present (Fraser et al., 2021). Overall, the two nvRNAPs had a similar shape, displaying the typical "crab-claw" architecture of bacterial msRNAPs but having a smaller appearance, probably due to the presence of split β/β'-like subunits. The AR9 nvRNAP also contains the conserved structural elements that govern the reaction cycle, except for the β' rudder.

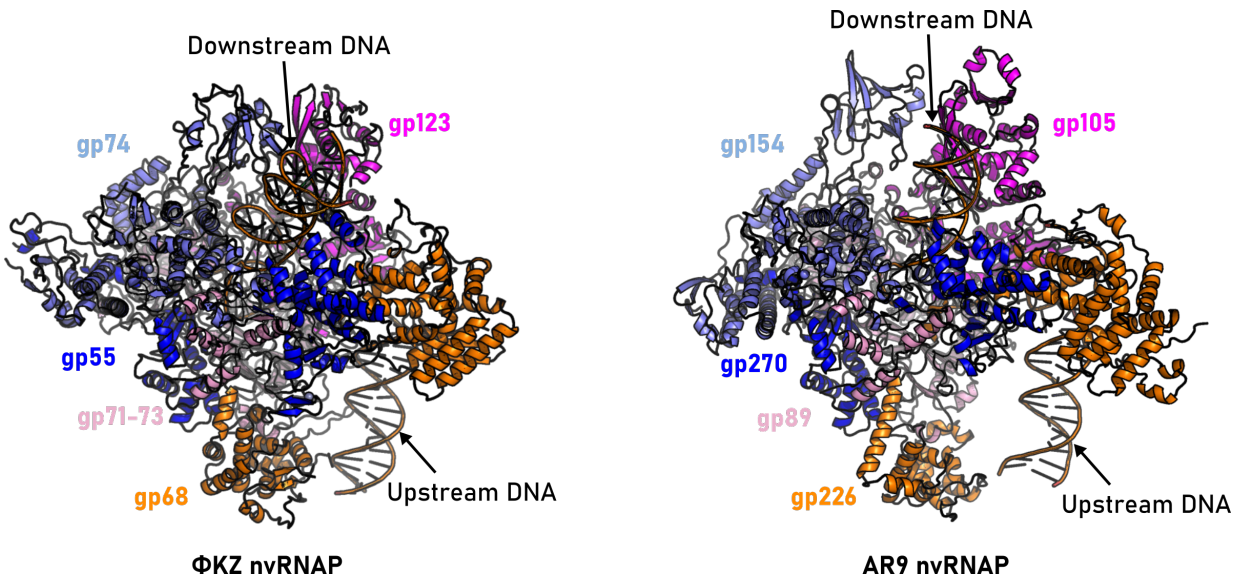


Figure 5.27: Structural comparison of the ΦKZ and AR9 nvRNAPs. (Left) Structure of the 3.3 Å ΦKZ nvRNAP complex bound to the P119L RNA/DNA template including a docked model of the upstream DNA and the N-terminus of gp68. Each subunit is coloured as indicated. (Right) Structure of the AR9 nvRNAP (Fraser et al., 2021). Each subunit is coloured following the same colour-code than for the ΦKZ nvRNAP for comparison.

The AR9 subunit homologous to the C-terminus of the bacterial β' subunit contained an insertion after the trigger loop (referred to as TL domain in the AR9 pre-print), as we observed for ΦKZ gp74. Superimposition of the two trigger loop insertion domains showed that they had a nearly identical fold (C_α RMSD 2.4 Å), with most secondary structure elements being found at the same place (Figure 5.28). We found that the gp74 insertion was in a different position to that in the AR9 structure, probably because both structures were obtained with DNA templates mimicking different transcription intermediates. Intriguingly, we observed that the Zinc finger domain located at the N-terminus of the β'-like subunit gp55 was absent in the AR9 equivalent subunit gp270.

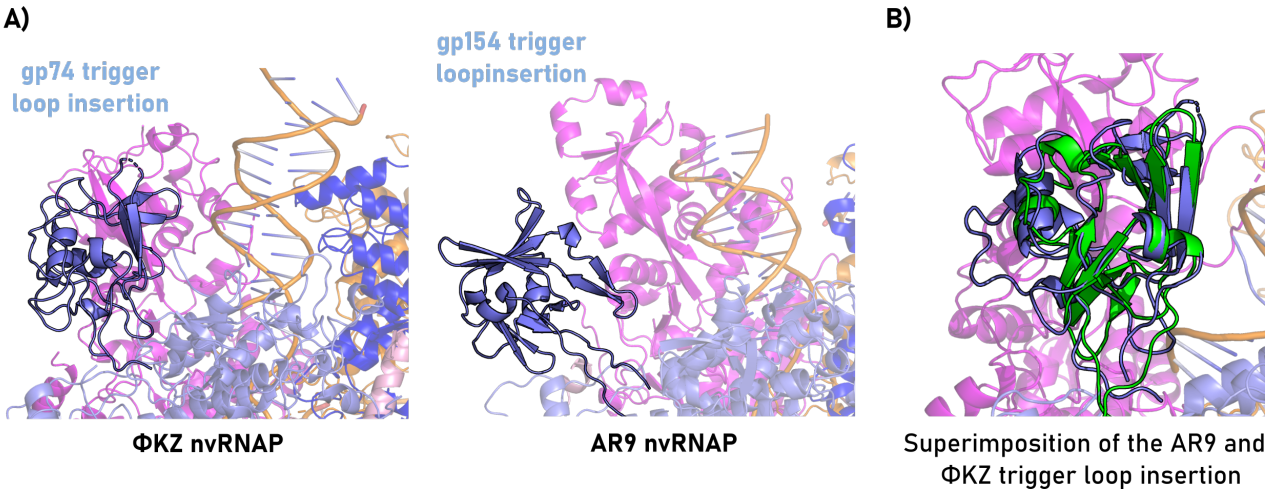


Figure 5.28: Structural comparison of the ΦKZ and AR9 nvRNAP trigger loop insertions. A) (Left) The ΦKZ nvRNAP structure bound to the P119L RNA/DNA template with the trigger loop insertion highlighted. (Right) The AR9 nvRNAP structure bound to its promoter with the trigger loop insertion highlighted. The figure shows how the position of the trigger loop insertion depends on the position of the DNA. B) Superimposition of the AR9 trigger loop insertion (395-511) (green) on the ΦKZ trigger loop insertion (396-533) (blue).

Although the percentage of sequence similarity between them was not extremely high (18%)(Figure 5.29), the promoter recognition subunit of AR9, gp226, had the same structural organisation as that of the ΦKZ gp68: two helical domains, the NTD and CTD, connected by a linker. The CTDs of the two promoter recognition subunits of the ΦKZ and AR9 nvRNAPs (gp68 and gp226, respectively) were found to be structurally similar (C_α RMSD 3.651 Å when residues 262-461 of gp226 were superimposed on residues 274-521 of gp68). Both domains were located in the equivalent position of region 4 of bacterial σ factors, however the bacteriophage subunits were more expanded as previously observed (see Chapter 4) (de Martín Garrido et al., 2021b).

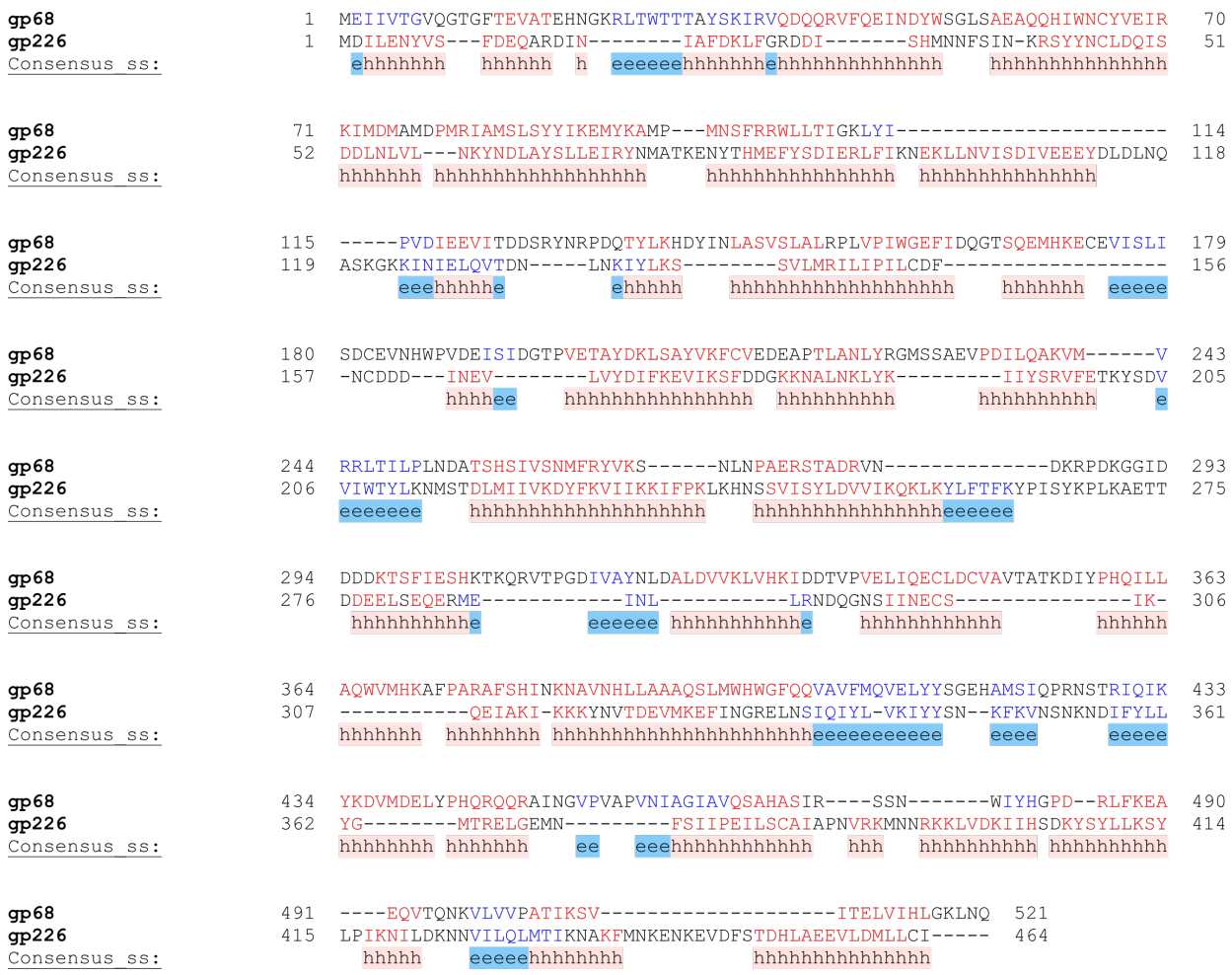


Figure 5.29: Sequence alignment of the ΦKZ and AR9 nvRNAP promoter recognition subunits. The results were computed with PROMALS3D. Residues in blue represent β-sheets whereas residues in red represent α helices. The last line represents the consensus secondary structure following the same colour code.

The CTD of AR9 gp226 contacted the DNA at approximately region -35 through a loop (386-395) (Fraser et al., 2021) in contrast to region 4 of bacterial σ factors that utilizes a HTH motif (Murakami et al., 2002; Campbell et al., 2002). The authors termed this CTD gp226 region as the pseudo-$^{-35}$element-binding motif (Fraser et al., 2021). The equivalent region in ΦKZ gp68 was a disordered loop rich in hydrophobic residues (414-428) that could potentially interact with the upstream DNA (Figure 5.30). Whereas in the AR9 nvRNAP structure the only DNA binding region of the promoter-recognition subunit CTD was the pseudo-$^{-35}$element-binding motif, in the ΦKZ gp68 CTD we could observe a second putative interface through the tip of an extended loop (442-462). Although we could not draw any conclusions about the nature of this interaction because the upstream the DNA is low resolution in our structure, the electrostatic potential of the loop supported a role in DNA binding (Figure 5.24).

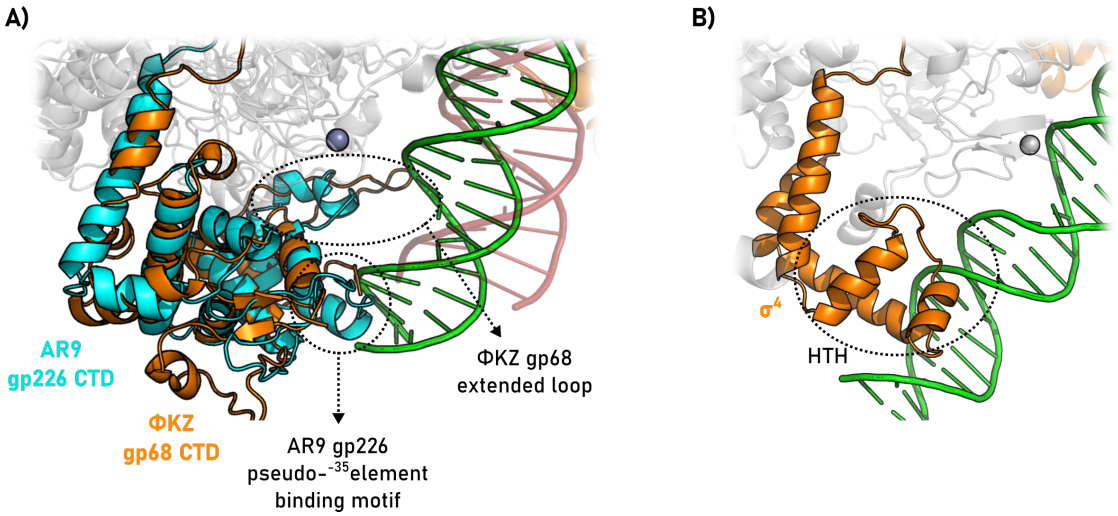


Figure 5.30: Structural comparison of the ΦKZ and AR9 nvRNAP promoter recognition subunit CTDs. A) The AR gp226 (cyan) CTD (Fraser et al., 2021) superimposed on the ΦKZ gp68 CTD (orange). The gp68 extended loop and the pseudo-$^{-35}$element-binding motif are encircled with black dotted line. The upstream DNA of the ΦKZ nvRNAP-DNA is depicted in green and the upstream DNA of the AR9 nvRNAP-DNA structure in firebrick red. B) σ^{70} region 4 (orange) of an *E.coli* transcribing complex. Upstream DNA is depicted in green.

The NTD of gp68 was not resolved at high-resolution in the ΦKZ structure bound to the P119L RNA/DNA template (Figure 5.14), however we were able to generate a model with Alphafold2 (Jumper et al., 2021) that we could accurately fit into the map. The overall structure of the model was similar to the AR9 gp226 NTD and was located in the same position within the rest of the complex. Although the sequence similarity between gp68 and gp226 is low (18%) (Figure 4.23), the resemblance of their three-dimensional structures suggested that gp68 NTD could also be involved in promoter recognition and bubble formation. The authors of the AR9 structure found that the closest structural homologue to the AR9 gp226 was the *E.coli* σ^{E} factor in region 2 and 4 of σ, as well as the σ finger, with an RMSD of 4.2 Å when superimposed onto residues 184-316 (Fraser et al., 2021). Most of the superimposed residues belonged to the NTD of gp68 which was not well-resolved in our structure, we therefore could not perform a direct comparison.

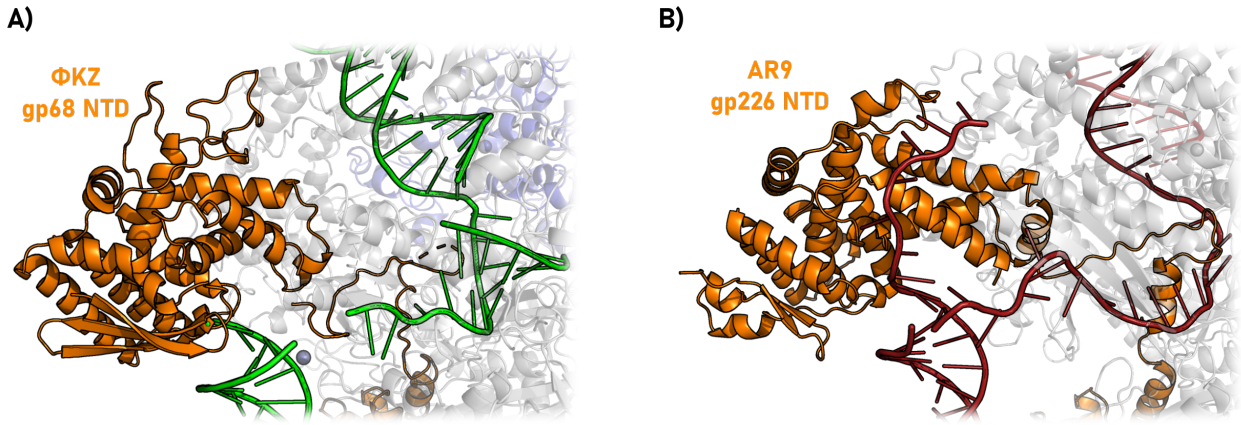


Figure 5.31: Structural comparison of the ΦKZ and AR9 nvRNAP promoter recognition subunit NTDs. A) The ΦKZ gp68 NTD structural model (orange). B) The AR9 gp226 NTD structure (Fraser et al., 2021) specifically interacting with the DNA and stabilising the transcription bubble.

5.8 Discussion

We have previously determined the structure of the ΦKZ nvRNAP revealing that the fifth subunit, essential for promoter recognition is embedded within the complex, and probably evolved from an ancestral σ factor (de Martín Garrido et al., 2021b)(Chapter 4). In bacteria, transcription initiation requires the binding of a σ factor to specifically initiate transcription from bacterial promoters (Saecker et al., 2011). Apart from recognising promoter elements, σ factors are also involved in the stabilisation of the open promoter complex, stimulation of RNA synthesis and interaction with other transcriptional activators. Bacteria usually encode a principal σ factor responsible for transcribing most of the constitutive genes and also contain alternative σ factors which transcribe genes involved in specific functions, such as virulence-associated genes or those involved in stress response (Paget and Helmann, 2003; Browning and Busby, 2004). During elongation, when RNA transcripts reach 12-15 nucleotides in length, σ factor dissociates from the msRNAP core ($\beta\beta'\alpha 2\omega$) (Mooney et al., 2005). In this chapter, we have presented a high-resolution structure of the ΦKZ nvRNAP bound to an RNA/DNA template based on its specific promoter P119L, revealing interesting insights and evolutionarily relationships of these non-canonical msRNAPs.

The structure shows that the DNA is loaded in a similar manner to bacterial msRNAPs, using universally conserved structural elements found in the split β/β'-like subunits. We noted that the only element that was not conserved in the ΦKZ nvRNAP was the rudder, found within the clamp and involved in the stabilisation of the unwound DNA in transcription elongation complexes, suggesting that other parts of the enzyme are probably involved in this function. The presence of nucleic acids revealed the structure of novel domains that were weakly-resolved in the ΦKZ nvRNAP structure without DNA (PDB 7OGP) (de Martín Garrido et al., 2021b), including the clamp, the gp74 trigger loop insertion and the gp123 middle domain, equivalent to the bacterial $\beta 2$ domain. All these regions are probably involved, to some extent, in the stabilisation of the open promoter complex. We therefore suggest that they probably transition between different conformations, adopting a closed conformation, being closer to the core of

the enzyme, that stabilises the transcription bubble in the presence of DNA. The gp123 middle domain structure was closer to the *M.tuberculosis* $\beta 2$ domain. *M.tuberulosis* is a member of the *Actinobacteria* class which, unlike *E.Coli*, lacks any lineage-specific insertion in the RNAP β subunit (Lane and Darst, 2010b; Hubin et al., 2017), hence explaining why gp123 middle domain is identified as structurally closer to *M.tuberculosis*. Conversely, similarly to *E.coli*, the ΦKZ nvRNAP gp74 subunit has a trigger loop insertion, however the structure of this insertion is not similar to any other bacterial msRNAP structure. These observations are in line with previous structural comparisons showing that each nvRNAP subunit was structurally similar to different bacterial subunits (see Chapter 4), hence suggesting that the four β/β'-like subunits are distant homologues of the bacterial subunits.

At the catalytic core of our ΦKZ nvRNAP structure bound to the P119L template, we could observe a 6 bp RNA transcript. The DNA/RNA template used to prepare nvRNAP bound to DNA initially contained an RNA scaffold of 5 bp (Figure 5.5), and we added ATP, CTP and GTP so that transcription will occur until the first A encountered at the template strand, incorporating +3 nucleotides and resulting in a final 8-nucleotide RNA transcript. We speculate that we could be capturing structures within two different scenarios. On the one hand, we could be capturing an intermediate of abortive initiation. In bacterial msRNAPs, it is known that short RNA/DNA hybrids ($<$ 10 nucleotide transcript) are unstable and usually result in abortive transcription due to accumulated stress (Yuhong Zuo et al., 2015; Goldman et al., 2009). However, abortive initiation would not yield a single addition at high levels, so we should be detecting other intermediates (+2 or +3 transcribed nucleotides) during sorting in single-particle EM analysis, which was not the case for ΦKZ-DNA complexes. On the other hand, in our structure it is noticeable that, unlike the downstream dsDNA, the upstream dsDNA and the N-terminus of gp68 were resolved at a lower resolution than the core of the enzyme, hence suggesting some flexibility in these regions. Moreover, we have previously observed thatΦKZ nvRNAP-DNA complexes were not sufficiently stable on the grid. During data processing, we noted that classes without the gp68 NTD had the same RNA length, we therefore speculate that it is more plausible that with our P119L RNA/DNA template, the nvRNAP is experiencing promoter release after a single nucleotide addition, and that the stabilisation of nvRNAP-DNA

complexes would need further exploration.

Although the resolution of the upstream DNA region and the N-terminus of gp68 was lower than that of the core of the enzyme, we could dock structural models for those two elements. The fifth subunit of the ΦKZ nvRNAP is formed by two helical domains, CTD and NTD separated by a flexible linker. The gp68 CTD preserves the same architecture with and without nucleic acids (see Chapter 4)(de Martín Garrido et al., 2021b) and probably contacts the upstream DNA in two different regions through a disordered loop and an extended loop. According to our structural model, the gp68 NTD is located in the same position within the complex as region 2 of other bacterial σ^{70}-like factors. Region 2 of bacterial σ factors specifically interacts with the -10 element of bacterial promoters, hence participating in promoter recognition and stabilisation of the open promoter complex. Considering both the structural model and the position within the complex of gp68, we suspect that the gp68 NTD would be involved in promoter recognition and transcription bubble formation. The gp68 linker and the NTD were completely disordered in the nvRNAP holoenzyme structure without DNA and, although the resolution was substantially improved, the NTD was only resolved to intermediate resolution in the ΦKZ nvRNAP structure bound to the RNA/DNA P119L-template, which suggested that it is probably a mobile element. Previous limited proteolysis results showed that full-length gp68 was protected from proteolysis in the presence of DNA and that such protection was stronger with longer dsDNAs (de Martín Garrido et al., 2021b) (Chapter 4). We therefore propose that the gp68 linker and NTD would transition from an open state, further from the nvRNAP core, to a closed conformation during promoter recognition, playing a crucial role in promoter recognition and bubble formation and stabilisation.

The overall architecture of the ΦKZ nvRNAP complex bound to the P119L RNA/DNA promoter resembles that of the AR9 nvRNAP bound to its specific promoter, which was recently released (Fraser et al., 2021). Both nvRNAPs have split subunits homologous to the two largest bacterial subunits, β and β', conserving most of the structural elements that are involved in the transcription reaction cycle. Notably, both nvRNAPs contain an insertion in the trigger loop with no homology to any bacterial msRNAP subunits. Moreover, the rudder is also absent in the AR9 nvRNAP. The fact that the split subunits of the AR9 and the ΦKZ nvRNAP are

similarly divergent from bacterial msRNAPs subunits, supports the notion that more giant bacteriophage may have evolved from a similar ancestor to encode their own non-canonical msRNAPs.

The promoter recognition subunit is also structurally similar in the ΦKZ and the AR9 nvRNAP. Both subunits are positioned in the same location within the complex and display the same domain organisation: a CTD and an NTD separated by a flexible linker. Similarly to our observations in the ΦKZ nvRNAP, the gp226 NTD was not visible in the AR9 nvRNAP structure without DNA, suggesting that this domain probably transitions from an open to closed conformation during promoter recognition. The consensus sequence of promoters recognised by AR9 nvRNAP is located between positions -11 and -8 relative to the transcription start site, while the consensus sequence of late promoters recognised by the ΦKZ nvRNAP overlaps with the transcription start site (5'-TATG-3'). Moreover, unlike ΦKZ, the AR9 genome contains uracils in the template strand. The authors showed that the AR9 promoter-recognition subunit recognises an A-rich motif of the non-template strand upstream of the consensus region and forms a specific pocket for the U-rich consensus region together with the β-like subunit gp105, providing a specific mechanism for promoter recognition and bubble formation (Fraser et al., 2021). Due to the low resolution of our map in the gp68 NTD region, including the connection of the transcription bubble with the upstream DNA, is still unclear how this subunit specifically recognises ΦKZ late promoters.

The CTD of gp68 has two putative binding interfaces with the DNA in comparison to the unique pseudo-$^{-35}$element-binding motif found in the promoter-recognition subunit of AR9. We could not observe an equivalent of the gp68 extended loop in the gp226 subunit of AR9, which suggested that this interaction was likely to be specific to the ΦKZ nvRNAP. Due to the low resolution of our map in the upstream DNA section, we could not not establish which specific nucleotides were involved in this interaction, however from the comparison with the AR9 structure we surmise that the gp68 extension loop would contact the DNA in a more downstream sequence than the pseudo-$^{-35}$element-binding motif of AR9 gp226. Unlike bacterial promoters, which have a conserved -35 element, neither AR9 promoters nor ΦKZ promoters recognised by their respective nvRNAPs show sequence conservation in regions interacting with

the promoter-recognition CTD, which suggests that the interaction is likely to be novel and sequence independent. We found only one distinct feature between the β/β'-like subunits of AR9 and ΦKZ nvRNAP. Unlike ΦKZ gp55, the β'-like subunit of AR9 gp270 lacks the Zn^{2+} binding domain conserved at the N-terminus of bacterial β' subunits. This Zn^{2+} binding domain is conserved in all multi-subunit msRNAPs and is involved in promoter recognition.

