

# **Structural and biochemical studies of the large terminase protein from thermophilic viruses**

Rui-Gang Xu

PhD

University of York

Chemistry

September 2017

## Abstract

In large dsDNA viruses, DNA is translocated into the procapsid by a powerful motor, formed by portal protein, large and small terminase proteins. An essential component of the motor, large terminase protein, is composed of an ATPase domain and a nuclease domain. In this thesis, large terminase proteins from several thermophilic bacteriophages: GVE2; GBSV1; phiOH2; 1-1; G20c and D6E; were cloned, expressed and purified. Of the soluble and functional proteins, the large terminases from G20c and D6E were chosen for further structural and mechanistic studies.

Crystal structures of the large terminase nuclease domain from *Thermus thermophilus* bacteriophage G20c exhibits the closest similarity with the RuvC endonucleases which belong to the RNase H-like superfamily. Biochemical and mutagenesis studies identified catalytically indispensable residues including D294, D429 and D347 at the nuclease active site. Structures in complex with different metals reveal that only one catalytic metal ion is found at the active site in the absence of DNA. Binding of the second metal ion is likely to be facilitated by the observed conformational variability of loop L<sub>1</sub>, which may bring the two catalytic aspartic acids D347 and D294 closer to each other. These data support a RuvC like two-metal catalysis mechanism for the G20c large terminase nuclease.

The crystal structure of the large terminase from the *Geobacillus stearothermophilus* bacteriophage D6E shows a unique inter-domain orientation of the N-terminal ATPase and C-terminal nuclease domains. AUC and SEC data demonstrated the presence of pentamers formed by the ATPase domain and a model pentamer was obtained by molecular docking from which a putative *trans*-arginine, R158, was predicted- and confirmed experimentally- as functioning as a *trans*-arginine finger. This residue is in an equivalent position to the trans-finger for the HerA/FtsK superfamily of ATPases, suggesting that large terminase proteins may be more closely related than previously thought.

# List of Contents

Abstract.....	2
List of Contents.....	3
List of Figures.....	9
List of Equations.....	15
Acknowledgements.....	16
Author's declaration.....	17
1 Introduction.....	18
1.1 Bacteriophage.....	18
1.1.1 Life cycles.....	18
1.1.2 Bacteriophage therapy.....	19
1.2 Thermophilic Bacteriophages.....	20
1.3 Viral DNA packaging systems.....	21
1.4 Bacteriophage DNA packaging motor.....	21
1.4.1 The small terminase protein.....	22
1.4.1.1 Structure and function.....	22
1.4.1.2 Small-large terminase interaction.....	22
1.4.2 The portal protein.....	23
1.4.2.1 Structure and function.....	23
1.4.2.2 Portal-large terminase interaction.....	24
1.4.3 The large terminase protein.....	24
1.4.3.1 The ATPase domain.....	25
1.4.3.2 The nuclease domain.....	28
1.5 DNA cleavage specificity.....	30
1.6 Proposed mechanisms for large terminase DNA translocation.....	30
1.6.1 Inch-worm (electrostatic force driven) mechanism.....	31
1.6.2 Chemo-mechanical coupling mechanism.....	31
1.6.3 Torsional compression mechanism.....	31

1.6.4	The “lever-like motion” mechanism.....	32
1.6	Aims of research.....	32
2	Material and methods.....	33
2.1	DNA Cloning.....	33
2.1.1	PCR amplification.....	33
2.1.2	Ligation Independent Cloning.....	34
2.1.3	Colony PCR.....	35
2.1.4	Site directed mutagenesis.....	36
2.2	Transformation.....	36
2.3	Plasmid DNA production and purification.....	37
2.4	E. coli protein expression.....	37
2.4.1	The pET expression system (89-91).....	37
2.4.2	Protein expression protocol.....	38
2.4.3	Selenomethionine labelled protein expression protocol.....	38
2.5	SDS-PAGE.....	39
2.6	Agarose gel electrophoresis.....	40
2.7	Expression and solubility tests.....	40
2.8	Protein purification.....	40
2.8.1	Nickel affinity chromatography.....	40
2.8.2	Size exclusion chromatography.....	41
2.8.3	Ion exchange chromatography.....	41
2.9	Thermal shift assay.....	42
2.9.1	Buffer type and pH optimisation.....	42
2.9.2	Additives screen.....	43
2.9.3	Protein-ligand interaction.....	43
2.10	Malachite green phosphate assay.....	43
2.11	In vitro nuclease assay.....	44
2.12	Analytical ultra-centrifugation.....	44

2.12.1	Sedimentation velocity theory.....	44
2.12.2	Protein oligomeric state analysis.....	45
2.12.3	Analysis of the Protein-DNA complex with Poly (T) ssDNA.....	45
2.12.4	Protein dsDNA-complex analysis.....	46
2.13	Micro-scale thermophoresis.....	46
2.13.1	Thermophoresis theory.....	46
2.13.2	Thermophoresis signal analysis.....	47
2.13.3	Determination of EC50 and binding cooperativity.....	48
2.14	Electrophoretic motility shift assay.....	48
2.15	Coupled ATPase assay.....	48
2.16	Calibration of Superdex 200 size exclusion column.....	49
2.17	Protein Crystallisation.....	50
2.17.1	The crystallisation phase diagram.....	51
2.18	Protocol of protein crystallisation using vapor diffusion.....	52
2.19	Seeding.....	52
2.20	Post-crystallisation manipulation.....	53
2.21	X-ray data collection and processing.....	53
2.22	SAD, MAD phasing and molecular replacement.....	54
2.23	Density modification, model building, refinement and validation.....	54
3	Isolation and characterisation of large terminases from several thermophiles.....	55
3.1	Structure based sequence alignment.....	55
3.2	Protein disorder prediction.....	55
3.3	Cloning, expression and solubility tests.....	55
3.4	Solubility screen.....	57
3.5	Purification.....	59
3.6	Protein thermal stability.....	61
3.7	C-terminal truncation of G20c large terminase.....	64
3.8	ATPase activity.....	66

3.9	Nuclease activity.....	67
3.10	Crystallisation.....	69
3