Considering the differences between the DNA binding interface in the promoter-recognition subunit, the absence of the N-terminal β' Zn^{2+} binding domain in the AR9 nvRNAP, and the low similarity between promoters recognised by the ΦKZ nvRNAP, we suspect that the promoter recognition mechanism of the ΦKZ and AR9 nvRNAPs must be divergent. The high level of similarity between the four β/β'-like subunits of the AR9 and ΦKZ nvRNAP, compared to the greater degree of divergence between their promoter recognition subunits both in sequence and, to a lesser extent, structure, is in line with our hypothesis of the σ subunit being transferred initially to the bacteriophage genome and being mutated to favour viral promoter recognition, hence being highly diverged from bacterial σ. Genes encoding the β/β'-like subunits would have been transferred later, thus explaining the greater degree of divergence of the promoter recognition subunit with respect to one other.

Bacteriophage AR9 infects Gram positive *B.subtilis*, which is not closely related to the ΦKZ host Gram negative *P.aeruginosa*. Moreover, AR9 contains uracil in their genomic DNA. The fact that these two apparently distant bacteriophages encode similar transcriptional machinery suggests that this behaviour could extend to more giant bacteriophages. However, we speculate that each promoter-recognition subunit would have probably evolved to meet the promoter requirements of each bacteriophage genome. Unlike AR9, bacteriophage ΦKZ forms a proteinacious "pseudo-nucleus" which protects the DNA during infection (Chaikeeratisak et al., 2017a,b; Danilova et al., 2020; Mendoza et al., 2020). This shield could promote a special rearrangement to facilitate promoter recognition which would also explain some divergences observed between their promoter-recognition subunits and why, unlike AR9 (Sokolova et al., 2017), the ΦKZ nvRNAP does not remain catalytically active in the absence of the fifth subunit. Moreover, we suspect that ΦKZ nvRNAP subunits might possibly be imported individually into the pseudonucleous where they would be further assembled into a functional complex. Unlike bacterial

σ factors, the fact that the promoter recognition subunit is unique and is integrated into the complex might provide an advantage for bacteriophage fitness, thus minimising the amount of DNA needing to be packed and the quantity of genes needing to be transcribed.

Overall, we suggest that AR9 and ΦKZ, and probably other giant bacteriophages, have developed a minimal but elegant five-subunit polymerase without canonical α and ω subunits that can transcribe from late phage promoters that only contain a small consensus region using a novel promoter-recognition subunit embedded in the complex. The fact that the fifth subunit is integrated into the complex and contacts the DNA in promoter regions that do not show sequence conservation, supports our previous hypothesis that, unlike in bacterial msRNAPs, the exchange of σ-like factors became redundant in bacteriophage ΦKZ and AR9 life cycles. The overall structure of the two non-canonical nvRNAPs is similar between them, conserving the main structural features of eubacterial msRNAPs, and having the promoter-recognition subunit in the same location. However, we observed that the CTD domain of the promoter-recognition of the ΦKZ nvRNAP establishes a second DNA binding interface absent in the subunit of AR9 through an extended loop. Furthermore, promoters recognised by these two non-canonical polymerases are very different. From these observations, we propose that despite having developed a very similar nvRNAP containing a novel promoter-recognition subunit with no homology to any other protein with known function, AR9 and ΦKZ nvRNAPs are likely to have slightly divergent mechanism for promoter recognition and transcription initiation. It would be interesting to design a DNA template from the ΦKZ P119L promoter that would mimick the current AR9-DNA structure to perform an exhaustive comparison between these interesting polymerases.

Chapter 6

Conclusions and outlook

6.1 Conclusions

6.1.1 The Folliculin complex and its interacting partners

FLCN is the gene responsible for BHD syndrome, an hereditary disease that predisposes carriers to the appearance of fibrofolliculomas, lung cysts, and an increased risked of developing kidney cancer. The FLCN protein forms complexes with two larger FNIP proteins, FNIP1 and FNIP2, and together they regulate Rag and Rab GTPases. The Rag GTPases are responsible for the recruitment of mTORC1 to the lysosome. mTORC1 is a central regulator of metabolism that ensures that cells only grow under optimal conditions. When nutrients are available mTORC1 is recruited to the lysosome by the Rag GTPase heterodimer where it is allosterically activated by the Rheb GTPase in the GTP-bound form. The regulation of the Rag GTPase nucleotide state involves several multiprotein complexes with GAP and GEF activity. The FLCN-FNIP2 is one of these complexes, as it possesses GAP activity for the Rag C/D GTPase, hence being a positive modulator of mTORC1.

The structure of the FLCN-FNIP2 complex bound to the Rag GTPase heterodimer and Ragulator, which anchors the GTPases at the lysosomal membrane, was published in 2019, revealing

the molecular mechanism by which the FLCN-FNIP complex binds to the Rag GTPases (Shen et al., 2019; Lawrence et al., 2019). In this work, we have presented a 5.7 Å 3D reconstruction of the *Sl* FLCN-FNIP2 complex and preliminary studies of the *Sl* FLCN-FNIP2 complex binding to the Rag GTPase heterodimer. We have demonstrated that the *Sl* FLCN-FNIP2 retains the same architecture as in the human complex, and that the complex preserves the same conformation in the absence of the Rag GTPase heterodimer. Failed attempts to obtain a stable *Sl* FLCN-FNIP2 complex bound to the Rag GTPases, reinforced the fact that the FLCN-FNIP2 interaction with the Rag GTPases is highly dependent on their nucleotide setting. Apart from the regulation of the Rag GTPases, FLCN is also involved in the modulation of the Rab GTPases, which are involved in membrane trafficking and lyososomal positioning. How FLCN coordinates all these functions is still unknown and might well contribute to understanding its role as a tumour suppressor.

6.1.2 The ΦKZ non-virion RNAP

The second part of my work was focused on the ΦKZ nvRNAP. Bacteriophage ΦKZ is a member of a family of giant bacteriophages which infects *P.aeruginosa*, a pathogen of great medical concern. The large genome of bacteriophage ΦKZ encodes many proteins that are usually provided by the host. Amongst these proteins encoded by the ΦKZ genome, there are those encoding two multi-subunit RNAPs; a virion RNAP which transcribes early bacteriophage genes, and a non-virion RNAP which transcribes late bacteriophage genes (Ceyssens et al., 2014). We have determined for the first time the structure of the ΦKZ nvRNAP on its own, and bound to a RNA/DNA template based on its specific promoter P119L.

The structure of the ΦKZ nvRNAP revealed that the overall architecture is very similar to that of bacterial msRNAPs, conserving most of the structural elements involved in the transcription reaction in four split subunits homologous to the bacterial β and β' subunits. However, the ΦKZ nvRNAP includes an extra fifth subunit, gp68, which occupies the same location within the complex as bacterial σ factors but is integrated into the complex. Biochemical studies have previously reported that gp68 is essential for promoter recognition and transcription initiation

(Yakunina et al., 2015; Orekhova et al., 2019). The structure of the ΦKZ nvRNAP bound to an RNA/DNA template based on the specific P119L promoter supports a putative role of gp68 in promoter recognition. The structures suggest that the gp68 NTD transitions between an open conformation, far from the nvRNAP core, to a closed conformation, close to the nvRNAP core, when the enzyme is bound to its specific promoter, suggesting that this domain is probably involved in promoter recognition.

Comparison of the ΦKZ nvRNAP structure bound to its promoter with that of AR9 nvRNAP bound to its specific promoter revealed that the overall architecture and the position within the complex of the promoter-recognition subunit is similar in these two non-canonical msRNAPs, suggesting that this transcriptional mechanism could be extended to more giant bacteriophages. The promoter-recognition subunit is integrated in both ΦKZ and AR9 nvRNAPs, suggesting that the exchange of alternative factors was probably redundant for the bacteriophage life-cycle. Although their promoter recognition subunits have a similar structure, they have a low percentage of sequence identity, and differences in their DNA binding interfaces. Therefore, we speculate that the σ subunit was initially acquired by the bacteriophage and was mutated substantially to specifically recognise viral promoters. Further studies of these non-canonical nvRNAPs can provide valuable information about the evolution of giant bacteriophages. Moreover, the study of the transcription machinery of bacteriophage ΦKZ will hopefully help to understand its life cycle in *P.aeruginosa*.

6.2 Future Work

6.2.1 The Folliculin complex and its interacting partners

With the publication of the human lysosomal FLCN complex structure, the molecular mechanism by which it regulates the Rag GTPase activity, which in turn modulates mTORC1, has been elucidated. However, the Rag GTPase heterodimer nucleotide state is regulated by several multiprotein complexes. How all these complexes coordinate to interact with the Rag GTPase heterodimer at the lysosomal membrane is still unknown and is probably more complicated than the current structures indicate. Tomographic studies of the lysosomal membrane under different nutrient conditions may be a suitable approach to study the Rag GTPase activation at the lysosomal membrane and to understand how they bind to all their regulators. Moreover, both FLCN and FNIP have phosphorylation sites, and the FLCN phosphorylation appears to affect complex formation, with both FNIP1 and FNIP2 preferentially binding phosphorylated FLCN (Baba et al., 2006; Takagi et al., 2008). FNIP2 contains large unstructured regions that are thought to be involved in phosphorylation events that are not visible in the cryo-EM structure (Shen et al., 2019; Lawrence et al., 2019). Generation of mutants of these phosphorylation sites could be an approach to further study their functional relevance in the FLCN-FNIP complex formation and interaction with the Rag GTPase heterodimer.

The FLCN-FNIP2 complex is also involved in other cellular functions, interacting with the Rab GTPases. As a future perspective, it would be really interesting to investigate FLCN-Rab complexes at a structural level to further understand how the FLCN-FNIP complex interacts with the Rab GTPases and coordinates this function with the regulation of the Rag GTPase heterodimer. Particularly, the FLCN interaction with RILP-Rab complexes places FLCN at the intersection between nutrient signalling and lysosomal positioning (Starling et al., 2016). Preliminary studies in our lab attempted to purify an *in vitro Sl* FLCN complex with RILP-Rab, however the complex could not be recapitulated in insect cells, suggesting that this behaviour might be not conserved in all organisms or, more likely, that there are more proteins involved

in the interaction. The *Sl* FLCN-FNIP2 complex retains the same conformation when is not bound to the Rag GTPases, a behaviour that is probably preserved in the human complex given their structural similarity. Structural studies of the FLCN interaction with RILP-Rab might reveal whether or not the same conformation is retained in the interaction with other FLCN interacting partners.

Germline mutations affecting the whole *FLCN* gene are known to be responsible for BHD syndrome. Most mutations found in BHD patients result in frameshifts (insertion/deletion), nonesense open reading frames or loss of proper mRNA splicing, causing protein truncation. The FLCN protein interacts with the Rag GTPases through its N-terminal longin domain, however mutations affecting BHD patients have been also reported in the FLCN DENN domain. To further understand how folliculin coordinates all its functions, one strategy could be the generation of FLCN mutants with most common mutations found in BHD patients to perform structural studies.

6.2.2 The ΦKZ non-virion RNAP

A complete mechanism for transcription initiation by the ΦKZ nvRNAP is still missing. We have determined the structure of this non-canonical msRNAP with and without DNA, revealing its architecture for the first time, however our structure of the ΦKZ nvRNAP bound to its specific promoter P119L still lacks high-resolution information for important parts of the complex, such as the N-terminus of gp68, the complete transcription bubble and the upstream dsDNA. Conversely, the AR9 nvRNAP structure bound to its specific promoter (Fraser et al., 2021) provides a molecular model for both the N-terminus of the promoter-recognition subunit and the upstream dsDNA.

In the future, it will be really interesting to determine the structure of the ΦKZ nvRNAP bound to its specific promoter using a DNA template as similar as possible to that used for the AR9 nvRNAP structure determination (Fraser et al., 2021) to perform a direct comparison of the two non-canonical nvRNAP structures. Moreover, it is also noticeable that we could only get the ΦKZ nvRNAP to add one nucleotide to the RNA/DNA P119L template. This observation,

combined with the flexibility of the N-terminus of gp68 and the upstream dsDNA, suggests that our structure might be capturing a promoter release intermediate. Therefore, further studies of the ΦKZ nvRNAP transcription requirements might be needed in order to obtain a stable structure bound to its promoter. The vRNAP transcribes early phage promoters, whereas the nvRNAP specifically transcribes phage late promoters (Ceyssens et al., 2014; Yakunina et al., 2015; Orekhova et al., 2019). Which polymerase transcribes middle phage promoters is still unknown, however it might be possible that a bacteriophage-encoded DNA-binding factor could specifically recruit one of the non-canonical polymerases to the middle promoters. Binding studies with ΦKZ-encoded DNA-binding factors might reveal new insights into the transcriptional machinery of bacteriophage ΦKZ.

Biochemical studies have demonstrated that gp68 is crucial for the assembly of a functional complex, revealing that when gp68 is added *in vitro* to 4s complex containing the four β/β'-like subunits, is not able to start transcription from an RNA/DNA template, unlike *in vivo* assembled 5s ΦKZ nvRNAP complexes. Structural analysis of the *in vitro* assembled complexes might reveal what happens when gp68 is added *a posteriori* and why these complexes are not functional. Furthermore, to provide a full mechanism of the transcription cycle, generation of gp68-CTD and gp68-NTD deletions and point mutations on residues that interact with the DNA would be essential to understand the role of each domain during transcription. Another approach to study the role of gp68 in transcription would be to link the NTD to the core of the polymerase to freeze transcription intermediates and study the transition from open to closed conformation of this domain.

Bibliography

Hans-W Ackermann. Tailed Bacteriophages: The Order Caudovirales. In *ADVANCES IN VIRUS RESEARCH*, volume 51, pages 135–201. 1998. doi: 10.1016/S0065-3527(08)60785-X.

Pavel V. Afonine, Ralf W. Grosse-Kunstleve, Nathaniel Echols, Jeffrey J. Headd, Nigel W. Moriarty, Marat Mustyakimov, Thomas C. Terwilliger, Alexandre Urzhumtsev, Peter H. Zwart, and Paul D. Adams. Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallographica Section D Biological Crystallography*, 68(4):352–367, 4 2012. ISSN 0907-4449. doi: 10.1107/S0907444912001308.

Byung-yoon Ahn and Bernard Moss. RNA polymerase-associated transcription specificity factor encoded by vaccinia virus (RNA polymerase subunit/a factor/RAP3O RNA polymeraebindlng protein/gene expression). *Biochemistry*, 89:3536–3540, 1992.

Madhanagopal Anandapadamanaban, Glenn R. Masson, Olga Perisic, Alex Berndt, Jonathan Kaufman, Chris M. Johnson, Balaji Santhanam, Kacper B. Rogala, David M. Sabatini, and Roger L. Williams. Architecture of human Rag GTPase heterodimers and their complex with mTORC1. *Science*, 366(6462):203–210, 10 2019. ISSN 0036-8075. doi: 10.1126/SCIENCE.AAX3939.

Christopher H S Aylett, Thierry Izoré, Linda A. Amos, and Jan Löwe. Structure of the Tubulin/FtsZ-like protein TubZ from pseudomonas bacteriophage ΦkZ. *Journal of Molecular Biology*, 425(12):2164–2173, 6 2013. ISSN 00222836. doi: 10.1016/j.jmb.2013.03.019.

Christopher H S Aylett, Evelyn Sauer, Stefan Imseng, Daniel Boehringer, Michael N Hall,

Nenad Ban, and Timm Maier. Architecture of human {mTOR} complex 1. *Science*, 351(6268):48–52, 1 2016a. ISSN 0036-8075. doi: 10.1126/science.aaa3870. URL http://www.sciencemag.org/cgi/doi/10.1126/science.aaa3870.

Christopher H.S. Aylett, Evelyn Sauer, Stefan Imseng, Daniel Boehringer, Michael N. Hall, Nenad Ban, and Timm Maier. Architecture of human mTOR complex 1. *Science*, 351(6268): 48–52, 2016b. ISSN 10959203. doi: 10.1126/science.aaa3870.

M. Baba, S.-B. Hong, N. Sharma, M. B. Warren, M. L. Nickerson, A. Iwamatsu, D. Esposito, W. K. Gillette, R. F. Hopkins, J. L. Hartley, M. Furihata, S. Oishi, W. Zhen, T. R. Burke, W. M. Linehan, L. S. Schmidt, and B. Zbar. Folliculin encoded by the BHD gene interacts with a binding protein, FNIP1, and AMPK, and is involved in AMPK and mTOR signaling. *Proceedings of the National Academy of Sciences*, 103(42):15552–15557, 2006. ISSN 0027-8424. doi: 10.1073/pnas.0603781103. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0603781103.

Masaya Baba, Mutsuo Furihata, Seung-Beom Hong, Lino Tessarollo, Diana C. Haines, Eileen Southon, Vishal Patel, Peter Igarashi, W. Gregory Alvord, Robert Leighty, Masahiro Yao, Marcelino Bernardo, Lilia Ileva, Peter Choyke, Michelle B. Warren, Berton Zbar, W. Marston Linehan, and Laura S. Schmidt. Kidney - Targeted Birt-Hogg-Dubé Gene Inactivation in a Mouse Model: Erk1/2 and Akt-mTOR Activation, Cell Hyperproliferation, and Polycystic Kidneys. *Journal of the National Cancer Institute*, 100:140–154, 2008. doi: 10.1093/jnci/djm288. URL https://academic.oup.com/jnci/article-abstract/100/2/140/1130264.

Liron Bar-Peled, Lawrence D Schweitzer, Roberto Zoncu, and David M Sabatini. Ragulator is a {GEF} for the rag {GTPases} that signal amino acid levels to {mTORC1}. *Cell*, 150(6):1196–1208, 9 2012. doi: 10.1016/j.cell.2012.07.032. URL http://dx.doi.org/10.1016/j.cell.2012.07.032.

G. Barry, C. L. Squires, and C. Squires. Control features within the rplJL-rpoBC transcription unit of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 76(10):4922, 1979. ISSN 00278424. doi: 10.1073/PNAS.76.10.4922.

Imre Berger, Daniel J. Fitzgerald, and Timothy J. Richmond. Baculovirus expression system for heterologous multiprotein complexes. *Nature Biotechnology*, 22(12):1583–1587, 2004. ISSN 10870156. doi: 10.1038/nbt1036.

Giuseppe Bertani. Lysogeny at Mid-Twentieth Century: P1, P2, and Other Experimental Systems. *Journal of Bacteriology*, 186(3):595–600, 2 2004. ISSN 0021-9193. doi: 10.1128/JB.186.3.595-600.2004.

Arthur R. Birt, Georgina R. Hogg, and W. James Dubé. Hereditary Multiple Fibrofolliculomas With Trichodiscomas and Acrochordons. *Archives of Dermatology*, 113(12):1674–1677, 1977. ISSN 15383652. doi: 10.1001/archderm.1977.01640120042005.

Hande Boyaci, James Chen, Rolf Jansen, Seth A. Darst, and Elizabeth A. Campbell. Structures of an RNA polymerase promoter melting intermediate elucidate DNA unwinding. *Nature*, 565(7739):382–385, 1 2019. ISSN 14764687. doi: 10.1038/s41586-018-0840-5. URL `https://doi.org/10.1038/s41586-018-0840-5`.

S. Brenner, F. Jacob, and M. Meselson. An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis. *Nature*, 190(4776):576–581, 1961. ISSN 1476-4687. doi: 10.1038/190576a0.

Eric J. Brown, Mark W. Albers, Tae Bum Shin, Kazuo ichikawa, Curtis T. Keith, William S. Lane, and Stuart L. Schreiber. A mammalian protein targeted by G1-arresting rapamycin–receptor complex. *Nature*, 369(6483):756–758, 6 1994. ISSN 0028-0836. doi: 10.1038/369756a0. URL `http://www.nature.com/articles/369756a0`.

Douglas F. Browning and Stephen J.W. Busby. The regulation of bacterial transcription initiation. *Nature reviews. Microbiology*, 2(1):57–65, 1 2004. ISSN 1740-1526. doi: 10.1038/nrmicro787.

Ben Burrowes, David R. Harper, Joseph Anderson, Malcolm McConville, and Mark C. Enright. Bacteriophage therapy: potential uses in the control of antibiotic-resistant pathogens. *Expert Review of Anti-infective Therapy*, 9(9):775–785, 9 2011. ISSN 1478-7210. doi: 10.1586/eri.11.90.

Elizabeth A. Campbell, Nataliya Korzheva, Arkady Mustaev, Katsuhiko Murakami, Satish Nair, Alex Goldfarb, and Seth A. Darst. Structural mechanism for rifampicin inhibition of bacterial RNA polymerase. *Cell*, 104(6):901–912, 2001. ISSN 00928674. doi: 10.1016/S0092-8674(01)00286-0.

Elizabeth A. Campbell, Oriana Muzzin, Mark Chlenov, Jing L. Sun, C. Anders Olson, Oren Weinman, Michelle L. Trester-Zedlitz, and Seth A. Darst. Structure of the bacterial RNA polymerase promoter specificity σ subunit. *Molecular Cell*, 9 (3):527–539, 2002. ISSN 10972765. doi: 10.1016/S1097-2765(02)00470-7. URL `http://www.molecule.org/cgi/content/full/9/3/`.

David Carling. The AMP-activated protein kinase cascade – a unifying system for energy control. *Trends in Biochemical Sciences*, 29(1):18–24, 1 2004. ISSN 0968-0004. doi: 10.1016/J.TIBS.2003.11.005.

Nicolas Cermakian, Tatsuya M Ikeda, Pedro Miramontes, B Franz Lang, Michael W Gray, and Robert Cedergren. On the Evolution of the Single-Subunit RNA Polymerases. *J Mol Evol*, 45:671–81, 1997.

P.-J. Ceyssens, L. Minakhin, A. Van den Bossche, M. Yakunina, E. Klimuk, B. Blasdel, J. De Smet, J.-P. Noben, U. Blasi, K. Severinov, and R. Lavigne. Development of Giant Bacteriophage KZ Is Independent of the Host Transcription Apparatus. *Journal of Virology*, 88(18):10501–10510, 9 2014. ISSN 0022-538X. doi: 10.1128/jvi.01347-14. URL `https://jvi.asm.org/content/88/18/10501` `https://jvi.asm.org/content/88/18/10501.abstract`.

Vorrapon Chaikeeratisak, Katrina Nguyen, Mackennon E Egan, Marcella L Erb, Anastasia Vavilina, and Joe Pogliano Correspondence. The Phage Nucleus and Tubulin Spindle Are Conserved among Large Pseudomonas Phages. *Cell Reports*, 20, 2017a. doi: 10.1016/j.celrep.2017.07.064. URL `http://dx.doi.org/10.1016/j.celrep.2017.07.064`.

Vorrapon Chaikeeratisak, Katrina Nguyen, Kanika Khanna, Axel F. Brilot, Marcella L. Erb, Joanna K.C. Coker, Anastasia Vavilina, Gerald L. Newton, Robert Buschauer, Kit Pogliano,

Elizabeth Villa, David A. Agard, and Joe Pogliano. Assembly of a nucleus-like structure during viral replication in bacteria. *Science*, 355(6321):194–197, 1 2017b. ISSN 10959203. doi: 10.1126/SCIENCE.AAL2130/SUPPL_FILE/CHAIKEERATISAK.SM.PDF.

Vorrapon Chaikeeratisak, Katrina Nguyen, Kanika Khanna, Axel F. Brilot, Marcella L. Erb, Joanna K.C. Coker, Anastasia Vavilina, Gerald L. Newton, Robert Buschauer, Kit Pogliano, Elizabeth Villa, David A. Agard, and Joe Pogliano. Assembly of a nucleus-like structure during viral replication in bacteria. *Science*, 355(6321):194–197, 1 2017c. ISSN 10959203. doi: 10.1126/SCIENCE.AAL2130/SUPPL_FILE/CHAIKEERATISAK.SM.PDF. URL `https://www.science.org/doi/abs/10.1126/science.aal2130`.

Vorrapon Chaikeeratisak, Kanika Khanna, Katrina T. Nguyen, Joseph Sugie, MacKennon E. Egan, Marcella L. Erb, Anastasia Vavilina, Poochit Nonejuie, Eliza Nieweglowska, Kit Pogliano, David A. Agard, Elizabeth Villa, and Joe Pogliano. Viral Capsid Trafficking along Treadmilling Tubulin Filaments in Bacteria. *Cell*, 177(7):1771–1780, 6 2019. ISSN 10974172. doi: 10.1016/j.cell.2019.05.032.

Anirban Chakraborty, Dongye Wang, Yon W. Ebright, You Korlann, Ekaterine Kortkhonjia, Taiho Kim, Saikat Chowdhury, Sivaramesh Wigneshweraraj, Herbert Irschik, Rolf Jansen, B. Tracy Nixon, Jennifer Knight, Shimon Weiss, and Richard H. Ebright. Opening and closing of the bacterial RNA polymerase clamp. *Science*, 337(6094):591–595, 8 2012. ISSN 10959203. doi: 10.1126/science.1218716. URL `https://pubmed.ncbi.nlm.nih.gov/22859489/`.

Jue Chen, Gang Lu, Jeffrey Lin, Amy L. Davidson, and Florante A. Quiocho. A tweezers-like motion of the ATP-binding cassette dimer in an ABC transport cycle. *Molecular Cell*, 12(3): 651–661, 2003. ISSN 10972765. doi: 10.1016/j.molcel.2003.08.004.

Nam Young Cho, Mieyoung Choi, and Lucia B. Rothman-Denes. The Bacteriophage N4-coded Single-stranded DNA-binding Protein (N4SSB) is the Transcriptional Activator ofEscherichia coliRNA Polymerase at N4 Late Promoters. *Journal of Molecular Biology*, 246(4):461–471, 3 1995. ISSN 0022-2836. doi: 10.1006/JMBI.1994.0098.

S Clark. Transcriptional specificity of a multisubunit RNA polymerase induced by Bacillus subtilis bacteriophage PBS2. *Journal of Virology*, 25(1), 1978.

Steve Clark, Richard Losick, and Janice Pero. New RNA polymerase from Bacillus subtilis infected with phage PBS2. *Nature*, 252(5478):21–24, 1974. ISSN 00280836. doi: 10.1038/252021a0.

Caterina Collodet, Marc Foretz, Maria Deak, Laurent Bultot, Sylviane Metairon, Benoit Viollet, Gregory Lefebvre, Frederic Raymond, Alice Parisi, Gabriele Civiletto, Philipp Gut, Patrick Descombes, and Kei Sakamoto. AMPK promotes induction of the tumor suppressor FLCN through activation of TFEB independently of mTOR. *The FASEB Journal*, 33(11):12374–12391, 8 2019. ISSN 0892-6638. doi: 10.1096/fj.201900841R. URL `https://www.fasebj.org/doi/10.1096/fj.201900841R`.

André M. Comeau, Graham F. Hatfull, Henry M. Krisch, Debbie Lindell, Nicholas H. Mann, and David Prangishvili. Exploring the prokaryotic virosphere. *Research in Microbiology*, 159(5):306–313, 6 2008. ISSN 09232508. doi: 10.1016/j.resmic.2008.05.001. URL `https://linkinghub.elsevier.com/retrieve/pii/S092325080800082X`.

Patrick Cramer. Multisubunit RNA polymerases. *Current Opinion in Colloid and Interface Science*, 12(1):89–97, 2002. doi: https://doi.org/10.1016/S0959-440X(02)00294-4.

F. H. C. Crick, Leslie Barnett, S. Brenner, and R. J. Watts-Tobin. General Nature of the Genetic Code for Proteins. *Nature*, 192(4809):1227–1232, 12 1961. ISSN 0028-0836. doi: 10.1038/1921227a0. URL `https://www.nature.com/articles/1921227a0`.

Yana A. Danilova, Viktoriia V. Belousova, Andrey V. Moiseenko, Innokentii E. Vishnyakov, Maria V. Yakunina, and Olga S. Sokolova. Maturation of Pseudo-Nucleus Compartment in P. aeruginosa, Infected with Giant phiKZ Phage. *Viruses*, 12(10):1197, 10 2020. ISSN 1999-4915. doi: 10.3390/v12101197. URL `https://www.mdpi.com/1999-4915/12/10/1197`.

Mariana E G de Araujo, Andreas Naschberger, Barbara G Fürnrohr, Taras Stasyk, Theresia Dunzendorfer-Matt, Stefan Lechner, Stefan Welti, Leopold Kremser, Giridhar Shivalingaiah, Martin Offterdinger, Herbert H Lindner, Lukas A Huber, and Klaus

Scheffzek. Crystal structure of the human lysosomal mTORC1 scaffold complex and its impact on signaling. *Science*, 358(6361):377–381, 10 2017. ISSN 0036-8075. doi: 10.1126/science.aao1583. URL `http://www.ncbi.nlm.nih.gov/pubmed/28935770` `https://www.science.org/doi/10.1126/science.aao1583`.

Maria De Luca, Laura Cogli, Cinzia Progida, Veronica Nisi, Roberta Pascolutti, Sara Sigismund, Pier Paolo Di Fiore, and Cecilia Bucci. RILP regulates vacuolar ATPase through interaction with the V1G1 subunit. *Journal of Cell Science*, 127(12):2697–2708, 2014. ISSN 14779137. doi: 10.1242/jcs.142604.

Natàlia de Martín Garrido and Christopher H.S. Aylett. Nutrient Signaling and Lysosome Positioning Crosstalk Through a Multifunctional Protein, Folliculin, 3 2020. ISSN 2296634X.

Natàlia de Martín Garrido, Michael A. Crone, Kailash Ramlaul, Paul A. Simpson, Paul S. Freemont, and Christopher H.S. Aylett. Bacteriophage MS2 displays unreported capsid variability assembling T = 4 and mixed capsids. *Molecular Microbiology*, 113(1):143–152, 1 2020. ISSN 13652958. doi: 10.1111/mmi.14406.

Natàlia de Martín Garrido, Wencheng Fu, Kailash Ramlaul, Zining Zhu, David Miller, Daniel Boehringer, and Christopher H.S. Aylett. Direct transfer of electron microscopy samples to wetted carbon and graphene films via a support floatation block. *Journal of Structural Biology*, 213(1), 2021a. ISSN 10958657. doi: 10.1016/j.jsb.2020.107677.

Natàlia de Martín Garrido, Mariia Orekhova, Yuen Ting Emilie Lai Wan Loong, Anna Litvinova, Kailash Ramlaul, Tatyana Artamonova, Alexei S Melnikov, Pavel Serdobintsev, Christopher H S Aylett, and Maria Yakunina. Structure of the bacteriophage PhiKZ non-virion RNA polymerase. *Nucleic Acids Research*, 49(13):7732–7739, 7 2021b. ISSN 0305-1048. doi: 10.1093/NAR/GKAB539.

Natàlia de Martín Garrido, Kailash Ramlaul, and Christopher H. S. Aylett. Preparation of Sample Support Films in Transmission Electron Microscopy using a Support Floatation Block. *JoVE (Journal of Visualized Experiments)*, 2021(170):e62321, 4 2021c. ISSN 1940-087X. doi: 10.3791/62321.

M. Delbru ck. Adsorption of Bacteriophage under Various Physiological Conditions of the Host. *Journal of General Physiology*, 23(5):631–642, 5 1940a. ISSN 1540-7748. doi: 10.1085/jgp.23.5.631.

M. Delbru ck. The Growth of Bacteriophage and Lysis of the Host. *Journal of General Physiology*, 23(5):643–660, 5 1940b. ISSN 1540-7748. doi: 10.1085/jgp.23.5.643.

Felix D'Herelle. Sur un microbe invisible antagoniste des bacillus dysenteriques. *Comptes rendus Acad Sci Paris*, 165:373–375, 1917.

Christian C Dibble, Winfried Elis, Suchithra Menon, Wei Qin, Justin Klekota, John M Asara, Peter M Finan, David J Kwiatkowski, Leon O Murphy, and Brendan D Manning. {TBC1D7} is a third subunit of the {TSC1}-{TSC2} complex upstream of {mTORC1}. *Mol Cell*, 47(4):535–546, 8 2012. doi: 10.1016/j.molcel.2012.06.009. URL http://dx.doi.org/10.1016/j.molcel.2012.06.009.

Mark P. Dodding. Folliculin–A tumor suppressor at the intersection of metabolic signaling and membrane traffic. *Small GTPases*, 8(2):100–105, 4 2017. ISSN 21541256. doi: 10.1080/21541248.2016.1204808.

Alexey Drozdetskiy, Christian Cole, James Procter, and Geoffrey J. Barton. JPred4: a protein secondary structure prediction server. *Nucleic Acids Research*, 43(W1): W389–W394, 7 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv332. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv332.

S ´ Ebastien Rodrigue, Roberta Provvedi, Pierre-E ´ Tienne Jacques, Luc Gaudreau, and Riccardo Manganelli. The sigma factors of Mycobacterium tuberculosis. *FEM*, 30:926–941, 2006. doi: 10.1111/j.1574-6976.2006.00040.x. URL https://academic.oup.com/femsre/article/30/6/926/494395.

Leeanna El-Houjeiri, Elite Possik, Tarika Vijayaraghavan, Mathieu Paquette, José A. Martina, Jalal M. Kazan, Eric H. Ma, Russell Jones, Paola Blanchette, Rosa Puertollano, and Arnim Pause. The Transcription Factors TFEB and TFE3 Link the FLCN-AMPK Signaling Axis

to Innate Immune Response and Pathogen Resistance. *Cell Reports*, 26(13):3613–3628, 3 2019. ISSN 22111247. doi: 10.1016/j.celrep.2019.02.102.

Emory L. Ellis and Max Delbru ck. The Growth of Bacteriophage. *Journal of General Physiology*, 22(3):365–384, 1 1939. ISSN 1540-7748. doi: 10.1085/jgp.22.3.365.

P. Emsley and K. Cowtan. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr*, 60(12):2126–2132, 11 2004. ISSN 0907-4449. doi: 10.1107/S0907444904019158.

P Emsley, B Lohkamp, W G Scott, and K Cowtan. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr*, 66:486–501, 2010. ISSN 0907-4449. doi: 10.1107/S0907444910007493. URL `http://www.python.org`.

Carola Engler, Romy Kandzia, Sylvestre Marillonnet, G Van Eldik, and J Botterman. A One Pot, One Step, Precision Cloning Method with High Throughput Capability. *PLoS ONE*, 3 (11):e3647, 2008. ISSN 1932-6203. doi: 10.1371/journal.pone.0003647.

Carola Engler, Ramona Gruetzner, Romy Kandzia, and Sylvestre Marillonnet. Golden Gate Shuffling: A One-Pot DNA Shuffling Method Based on Type IIs Restriction Enzymes. *PLoS ONE*, 4(5):e5553, 5 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0005553.

S C Falco, K Vander Laan, and L B Rothman-Denes. Virion-associated RNA polymerase required for bacteriophage N4 development. *Biochemistry*, 74(2):520–523, 1977.

S Carl Falco, William Zehring, and Lucia B Rothman-Denes+. DNA-dependent RNA polymerase from bacteriophage N4 virions. Purification and characterization. *THE JOURNAL OF BIOLOGICAL CHEMISTRY*, 255(9), 1980. doi: 10.1016/S0021-9258(19)85670-3.

Alan Fauconnier. Guidelines for Bacteriophage Product Certification. *Methods in Molecular Biology*, 1693:253–268, 2018. ISSN 10643745. doi: 10.1007/978-1-4939-7395-8_19.

Przemyslaw A. Filipek, Mariana E.G. de Araujo, Georg F. Vogel, Cedric H. De Smet, Daniela Eberharter, Manuele Rebsamen, Elena L. Rudashevskaya, Leopold Kremser, Teodor Yordanov, Philipp Tschaikner, Barbara G. Fürnrohr, Stefan Lechner, Theresia Dunzendorfer-

Matt, Klaus Scheffzek, Keiryn L. Bennett, Giulio Superti-Furga, Herbert H. Lindner, Taras Stasyk, and Lukas A. Huber. LAM TOR/Ragulator is a negative regulator of Arl8band BORC-dependent late endosomal positioning. *Journal of Cell Biology*, 216(12):4199–4215, 2017. ISSN 15408140. doi: 10.1083/jcb.201703061.

Andrei Fokine, Victor A. Kostyuchenko, Andrey V. Efimov, Lidia P. Kurochkina, Nina N. Sykilinda, Johan Robben, Guido Volckaert, Andreas Hoenger, Paul R. Chipman, Anthony J. Battisti, Michael G. Rossmann, and Vadim V. Mesyanzhinov. A three-dimensional cryo-electron microscopy structure of the bacteriophage ϕKZ head. *Journal of Molecular Biology*, 352(1):117–124, 9 2005. ISSN 00222836. doi: 10.1016/j.jmb.2005.07.018.

Andrei Fokine, Anthony J Battisti, Valorie D Bowman, Andrei V Efimov, Lidia P Kurochkina, Paul R Chipman, Vadim V Mesyanzhinov, and Michael G Rossmann. Cryo-EM Study of the Pseudomonas Bacteriophage PhiKZ. *Structure*, 15:1099–1104, 2007. doi: 10.1016/j.str.2007.07.008.

Alec Fraser, Maria L Sokolova, Arina V Drobysheva, Julia V Gordeeva, and Sergei Borukhov. Template strand deoxyuridine promoter recognition by a viral RNA polymerase. *bioRxiv*, 2021.

M. Frech, A. Wittinghofer, T. A. Darden, L. G. Pedersen, M. W. Anderson, C. K. Foley, and P. S. Charifson. Role of Glutamine-61 in the Hydrolysis of GTP by p21H-ras: An Experimental and Theoretical Study. *Biochemistry*, 33(11):3237–3244, 1994. ISSN 15204995. doi: 10.1021/bi00177a014.

Jason Gill and Paul Hyman. Phage Choice, Isolation, and Preparation for Phage Therapy. *Current Pharmaceutical Biotechnology*, 11(1):2–14, 1 2010. ISSN 13892010. doi: 10.2174/138920110790725311.

Michael L. Gleghorn, Elena K. Davydova, Lucia B. Rothman-Denes, and Katsuhiko S. Murakami. Structural Basis for DNA-Hairpin Promoter Recognition by the Bacteriophage N4 Virion RNA Polymerase. *Molecular Cell*, 32(5):707–717, 12 2008. ISSN

10972765. doi: 10.1016/J.MOLCEL.2008.11.010/ATTACHMENT/FF4E8E7B-32C0-4456-A5D9-FF81A432652B/MMC2.MPG.

Robert Glyde, Fuzhou Ye, Milija Jovanovic, Ioly Kotta-Loizou, Martin Buck, and Xiaodong Zhang. Structures of Bacterial RNA Polymerase Complexes Reveal the Mechanism of DNA Loading and Transcription Initiation. *Molecular Cell*, 70(6):1111–1120, 6 2018. ISSN 10974164. doi: 10.1016/j.molcel.2018.05.021.

Seth R Goldman, Richard H Ebright, and Bryce E Nickels. Direct Detection of Abortive RNA Transcripts in Vivo NIH Public Access. *Science*, 324(5929):927–928, 2009. doi: 10.1126/science.1169237.

Jonathan M. Goodwin, Ward G. WalkupIV, Kirsty Hooper, Taoyingnan Li, Chieko Kishi-Itakura, Aylwin Ng, Timothy Lehmberg, Archana Jha, Sravya Kommineni, Katherine Fletcher, Jorge Garcia-Fortanet, Yaya Fan, Qing Tang, Menghao Wei, Asmita Agrawal, Sagar R. Budhe, Sreekanth R. Rouduri, Dan Baird, Jeff Saunders, Janna Kiselar, Mark R. Chance, Andrea Ballabio, Brent A. Appleton, John H. Brumell, Oliver Florey, and Leon O. Murphy. GABARAP sequesters the FLCN-FNIP tumor suppressor complex to couple autophagy with lysosomal biogenesis. *Science Advances*, 7(40), 10 2021. doi: 10.1126/SCIADV.ABJ2485.

Tanja M. Gruber and Carol A. Gross. Multiple Sigma Subunits and the Partitioning of Bacterial Transcription Space. *Annual Review of Microbiology*, 57(1):441–466, 10 2003. ISSN 0066-4227. doi: 10.1146/annurev.micro.57.030502.090913.

David A. Guertin and David M. Sabatini. Defining the Role of mTOR in Cancer. *Cancer Cell*, 12(1):9–22, 7 2007. ISSN 1535-6108. doi: 10.1016/J.CCR.2007.05.008.

Dana M. Gwinn, David B. Shackelford, Daniel F. Egan, Maria M. Mihaylova, Annabelle Mery, Debbie S. Vasquez, Benjamin E. Turk, and Reuben J. Shaw. AMPK Phosphorylation of Raptor Mediates a Metabolic Checkpoint. *Molecular Cell*, 30(2):214–226, 4 2008. ISSN 10972765. doi: 10.1016/j.molcel.2008.03.003.

Emilie Vander Haar, Seong il Lee, Sricharan Bandhakavi, Timothy J. Griffin, and Do Hyung Kim. Insulin signalling to mTOR mediated by the Akt/PKB substrate PRAS40. *Nature Cell Biology 2007 9:3*, 9(3):316–323, 2 2007. ISSN 1476-4679. doi: 10.1038/ncb1547. URL https://www.nature.com/articles/ncb1547.

Alex R. Hall, Daniel De Vos, Ville Petri Friman, Jean Paul Pirnay, and Angus Buckling. Effects of sequential and simultaneous applications of bacteriophages on populations of Pseudomonas aeruginosa in vitro and in wax moth larvae. *Applied and Environmental Microbiology*, 78(16): 5646–5652, 2012. ISSN 00992240. doi: 10.1128/AEM.00757-12.

Kenta Hara, Yoshiko Maruki, Xiaomeng Long, Ken-ichi Yoshino, Noriko Oshiro, Sujuti Hidayat, Chiharu Tokunaga, Joseph Avruch, and Kazuyoshi Yonezawa. Raptor, a Binding Partner of Target of Rapamycin (TOR), Mediates TOR Action. *Cell*, 110 (2):177–189, 7 2002. ISSN 00928674. doi: 10.1016/S0092-8674(02)00833-4. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867402008334.

James R. Harris. The production of paracrystalline two-dimensional monolayers of purified protein molecules. *Micron (1969)*, 13(2):147–168, 1 1982. ISSN 0047-7206. doi: 10.1016/0047-7206(82)90081-4.

Hisashi Hasumi, Masaya Baba, Seung-Beom Hong, Yukiko Hasumi, Ying Huang, Masahiro Yao, Vladimir A Valera, W Marston Linehan, and Laura S Schmidt. Identification and characterization of a novel folliculin-interacting protein {FNIP2}. *Gene*, 415(1-2):60–67, 5 2008a. doi: 10.1016/j.gene.2008.02.022. URL http://dx.doi.org/10.1016/j.gene.2008.02.022.

Hisashi Hasumi, Masaya Baba, Seung-Beom Hong, Yukiko Hasumi, Ying Huang, Masahiro Yao, Vladimir A Valera, W Marston Linehan, and Laura S Schmidt. Identification and characterization of a novel folliculin-interacting protein {FNIP2}. *Gene*, 415(1-2):60–67, 5 2008b. doi: 10.1016/j.gene.2008.02.022. URL http://dx.doi.org/10.1016/j.gene.2008.02.022.

Hisashi Hasumi, Masaya Baba, Yukiko Hasumi, Martin Lang, Ying Huang, Hyoung Bin F. Oh, Masayuki Matsuo, Maria J. Merino, Masahiro Yao, Yusuke Ito, Mitsuko Furuya, Yasuhiro Iribe, Tatsuhiko Kodama, Eileen Southon, Lino Tessarollo, Kunio Nagashima, Diana C.

Haines, W. Marston Linehan, and Laura S. Schmidt. Folliculin-interacting proteins Fnip1 and Fnip2 play critical roles in kidney tumor suppression in cooperation with Flcn. *Proceedings of the National Academy of Sciences of the United States of America*, 112(13):E1624–E1631, 3 2015. ISSN 10916490. doi: 10.1073/pnas.1419502112.

Yukiko Hasumi, Masaya Baba, Rieko Ajima, Hisashi Hasumi, Vladimir A. Valera, Mara E. Klein, Diana C. Haines, Maria J. Merino, Seung Beom Hong, Terry P. Yamaguchi, Laura S. Schmidt, and W. Marston Linehan. Homozygous loss of BHD causes early embryonic lethality and kidney tumor development with activation of mTORC1 and mTORC2. *Proceedings of the National Academy of Sciences of the United States of America*, 106(44):18722–18727, 11 2009. ISSN 00278424. doi: 10.1073/pnas.0908853106.

Graham F. Hatfull, Marisa L. Pedulla, Deborah Jacobs-Sera, Pauline M. Cichon, Amy Foley, Michael E. Ford, Rebecca M. Gonda, Jennifer M. Houtz, Andrew J. Hryckowian, Vanessa A. Kelchner, Swathi Namburi, Kostandin V. Pajcini, Mark G. Popovich, Donald T. Schleicher, Brian Z. Simanek, Alexis L. Smith, Gina M. Zdanowicz, Vanaja Kumar, Craig L. Peebles, William R. Jacobs, Jeffrey G. Lawrence, and Roger W. Hendrix. Exploring the Mycobacteriophage Metaproteome: Phage Genomics as an Educational Platform. *PLOS Genetics*, 2 (6):e92, 2006. ISSN 1553-7404. doi: 10.1371/JOURNAL.PGEN.0020092.

J Heitman, N R Movva, and M N Hall. Targets for cell cycle arrest by the immunosuppressant rapamycin in yeast. *Science*, 253(5022):905–909, 8 1991. doi: 10.1126/science.1715094. URL `http://dx.doi.org/10.1126/science.1715094`.

Tomasz Heyduk, Ewa Heyduk, Konstantin Severinov, Hong Tang, and Richard H. Ebright. Determinants of RNA polymerase α subunit for interaction with β, β, and σ subunits: Hydroxyl-radical protein footprinting. *Proceedings of the National Academy of Sciences of the United States of America*, 93(19):10162–10166, 9 1996. ISSN 00278424. doi: 10.1073/pnas.93.19.10162.

L Holm and C Sander. Dali: a network tool for protein structure comparison. *Trends in biochemical sciences*, 20(11):478–480, 1995. ISSN 0968-0004. doi: 10.1016/S0968-0004(00)89105-7. URL `https://pubmed.ncbi.nlm.nih.gov/8578593/`.

Elizabeth A Hubin, Mirjana Lilic, Seth A Darst, and Elizabeth A Campbell. Structural insights into the mycobacteria transcription initiation complex from analysis of X-ray crystal structures. *Nature Communications*, 8:16072, 2017. doi: 10.1038/ncomms16072. URL `www.nature.com/naturecommunications`.

William S. Hummers and Richard E. Offeman. Preparation of Graphitic Oxide. *Journal of the American Chemical Society*, 80(6):1339–1339, 3 1958. ISSN 0002-7863. doi: 10.1021/ja01539a017.

Ken Inoki, Yong Li, Tianquan Zhu, Jun Wu, and Kun-Liang Guan. {TSC2} is phosphorylated and inhibited by Akt and suppresses {mTOR} signalling. *Nat Cell Biol*, 4(9):648–657, 9 2002. doi: 10.1038/ncb839. URL `http://dx.doi.org/10.1038/ncb839`.

Lakshminarayan M. Iyer, Eugene V. Koonin, and L. Aravind. Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases. *BMC Structural Biology*, 3:1–23, 2003. ISSN 14726807. doi: 10.1186/1472-6807-3-1.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873*, 596(7873):583–589, 7 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2.

Athanasios Kakasis and Gerasimia Panitsa. Bacteriophage therapy as an alternative treatment for human infections. A comprehensive review. *International Journal of Antimicrobial Agents*, 53(1):16–21, 1 2019. ISSN 0924-8579. doi: 10.1016/J.IJANTIMICAG.2018.09.004.

Berthold Kastner, Niels Fischer, Monika Mariola Golas, Bjoern Sander, Prakash Dube, Daniel

Boehringer, Klaus Hartmuth, Jochen Deckert, Florian Hauer, Elmar Wolf, Hannes Uchtenhagen, Henning Urlaub, Franz Herzog, Jan Michael Peters, Dietmar Poerschke, Reinhard Lührmann, and Holger Stark. GraFix: Sample preparation for single-particle electron cryomicroscopy. *Nature Methods*, 5(1):53–55, 1 2008. ISSN 15487091. doi: 10.1038/nmeth1139.

Anthony D. Keefe, David S. Wilson, Burckhard Seelig, and Jack W. Szostak. One-Step Purification of Recombinant Proteins Using a Nanomolar-Affinity Streptavidin-Binding Peptide, the SBP-Tag. *Protein Expression and Purification*, 23(3):440–446, 12 2001. ISSN 1046-5928. doi: 10.1006/PREP.2001.1515.

Damir Khabibullin, Douglas A. Medvetz, Miguel Pinilla, Venkatesh Hariharan, Chenggang Li, Anja Hergrueter, Maria Laucho Contreras, Erik Zhang, Andrey Parkhitko, Jane J. Yu, Caroline A. Owen, Hayden Huang, Rebecca M. Baron, and Elizabeth P. Henske. Folliculin regulates cell-cell adhesion, AMPK, and mTORC1 in a cell-type-specific manner in lung-derived cells. *Physiological Reports*, 2(8):1–12, 2014. ISSN 2051817X. doi: 10.14814/phy2.12107.

Maryam Khoshouei, Mazdak Radjainia, Wolfgang Baumeister, and Radostin Danev. Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate. *Nature Communications*, 8:16099, 2017. doi: 10.1038/ncomms16099.

Do-Hyung Kim, D D Sarbassov, Siraj M Ali, Jessie E King, Robert R Latek, Hediye Erdjument-Bromage, Paul Tempst, and David M Sabatini. {mTOR} interacts with raptor to form a nutrient-sensitive complex that signals to the cell growth machinery. *Cell*, 110(2):163–175, 7 2002. ISSN 0092-8674. doi: 10.1016/S0092-8674(02)00808-5. URL `http://dx.doi.org/10.1016/S0092-8674(02)00808-5`.

Do-Hyung Kim, Dos D Sarbassov, Siraj M Ali, Robert R Latek, Kalyani V P Guntur, Hediye Erdjument-Bromage, Paul Tempst, and David M Sabatini. {G\betaL}, a Positive Regulator of the Rapamycin-Sensitive Pathway Required for the Nutrient-Sensitive Interaction between Raptor and {mTOR}. *Molecular cell*, 11(4):895–904, 4 2003. ISSN 10972765. doi: 10.1016/S1097-2765(03)00114-X.

Joungmok Kim and Kun-Liang Guan. mTOR as a central hub of nutrient signalling and cell

growth. *Nature Cell Biology*, 21(1):63–71, 1 2019. ISSN 1465-7392. doi: 10.1038/s41556-018-0205-1. URL http://www.nature.com/articles/s41556-018-0205-1.

Dari Kimanius, Liyi Dong, Grigory Sharov, Takanori Nakane, and Sjors H. W. Scheres. New tools for automated cryo-EM single-particle analysis in RELION-4.0. *Biochemical Journal*, 478(24):4169–4185, 12 2021. ISSN 0264-6021. doi: 10.1042/BCJ20210708.

Lisa N. Kinch and Nick V. Grishin. Longin-like folds identified in CHiPS and DUF254 proteins: Vesicle trafficking complexes conserved in eukaryotic evolution. *Protein Science*, 15(11):2669–2674, 11 2006. ISSN 09618368. doi: 10.1110/ps.062419006. URL http://doi.wiley.com/10.1110/ps.062419006.

Viktor I. Korolchuk, Shinji Saiki, Maike Lichtenberg, Farah H. Siddiqi, Esteban A. Roberts, Sara Imarisio, Luca Jahreiss, Sovan Sarkar, Marie Futter, Fiona M. Menzies, Cahir J. O'Kane, Vojo Deretic, and David C. Rubinsztein. Lysosomal positioning coordinates cellular nutrient responses. *Nature Cell Biology*, 13(4):453–462, 2011. ISSN 14657392. doi: 10.1038/ncb2204.

Detlev H Kroger and Cornelia Schroeder. Bacteriophage T3 and Bacteriophage T7 Virus-Host Cell Interactions. Technical Report 1, 1981.

V. N. Krylov and I. Zhazykov Zh. Pseudomonas bacteriophage phiKZ as a model for studying genetical control of morphogenesis. *Genetika*, 14(4):678–685, 1978. ISSN 00166758. URL https://pubmed.ncbi.nlm.nih.gov/95987/.

V. N. Krylov, T. A. Smirnova, I. B. Minenkova, T. G. Plotnikova, I. Z. Zhazikov, and E. A. Khrenova. Pseudomonas bacteriophage ΦKZ contains an inner body in its capsid. *Canadian Journal of Microbiology*, 30(6):758–762, 6 1984. ISSN 0008-4166. doi: 10.1139/m84-116. URL http://www.nrcresearchpress.com/doi/10.1139/m84-116.

V. N. Krylov, D. M. Dela Cruz, K. Hertveldt, and H. W. Ackermann. "ϕKZ-like viruses", a proposed new genus of myovirus bacteriophages. *Archives of Virology*, 152(10):1955–1959, 10 2007. ISSN 03048608. doi: 10.1007/s00705-007-1037-7.

Konstantin Kuznedelov, Nataliya Korzheva, Arkady Mustaev, and Konstantin Severinov.

Structure-based analysis of RNA polymerase function: The largest subunit's rudder contributes critically to elongation complex stability and is not involved in the maintenance of RNA-DNA hybrid length. *EMBO Journal*, 21(6):1369–1378, 2002. ISSN 02614189. doi: 10.1093/emboj/21.6.1369.

Tony Kwan, Jing Liu, Michael DuBow, Philippe Gros, and Jerry Pelletier. The complete genomes and proteomes of 27 Staphylococcus aureus bacteriophages. *Proceedings of the National Academy of Sciences*, 102(14):5174–5179, 4 2005. ISSN 0027-8424. doi: 10.1073/PNAS.0501140102.

William J. Lane and Seth A. Darst. Molecular Evolution of Multisubunit RNA Polymerases: Structural Analysis. *Journal of Molecular Biology*, 395(4):686–704, 1 2010a. ISSN 00222836. doi: 10.1016/j.jmb.2009.10.063.

William J. Lane and Seth A. Darst. Molecular Evolution of Multisubunit RNA Polymerases: Sequence Analysis. *Journal of Molecular Biology*, 395(4):671–685, 1 2010b. ISSN 00222836. doi: 10.1016/j.jmb.2009.10.062.

Mathieu Laplante and David M. Sabatini. mTOR Signaling in Growth Control and Disease. *Cell*, 149(2):274–293, 4 2012. ISSN 0092-8674. doi: 10.1016/J.CELL.2012.03.017.

Laura A. Laviolette, Julien Mermoud, Isabel A. Calvo, Nicholas Olson, Myriam Boukhali, Ortrud K. Steinlein, Elisabeth Roider, Elke C. Sattler, Dachuan Huang, Bin Tean Teh, Mo Motamedi, Wilhelm Haas, and Othon Iliopoulos. Negative regulation of EGFR signalling by the human folliculin tumour suppressor protein. *Nature Communications*, 8(1):15866, 8 2017. ISSN 2041-1723. doi: 10.1038/ncomms15866.

Daria Lavysh, Maria Sokolova, Leonid Minakhin, Maria Yakunina, Tatjana Artamonova, Sergei Kozyavkin, Kira S. Makarova, Eugene V. Koonin, and Konstantin Severinov. The genome of AR9, a giant transducing Bacillus phage encoding two multisubunit RNA polymerases. *Virology*, 495:185–196, 8 2016. ISSN 10960341. doi: 10.1016/j.virol.2016.04.030.

Rosalie E. Lawrence, Simon A. Fromm, Yangxue Fu, Adam L. Yokom, Do Jin Kim, Ashley M. Thelen, Lindsey N. Young, Chun-Yan Lim, Avi J. Samelson, James H. Hurley, and Roberto

Zoncu. Structural mechanism of a Rag GTPase activation checkpoint by the lysosomal folliculin complex. *Science*, 366(6468):971–977, 10 2019. ISSN 0036-8075. doi: 10.1126/science.aax0364.

Bryan R. Lenneman and Lucia B. Rothman-Denes. Structural and biochemical investigation of bacteriophage N4-encoded RNA polymerases, 4 2015. ISSN 2218273X.

Timothy P Levine, Rachel D Daniels, Louise H Wong, Alberto T Gatta, Andreas Gerondopoulos, and Francis A Barr. Discovery of new Longin and Roadblock domains that form platforms for small {GTPases} in Ragulator and {TRAPP}-{II}. *Small {GTPases}*, 4(2): 62–69, 6 2013. doi: 10.4161/sgtp.24262. URL `http://dx.doi.org/10.4161/sgtp.24262`.

Emilie Levivier, Bruno Goud, Michel Souchet, Thierry P.G. Calmels, Jean Paul Mornon, and Isabelle Callebaut. uDENN, DENN, and dDENN: Indissociable domains in Rab and MAP kinases signaling pathways. *Biochemical and Biophysical Research Communications*, 287(3): 688–695, 2001. ISSN 0006291X. doi: 10.1006/bbrc.2001.5652.

Yi-Lynn Liang, Maryam Khoshouei, Mazdak Radjainia, Yan Zhang, Alisa Glukhova, Jeffrey Tarrasch, David M. Thal, Sebastian G. B. Furness, George Christopoulos, Thomas Coudrat, Radostin Danev, Wolfgang Baumeister, Laurence J. Miller, Arthur Christopoulos, Brian K. Kobilka, Denise Wootten, Georgios Skiniotis, and Patrick M. Sexton. Phase-plate cryo-EM structure of a class B GPCR–G-protein complex. *Nature 2017 546:7656*, 546(7656):118–123, 4 2017. ISSN 1476-4687. doi: 10.1038/nature22327.

Derek H.K. Lim, Pauline K. Rehal, Michael S. Nahorski, Fiona Macdonald, Tijs Claessens, Michel Van Geel, Lieke Gijezen, Johan J.P. Gille, Sophie Giraud, Stephane Richard, Maurice van Steensel, Fred H. Menko, and Eamonn R. Maher. A new locus-specific database (LSDB) for mutations in the folliculin (¡i¿FLCN¡/i¿) gene. *Human Mutation*, 31(1):E1043–E1051, 1 2010. ISSN 10597794. doi: 10.1002/humu.21130. URL `http://doi.wiley.com/10.1002/humu.21130`.

Bin Liu, Chuan Hong, Rick K Huang, Zhiheng Yu, and Thomas A Steitz. Structural basis of bacterial transcription activation. *Science*, 358:947–951, 2017.

Robbie Loewith, Estela Jacinto, Stephan Wullschleger, Anja Lorberg, José L Crespo, Débora Bonenfant, Wolfgang Oppliger, Paul Jenoe, and Michael N Hall. Two {TOR} complexes, only one of which is rapamycin sensitive, have distinct roles in cell growth control. *Mol Cell*, 10(3):457–468, 9 2002. doi: 10.1016/S1097-2765(02)00636-6. URL http://dx.doi.org/10.1016/S1097-2765(02)00636-6.

M. Lonetto, M. Gribskov, and C. A. Gross. The σ70 family: Sequence conservation and evolutionary relationships, 6 1992. ISSN 00219193. URL http://jb.asm.org/.

Xiaomeng Long, Yenshou Lin, Sara Ortiz-Vega, Kazuyoshi Yonezawa, and Joseph Avruch. Rheb binds and regulates the {mTOR} kinase. *Curr Biol*, 15(8):702–713, 4 2005. ISSN 0960-9822. doi: 10.1016/j.cub.2005.02.053. URL http://dx.doi.org/10.1016/j.cub.2005.02.053.

Simon C. Lovell, Ian W. Davis, W. Bryan Arendall, Paul I.W. De Bakker, J. Michael Word, Michael G. Prisant, Jane S. Richardson, and David C. Richardson. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins*, 50(3):437–450, 2 2003. ISSN 1097-0134. doi: 10.1002/PROT.10286.

R. Matteoni and T. E. Kreis. Translocation and clustering of endosomes and lysosomes depends on microtubules. *Journal of Cell Biology*, 105(3):1253–1265, 1987. ISSN 00219525. doi: 10.1083/jcb.105.3.1253.

Turk J Med Sci, Kübra Can, Uğur Aksu, and Osman Şadi YENEN. Investigation of PhiKZ phage therapy against Pseudomonas aeruginosa in mouse pneumonia model. *Turkish Journal of Medical Sciences*, 48:670–678, 2018. doi: 10.3906/sag-1711-22. URL http://journals.tubitak.gov.tr/medical/.

Senén D. Mendoza, Eliza S. Nieweglowska, Sutharsan Govindarajan, Lina M. Leon, Joel D. Berry, Anika Tiwari, Vorrapon Chaikeeratisak, Joe Pogliano, David A. Agard, and Joseph Bondy-Denomy. A bacteriophage nucleus-like compartment shields DNA from CRISPR nucleases. *Nature*, 577(7789):244–248, 1 2020. ISSN 14764687. doi: 10.1038/s41586-019-1786-y.

Jin Meng and Shawn M Ferguson. GATOR1-dependent recruitment of FLCN-FNIP to lyso-

somes coordinates Rag GTPase heterodimer nucleotide status in response to amino acids. *Journal of Cell Biology*, 217(8):2765–2776, 2018. ISSN 15408140. doi: 10.1083/jcb.201712177.

Vadim V. Mesyanzhinov, Johan Robben, Barbara Grymonprez, Victor A. Kostyuchenko, Maria V. Bourkaltseva, Nina N. Sykilinda, Victor N. Krylov, and Guido Volckaert. The genome of bacteriophage ϕKZ of Pseudomonas aeruginosa. *Journal of Molecular Biology*, 317(1):1–19, 2002. ISSN 00222836. doi: 10.1006/jmbi.2001.5396.

‡ Michael Spoerner, ‡ Andrea Nuehs, ‡ Petra Ganser, § Christian Herrmann, § Alfred Wittinghofer, , and ‡ Hans Robert Kalbitzer*. Conformational States of Ras Complexed with the GTP Analogue GppNHp or GppCH2p: Implications for the Interaction with Effector Proteins†. *Biochemistry*, 44(6):2225–2236, 2 2005. doi: 10.1021/BI0488000.

L Minakhin, S Bhagat, A Brunning, E A Campbell, S A Darst, R H Ebright, and K Severinov. Bacterial RNA polymerase subunit omega and eukaryotic RNA polymerase subunit RPB6 are sequence, structural, and functional homologs and promote RNA polymerase assembly. *Proceedings of the National Academy of Sciences of the United States of America*, 98(3): 892–7, 1 2001. ISSN 0027-8424. doi: 10.1073/pnas.98.3.892.

K. A. Miroshnikov, N. M. Faizullina, N. N. Sykilinda, and V. V. Mesyanzhinov. Properties of the endolytic transglycosylase encoded by gene 144 of Pseudomonas aeruginosa bacteriophage phiKZ. *Biochemistry (Moscow) 2006 71:3*, 71(3):300–305, 3 2006. ISSN 1608-3040. doi: 10.1134/S0006297906030102.

Ashwini K. Mishra and David G. Lambright. Invited review: Small GTPases and their GAPs. *Biopolymers*, 105(8):431–448, 8 2016. ISSN 00063525. doi: 10.1002/bip.22833.

Ha Yoon Mo, Hyun Ji Son, Eun Ji Choi, Nam Jin Yoo, Chang Hyeok An, and Sug Hyung Lee. Somatic mutations of candidate tumor suppressor genes folliculin-interacting proteins FNIP1 and FNIP2 in gastric and colon cancers. *Pathology - Research and Practice*, 215:152646, 9 2019. ISSN 03440338. doi: 10.1016/j.prp.2019.152646.

Subhanjan Mondal, Kevin Hsiao, and Said A Goueli. A Homogenous Bioluminescent System for Measuring GTPase, GTPase Activating Protein, and Guanine Nucleotide Exchange

Factor Activities. *Assay and Drug Development Technologies*, 13(8):444–455, 2015. doi: 10.1089/adt.2015.643.

Rachel Anne Mooney, Seth A. Darst, and Robert Landick. Sigma and RNA polymerase: an on-again, off-again relationship? *Molecular cell*, 20(3):335–345, 11 2005. ISSN 1097-2765. doi: 10.1016/J.MOLCEL.2005.10.015.

David Moreira and Purificación López-García. Evolution of viruses and cells: do we need a fourth domain of life to explain the origin of eukaryotes? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370:20140327., 2015. doi: 10.1098/rstb.2014.0327.

Dirk Mossmann, Sujin Park, and Michael N. Hall. mTOR signalling and cellular metabolism are mutual determinants in cancer. *Nature Reviews Cancer*, 18 (12):744–757, 12 2018. ISSN 1474-175X. doi: 10.1038/s41568-018-0074-8. URL `http://www.nature.com/articles/s41568-018-0074-8`.

Katsuhiko S. Murakami, Shoko Masuda, and Seth A. Darst. Structural basis of transcription initiation: RNA polymerase holoenzyme at 4 Å resolution. *Science*, 296(5571):1280–1284, 5 2002. ISSN 00368075. doi: 10.1126/SCIENCE.1069594/FORMAT/PDF. URL `https://www.science.org`.

Shigeyki Nada, Akihiro Hondo, Atsuko Kasai, Masato Koike, Kazunobu Saito, Yasuo Uchiyama, and Masato Okada. The novel lipid raft adaptor p18 controls endosome dynamics by anchoring the MEK–ERK pathway to late endosomes. *The EMBO Journal*, 28(5):477–489, 3 2009. ISSN 0261-4189. doi: 10.1038/emboj.2008.308. URL `http://emboj.embopress.org/cgi/doi/10.1038/emboj.2008.308`.

Sergei Nechaev and Konstantin Severinov. BACTERIOPHAGE-INDUCED MODIFICATIONS OF HOST RNA POLYMERASE. *Annu. Rev. Microbiol*, 57:301–323, 2003. doi: 10.1146/annurev.micro.57.030502.090942. URL `www.annualreviews.org`.

Michael L. Nickerson, Michelle B. Warren, Jorge R. Toro, Vera Matrosova, Gladys Glenn, Maria L. Turner, Paul Duray, Maria Merino, Peter Choyke, Christian P. Pavlovich, Nirmala Sharma, McClellan Walther, David Munroe, Rob Hill, Eamonn Maher, Cheryl Greenberg,

Michael I. Lerman, W. Marston Linehan, Berton Zbar, and Laura S. Schmidt. Mutations in a novel gene lead to kidney tumors, lung wall defects, and benign tumors of the hair follicle in patients with the Birt-Hogg-Dubé syndrome. *Cancer Cell*, 2(2):157–164, 2002. ISSN 15356108. doi: 10.1016/S1535-6108(02)00104-6.

Ravi K. Nookala, Lars Langemeyer, Angela Pacitto, Bernardo Ochoa-Montaño, Jane C. Donaldson, Beata K. Blaszczyk, Dimitri Y. Chirgadze, Francis A. Barr, J. Fernando Bazan, and Tom L. Blundell. Crystal structure of folliculin reveals a hidDENN function in genetically inherited renal cancer. *Open Biology*, 2(AUG), 2012. ISSN 20462441. doi: 10.1098/rsob.120071.

Won Jun Oh and Estela Jacinto. mTOR complex 2 signaling and functions. *Cell Cycle*, 10(14):2305–2316, 2011. ISSN 15514005. doi: 10.4161/cc.10.14.16586. URL `http://www.tandfonline.com/action/journalInformation?journalCode=kccy20`.

Mariia Orekhova, Alevtina Koreshova, Tatyana Artamonova, Mikhail Khodorkovskii, and Maria Yakunina. The study of the phiKZ phage non-canonical non-virion RNA polymerase. *Biochemical and Biophysical Research Communications*, 511(4):759–764, 4 2019. ISSN 10902104. doi: 10.1016/j.bbrc.2019.02.132.

Noriko Oshiro, Rinako Takahashi, Ken Ichi Yoshino, Keiko Tanimura, Akio Nakashima, Satoshi Eguchi, Takafumi Miyamoto, Kenta Hara, Kenji Takehana, Joseph Avruch, Ushio Kikkawa, and Kazuyoshi Yonezawa. The proline-rich Akt substrate of 40 kDa (PRAS40) is a physiological substrate of mammalian target of rapamycin complex 1. *Journal of Biological Chemistry*, 282(28):20329–20339, 7 2007. ISSN 00219258. doi: 10.1074/JBC.M702636200/ATTACHMENT/2354D7A4-A639-49A4-A8E3-F05163553C84/MMC1.PDF.

Angela Pacitto, David B Ascher, Louise H Wong, Beata K Blaszczyk, Ravi K Nookala, Nianshu Zhang, Svetlana Dokudovskaya, Tim P Levine, and Tom L Blundell. Lst4, the yeast Fnip1/2 orthologue, is a {DENN}-family protein. *Open Biol*, 5(12):150174, 12 2015. doi: 10.1098/rsob.150174. URL `http://dx.doi.org/10.1098/rsob.150174`.

Mark S. Paget. Bacterial sigma factors and anti-sigma factors: Structure, function and distribution, 7 2015. ISSN 2218273X. URL www.mdpi.com/journal/biomolecules/.

Mark Sb Paget and John D Helmann. The 70 family of sigma factors. *Genome Biology*, 4:203, 2003.

Marie-Pierre Péli-Gulli, Alessandro Sardu, Nicolas Panchaud, Serena Raucci, and Claudio De Virgilio. Amino Acids Stimulate {TORC1} through Lst4-Lst7, a {GTPase}-Activating Protein Complex for the Rag Family {GTPase} Gtr2. *Cell Rep*, 13(1):1–7, 10 2015. doi: 10.1016/j.celrep.2015.08.059. URL http://dx.doi.org/10.1016/j.celrep.2015.08.059.

Samuel Peña-Llopis, Silvia Vega-Rubin-de Celis, Jacob C Schwartz, Nicholas C Wolff, Tram Anh T Tran, Lihua Zou, Xian-Jin Xie, David R Corey, James Brugarolas, Michel C Vellard, Valeria Facchinetti, David M Sabatini, Andrea Ballabio, Jacob C Schwarts, Nicholas C Wolff, Tram Anh T Tran, Lihua Zou, Xian-Jin Xie, David R Corey, James Brugarolas, Jacob C Schwartz, Nicholas C Wolff, Tram Anh T Tran, Lihua Zou, Xian-Jin Xie, David R Corey, and James Brugarolas. Regulation of TFEB and V-ATPases by mTORC1. *The EMBO journal*, 30(16):3242–58, 7 2011. ISSN 1460-2075. doi: 10.1038/emboj.2011.257. URL http://emboj.embopress.org/cgi/doi/10.1038/emboj.2011.257 https://www.embopress.org/doi/pdf/10.1038/emboj.2011.257 http://www.ncbi.nlm.nih.gov/pubmed/21804531 http://www.pubmedcentral.nih.gov/articlere

Wei Peng and Jenna L Jewell. Amino Acid Sensing: Architecture of mTORC1 on the Lysosome Surface. *Current Biology*, 30(2):R89–R91, 1 2020. ISSN 09609822. doi: 10.1016/j.cub.2019.11.087. URL https://doi.org/10.1016/j.cub.2019.11.087 https://linkinghub.elsevier.com/retrieve/pii/S0960982219315921.

Timothy R. Peterson, Mathieu Laplante, Carson C. Thoreen, Yasemin Sancak, Seong A. Kang, W. Michael Kuehl, Nathanael S. Gray, and David M. Sabatini. DEPTOR Is an mTOR Inhibitor Frequently Overexpressed in Multiple Myeloma Cells and Required for Their Survival. *Cell*, 137(5):873–886, 5 2009. ISSN 00928674. doi: 10.1016/J.CELL.2009.03.046/ATTACHMENT/DC565DD9-70D9-42AE-86CF-3A65F80DCBDD/MMC2.XLS.

Constance S. Petit, Agnes Roczniak-Ferguson, and Shawn M. Ferguson. Recruitment of folliculin to lysosomes supports the amino acid-dependent activation of rag gtpases. *Journal of Cell Biology*, 202(7):1107–1122, 2013a. ISSN 00219525. doi: 10.1083/jcb.201307084.

Constance S. Petit, Agnes Roczniak-Ferguson, and Shawn M. Ferguson. Recruitment of folliculin to lysosomes supports the amino acid-dependent activation of rag gtpases. *Journal of Cell Biology*, 202(7):1107–1122, 2013b. ISSN 00219525. doi: 10.1083/jcb.201307084.

Diana P. Pires, Diana Vilas Boas, Sanna Sillankorva, and Joana Azeredo. Phage Therapy: a Step Forward in the Treatment of Pseudomonas aeruginosa Infections. *Journal of Virology*, 89(15):7449–7456, 8 2015. ISSN 0022-538X. doi: 10.1128/jvi.00385-15.

Elite Possik, Zahra Jalali, Yann Nouët, Ming Yan, Marie Claude Gingras, Kathrin Schmeisser, Lorena Panaite, Fanny Dupuy, Dmitri Kharitidi, Laëtitia Chotard, Russell G. Jones, David H. Hall, and Arnim Pause. Folliculin Regulates Ampk-Dependent Autophagy and Metabolic Stress Survival. *PLoS Genetics*, 10(4), 2014. ISSN 15537404. doi: 10.1371/journal.pgen.1004273.

Elite Possik, Andrew Ajisebutu, Sanaz Manteghi, Marie Claude Gingras, Tarika Vijayaraghavan, Mathieu Flamand, Barry Coull, Kathrin Schmeisser, Thomas Duchaine, Maurice van Steensel, David H. Hall, and Arnim Pause. FLCN and AMPK Confer Resistance to Hyperosmotic Stress via Remodeling of Glycogen Stores. *PLoS Genetics*, 11(10), 2015. ISSN 15537404. doi: 10.1371/journal.pgen.1005520.

B. V. Venkataram Prasad and Michael F Schmid. Principles of Virus Structural Organization. In *Rossmann M., Rao V. (eds) Viral Molecular Machines. Advances in Experimental Medicine and Biology*, volume 726, pages 17–47. 2012. ISBN 978-1-4614-0980-9. doi: 10.1007/978-1-4614-0980-9_3.

Jing Pu, Tal Keren-Kaplan, and Juan S. Bonifacino. A Ragulator-BORC interaction controls lysosome positioning in response to amino acid availability. *Journal of Cell Biology*, 216(12): 4183–4197, 2017. ISSN 15408140. doi: 10.1083/jcb.201703094.

Yoana Rabanal-Ruiz and Viktor Korolchuk. mTORC1 and Nutrient Homeostasis: The Central Role of the Lysosome. *International Journal of Molecular Sciences*, 19(3):818, 3 2018. ISSN 1422-0067. doi: 10.3390/ijms19030818.

Kailash Ramlaul and Christopher H.S. S. Aylett. Signal integration in the (m)TORC1 growth pathway, 7 2018. ISSN 16747992.

Kailash Ramlaul, Wencheng Fu, Hua Li, Natàlia de Martin Garrido, Lin He, Manjari Trivedi, Wei Cui, Christopher H.S. Aylett, and Geng Wu. Architecture of the Tuberous Sclerosis Protein Complex. *Journal of Molecular Biology*, 433(2):166743, 1 2021. ISSN 00222836. doi: 10.1016/j.jmb.2020.166743.

Sara Reardon. Phage therapy gets revitalized. *Nature*, 510(7503):15–16, 2014. ISSN 14764687. doi: 10.1038/510015A.

David M. Rees, Oliver Willhoft, Chia Liang Lin, Rohan Bythell-Douglas, and Dale B. Wigley. Production and Assay of Recombinant Multisubunit Chromatin Remodeling Complexes. *Methods in enzymology*, 592:27–47, 2017. ISSN 1557-7988. doi: 10.1016/BS.MIE.2017.03.015.

Kacper B. Rogala, Xin Gu, Jibril F. Kedir, Monther Abu-Remaileh, Laura F. Bianchi, Alexia M. S. Bottino, Rikke Dueholm, Anna Niehaus, Daan Overwijn, Ange-Célia Priso Fils, Sherry X. Zhou, Daniel Leary, Nouf N. Laqtom, Edward J. Brignole, and David M. Sabatini. Structural basis for the docking of mTORC1 on the lysosomal surface. *Science*, page eaay0166, 10 2019. ISSN 0036-8075. doi: 10.1126/science.aay0166.

Alexis Rohou and Nikolaus Grigorieff. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *Journal of Structural Biology*, 192(2):216–221, 11 2015. ISSN 10958657. doi: 10.1016/j.jsb.2015.08.008.

David M. Sabatini, Hediye Erdjument-Bromage, Mary Lui, Paul Tempst, and Solomon H. Snyder. RAFT1: A mammalian protein that binds to FKBP12 in a rapamycin-dependent fashion and is homologous to yeast TORs. *Cell*, 78(1):35–43, 7 1994. ISSN 0092-8674. doi: 10.1016/0092-8674(94)90570-3.

Ruth M. Saecker, M. Thomas Record, and Pieter L. Dehaseth. Mechanism of Bacterial Transcription Initiation: RNA Polymerase - Promoter Binding, Isomerization to Initiation-Competent Open Complexes, and Initiation of RNA Synthesis. *Journal of Molecular Biology*, 412(5):754–771, 10 2011. ISSN 0022-2836. doi: 10.1016/J.JMB.2011.01.018.

Yasemin Sancak, Timothy R. Peterson, Yoav D. Shaul, Robert A. Lindquist, Carson C. Thoreen, Liron Bar-Peled, and David M. Sabatini. The Rag GTPases bind raptor and mediate amino acid signaling to mTORC1. *Science*, 320(5882):1496–1501, 6 2008. ISSN 1095-9203. doi: 10.1126/science.1157535.

Yasemin Sancak, Liron Bar-Peled, Roberto Zoncu, Andrew L. Markhard, Shigeyuki Nada, and David M. Sabatini. Ragulator-Rag complex targets {mTORC1} to the lysosomal surface and is necessary for its activation by amino acids. *Cell*, 141(2):290–303, 4 2010. ISSN 1097-4172. doi: 10.1016/j.cell.2010.02.024.

Dos D Sarbassov, Siraj M Ali, Do-Hyung Kim, David A Guertin, Robert R Latek, Hediye Erdjument-Bromage, Paul Tempst, and David M Sabatini. Rictor, a Novel Binding Partner of mTOR, Defines a Rapamycin-Insensitive and Raptor-Independent Pathway that Regulates the Cytoskeleton. *Current Biology*, 14(14):1296–1302, 7 2004. ISSN 0960-9822. doi: 10.1016/J.CUB.2004.06.054.

Marco Sardiello, Michela Palmieri, Alberto Di Ronza, Diego Luis Medina, Marta Valenza, Vincenzo Alessandro Gennarino, Chiara Di Malta, Francesca Donaudy, Valerio Embrione, Roman S. Polishchuk, Sandro Banfi, Giancarlo Parenti, Elena Cattaneo, and Andrea Ballabio. A gene network regulating lysosomal biogenesis and function. *Science*, 325(5939): 473–477, 7 2009. ISSN 10959203. doi: 10.1126/science.1174447.

Robert A. Saxton and David M. Sabatini. mTOR Signaling in Growth, Metabolism, and Disease. *Cell*, 168(6):960–976, 3 2017. ISSN 0092-8674. doi: 10.1016/J.CELL.2017.02.004.

Laura S. Schmidt and W. Marston Linehan. FLCN: The causative gene for Birt-Hogg-Dubé syndrome. *Gene*, 640:28–42, 2018. ISSN 18790038. doi: 10.1016/j.gene.2017.09.044.

Annette Schürmann, Andreas Brauers, Silke Maßmann, Walter Becker, and Hans-Georg Joost. Cloning of a Novel Family of Mammalian GTP-binding Proteins (RagA, RagBs, RagB1) with Remote Similarity to the Ras-related GTPases. *Journal of Biological Chemistry*, 270 (48):28982–28988, 12 1995. ISSN 00219258. doi: 10.1074/jbc.270.48.28982.

Takeshi Sekiguchi, Eiji Hirose, Nobutaka Nakashima, Miki Ii, and Takeharu Nishimoto. Novel G Proteins, Rag C and Rag D, Interact with GTP-binding Proteins, Rag A and Rag B. *Journal of Biological Chemistry*, 276(10):7246–7257, 3 2001. ISSN 00219258. doi: 10.1074/jbc.M004389200.

Ekaterina Semenova, Marko Djordjevic, Boris Shralman, and Konstantin Severinov. The tale of two RNA polymerases: transcription profiling and gene expression strategy of bacteriophage Xp10. *Molecular Microbiology*, 55(3):764–777, 2 2005. ISSN 1365-2958. doi: 10.1111/J.1365-2958.2004.04442.X.

Carmine Settembre, Chiara Di Malta, Vinicia Assunta Polito, Moises Garcia Arencibia, Francesco Vetrini, Serkan Erdin, Serpil Uckac Erdin, Tuong Huynh, Diego Medina, Pasqualina Colella, Marco Sardiello, David C. Rubinsztein, and Andrea Ballabio. TFEB links autophagy to lysosomal biogenesis. *Science*, 332(6036):1429–1433, 6 2011. ISSN 00368075. doi: 10.1126/science.1204592.

David B. Shackelford and Reuben J. Shaw. The LKB1–AMPK pathway: metabolism and growth control in tumour suppression. *Nature Reviews Cancer*, 9(8):563–575, 8 2009. ISSN 1474-175X. doi: 10.1038/nrc2676.

Richard Sharp. Bacteriophages: biology and history. *Journal of Chemical Technology & Biotechnology*, 76(7):667–672, 7 2001. ISSN 1097-4660. doi: 10.1002/JCTB.434.

Kuang Shen and David M Sabatini. Ragulator and SLC38A9 activate the Rag GTPases through noncanonical GEF mechanisms. *Proceedings of the National Academy of Sciences*, 115(38): 9545–9550, 9 2018. ISSN 0027-8424. doi: 10.1073/pnas.1811727115.

Kuang Shen, Abigail Choe, and David M. Sabatini. Intersubunit Crosstalk in the Rag GTPase

Heterodimer Enables mTORC1 to Respond Rapidly to Amino Acid Availability. *Molecular Cell*, 68(3):552–565, 11 2017. ISSN 10972765. doi: 10.1016/j.molcel.2017.09.026.

Kuang Shen, Rick K. Huang, Edward J. Brignole, Kendall J. Condon, Max L. Valenstein, Lynne Chantranupong, Aimaiti Bomaliyamu, Abigail Choe, Chuan Hong, Zhiheng Yu, and David M. Sabatini. Architecture of the human GATOR1 and GATOR1-Rag GTPases complexes. *Nature*, 556(7699):64–69, 2018. ISSN 14764687. doi: 10.1038/nature26158.

Kuang Shen, Kacper B. Rogala, Hui-Ting Chou, Rick K. Huang, Zhiheng Yu, and David M. Sabatini. Cryo-EM Structure of the Human FLCN-FNIP2-Rag-Ragulator Complex. *Cell*, 179:1–11, 11 2019. ISSN 00928674. doi: 10.1016/j.cell.2019.10.036.

Mitsugu Shimobayashi and Michael N. Hall. Making new contacts: The mTOR network in metabolism and signalling crosstalk. *Nature Reviews Molecular Cell Biology*, 15(3):155–162, 3 2014. ISSN 14710072. doi: 10.1038/nrm3757.

Fabian Sievers, Andreas Wilm, David Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, Julie D. Thompson, and Desmond G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1):539, 1 2011. ISSN 1744-4292. doi: 10.1038/msb.2011.75.

Owen M. Siggs, Alexander Stockenhuber, Mukta Deobagkar-Lele, Katherine R. Bull, Tanya L. Crockford, Bethany L. Kingston, Greg Crawford, Consuelo Anzilotti, Violetta Steeples, Sahar Ghaffari, Gabor Czibik, Mohamed Bellahcene, Hugh Watkins, Houman Ashrafian, Benjamin Davies, Angela Woods, David Carling, Arash Yavari, Bruce Beutler, and Richard J. Cornall. Mutation of Fnip1 is associated with B-cell deficiency, cardiomyopathy, and elevated AMPK activity. *Proceedings of the National Academy of Sciences of the United States of America*, 113(26):E3706–E3715, 6 2016. ISSN 10916490. doi: 10.1073/pnas.1607592113.

TA Smirnova, IB Minenkova, EA Khrenova, TG Plotnikova, and V N Krylov. Electron microscope study of the intracellular development of the Pseudomonas aeruginosa bacteriophage phi KZ. *Zhurnal Mikrobiol. Epidemiol. Immunobiol*, 5:25–8, 1983.

Maria Sokolova, Sergei Borukhov, Daria Lavysh, Tatjana Artamonova, Mikhail Khodorkovskii, and Konstantin Severinov. A non-canonical multisubunit RNA polymerase encoded by the AR9 phage recognizes the template strand of its uracil-containing promoters. *Nucleic Acids Research*, 45(10):5958–5967, 6 2017. ISSN 13624962. doi: 10.1093/nar/gkx264. URL `https://academic.oup.com/nar/article/45/10/5958/3603318`.

Maria L. Sokolova, Inna Misovetc, and Konstantin V. Severinov. Multisubunit RNA Polymerases of Jumbo Bacteriophages. *Viruses*, 12(10):1064, 9 2020. ISSN 1999-4915. doi: 10.3390/v12101064.

Michael Spoerner, Andrea Nuehs, Christian Herrmann, Guido Steiner, and Hans Robert Kalbitzer. Slow conformational dynamics of the guanine nucleotide-binding protein Ras complexed with the GTP analogue GTPγS. *The FEBS Journal*, 274(6):1419–1433, 3 2007. ISSN 1742-4658. doi: 10.1111/J.1742-4658.2007.05681.X.

Holger Stark. GraFix: Stabilization of Fragile Macromolecular Complexes for Single Particle Cryo-EM. *Methods in Enzymology*, 481:109–126, 1 2010. ISSN 0076-6879. doi: 10.1016/S0076-6879(10)81005-5.

Georgina P Starling, Yan Y Yip, Anneri Sanger, Penny E Morton, Emily R Eden, and Mark P Dodding. Folliculin directs the formation of a Rab34–RILP complex to control the nutrient-dependent dynamic distribution of lysosomes. *EMBO reports*, 2016. ISSN 1469-221X. doi: 10.15252/embr.201541382.

Thomas A. Steitz. The structural basis of the transition from initiation to elongation phases of transcription, as well as translocation and strand separation, by T7 RNA polymerase. *Current Opinion in Structural Biology*, 14(1):4–9, 2 2004. ISSN 0959-440X. doi: 10.1016/J.SBI.2004.01.006.

C. K. Stover, X. Q. Pham, A. L. Erwin, S. D. Mizoguchi, P. Warrener, M. J. Hickey, F. S.L. Brinkman, W. O. Hufnagle, D. J. Kowallk, M. Lagrou, R. L. Garber, L. Goltry, E. Tolentino, S. Westbrock-Wadman, Y. Yuan, L. L. Brody, S. N. Coulter, K. R. Folger, A. Kas, K. Larbig, R. Lim, K. Smith, D. Spencer, G. K.S. Wong, Z. Wu, I. T. Paulsen, J. Relzer, M. H. Saler,

R. E.W. Hancock, S. Lory, and M. V. Olson. Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. *Nature*, 406(6799):959–964, 8 2000. ISSN 00280836. doi: 10.1038/35023079.

Edward Stuttfeld, Christopher H.S. Aylett, Stefan Imseng, Daniel Boehringer, Alain Scaiola, Evelyn Sauer, Michael N. Hall, Timm Maier, and Nenad Ban. Architecture of the human mTORC2 core complex. *eLife*, 7, 2 2018. ISSN 2050084X. doi: 10.7554/eLife.33101.

Ming-Yuan Su, Kyle L Morris, Do Jin Kim, Yangxue Fu, Rosalie Lawrence, Goran Stjepanovic, Roberto Zoncu, and James H Hurley. Hybrid Structure of the {RagA}/C-Ragulator {mTORC1} Activation Complex. *Mol Cell*, 68(5):835–846, 12 2017. doi: 10.1016/j.molcel.2017.10.016. URL `http://dx.doi.org/10.1016/j.molcel.2017.10.016`.

Catherine Sutherland and Katsuhiko S. Murakami. An Introduction to the Structure and Function of the Catalytic Core Enzyme of Escherichia coli RNA Polymerase . *EcoSal Plus*, 8(1), 2018. ISSN 2324-6200. doi: 10.1128/ecosalplus.esp-0004-2018.

Doug Sweetser, Michael Nonet, and Richard A Young. Prokaryotic and eukaryotic RNA polymerases have homologous core subunits (Saccharomyces cerevisiae/RPB2). Technical report, 1987.

Y Takagi, T Kobayashi, M Shiono, L Wang, X Piao, G Sun, D Zhang, M Abe, Y Hagiwara, K Takahashi, and O Hino. Interaction of folliculin (Birt-Hogg-Dubé gene product) with a novel Fnip1-like (FnipL/Fnip2) protein. *Oncogene*, 27(40):5339–5347, 9 2008. ISSN 0950-9232. doi: 10.1038/onc.2008.261.

Guang Tang, Liwei Peng, Philip R. Baldwin, Deepinder S. Mann, Wen Jiang, Ian Rees, and Steven J. Ludtke. EMAN2: An extensible image processing suite for electron microscopy. *Journal of Structural Biology*, 157(1):38–46, 1 2007. ISSN 10478477. doi: 10.1016/j.jsb.2006.05.009.

Andrew R Tee, Diane C Fingar, Brendan D Manning, David J Kwiatkowski, Lewis C Cantley, and John Blenis. Tuberous sclerosis complex-1 and -2 gene products function together to inhibit mammalian target of rapamycin ({mTOR})-mediated downstream sig-

naling. *Proc Natl Acad Sci {USA}*, 99(21):13571–13576, 10 2002. ISSN 0027-8424. doi: 10.1073/pnas.202476899. URL http://dx.doi.org/10.1073/pnas.202476899.

Romana Tomasoni and Anna Mondino. The tuberous sclerosis complex: balancing proliferation and survival. *466 Biochemical Society Transactions*, 39(2), 2011. doi: 10.1042/BST0390466.

Jorge R. Toro, Gladys Glenn, Paul Duray, Thomas Darling, Gregor Weirich, Berton Zbar, Marston Linehan, and Maria L. Turner. No Title. *Archives of Dermatology*, 135(10):1195–1202, 10 1999. doi: 10.1001/archderm.135.10.1195. URL http://archderm.jamanetwork.com/article.aspx?doi=10.1001/archderm.135.10.1195.

Zhi Yang Tsun, Liron Bar-Peled, Lynne Chantranupong, Roberto Zoncu, Tim Wang, Choah Kim, Eric Spooner, and David M. Sabatini. The folliculin tumor suppressor is a GAP for the RagC/D GTPases that signal amino acid levels to mTORC1. *Molecular Cell*, 52(4):495–505, 2013. ISSN 10972765. doi: 10.1016/j.molcel.2013.09.016.

F. W. Twort. AN INVESTIGATION ON THE NATURE OF ULTRA-MICROSCOPIC VIRUSES. *The Lancet*, 186(4814):1241–1243, 12 1915. ISSN 0140-6736. doi: 10.1016/S0140-6736(01)20383-3.

Alessandro Vannini and Patrick Cramer. Conservation between the RNA Polymerase I, II, and III Transcription Initiation Machineries, 2 2012. ISSN 10972765.

Dmitry G Vassylyev, Marina N Vassylyeva, Jinwei Zhang, Murali Palangat, Irina Artsimovitch, and Robert Landick. Structural basis for substrate loading in bacterial RNA polymerase. *Nature*, 448:163–168, 2007. doi: 10.1038/nature05931.

Cathy D. Vocke, Youfeng Yang, Christian P. Pavlovich, Laura S. Schmidt, Michael L. Nickerson, Carlos A. Torres-Cabala, Maria J. Merino, McClellan M. Walther, Berton Zbar, and W. Marston Linehan. High frequency of somatic frameshift BHD gene mutations in Birt-Hogg-Dubé-associated renal tumors. *Journal of the National Cancer Institute*, 97(12):931–935, 2005. ISSN 00278874. doi: 10.1093/jnci/dji154.

Dong Wang, David A. Bushnell, Kenneth D. Westover, Craig D. Kaplan, and Roger D. Kornberg. Structural Basis of Transcription: Role of the Trigger Loop in Substrate Specificity and Catalysis. *Cell*, 127(5):941–954, 12 2006. ISSN 00928674. doi: 10.1016/j.cell.2006.11.023.

Hong Wei Wang and Xiao Fan. Challenges and opportunities in cryo-EM with phase plate, 10 2019. ISSN 1879033X.

Andrew Waterhouse, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian T Heer, Tjaart A P De Beer, Christine Rempfer, Lorenza Bordoli, Rosalba Lepore, and Torsten Schwede. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46(W1):W296–W303, 2018. doi: 10.1093/nar/gky427.

Finn Werner. Structure and function of archaeal RNA polymerases, 9 2007. ISSN 0950382X.

Finn Werner and Dina Grohmann. Evolution of multisubunit RNA polymerases in the three domains of life, 2 2011. ISSN 17401526. URL `www.nature.com/reviews/micro`.

Weimin Wu, Julie A. Thomas, Naiqian Cheng, Lindsay W. Black, and Alasdair C. Steven. Bubblegrams reveal the inner body of bacteriophage ϕKZ, 1 2012. ISSN 10959203.

Maria Yakunina, Tatyana Artamonova, Sergei Borukhov, Kira S Makarova, Konstantin Severinov, and Leonid Minakhin. A non-canonical multisubunit RNA polymerase encoded by a giant bacteriophage. *Nucleic Acids Research*, 43(21):10411–10420, 2015. doi: 10.1093/nar/gkv1095.

Ming Yan, Marie Claude Gingras, Elaine A. Dunlop, Yann Nouët, Fanny Dupuy, Zahra Jalali, Elite Possik, Barry J. Coull, Dmitri Kharitidi, Anders Bondo Dydensborg, Brandon Faubert, Miriam Kamps, Sylvie Sabourin, Rachael S. Preston, David Mark Davies, Taren Roughead, Laëtitia Chotard, Maurice A.M. Van Steensel, Russell Jones, Andrew R. Tee, and Arnim Pause. The tumor suppressor folliculin regulates AMPK-dependent metabolic transformation. *Journal of Clinical Investigation*, 124(6):2640–2650, 6 2014. ISSN 15588238. doi: 10.1172/JCI71749.

Haijuan Yang, Xiaolu Jiang, Buren Li, Hyo J. Yang, Meredith Miller, Angela Yang, Ankita

Dhar, and Nikola P. Pavletich. Mechanisms of mTORC1 activation by RHEB and inhibition by PRAS40. *Nature*, 552(7685):368–373, 12 2017. ISSN 0028-0836. doi: 10.1038/nature25023.

Huirong Yang, Jia Wang, Mengjie Liu, Xizi Chen, Min Huang, Dan Tan, Meng Qiu Dong, Catherine C.L. Wong, Jiawei Wang, Yanhui Xu, and Hong Wei Wang. 4.4 Å Resolution Cryo-EM structure of human mTOR Complex 1. *Protein & Cell*, 7(12):878, 12 2016. ISSN 16748018. doi: 10.1007/S13238-016-0346-6.

Shigeyuki Yokoyama, Shigeyuki Yokoyama, Marina N. Vassylyeva, and Shigeyuki Yokoyama. Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6. Å resolution. *Nature*, 417(6890):712–719, 6 2002. ISSN 00280836. doi: 10.1038/nature752.

Ryo Yonehara, Shigeyuki Nada, Tomokazu Nakai, Masahiro Nakai, Ayaka Kitamura, Akira Ogawa, Hirokazu Nakatsumi, Keiichi I. Nakayama, Songling Li, Daron M. Standley, Eiki Yamashita, Atsushi Nakagawa, and Masato Okada. Structural basis for the assembly of the Ragulator-Rag GTPase complex. *Nature Communications*, 8(1):1–10, 2017. ISSN 20411723. doi: 10.1038/s41467-017-01762-3. URL `http://dx.doi.org/10.1038/s41467-017-01762-3`.

Shin Ichiro Yoshimura, Andreas Gerondopoulos, Andrea Linford, Daniel J. Rigden, and Francis A. Barr. Family-wide characterization of the DENN domain Rab GDP-GTP exchange factors. *Journal of Cell Biology*, 191(2):367–381, 2010. ISSN 00219525. doi: 10.1083/jcb.201008051.

Authors Yuhong Zuo, Thomas A Steitz Correspondence, Yuhong Zuo, and Thomas A Steitz. Crystal Structures of the E. coli Transcription Initiation Complexes with a Complete Bubble. *Molecular Cell*, 58:534–540, 2015. doi: 10.1016/j.molcel.2015.03.010.

Dapeng Zhang, Lakshminarayan M. Iyer, Fang He, and L. Aravind. Discovery of novel DENN proteins: Implications for the evolution of eukaryotic intracellular membrane structures and human disease. *Frontiers in Genetics*, 3(DEC):1–10, 2012. ISSN 16648021. doi: 10.3389/fgene.2012.00283.

Gongyi Zhang, Elizabeth A. Campbell, Leonid Minakhin, Catherine Richter, Konstantin Severinov, and Seth A. Darst. Crystal structure of thermus aquaticus core RNA polymerase at 3.3 å resolution. *Cell*, 98(6):811–824, 9 1999. ISSN 00928674. doi: 10.1016/S0092-8674(00)81515-9.

Yong Zhang, Xinsheng Gao, Leslie J. Saucedo, Binggen Ru, Bruce A. Edgar, and Duojia Pan. Rheb is a direct target of the tuberous sclerosis tumour suppressor proteins. *Nature Cell Biology*, 5(6):578–581, 6 2003. ISSN 1465-7392. doi: 10.1038/ncb999.

Lingling Zhao, Xin Ji, Xiangxiang Zhang, Lin Li, Yaping Jin, and Wei Liu. FLCN is a novel Rab11A-interacting protein that is involved in the Rab11A-mediated recycling transport. *Journal of Cell Science*, 131:jcs218792, 2018. doi: 10.1242/jcs.218792.

Shawn Q Zheng, Eugene Palovcak, Jean-Paul Armache, Kliment A Verba, Yifan Cheng, and David A Agard. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nature Methods*, 14(4):331–332, 4 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4193.

Jasenko Zivanov, Takanori Nakane, Björn O. Forsberg, Dari Kimanius, Wim J.H. Hagen, Erik Lindahl, and Sjors H.W. Scheres. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife*, 7, 11 2018. doi: 10.7554/ELIFE.42166.

Appendices

Appendix A: Principles of Transmission Electron Microscopy

1 Introduction

A central topic of study in molecular biology is the understanding of how macromolecules develop their role within the cell. A discipline that has substantially contributed to the major molecular biology knowledge is structural biology. The aim of structural biology is to elucidate *three-dimensional* (3D) structures of macromolecular complexes to high enough resolution to provide an atomic model to further understand their function in biological processes. There are different approaches to determine the 3D structure of macromolecules. Throughout history, the most popular technique has been X-ray crystallography, which yields high-resolution structures providing that the complex of interest can crystallise. Another technique which can provide exclusive information about dynamics and interactions is *nuclear magnetic resonance* (NMR), however it is very complicated for larger molecules. Apart from their own limitations, both techniques require a high concentration of the pure complex, which can be difficult to achieve with certain samples. Using *cryogenic electron microscopy* (cryo-EM) the structure of multiprotein complexes that were unable to crystallise have been determined, and over the last decades the number of structures deposited in the *Electron Microscopy Data Bank* (EMDB) has dramatically increased, which is an example of the popularity of this technique.

The first *transmission electron microscope* (TEM) was built in 1932 by the German physicist Ernst Ruska (Knoll and Ruska, 1932) who was later awarded with the Nobel Prize in Physics in 1986. The main challenge of observing biological samples under the TEM is to preserve their integrity under the electron microscope vacuum requirement. Initially, specimens were stained with a solution of heavy metal salt, such as uranyl acetate or uranyl molybdate, in order to visualise them under the electron microscope (Brenner and Horne, 1959). However, high resolution features and interior details were lost with this technique. Although specimen fixation is still used nowadays, especially at initial phases of structural analysis, it was vitrification that allowed high-resolution structure determination. Firstly, Taylor and Glaeser (Taylor and Glaeser, 1974) accomplished 3 Å resolution diffraction patterns with frozen hydrated catalase crystals, demonstrating that freezing could preserve sample integrity under the TEM vacuum. However, it was the implementation of plunge freezing by Dubochet and colleagues (Adrian et al., 1984; Dubochet et al., 1982) that implemented a practical methodology for freezing of hydrated samples in vitreous ice, therefore introducing cryo-EM.

Combination of plunge freezing with the idea of computationally averaging many TEM images of individual molecules in equal or similar conformations (Frank et al., 1978), became what we now known as single-particle cryo-EM. In *single-particle analysis* (SPA) micrographs are interpreted as *two-dimensional* (2D) projections from an object. Combining such projections and taking a sufficiently large angular range, the object of study can be eventually recovered in 3D reconstruction (De Rosier

and Klug, 1968). Such reconstruction is a 3D map revealing not only the shape of the object but also density variations in the interior of the object. In order to limit radiation damage, microscopes are used at low dose, hence several tens of thousands of particles need to be averaged to increase the *signal-to-noise* (SNR) and to provide several different views needed for the 3D reconstruction.

Since the development of SPA, fast progress was made with symmetric particles, such as icosahedral viruses. In 1997, SPA of Hepatitis B virus resolved α- helices (Böttcher et al., 1997; Conway et al., 1997), and in 2010 a *de novo* atomic reconstruction was built for aquareovirus and adenovirus (Liu et al., 2010; Zhang et al., 2010). Progress on SPA for asymmetric complexes have been slower because averaging and reconstruction is more complicated. First developments in image processing for asymmetric particles were made in 1980 with study of negatively-stained ribosomes (Radermacher et al., 1987a; Verschoor et al., 1984). Such techniques were gradually improved and adapted for cryo-EM, until the structure of the ribosome was determined near to 10 Å resolution (Gabashvili et al., 2000). Since 2013 evolution in cryo-EM has been that fast that it is considered a revolution (Kühlbrandt, 2014), mainly by virtue of the development of direct electron detectors (McMullan et al., 2016) and new powerful computational algorithms, which have facilitated the generation of reliable and high enough resolution structures to obtain atomic level details with cryo-EM. Nowadays, SPA techniques have improved to a point where high-resolution 3D maps in the range of 2-4 Å can be routinely achieved and then used to build atomic models, which shows that this technique can currently reach similar resolutions to those obtained by X-ray Crystallography.

2 The electron microscope

A TEM is an analytical tool that allows visualisation of specimens in the microspace and nanospace order by virtue of high energy electrons (100 keV or higher). Electrons have a much shorter wavelength compared to light, hence, electron microscopy can reveal levels of detail unreachable by light microscopy, allowing detailed structural analysis at high resolution. Similar to the conventional light microscope, the electron microscope is formed by a radiation source, a system of lenses and a detection and recording system. Since electrons are also scattered by molecules in the air, the entire column of the microscope containing the beam needs to be at vacuum. To preserve the vacuum during sample insertion, there are airlock mechanisms. The requirement of vacuum, however, is incompatible with the naturally hydrated state of biological samples, which was resolved by the introduction of vitrification to prepare cryo-EM samples (Adrian et al., 1984; Dubochet et al., 1982).

The resolution of electron microscopes have improved up to the point that it is currently possible to resolve single atoms. In order to obtain such high-resolution information it is essential to have

a highly coherent electron source, a well-aligned system, a precise detector and a convenient sample (Franken et al., 2020). And yet, even if all those conditions are met some samples do not enable high resolution, such as extremely radiation sensitive specimens or those with high tendency to aggregate.

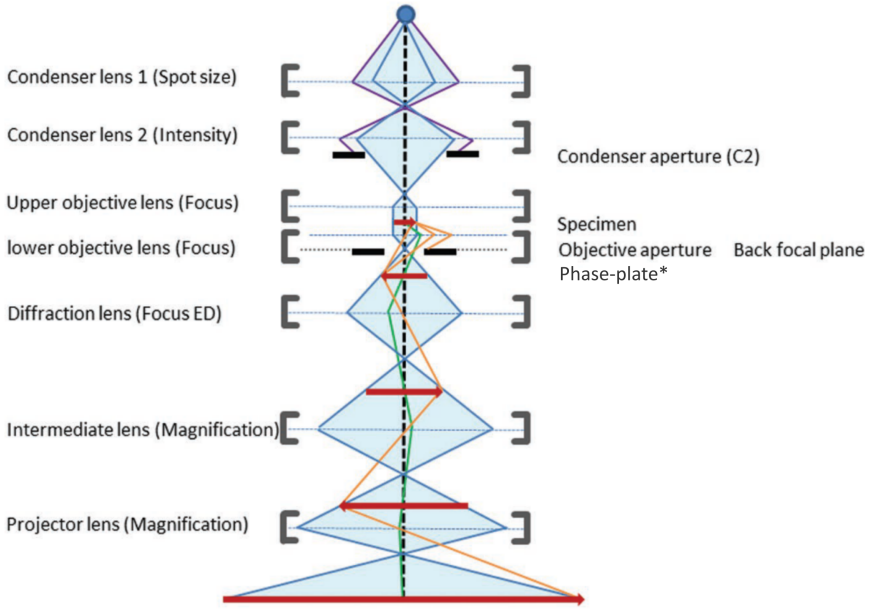


Figure 1: Schematic diagram of the electron microscope internal structure. The beam is shown in blue, grey brackets represent lenses, label with their name and the microscope function that controls them, black bars are apertures, and the red arrows is the specimen and its intermediate images. The paths of scattered are drawn in orange and green. Phase plate is indicated with an * as it is optional. Figure adapted from (Franken et al., 2020) under a Creative Commons license.

2.1 Electron sources

In an electron microscope, electrons are emitted by the cathode, kept at high negative voltage. There are two types of electron sources that are commonly used: thermionic cathode and *field emission gun* (FEG). The principle of thermionic cathodes is that when metals are heated at sufficiently high temperatures, electrons will be released from their surface. The standard thermionic cathode is a v-shaped wire made of tungsten heated to 2 500-3 000 K. Another typical material used as electron source is *lanthanum hexaboride* (LaB_6), which produces electrons from a smaller area at the crystal vertex and have a lower work function, therefore producing a more coherent beam and leading to higher brightness than that of the tungsten filament. Additionally, LaB_6 cathodes have a longer lifespan because they can be heated at lower temperatures than those made of tungsten, approximately 1 400-2 000 K. The cathode filament is surrounded by a cup-shaped electrode with an aperture of ∼1 mm and kept at a slightly more negative potential than that of the cathode (∼500 V). This metallic

disc is called Wehnelt cylinder and allows the electrons from the cathode to exit towards a restricted area, thus focusing the emitted electrons towards the optic axis (Frank, 2006).

In FEGs the exit of electrons is enhanced and directionally focused by the application of an intense electric field (in the region of 2 kV) to a very thin tip of tungsten, 15-25 nm radius. The restriction of emission from the cathode to a very small angular range by means of the extremely thin tip, produces extraordinary brightness and beam coherence. Moreover, all the electrons have the same energy, which ultimately defines their wavelength. At present FEGs are the current choice for high-electron microscopes, however they are more susceptible to contamination and require a much higher voltage. Regardless the type of cathode used as electron source, the emitted electrons are then accelerated through high energy voltages in the range of 50-300 kV, depending on the microscope.

2.2 Lens system

After being emitted, electrons travel 1-2 m through the column, kept at vacuum, from the top to the bottom. Unlike photons, electrons cannot be focused through glass lenses and, instead, the column contains a set of electromagnetic lenses that contribute to the magnified image formation of the object. Each lens is formed by a coil of copper wire with current running throughout and a hole in the centre through which the beam travels.

After the electron source there are two condenser lenses that control the spot size (C1) and the intensity (C2) (Figure 1). These lenses are part of the illumination system and define how the beam impacts on the specimen. The first lens, C1, is normally located immediately below the Wehnelt cylinder and forms a demagnified image of the source, whereas the second condenser is used to define the physical spot size and the angle of convergence that illuminates the specimen. Additionally, the angle of convergence can be controlled by the C2 aperture. The C2 aperture removes electrons with large angular spread, hence improving beam coherence. The beam coherence determines how different electrons from the same source are scattered and generate a different diffraction pattern, and depends on how small is the source and also the microscope optics. Smaller apertures are usually preferable but they also reduce beam intensity, therefore the smallest possible C2 aperture that still provides sufficient beam intensity should be used. In EM it is important to illuminate the specimen using a parallel electron beam in order to prevent image deterioration. That means that the angle with which the electron beam interacts with the specimen is zero, which is known as parallel illumination (Glaeser et al., 2011). In microscopes with two condenser lenses, parallel illumination can only be achieved for a single beam intensity. For this reason, modern microscopes usually have more condensers, allowing parallel illumination at a wide range of C1 and C2 positions (Herzik, 2021).

Below the condenser lenses sit the objective lens, which are the first lens contributing to image formation. The specimen is normally located in the middle of the objective lens, immersed between the magnetic field, which makes the design of mechanisms of insertion quite complex. After interacting with the sample, electrons are focused by the objective lens, which also provide the primary magnification (20-50K). Parallel beams are focused at the the back focal plane with electrons scattered at the same angle being focused at the same point. Therefore, the focusing of incident beams is achieved at the back focal plane, where a diffraction pattern is formed. The objective aperture is placed at the back focal plane and determines what portion of the scattered rays contribute in the image formation, excluding electrons scattered at high angles, therefore improving image contrast and acting as a low-pass filter. Low energy electrons are usually caused by inelastic scattering. Removing electrons scattered at high angles with the objective aperture increases amplitude contrast, however it also limits the resolution. Therefore, a relatively large objective aperture is recommended for high-resolution data collections (Cheng et al., 2015).

After the objective lenses the beam goes through the intermediate lenses, located between the object and the projector lenses. There can be one or several intermediate lenses that can be tuned in strength to produce a wide range of magnifications. In some cases there are particular settings to change from imaging to diffraction mode. Finally, the projector lenses magnify the image and project it to the a fluorescent screen (usually phosphor) or another system to acquire and store images. The fluorescent screen is normally a metal plane coated with a layer of fluorescent material that converts the electrons into light, thus allowing direct visualisation of the object image. Regions where electrons do not pass through the sample appear dark in the image while regions where electrons are strongly scattered appear brighter in the image, with a range of greys depending on the way electrons interact with the sample and how they are scattered.

Electromagnetic lenses lead to a series of aberrations that affect the focal point within the lens, hence affecting image quality and, ultimately, limiting resolution: *spherical aberration* (Cs), *chromatic aberration* (Cc), astigmatism and coma (Figure 2) (Hetherington, 2004). These aberrations are especially important in the condenser lenses, which determine the beam quality, and the objective lens, where aberrations can be magnified. Spherical aberration is the effect of the electromagnetic field in the lens being variable, which means that beams passing further away from the lens centre will be focused more strongly than electrons travelling close to the lens centre. As a consequence the focal length is variable the image point is blurred (Figure 2). Spherical aberration can be corrected with small objective apertures or with asymmetric spherical aberration correctors (Hosokawa et al., 2013).

Chromatic aberration arises when the lens focuses electrons depending on their energies. Electrons

with lower energies are bent more strongly by the lenses, which results in blurring of the final image (Figure 2 C). Variations in electron energy are usually associated with high voltage fluctuations, electron energy spread in the primary beam and loss of energy during inelastic scattering events. The development of FEG electron sources has substantially improved the beam coherence, therefore decreasing chromatic aberration. Furthermore, chromatic aberration can be reduced using a monochromator, which select a particular energy spread (Tiemeijer et al., 2012).

Astigmatism in the electron microscope appears when there is an asymmetric magnetic field in the lens, which usually results in the lens being stronger in one direction than in the perpendicular direction. The appearance of astigmatism results in the focus point varying within x and y position in the lens, which usually results in the image of a point becoming an ellipse instead of a circle. The majority of microscopes include special corrector elements, stigmators, underneath the objective lens to compensate astigmatism (Orlova and Saibil, 2011) and, ultimately, small remaining levels can be corrected computationally.

Coma aberration appears when the direction of illumination is different to the direction of the true optical axis, resulting on a point being focused in different positions. The extent to which coma is present is usually determined measuring the defocus of images taken at different beam tilts. The axis with respect to the objective lens by the electron beam where coma become negligible is called the coma-free axis. Coma needs to be minimized when setting up high-resolution collections (Glaeser et al., 2011). Originally, a method for coma-free alignment was proposed by Zemlin et al. (1978), however the procedure was based in the evaluation of several optical diffractograms, which was quite laborious. Currently, in modern electron microscopes and detectors coma-free alignment can be performed automatically.

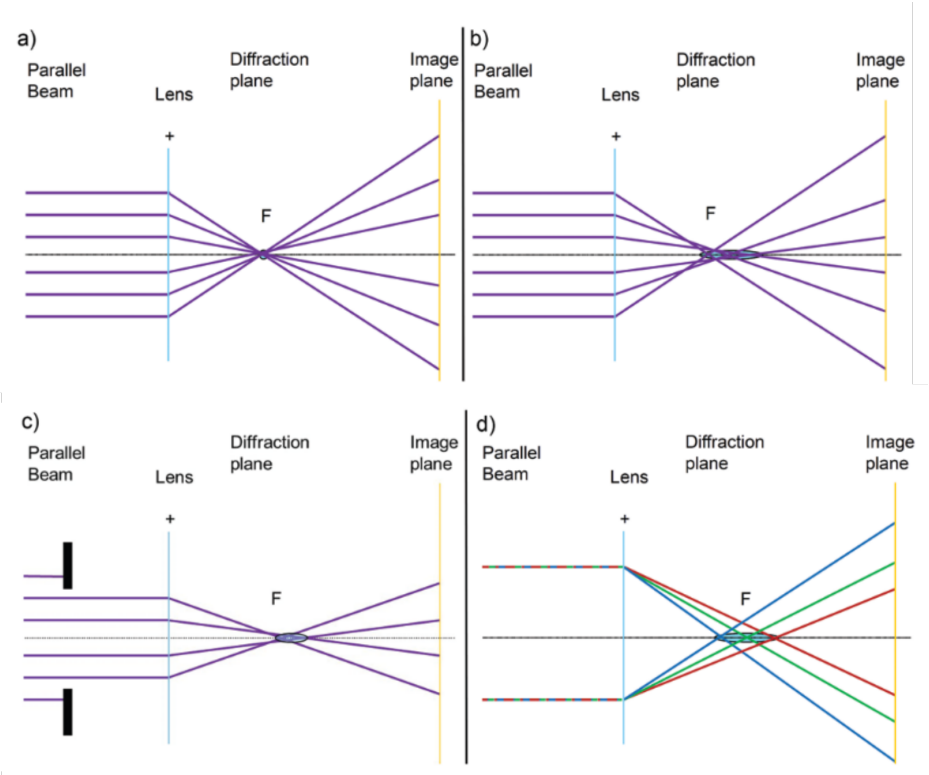


Figure 2: Lens aberrations. a) A perfect lens has a focal point (F). b) Spherical aberration originates from the fact that electrons that go through the lens experience a stronger electromagnetic field when passing closer to the coil compared to electrons that pass through the center of the lens, resulting in a focal area of least confusion, rather than a point. c) A C2 aperture can reduce the size of this area. d) Chromatic aberration results from variations in electron speed. Slower electrons (blue) will be focused more compared to faster electrons (red) and the electron magnetic lens will have a focus range rather than a single focal point. Figure adapted from (Franken et al., 2020) under a Creative Commons License.

2.3 Detector systems

The last component of the electron microscope that have a great impact on the image quality is the detection system, located at the image plane. Originally, photographic films or sctintillator-based digital cameras, such as *charge-coupled device* (CCD) and *complementary metal oxide semiconductor* (CMOS) were the standard detectors for recording images of frozen specimens under the electron microscopy. Nevertheless, the development of *direct electron devices* (DED) (Bammes et al., 2012; McMullan et al., 2016) caused a resolution revolution in the the cryo-EM field (Kühlbrandt, 2014). The performance of detectors can be specified with two parameters: the *detective quantum efficiency* (DQE) and the *modulation transfer function* (MTF) (McMullan et al., 2009a). DQE provides a measurement of how much the detector degrades the original signal of the image by comparing the

SNR in the output with that in the input, as defined in Equation 1:

$$DQE = \frac{(S/N)^2_{\mathrm{OUT}}}{(S/N)^2_{\mathrm{IN}}} \tag{1}$$

DQE is normally plotted as a function of spatial frequency. A detector with a DQE of 1 would be a perfect detector with zero addition of noise. For a pixelated detector, the maximum spatial frequency that can be recorded by the detector depends on the pixel size. This defines the Nyquist limit, which corresponds to a wavelength of 2 pixels and represents the largest spatial frequency that can be obtained by a pixelated image. Thus, the Nyquist limit sets the maximum resolution achievable at twice the pixel size, spatial frequencies above the Nyquist limit suffer aliasing. The MTF, which is also a function of spatial frequency, measures how the detector responds to a signal of fixed amplitude and is set by different factors, such as the pixel size or the spreading of electrons in the detector. A detector with MTF of 1 would retain perfectly the relative amplitude of that spatial frequency.

Photographic films were the most popular detector system in cryo-EM for several years. The basic principle of photographic films is that when electrons impact on a photographic emulsion of silver halide crystals, just as light, they make silver halide grains which are later developed into visible silver grains. Films provide decent resolution, however they are associated with laborious and tedious procedures that prevent processing of large number of images and automated data acquisition. Scintillator-based digital cameras, such as *charge-coupled device* (CCD) or *complementary metal oxide semiconductor* records electron indirectly by transforming them into photons with a phosphor scintillator layer. The CCD was invented by W.S. Boyle and G. E. Smith, which were later awarded with a Nobel Prize (Boyle and Smith, 1970), and soon gained popularity due to the implementation of automated data collection systems (Suloway et al., 2005). Compared to films, scintillator-based detectors have the ability to directly record digital images, allowing high-throughput data collection and control of imaging parameters, however, they have a low DQE at high spatial frequencies close to Nyquist because they rely on an indermediate process and usually suffer delocalisation. Although films provided higher resolutions than CCD cameras, the latter were more popular due to their power to record many images automatically.

CCD cameras were the most common detector until the development of DED (Bammes et al., 2012; Clough et al., 2014; McMullan et al., 2016). Direct electron detection using a CMOS-based active pixel was a promising alternative to CCD cameras. These cameras directly detect electrons without the need for a scintillator. Moreover, they provide immediate access to digital images, while having a comparable DQE at high spatial frequencies to that of photographic films (McMullan et al., 2009a,b), being able to reach resolutions close to Nyquist (Bammes et al., 2012). Beam-induced movement is another limiting factor to achieve high resolution in cryo-EM (Brilot et al., 2012). The high frame

rate of CMOS-based DED allowed for correction of such movement by recording a movie throughout exposure (McMullan et al., 2016). A single exposure is fractionated into a set of frames, which can then be aligned to correct for movement of the sample and dose-weighted to account for radiation damage of the specimen.

DED can be operated in integrating mode or counting mode. In counting mode, a low electron dose is used to illuminate the sample so that individual electron impacts can be identified. The low dose minimises the probability of two electrons hitting the same point being detected as a single one, a phenomenon known as conicidence loss (McMullan et al., 2016). Electron counting further enhances the DQE of DED. Apart from being operated at low dose, the detection of individual electrons requires a rapid frame rate, to minimise coincidence loss. The new generation of detectors are known as super-resolution DED, and they usually have a high frame rate and are able to find the centre of the peak to sub-pixel accuracy (Li et al., 2013b).

3 Principles of image formation

Unlike X-rays, electrons interact strongly with matter by scattering, and they can be focused by electromagnetic lenses, hence allowing the study of matter by imaging rather than only by diffraction, which includes structural phase information. The wavelength of an electron is a function of its speed, therefore it decreases with increasing acceleration voltage. The lower the wavelength, the higher the resolving power (Franken et al., 2020). For these properties electrons are quite optimal to study biological specimens at high resolutions, however there are two main aspects to be considered during imagining: samples must be analysed at vacuum and they can suffer radiation damage.

A TEM image is formed by the interaction of electron waves (beam) with the sample. After electrons interact with the sample the resulting image is magnified through a series of lenses (see section 2.2). Upon sample interaction, electrons can be elastically or inelastically scattered. Elastic scattering involves no transfer of energy and has a wide angular distribution, hence providing high-resolution information. On the other hand, inelastic scattering involves transfer of energy. Inelastically scattered electrons can either return to a lower energy state or can be ejected as secondary electrons generating highly reactive free radical species (Frank, 2006). Both low energy electrons and free radicals can damage the sample, hence creating radiation damage, and increasing undesired background. Their effect can be attenuated using energy filters that prevent low energy electrons from arriving to the detector, which results in an increase of the amplitude contrast.

After the beam interacts with the sample, each point produces a spherical wave (scattered) and a plane wave (unscattered). Electrons that are elastically scattered contribute to the contrast of the

image. On the one hand, the scattered electrons at high angles are removed by placing an aperture at the back focal plane of the objective lens, which generates amplitude contrast. On the other hand, the phase difference between elastically scattered electrons and unscattered electrons when they are combined at the image plane, creates phase contrast. Biological specimens are usually approximated as weak phase objects, which means that they do not change the amplitude of the incident electron wave, but alter the phase slightly. For this reason, there is hardly any contrast. Phase contrast,is generated by going out of focus by allowing the scattered wave (spherical) to propagate a away from the initial point. Alternatively, phase contrast can be generated by the introduction of a phase plate at the the back focal plane of the objective lens (see below) (Franken et al., 2020).

The *contrast transfer function* (CTF) is a mathematical expression that defines how the transmission electron microscope transfer information to the image as a function of spatial frequency (k) (Frank, 2006). This mathematical expression is a quantitative relationship between the wave function of the electron beam after sample interaction and the contrast of the final image. The CTF includes a sin and cos terms as per Equation 2:

$$CTF(k) = 2[sin(\gamma(k)) - Wcos(\gamma(k))] \tag{2}$$

where W is a constant related to the ratio of amplitude contrast to phase contrast, and, in the absence of astigmatism, γ(k) is represented by Equation 3:

$$\gamma(k) = 2\pi(-0.5\Delta z\lambda k^2 + 0.25C_s\lambda^3 k^4) \tag{3}$$

where λ represents the wavelength of electrons, C_s is the spherical aberration coefficient and Δz is the defocus (distance of the specimen to the focal plane). Defocus values are negative when the focal plane of the objective lens is below the sample (underfocuse) and positive when the focal plane is above the sample (overfocus).

According the Equations 2 and 3, contrast provided by features of different spatial frequencies follows a pattern of oscillations that includes positive and negative values between -1 and 1, and zero values (Figure 3). When the CTF is zero, no contrast is transferred and the information corresponding to those spatial frequencies is lost. The position where the CTF is zero can be modified by the defocus value, therefore at high-resolution collections we usually collect at different defocus values to recover the information lost for each value. Due to the imperfections on the image system, the CTF is attenuated at high spatial frequencies (Figure 3). This effect can be modelled by an envelope function which includes terms accounting for the beam incoherence, beam energy spread and sample drift. This envelope function is also dependent on the defocus value (Figure 3). Therefore, it is possible to

tune the defocus value to improve contrast in high resolution spatial frequencies, using large defocus values. However at this values the decay of the CTF at high spatial frequencies is also higher.

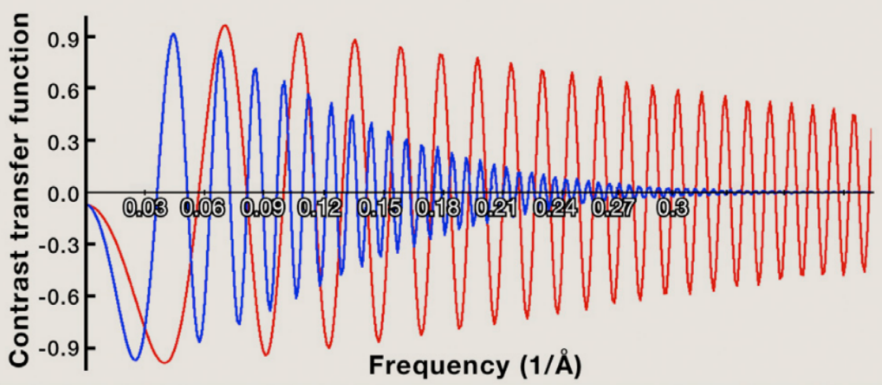


Figure 3: The contrast transfer function. Simulations of CTF at 1.2 μm (red) and 3.0 μm (blue) defocuses , with an acceleration voltage of 200 kV and angular spread of 0.07 mrad. Note that 3 μm defocus generates sufficient contrast for particles with a molecular weight of ∼300 kDa, CTF envelope drops to close to zero at 3-4 Å resolution. Figure adapted from (Cheng, 2015).

The CTF imposes another problem and that is the change of sign of the contrast (Figure 3). For this reason images must be corrected to obtain an undistorted image of the object, where contrast is not affected by the changes in intensity or the sign of the CTF and envelope function. The simplest solution is to apply a phase correction to make all the contrast values positive by phase flipping, which is achieved by multiplying by -1 the experimental images in Fourier space at the spatial frequencies where the contrast is negative (Frank, 2006). However, this approach does not correct for the decay of intensity at high spatial frequencies caused by the envelope function. To correct for the fading of contrast at high spatial frequencies a more sophisticated approach such as the Wiener filter is needed (Frank, 2006).

A simplified expression for image formation in reciprocal sapce (k) is given by Equation 4:

$$I(k) = CTF(k)\Phi(k) + N(k) \tag{4}$$

Where CTF(k) is the CTF (including the envelope function), Φ(k) is the Fourier transform of the projection of the imaged object, and N(k) represents the noise, which is usually dependent on the spatial frequency. The Wiener filter is a function F(k) that when is multiplied by the Fourier transform of the experimental image in reciprocal space, results in an approximation of the Fourier transform of the projection of the object unaffected by the CTF and the addition of noise (Frank, 2006). For the case of a single 2D image of a particle, it can be demonstrated that the Wiener filter allowing for the

optimal approximation with minimized mean-square error is (Equation 5):

$$F(k) = \frac{CTF(k)}{CTF(k)^2 + \frac{1}{SNR(k)}} \tag{5}$$

A good estimation of the SNR is required to apply a Wiener filter to avoid excessive amplification of the noise, and enable recovery of high-resolution information. Despite the application of the Wiener filter, during data collection it is still required to collect data at different defocus values to recover information at the spatial frequencies values where the CTF is zero. Furthermore, to apply The Wiener filter it is necessary to estimate some parameters such as the spherical aberration, the defocus values or the astigmatism. This is normally done by mathematical approaches where different values are fitted to the power spectrum of the experimental images in a procedure named CTF estimation (Mindell and Grigorieff, 2003; Rohou and Grigorieff, 2015; Zhang, 2016).

Alternatively to the defocus scheme, the introduction of a phase plate at the back focal plane causes a phase shift between the scattered and unscattered beam. Phase plates are well-implemented in light microscopy (Zernike, 1955). In electron microscopy, the newest phase plate, also known as *Volta phase plate* (VPP), was developed by the Baumeister group (Danev and Baumeister, 2016; Danev et al., 2014). The VPP consists of a thin carbon film heated to 200-240 °C which builds up a charge at the back focal plane when interacting with the unscattered beam. This results in a phase shift of the unscattered beam from the scattered beam, which are combined again where the image is formed, generating a phase delay. In TEM, such phase shift converts the CTF from a sine-shape function to a cosine-shape function, hence preserving the signal at low spatial frequencies and enhancing contrast (Wang and Fan, 2019). The phase shift contributes to the CTF of the image and can be determined during CTF estimation (Rohou and Grigorieff, 2015; Zhang, 2016). The introduction of a phase plate suppresses the requirement of imaging out of focus, however, imaging exactly at focus is time-consuming and implies a lack of CTF information. Therefore, phase plate collections are usually performed at a very small amount of defocus, making CTF correction possible. High resolution data collection using a VPP introduction a small defocus has been successful for the structure determination of small biological samples, such as the 64 kDa haemoglobin (Khoshouei et al., 2017) and the GPCR-G-protein complex (Liang et al., 2017) at atomic resolution.

4 Sample preparation

One of the main challenges of cryo-EM is to prepare specimens so that they are preserved and hydrated inside the vacuum of the microscope. Furthermore, biological specimens are usually low contrast which complicates their direct observation. Initially, negative staining with heavy salt metals, such as uranyl

acetate, was the solution to provide high enough contrast, however only the shape of the molecule could be reconstructed, without being able to observe the interior of molecules.

When sophisticated averaging methods were developed, it was possible to use other methods which provided less contrast but preserved molecules in a aqueous-like environment: glucose embedding, tannic acid, vitreous ice and cryo-negative staining (Frank, 2006). Ice embedding has proved to be the best method to preserve the interior features of molecules, being able to achieve atomic resolution and resulting in a really popular method in structural biology, cryo-electron microscopy (cryo-EM). Ideally, after vitrification macromolecules preserve the same particle distribution, orientation and structure that they previously had in solution. However, this is not the case for most of the macromolecules of interest, with main issues being adsorption to the air-water interface, decreased number of particles and preferential orientation (Glaeser, 2018). There are many different approaches to overcome these problems which are normally used in parallel, such as optimisation of buffer components or cross-linking to stabilise the sample, the use of detergents to cover the air-water interface, use of a rapid plunge-freezing robot (Noble et al., 2018), or immobilisation of the sample through an additional support film (de Martín Garrido et al., 2021).

4.1 Biochemical considerations

Single-particle analysis requires a homogeneous and stable sample for the determination of the 3D structure through computational averaging of images of the same molecule of study under identical or similar conformations (Frank et al., 1978). Therefore, protein purification procedures must be extensively optimised to provide a homogeneous sample suitable for single-particle cryo-EM.

The first step of sample preparation is the optimisation of the buffer components to try to stabilise sample in solution. Most popular strategies include the addition of stabilising co-solutes (polyamines, glycerol or trehalose), pH optimisation and salt concentration for the protein of study. However, some of these buffer additives are incompatible with freezing (de Martín Garrido et al., 2021). To protect macromolecules from hydrophobic interfaces which will cause sample denaturation, buffers are sometimes supplemented with detergents (Chen et al., 2019; Cheng et al., 2015), however these approaches require extensive screening and usually complicate optimisation of ice thickness (Glaeser et al., 2016). Moreover, many macromolecular complexes are incompatible with some detergents.

When the macromolecule in question is a multiprotein complex sample stabilisation usually requires further stabilisation. Indeed, most proteins regulating important cellular processes form weak and dynamic multiprotein complexes that usually oscillate between different conformations (Alberts, 1998), hence compromising their stabilisation in solution. Both the large size of these complexes and their

flexibility depending on their functional state complicates their study by X-ray crystallography and leaves cryo-EM as a suitable technique for their structural analysis. Each complex is stable in an aqueous solution under specific conditions. If the dissociation constant (K_d) of the complex in question is available, it is possible to estimate the concentration where the complex would remain intact. However, if this concentration is substantially high, the use of cross-linking agents might be necessary to stabilise the complex. The addition of a cross-linking agent is usually beneficial to hold together its subunits, however it usually introduces artifacts caused by non-specific interactions, inter-particle fixation and aggregation (de Martín Garrido et al., 2021). Density gradient cross-linking, which was initially investigated by Kastner an colleagues as "GraFix", (Kastner et al., 2008; Stark, 2010), successfully stabilise dynamic multiprotein complexes while diluting large aggregates and artifacts. However, samples derived from density gradients usually require buffer exchange prior to cryo-EM sample preparation, which can be problematic for some complexes.

4.2 Support choice

The standard support for TEM specimen preparation is a perforated amorphous carbon foil, with 1-2 µm diameter holes placed across a 3 mm metal mesh grid (Russo and Passmore, 2016) (Figure 4). The grid requires a small size due to the small column diameter (less than 1 cm diameter) and to avoid large beam movements. Upon vitrification, sample is spanned across the hole which can be distributed with irregular spacing and size hole, or located in regular arrays. An even distribution of the holes is more commonly used, as it facilitates automated data acquisition within the holes.

Grids are made of metal as they have conductivity for electrons and heat and they are malleable. The most popular metal for the grid is copper, although other materials such as gold are also commonly used (Figure 4). Samples can be deposited directly onto the grid to form a thin layer of liquid throughout the holes, or onto an additional thin layer of another material to provide adherent surface and minimise contact with the air-water interface (Russo and Passmore, 2016). Amorphous carbon is usually the standard material for perforated foils because it is relatively electron transparent, cheap and easy to produce. Thin and flat carbon films can be easily prepared by deposition onto mica using a carbon evaporator at high vacuum (Fujiyoshi, 1998). However, amorphous carbon foils also have some limitations, such as contribution to beam-induced movement (Brilot et al., 2012) and contribution to specimen charging (Russo and Passmore, 2014b) (Figure 4).

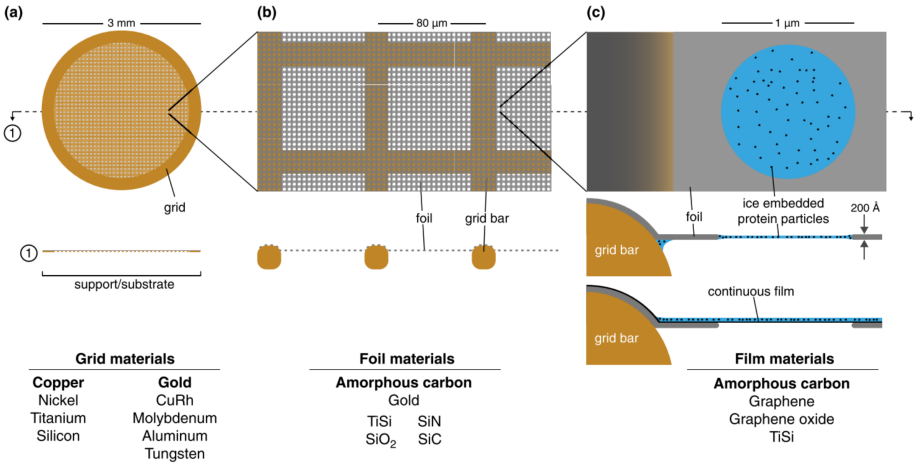


Figure 4: Types and design of cryo-EM specimen support: Top view and section diagrams of typical specimen support geometries, comprising a perforated foil on a metal mesh grid. Sometimes an additional thin continuous film is added to the foil to change its surface properties. Three different magnifications are shown (a)–(c) along with lists of materials used for each component of the support. The most commonly used materials are in bold. Figure adapted from Russo and Passmore (2016) under a Creative Commons License.

Immobilisation of particles to an extra layer of an appropriate support film onto the grid is a popular method to minimise interactions between macromolecules and the air water interface. An extra layer of amorphous carbon film, usually reduces charge-induced movement and the concentration of sample required (Grassucci et al., 2007), but it increases background noise, hence reducing contrast. There are some alternative supports, such as *titanium silicon* (TiSi) (Rhinow and Kühlbrandt, 2008) or graphene (Geim, 2009). Graphene is a single atom thick layer of sp^2 carbon which is nearly transparent at the spatial frequencies of interest of cryo-EM (Meyer et al., 2007; Pantelic et al., 2011), therefore having a low noise contribution which can be removed by Fourier masking. Moreover, it has high conductivity and mechanical strength. Due to its extraordinary properties graphene has been proposed as a theoretically ideal support for cryo-EM specimen preparation. There are well-established protocols for the direct transfer of Cu-grown graphene monolayers to cryo-EM grids (Naydenova et al., 2019; Regan et al., 2010), however the application of sample to these grids usually involves laborious handling steps which contribute to film contamination from the air, and expose protein samples to the air-water interface. Graphene monolayers are susceptible to airborne contaminants, which increase their hydrophobicity, hence preventing recruitment of macromolecules. Nevertheless, graphene can be chemically modified to render it hydrophilic (e.g converting it into graphene oxide) (Palovcak et al., 2018; Pantelic et al., 2010), or treated with hydrogen-plasma cleaning to convert it to its partially hydrogenated form (Russo and Passmore, 2014a), however the former usually causes the loss of graphene

properties and the latter requires expensive equipment. Recently, the direct wetted preparation of graphene supported cryo-EM samples using a support floatation block has been reported, avoiding airborne contamination and the requirement of further treatment (de Martín Garrido et al., 2021).

Prior to sample application, grids are usually washed with water and organic solvents (i.e. ethyl acetate) to remove residual organic contamination that has accrued during the grid production process. Further treatment of the grids is needed to remove any contamination that has not been washed off and to render them hydrophilic for proteins to adsorb onto them. This processes can be achieved generating a plasma with air (glow-discharge) or using more sophisticated equipment that uses amines, argon or hydrogen.

4.3 Negative staining

Negative staining was firstly introduced by Brenner and Horne (1959) and has been widely used to visualise specimens under the TEM at high contrast. Typically, to prepare negatively-stained specimens, the sample is initially applied to a carbon-coated grid, followed by application of a 2 % uranyl acetate solution. Then the excess of liquid is blotted away and grids are left to air-dry. Although uranyl acetate is the most commonly used stain because is particularly electron dense and it decays at a very low rate, other heavy salt metals, such as uranyl formate or ammonium molybdate can be used.

Under the TEM, heavy salt metals produce high-contrast, whereas biological molecules are composed of mostly relatively light atoms. Therefore, in a micrograph of a negatively-stained specimen the biological specimens appear lighter than the background, as they are scattered less strongly. Although heavy salt metals can go into specimen cavities to some extent, the structural information provided by negative staining is usually limited to the shape of the molecule. Furthermore, particles of a certain size can usually be stained unevenly, with sections far away from the carbon being under-stained and without contributing to the projection of the image, and the stain and air-drying may cause specimen distortion due to dehydration and flattening (Frank, 2006). For these reasons, the achievable resolution by negative staining is limited, however this technique is still currently used at early stages of structural analysis to assess sample quality through a quick and straightforward procedure.

4.4 Vitrified specimens

Flash-freezing of purified suspensions of macromolecular complexes prevents the crystallisation of water during the freezing process, resulting in vitrified complexes that preserve their native structures

(de Martín Garrido et al., 2021). This process is usually done in a chamber with controlled temperature and humidity (Adrian et al., 1984; Dobro et al., 2010; Dubochet et al., 1988; McDowall et al., 1983). After blotting, a thin layer remains in the hole, which is subsequently flash-frozen into low temperature, usually liquid ethane. The specimen must be frozen very quickly to prevent formation of crystalline ice, which would generate strong scattering on its own. For cryo-EM purposes, grids need to be made of a material with a low heat capacity (the amount of heat required to increase the temperature of one mole of substance by 1 K). Conversely, the cryogen needs to have a high heat capacity and has to be able to remain liquid in the widest range of temperature possible. Liquid ethane is usually used to flash-freeze biological specimens as it evaporates very slowly, hence preventing ice crystallisation. A mixture of liquid ethane and propane is also commonly used. After vitrification, frozen samples must be kept at liquid nitrogen to prevent melting and formation of crystalline ice.

Vitrified samples for electron microscopy must be thin to preserve phase contrast and prevent multiple scattering events (Grimm et al., 1996; Leapman and Ornberg, 1988), ideally no more than 100-200 nm and as thin as 20-30 nm (these values might change for each sample and the technique that is being used). The requirement of a thin layer of vitreous ice means that complexes are close to the air-water interface (Noble et al., 2018), which causes a wide range of damaging effects to the sample, including denaturation and preferential orientation. Moreover, when specimens are irradiated by the electron beam they move and build-up charge (Berriman and Rosenthal, 2012; Brink et al., 1998; Henderson and Glaeser, 1985) which ultimately limits resolution. The sample support itself can contribute to this movement (Russo and Passmore, 2014b), therefore the support choice is key for cryo-EM sample preparation. Currently, preparation of vitrified specimens is usually done with automated equipment that allow fine control of blotting parameters and control of the environmental conditions, such as temperature and humidity (Frederik and Hubert, 2005).

Ideally, macromolecules preserve the same identical particle distribution, orientation and structure than they had when they were in solution prior to freezing. However, many macromolecules of interest tend to be problematic to prepare in the form of vitrified samples, adsorbing to the air-water interface, suffering preferential orientations and decreasing the number of particles during vitrification. Preferential orientations may result in limitation of high resolution of the final 3D reconstruction (Naydenova and Russo, 2017). Although in some cases, a tilt collection may improve resolution (Zi Tan et al., 2017) in these cases, it is desirable to know the origin of such preferences. Sometimes the shape of the particle can be prone to preferential orientations, for example a ring-shape is more likely to fall on a top view than a side one, however adsorption to the air-water interface also increases preferential orientations.

5 Data processing

In single-particle electron microscopy, micrographs are interpreted as projections from an object. Hence, combining such projections and taking a sufficiently large angular range the object can be eventually recovered in 3D. To generate a 3D reconstruction, many images from identical or similar molecules are averaged to enhance the SNR, as well as to provide different views needed to obtain a complete 3D reconstruction. Providing that there are a sufficient number of images containing high-resolution information and that those images are aligned and classified appropriately, single-particle cryo-EM can generate a 3D reconstruction at atomic resolution. An atomic model can then be built *de novo* after sequence assignment on the 3D map.

A significant amount of time in single-particle cryo-EM analysis is taken by data processing. The main steps of data processing will be discussed in the following sections including motion correction, CTF estimation, selection and particle extraction, generation of an initial structure and its refinement and assessment of resolution.

5.1 Micrograph pre-processing

5.1.1 Motion correction

One of the main barriers limiting resolution in cryo-EM is noise. Noise can come from different sources:

- Background noise arising from the solvent (ice) or the support. The contrast of the proteins is only slightly larger than the contrast of water, therefore biological specimens produce really low contrast.
- Shot noise, caused by statistical variations in the number of electrons hit the detector.
- Radiation damage.
- Beam-induced specimen movement.
- Conformation and compositional changes in the sample (structural noise).

The combination of all these types of noise impose a limit in the achievable resolution, thus the signal-to-noise ratio (SNR) is an important parameter to measure the limits of resolution. SNR is importantly affected by radiation damage, which limits the amount of useful information per electron when exposure time increases. New direct electron detectors benefit from a high frame rate which can correct beam-induced specimen movement (blurring) by recording a movie throughout exposure, which means that a single exposure is fractionated into subframes of very short exposure, so that the

motion is diminished and the noise remains constant. Then, movies can be corrected for the beam-induced motion during image acquisition by aligning the frames (by registering identical features) and summing the registered subframes to produce a motion-corrected image.

The first motion correction alogrithms were MotionCorr from the lab of Yifan Cheng (Li et al., 2013a), and Unblur, developed by the Griegoreff lab (Grant and Grigorieff, 2015). Unblur, averages groups of adjacent subframes to an optimum dose to improve the SNR (it was done using rotavirus particles). In MotionCorr, they diagnosed that beam-induced motions were higher in the first few frames, where the amount of high-resolution information was at its maximum, so discarding these initial frames, resolution was improved. The two current approaches for motion correction are a Bayesian particle polishing approach in RELION (Zivanov et al., 2019) and MotionCorr2 (Zheng et al., 2017), the successor of MotionCorr, which overcome the particle size limit imposed by the previous algorithm.

MotionCorr2 algorithm is based on an experimental model that describes that local motion varies smoothly throughout exposure and that nonuniform motions are projections of complex 3D sample deformation onto the image plane. The algorithm uses a 2D polynomial function depending on time to describe these projections. The image is firstly divided into patches, and motions within each patch are determined iteratively by cross-correlation. These motions are then used to fit the 2D polynomial function at each individual pixel and then summed with/without radiation-damage weighting. Although particle polishing shows greater accuracy when correcting beam-induced movement, it requires a 3D reconstruction to use as a reference to align noisy experimental images and weight them.

The amount of noise present in all frames is constant, however the signal fades towards the latest frames, therefore decreasing the SNR. Baker et al modelled the relationship between electron exposure and SNR as a function of spatial frequency, and established a theoretical optimal exposure for each spatial frequency (Baker and Rubinstein, 2010). These optimal exposures could then be used to calculate weights for each frame. Since these theoretical studies, motion correction algorithms usually include a dose-weighting scheme, which usually down weights latter frames of a movie with low SNR, thus improving significantly resolution. The RELION approach particle polishing is based in modelling signal deterioration using B-factors and then dose-weighting using the B-factors model (Zivanov et al., 2019).

5.1.2 CTF estimation

To correct for the CTF decay at high spatial frequencies a more sophisticated approach such as the Wiener filter is needed. To apply this filter it is necessary to provide some parameters such as spherical aberration or astigmatism, which are intrinsic from each microscope, or the defocus values, which have

to be determined for each data collection.

Estimation of CTF parameters is usually done per each micrograph. Usually, the estimation of the CTF parameters is achieved by calculating the experimental power spectrum of the images and then comparing to the changes in contrast produced by different values of the parameters. Values yielding the best cross-correlation between the theoretical and experimental data are used for the determination of the CTF (Mindell and Grigorieff, 2003; Rohou and Grigorieff, 2015; Zhang, 2016). The cross-correlation coefficient and the estimated resolution are usually used to estimate the goodness of the fit. Other parameters estimated by the CTF include the defocus value, the angle of astigmatism and the phase shift if a phase-plate has been used.

After obtaining a 3D reconstruction which will have enhanced SNR than the individual images, CTF parameters can be refined for each individual particle, which allow for more accurate CTF determination and, therefore, better recovery of high-resolution information (Zivanov et al., 2018). Furthermore, CTF refinement can also estimate other aberrations, such as beam tilt or anisotropic magnification, however this would only work with a high-resolution reconstruction.

5.2 Particle pre-processing and extraction

After micrograph pre-processing, coordinates for the particles have to be determined on each micrograph. There are different approaches for particle selection. Manual particle selection by visual micrograph inspection can be very accurate, however it is usually very time-consuming and impractical for large data sets. An alternative is to select particles semi-automatically by generating a template after manual selection of few hundreds of particles. This approach is very dependent on the template quality and one has to be careful because you can end up selecting noise instead of signal. Moreover, if the template is generated with external data it is important to consider that the grey scale might be different. Template-based approaches are implemented in BOXER (Tang et al., 2007) and RELION (Scheres, 2015) and they usually result in several false positives that need further classification. Latest software based on deep-learning train neural networks to identify particles of interest (Bepler et al., 2019; Wagner et al., 2019). These algorithms are usually able to automatically pick unusually shaped proteins (see (Ramlaul et al., 2021) for an example).

After coordinates have been assigned for a good set of particles, those must be extracted into smaller boxes and normalised to the same grey scale. The box size must be slightly larger than the appearance of the particle on the micrograph due to the effects of the CTF. Initially, to accelerate extraction and initial classification steps, during particle extraction is common to average sets of pixels together to obtain an image with enhanced contrast and of smaller size. However, after particle sorting, particles

need to be re-scaled to their original size in order to obtain a high-resolution reconstruction.

5.3 Two-dimensional and three-dimensional classification

After extraction, particles are usually cleaned up from false positives, contaminants, ice crystals or particles without high-resolution information by 2D classification. Briefly, in this process similar particles are group, aligned and averaged into different 2D classes, which improves the SNR and allows the generation of well-defined views of the particles. Most modern 2D classification techniques combine the assignment of particles to class averages with particle alignment to the generated classes.

Usually, we distinguish between supervised and unsupervised classification. The former groups classes according to their similarity to a prototype or reference, while the second one groups particles depending on their similarity between them without any further guidance. Various strategies of unsupervised classification have been proposed to align and classify large sets of single-particle EM images, most of them based on the K-means clustering algorithm (Joyeux and Penczek, 2002). In this algorithm, the user can provide a number of classes (K) or, alternatively, initial K classes are generated by random distribution of particles. Then particle images are compared with the set of classes using cross-correlation for class assignment. The process is iterative, hence based on the new class assignments, new class averages are calculated and the whole procedure is repeated until a stable classification is achieved (Cheng et al., 2015).

Alternatively, RELION has implemented an expectation-maximisation *a posteriori* approach based on regularised likelihood optimisation (Scheres, 2012). In this approach, the user provides the set of particles with respective coordinates, the CTF estimated parameters, a tau value, which is an overestimation of the signal for regularisation, a particle diameter, and the number of classes for classification (K). Particles are initially distributed in random K classes. During expectation, particles are rotated to be similar to the model, whereas during maximisation the class is improved based on the better matches. Instead of assigning particles to a class it assigns a weight (or probability) based on the likelihood of this particles to belong to a particular class. To calculate the SNR, RELION initially assumes that the signal is correct and calculates the noise by subtraction, and the process is repeated iteratively.

After cleaning the data set by 2D classification, the aim of single-particle is to reconstruct the 3D object of study from the structural information contained in its 2D images. The extracted classes after 2D classification represent a set of 2D images corresponding to projections of the imaged 3D object. 3D reconstruction is usually achieved by back-projection: several 2D images representing different views of the object are back-projected from different angles to reconstitute the 3D object

(Penczek, 2010). An accurate reconstruction of the original 3D object can only be achieved if there are projections from several views.

For 3D reconstruction is essential to accurately determine the three Euler angles. If a previous 3D structure is available (e.g a low-resolution negative-stain EM 3D reconstruction, a filtered X-ray model or a filtered EM map of an homologue), it is possible to use it as a template for assignment of the three Euler angles by projection matching, however one has to be cautious as the grey scale and the SNR will be different. If any previous structure is available, *ab initio* 3D structure determination might be required. Theoretically, determination of a *de novo* 3D reconstruction could be done by serial tomography tilts, as the tilt angle is known *a priori*, however this process is very time-consuming. A experimental method for *ab initio* 3D structure determination which was very popular until the appearance of reliable computational algorithms, is the *random conical tilt* (RCT) approach (Radermacher et al., 1987b). This method is based on the fact that one of the orientation parameters is already known (tilt angle) and the others can be computed in a robust manner. RCT generates reliable structures, however it is time-consuming and requires additional experimental data. Alternatively, computational *de novo* structure determination algorithms can be used. These approaches determine five orientation parameters for each projection (the three Euler angles and two translations) to iteratively improve the 3D structure based on a mathematical criterion. Currently, *ab initio* structure determination algorithms based on *stochastic gradient descent* (SGD) are implemented in both RELION (Zivanov et al., 2018) and cryoSPARC (Punjani et al., 2017). In this approach, an initial structure is calculated with random set of particles. Then the process is repeated iteratively by aligning different subsets of particles to the current models and updating the models with the assigned orientations. The gradient represents the difference between the previous structure and the newly generated one. The final structure would be the one with the minimum gradient.

After obtaining an initial map, the structure is refined by 3D projection matching which consists in the modification of the orientation parameters of each projection to achieved a better match with back projections of the initial structure (Penczek, 2010).

5.4 Resolution and Post-processing

Refinement is usually monitored by the *Fourier shell correlation* (FSC) curve, which is used to define the resolution of the 3D reconstruction (Cheng et al., 2015). To obtain the FSC, data is randomly split in two half-sets. A 3D refinement is performed for particles belonging to each half-set independently against the same 3D template, resulting in two half-maps. The similarity between the 3D Fourier transforms of both maps is then compared by cross-correlation at each spatial frequency. The FSC curve represents the cross-correlation values of the two half-maps plotted as a function of spatial

frequency or resolution. A cut-off value for the FSC to determine the resolution of the reconstruction must be chosen to determine the resolution of the reconstruction obtained when both half-maps are combined. There are several criterion, however, for cryo-EM maps a threshold of 0.143 is usually set, based on the relationship between EM results and X-ray crystallography (Rosenthal and Henderson, 2003).

After refinement, RELION implements a post-processing algorithm which masks the final map, re-calculates the FSC with the mask (which can be automatically generated or provided by the user), weights the FSC (as it comes from two half-maps), and applies a B-factor. The falloff at high spatial frequencies in EM is usually represented by a Gaussian term called B-factor, which can be considered analogous to the temperature factor in X-ray crystallography (Rosenthal and Henderson, 2003). The most common procedure to estimate the B-factor in cryo-EM is from the Guinier plot, which represents the logarithm of the average structure (amplitude scattering) *versus* the square of the frequency ($1/d^2$). A Guinier plot normally consists of two distinct regions: (i) a pronounced decay dependent on protein shape and solvent and (ii) a relatively flat region corresponding to secondary structure and atomic features at high resolution. In the Guinier plot the scattering amplitude is normally described by a factor that includes the difference between scattering from the protein and from the solvent $(\rho - \rho_0)$(Rosenthal and Henderson, 2003). By applying a negative B-factor the falloff of the signal is sharpened so that all spatial frequencies contribute equally.

References

Marc Adrian, Jacques Dubochet, Jean Lepault, and Alasdair W. McDowall. Cryo-electron microscopy of viruses. *Nature*, 308(5954):32–36, 3 1984. ISSN 0028-0836. doi: 10.1038/308032a0.

Bruce Alberts. The cell as a collection of protein machines: Preparing the next generation of molecular biologists, 2 1998. ISSN 00928674.

Lindsay A. Baker and John L. Rubinstein. Radiation Damage in Electron Cryomicroscopy. *Methods in Enzymology*, 481(C):371–388, 1 2010. ISSN 0076-6879. doi: 10.1016/S0076-6879(10)81015-8.

Benjamin E. Bammes, Ryan H. Rochat, Joanita Jakana, Dong Hua Chen, and Wah Chiu. Direct electron detection yields cryo-EM reconstructions at resolutions beyond 3/4 Nyquist frequency. *Journal of Structural Biology*, 177(3):589–601, 3 2012. ISSN 10478477. doi: 10.1016/j.jsb.2012.01.008.

Tristan Bepler, Andrew Morin, Micah Rapp, Julia Brasch, Lawrence Shapiro, Alex J. Noble, and Bonnie Berger. Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nature Methods 2019 16:11*, 16(11):1153–1160, 10 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0575-8.

John A. Berriman and Peter B. Rosenthal. Paraxial charge compensator for electron cryomicroscopy. *Ultramicroscopy*, 116:106–114, 5 2012. ISSN 03043991. doi: 10.1016/j.ultramic.2012.03.006.

B. Böttcher, S. A. Wynne, and R. A. Crowther. Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy. *Nature*, 386(6620):88–91, 3 1997. ISSN 0028-0836. doi: 10.1038/386088a0.

W. S. Boyle and G. E. Smith. Charge Coupled Semiconductor Devices. *Bell System Technical Journal*, 49(4):587–593, 4 1970. ISSN 00058580. doi: 10.1002/j.1538-7305.1970.tb01790.x.

S. Brenner and R. W. Horne. A negative staining method for high resolution electron microscopy of viruses. *Biochimica et Biophysica Acta*, 34(C):103–110, 1 1959. ISSN 0006-3002. doi: 10.1016/0006-3002(59)90237-9.

Axel F. Brilot, James Z. Chen, Anchi Cheng, Junhua Pan, Stephen C. Harrison, Clinton S. Potter, Bridget Carragher, Richard Henderson, and Nikolaus Grigorieff. Beam-induced motion of vitrified specimen on holey carbon film. *Journal of Structural Biology*, 177(3):630–637, 3 2012. ISSN 10478477. doi: 10.1016/j.jsb.2012.02.003.

Jacob Brink, Michael B. Sherman, John Berriman, and Wah Chiu. Evaluation of charging on macro-

molecules in electron cryomicroscopy. *Ultramicroscopy*, 72(1-2):41–52, 4 1998. ISSN 03043991. doi: 10.1016/S0304-3991(97)00126-5.

James Chen, Alex J. Noble, Jin Young Kang, and Seth A. Darst. Eliminating effects of particle adsorption to the air/water interface in single-particle cryo-electron microscopy: Bacterial RNA polymerase and CHAPSO. *Journal of Structural Biology: X*, 1:100005, 1 2019. ISSN 25901524. doi: 10.1016/j.yjsbx.2019.100005.

Yifan Cheng. Single-particle Cryo-EM at crystallographic resolution, 4 2015. ISSN 10974172.

Yifan Cheng, Nikolaus Grigorieff, Pawel A. Penczek, and Thomas Walz. A primer to single-particle cryo-electron microscopy, 4 2015. ISSN 10974172.

R N Clough, G Moldovan, and A I Kirkland. Direct Detectors for Electron Microscopy. *Journal of Physics:*, Conf.Ser.:012046, 2014. doi: 10.1088/1742-6596/522/1/012046.

J. F. Conway, N. Cheng, A. Zlotnick, P. T. Wingfield, S. J. Stahl, and A. C. Steven. Visualization of a 4-helix bundle in the hepatitis B virus capsid by cryo-electron microscopy. *Nature*, 386(6620): 91–94, 3 1997. ISSN 0028-0836. doi: 10.1038/386091a0.

Radostin Danev and Wolfgang Baumeister. Cryo-EM single particle analysis with the Volta phase plate. *eLife*, 5(MARCH2016), 3 2016. ISSN 2050084X. doi: 10.7554/ELIFE.13046.

Radostin Danev, Bart Buijsse, Maryam Khoshouei, Jürgen M. Plitzko, and Wolfgang Baumeister. Volta potential phase plate for in-focus phase contrast transmission electron microscopy. *Proceedings of the National Academy of Sciences of the United States of America*, 111(44):15635–15640, 11 2014. ISSN 1091-6490. doi: 10.1073/PNAS.1418377111.

Natàlia de Martín Garrido, Wencheng Fu, Kailash Ramlaul, Zining Zhu, David Miller, Daniel Boehringer, and Christopher H.S. Aylett. Direct transfer of electron microscopy samples to wetted carbon and graphene films via a support floatation block. *Journal of Structural Biology*, 213(1), 2021. ISSN 10958657. doi: 10.1016/j.jsb.2020.107677.

D.J. De Rosier and A. Klug. Reconstruction of Three Dimensional Structures from Electron Micrographs. *Nature*, 217(5124):130–134, 1 1968. ISSN 0028-0836. doi: 10.1038/217130a0.

Megan J. Dobro, Linda A. Melanson, Grant J. Jensen, and Alasdair W. McDowall. Plunge freezing for electron cryomicroscopy. *Methods in enzymology*, 481(C):63–82, 2010. ISSN 1557-7988. doi: 10.1016/S0076-6879(10)81003-1.

J. Dubochet, J. Lepault, R. Freeman, J. A. Berriman, and J.-C. Homo. Electron microscopy of frozen

water and aqueous solutions. *Journal of Microscopy*, 128(3):219–237, 12 1982. ISSN 00222720. doi: 10.1111/j.1365-2818.1982.tb04625.x.

J. Dubochet, Marc Adrian, Jiin-Ju Chang, J. -C Homo, J. Lepault, Alasdair W. McDowall, and Patrick Shultz. Cryo-electron microscopy of vitrified specimens. *Quarterly Reviews of Biophysics*, 21(2):129–228, 1988. ISSN 14698994. doi: 10.1017/S0033583500004297.

J. Frank, W. Goldfarb, D. Eisenberg, and T.S. Baker. Reconstruction of glutamine synthetase using computer averaging. *Ultramicroscopy*, 3(C):283–290, 1 1978. ISSN 03043991. doi: 10.1016/S0304-3991(78)80038-2.

Joachim Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Oxford University Press, 3 2006. ISBN 9780195182187. doi: 10.1093/acprof:oso/9780195182187.001.0001.

Linda E. Franken, Kay Grünewald, Egbert J. Boekema, and Marc C. A. Stuart. A Technical Introduction to Transmission Electron Microscopy for Soft-Matter: Imaging, Possibilities, Choices, and Technical Developments. *Small*, 16(14):1906198, 4 2020. ISSN 1613-6810. doi: 10.1002/smll.201906198.

Peter M. Frederik and D. H.W. Hubert. Cryoelectron microscopy of liposomes. *Methods in enzymology*, 391(SPEC. ISS.):431–448, 2005. ISSN 0076-6879. doi: 10.1016/S0076-6879(05)91024-0.

Yoshinori Fujiyoshi. The structural study of membrane proteins by electron crystallography. *Advances in Biophysics*, 35:25–80, 1 1998. ISSN 0065227X. doi: 10.1016/S0065-227X(98)80003-8.

Irene S. Gabashvili, Rajendra K. Agrawal, Christian M.T. Spahn, Robert A. Grassucci, Dmitri I. Svergun, Joachim Frank, and Pawel Penczek. Solution Structure of the E. coli 70S Ribosome at 11.5 Å Resolution. *Cell*, 100(5):537–549, 3 2000. ISSN 00928674. doi: 10.1016/S0092-8674(00)80690-X.

A. K. Geim. Graphene: Status and prospects, 6 2009. ISSN 00368075.

Robert M. Glaeser. Proteins, interfaces, and cryo-EM grids, 3 2018. ISSN 18790399.

Robert M. Glaeser, Dieter Typke, Peter C. Tiemeijer, James Pulokas, and Anchi Cheng. Precise beam-tilt alignment and collimation are required to minimize the phase error associated with coma in high-resolution cryo-EM, 4 2011. ISSN 10478477.

Robert M. Glaeser, Bong Gyoon Han, Roseann Csencsits, Alison Killilea, Arto Pulk, and Jamie H.D. Cate. Factors that Influence the Formation and Stability of Thin, Cryo-EM Specimens, 2 2016. ISSN 15420086.

Timothy Grant and Nikolaus Grigorieff. Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. *eLife*, 4(MAY), 5 2015. doi: 10.7554/ELIFE.06980.

Robert A. Grassucci, Derek J. Taylor, and Joachim Frank. Preparation of macromolecular complexes for cryo-electron microscopy. *Nature Protocols*, 2(12):3239–3246, 12 2007. ISSN 17502799. doi: 10.1038/nprot.2007.452.

Rudo Grimm, Dieter Typke, Michael Bärmann, and Wolfgang Baumeister. Determination of the inelastic mean free path in ice by examination of tilted vesicles and automated most probable loss imaging. *Ultramicroscopy*, 63(3-4):169–179, 7 1996. ISSN 03043991. doi: 10.1016/0304-3991(96)00035-6.

Richard Henderson and Robert M. Glaeser. Quantitative analysis of image contrast in electron micrographs of beam-sensitive crystals. *Ultramicroscopy*, 16(2):139–150, 1 1985. ISSN 03043991. doi: 10.1016/0304-3991(85)90069-5.

Mark A. Herzik. Setting Up on the Talos Arctica for High-Resolution Data Collection. In *Methods in Molecular Biology*, volume 2215, pages 125–144. Humana, New York, NY, 2021. doi: 10.1007/978-1-0716-0966-8_6.

Crispin Hetherington. Aberration correction for TEM. *Materials Today*, 7(12):50–55, 12 2004. ISSN 13697021. doi: 10.1016/S1369-7021(04)00571-1.

Fumio Hosokawa, Hidetaka Sawada, Yukihito Kondo, Kunio Takayanagi, and Kazutomo Suenaga. Development of Cs and Cc correctors for transmission electron microscopy. *Microscopy*, 62(1): 23–41, 2 2013. ISSN 2050-5698. doi: 10.1093/jmicro/dfs134.

Laurent Joyeux and Pawel A. Penczek. Efficiency of 2D alignment methods. *Ultramicroscopy*, 92(2): 33–46, 7 2002. ISSN 0304-3991. doi: 10.1016/S0304-3991(01)00154-1.

Berthold Kastner, Niels Fischer, Monika Mariola Golas, Bjoern Sander, Prakash Dube, Daniel Boehringer, Klaus Hartmuth, Jochen Deckert, Florian Hauer, Elmar Wolf, Hannes Uchtenhagen, Henning Urlaub, Franz Herzog, Jan Michael Peters, Dietmar Poerschke, Reinhard Lührmann, and Holger Stark. GraFix: Sample preparation for single-particle electron cryomicroscopy. *Nature Methods*, 5(1):53–55, 1 2008. ISSN 15487091. doi: 10.1038/nmeth1139.

Maryam Khoshouei, Mazdak Radjainia, Wolfgang Baumeister, and Radostin Danev. Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate. *Nature Communications*, 8: 16099, 2017. doi: 10.1038/ncomms16099.

M. Knoll and E. Ruska. Das Elektronenmikroskop. *Zeitschrift fr Physik*, 78(5-6):318–339, 5 1932. ISSN 1434-6001. doi: 10.1007/BF01342199.

Werner Kühlbrandt. The Resolution Revolution. *Science*, 343(6178):1443–1444, 3 2014. ISSN 0036-8075. doi: 10.1126/science.1251652.

R. D. Leapman and R. L. Ornberg. Quantitative electron energy loss spectroscopy in biology. *Ultramicroscopy*, 24(2-3):251–268, 1 1988. ISSN 03043991. doi: 10.1016/0304-3991(88)90314-2.

Xueming Li, Paul Mooney, Shawn Zheng, Christopher R Booth, Michael B Braunfeld, Sander Gubbens, David A Agard, and Yifan Cheng. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods 2013 10:6*, 10(6): 584–590, 5 2013a. ISSN 1548-7105. doi: 10.1038/nmeth.2472.

Zhiting Li, Yongjin Wang, Andrew Kozbial, Ganesh Shenoy, Feng Zhou, Rebecca McGinley, Patrick Ireland, Brittni Morganstein, Alyssa Kunkel, Sumedh P. Surwade, Lei Li, and Haitao Liu. Effect of airborne contaminants on the wettability of supported graphene and graphite. *Nature Materials*, 12(10):925–931, 2013b. ISSN 14761122. doi: 10.1038/nmat3709.

Yi-Lynn Liang, Maryam Khoshouei, Mazdak Radjainia, Yan Zhang, Alisa Glukhova, Jeffrey Tarrasch, David M. Thal, Sebastian G. B. Furness, George Christopoulos, Thomas Coudrat, Radostin Danev, Wolfgang Baumeister, Laurence J. Miller, Arthur Christopoulos, Brian K. Kobilka, Denise Wootten, Georgios Skiniotis, and Patrick M. Sexton. Phase-plate cryo-EM structure of a class B GPCR–G-protein complex. *Nature 2017 546:7656*, 546(7656):118–123, 4 2017. ISSN 1476-4687. doi: 10.1038/nature22327.

Hongrong Liu, Lei Jin, Sok Boon S. Koh, Ivo Atanasov, Stan Schein, Lily Wu, and Z. Hong Zhou. Atomic Structure of Human Adenovirus by Cryo-EM Reveals Interactions Among Protein Networks. *Science*, 329(5995):1038–1043, 8 2010. ISSN 0036-8075. doi: 10.1126/science.1187433.

A. W. McDowall, J. -J Chang, R. Freeman, J. Lepault, C. A. Walter, and J. Dubochet. Electron microscopy of frozen hydrated sections of vitreous ice and vitrified biological samples. *Journal of Microscopy*, 131(1):1–9, 1983. ISSN 13652818. doi: 10.1111/j.1365-2818.1983.tb04225.x.

G. McMullan, S. Chen, R. Henderson, and A.R. Faruqi. Detective quantum efficiency of electron area detectors in electron microscopy. *Ultramicroscopy*, 109(9):1126–1143, 8 2009a. ISSN 03043991. doi: 10.1016/j.ultramic.2009.04.002.

G. McMullan, A. R. Faruqi, R. Henderson, N. Guerrini, R. Turchetta, A. Jacobs, and G. van Hoften. Experimental observation of the improvement in MTF from backthinning a CMOS direct electron detector. *Ultramicroscopy*, 109(9-3):1144, 8 2009b. ISSN 03043991. doi: 10.1016/J.ULTRAMIC.2009.05.005.

G. McMullan, A.R. Faruqi, and R. Henderson. Direct Electron Detectors. In *Methods in Enzymology*, volume 579, pages 1–17. Academic Press Inc., 1 2016. ISBN 9780128053829. doi: 10.1016/bs.mie.2016.05.056.

Jannik C. Meyer, A. K. Geim, M. I. Katsnelson, K. S. Novoselov, T. J. Booth, and S. Roth. The structure of suspended graphene sheets. *Nature*, 446(7131):60–63, 3 2007. ISSN 00280836. doi: 10.1038/nature05545.

Joseph A. Mindell and Nikolaus Grigorieff. Accurate determination of local defocus and specimen tilt in electron microscopy. *Journal of Structural Biology*, 142(3):334–347, 6 2003. ISSN 10478477. doi: 10.1016/S1047-8477(03)00069-8.

Katerina Naydenova and Christopher J. Russo. Measuring the effects of particle orientation to improve the efficiency of electron cryomicroscopy. *Nature Communications*, 8(1):1–5, 12 2017. ISSN 20411723. doi: 10.1038/s41467-017-00782-3.

Katerina Naydenova, Mathew J. Peet, and Christopher J. Russo. Multifunctional graphene supports for electron cryomicroscopy. *Proceedings of the National Academy of Sciences of the United States of America*, 116(24):11718–11724, 2019. ISSN 10916490. doi: 10.1073/pnas.1904766116.

Alex J. Noble, Hui Wei, Venkata P. Dandey, Zhening Zhang, Yong Zi Tan, Clinton S. Potter, and Bridget Carragher. Reducing effects of particle adsorption to the air–water interface in cryo-EM. *Nature Methods*, 15(10):793–795, 2018. ISSN 15487105. doi: 10.1038/s41592-018-0139-3. URL `http://dx.doi.org/10.1038/s41592-018-0139-3`.

E. V. Orlova and H. R. Saibil. Structural Analysis of Macromolecular Assemblies by Electron Microscopy. *Chemical Reviews*, 111(12):7710–7748, 12 2011. ISSN 0009-2665. doi: 10.1021/cr100353t.

Eugene Palovcak, Feng Wang, Shawn Q. Zheng, Zanlin Yu, Sam Li, Miguel Betegon, David Bulkley, David A. Agard, and Yifan Cheng. A simple and robust procedure for preparing graphene-oxide cryo-EM grids. *Journal of Structural Biology*, 204(1):80–84, 10 2018. ISSN 10958657. doi: 10.1016/j.jsb.2018.07.007.

Radosav S. Pantelic, Jannik C. Meyer, Ute Kaiser, Wolfgang Baumeister, and Jürgen M. Plitzko. Graphene oxide: A substrate for optimizing preparations of frozen-hydrated samples. *Journal of Structural Biology*, 170(1):152–156, 4 2010. ISSN 10478477. doi: 10.1016/j.jsb.2009.12.020.

Radosav S. Pantelic, Ji Won Suk, Carl W. Magnuson, Jannik C. Meyer, Philipp Wachsmuth, Ute Kaiser, Rodney S. Ruoff, and Henning Stahlberg. Graphene: Substrate preparation and

introduction. *Journal of Structural Biology*, 174(1):234–238, 4 2011. ISSN 10478477. doi: 10.1016/j.jsb.2010.10.002.

Pawel A. Penczek. Fundamentals of three-dimensional reconstruction from projections. *Methods in enzymology*, 482(C):1, 2010. ISSN 00766879. doi: 10.1016/S0076-6879(10)82001-4.

Ali Punjani, John L Rubinstein, David J Fleet, and Marcus A Brubaker. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods*, 14(3):290–296, 3 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4169.

M Radermacher, T Wagenknecht, A Verschoor, and J Frank. Three-dimensional structure of the large ribosomal subunit from Escherichia coli. *The EMBO Journal*, 6(4):1–107, 1987a.

M. Radermacher, T. Wagenknecht, A. Verschoor, and J. Frank. Three-dimensional reconstruction from a single-exposure, random conical tilt series applied to the 50S ribosomal subunit of Escherichia coli. *Journal of microscopy*, 146(Pt 2):113–136, 1987b. ISSN 0022-2720. doi: 10.1111/J.1365-2818.1987.TB01333.X.

Kailash Ramlaul, Wencheng Fu, Hua Li, Natàlia de Martin Garrido, Lin He, Manjari Trivedi, Wei Cui, Christopher H.S. Aylett, and Geng Wu. Architecture of the Tuberous Sclerosis Protein Complex. *Journal of Molecular Biology*, 433(2):166743, 1 2021. ISSN 00222836. doi: 10.1016/j.jmb.2020.166743.

William Regan, Nasim Alem, Benjamín Alemán, Baisong Geng, Çağlar Girit, Lorenzo Maserati, Feng Wang, Michael Crommie, and A. Zettl. A direct transfer of layer-area graphene. *Applied Physics Letters*, 96(11):2008–2011, 2010. ISSN 00036951. doi: 10.1063/1.3337091.

Daniel Rhinow and Werner Kühlbrandt. Electron cryo-microscopy of biological specimens on conductive titanium-silicon metal glass films. *Ultramicroscopy*, 108(7):698–705, 6 2008. ISSN 03043991. doi: 10.1016/j.ultramic.2007.11.005.

Alexis Rohou and Nikolaus Grigorieff. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *Journal of Structural Biology*, 192(2):216–221, 11 2015. ISSN 10958657. doi: 10.1016/j.jsb.2015.08.008.

Peter B. Rosenthal and Richard Henderson. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *Journal of Molecular Biology*, 333(4):721–745, 10 2003. ISSN 00222836. doi: 10.1016/j.jmb.2003.07.013.

Christopher J Russo and Lori A Passmore. Controlling protein adsorption on graphene for cryo-EM

using low-energy hydrogen plasmas. *Nature Methods*, 11(6):649–652, 6 2014a. ISSN 15487105. doi: 10.1038/nmeth.2931.

Christopher J. Russo and Lori A. Passmore. Ultrastable gold substrates for electron cryomicroscopy. *Science*, 346(6215):1377–1380, 12 2014b. ISSN 10959203. doi: 10.1126/science.1259530.

Christopher J. Russo and Lori A. Passmore. Progress towards an optimal specimen support for electron cryomicroscopy. *Current Opinion in Structural Biology*, 37:81–89, 4 2016. ISSN 1879033X. doi: 10.1016/j.sbi.2015.12.007.

Sjors H.W. Scheres. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *Journal of Structural Biology*, 180(3):519–530, 12 2012. ISSN 10478477. doi: 10.1016/j.jsb.2012.09.006.

Sjors H.W. Scheres. Semi-automated selection of cryo-EM particles in RELION-1.3. *Journal of Structural Biology*, 189(2):114–122, 2 2015. ISSN 1047-8477. doi: 10.1016/J.JSB.2014.11.010.

Holger Stark. GraFix: Stabilization of Fragile Macromolecular Complexes for Single Particle Cryo-EM. *Methods in Enzymology*, 481:109–126, 1 2010. ISSN 0076-6879. doi: 10.1016/S0076-6879(10)81005-5.

Christian Suloway, James Pulokas, Denis Fellmann, Anchi Cheng, Francisco Guerra, Joel Quispe, Scott Stagg, Clinton S. Potter, and Bridget Carragher. Automated molecular microscopy: The new Leginon system. *Journal of Structural Biology*, 151(1):41–60, 7 2005. ISSN 10478477. doi: 10.1016/j.jsb.2005.03.010.

Guang Tang, Liwei Peng, Philip R. Baldwin, Deepinder S. Mann, Wen Jiang, Ian Rees, and Steven J. Ludtke. EMAN2: An extensible image processing suite for electron microscopy. *Journal of Structural Biology*, 157(1):38–46, 1 2007. ISSN 10478477. doi: 10.1016/j.jsb.2006.05.009.

Kenneth A. Taylor and Robert M. Glaeser. Electron Diffraction of Frozen, Hydrated Protein Crystals. *Science*, 186(4168):1036–1037, 12 1974. ISSN 0036-8075. doi: 10.1126/science.186.4168.1036.

P.C. Tiemeijer, M. Bischoff, B. Freitag, and C. Kisielowski. Using a monochromator to improve the resolution in TEM to below 0.5Å. Part I: Creating highly coherent monochromated illumination. *Ultramicroscopy*, 114:72–81, 3 2012. ISSN 03043991. doi: 10.1016/j.ultramic.2012.01.008.

A. Verschoor, J. Frank, M. Radermacher, T. Wagenknecht, and M. Boublik. Three-dimensional reconstruction of the 30 S ribosomal subunit from randomly oriented particles. *Journal of Molecular Biology*, 178(3):677–695, 9 1984. ISSN 00222836. doi: 10.1016/0022-2836(84)90245-6.

Thorsten Wagner, Felipe Merino, Markus Stabrin, Toshio Moriya, Claudia Antoni, Amir Apelbaum,

Philine Hagel, Oleg Sitsel, Tobias Raisch, Daniel Prumbaum, Dennis Quentin, Daniel Roderer, Sebastian Tacke, Birte Siebolds, Evelyn Schubert, Tanvir R Shaikh, Pascal Lill, Christos Gatsogiannis, and Stefan Raunser. SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. *Communications Biology*, 2(1):218, 12 2019. ISSN 2399-3642. doi: 10.1038/s42003-019-0437-z.

Hong Wei Wang and Xiao Fan. Challenges and opportunities in cryo-EM with phase plate, 10 2019. ISSN 1879033X.

F. Zemlin, K. Weiss, P. Schiske, W. Kunath, and K.-H. Herrmann. Coma-free alignment of high resolution electron microscopes with the aid of optical diffractograms. *Ultramicroscopy*, 3(C):49–60, 1 1978. ISSN 03043991. doi: 10.1016/S0304-3991(78)80006-0.

F. Zernike. How I Discovered Phase Contrast. *Science*, 121(3141):345–349, 3 1955. ISSN 0036-8075. doi: 10.1126/science.121.3141.345.

Kai Zhang. Gctf: Real-time CTF determination and correction. *Journal of Structural Biology*, 193 (1):1–12, 1 2016. ISSN 1047-8477. doi: 10.1016/J.JSB.2015.11.003.

Xing Zhang, Lei Jin, Qin Fang, Wong H. Hui, and Z. Hong Zhou. 3.3 Å Cryo-EM structure of a nonenveloped virus reveals a priming mechanism for cell entry. *Cell*, 141(3):472–482, 4 2010. ISSN 00928674. doi: 10.1016/j.cell.2010.03.041.

Shawn Q Zheng, Eugene Palovcak, Jean-Paul Armache, Kliment A Verba, Yifan Cheng, and David A Agard. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nature Methods*, 14(4):331–332, 4 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4193.

Yong Zi Tan, Philip R. Baldwin, Joseph H. Davis, James R. Williamson, Clinton S. Potter, Bridget Carragher, and Dmitry Lyumkis. Addressing preferred specimen orientation in single-particle cryo-EMthrough tilting. *Nature Methods*, 14(8):793–796, 7 2017. ISSN 15487105. doi: 10.1038/nmeth.4347.

Jasenko Zivanov, Takanori Nakane, Björn O. Forsberg, Dari Kimanius, Wim J.H. Hagen, Erik Lindahl, and Sjors H.W. Scheres. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife*, 7, 11 2018. doi: 10.7554/ELIFE.42166.

Jasenko Zivanov, Takanori Nakane, and Sjors H.W. Scheres. A Bayesian approach to beam-induced motion correction in cryo-EM single-particle analysis. *urn:issn:2052-2525*, 6(1):5–17, 1 2019. ISSN 2052-2525. doi: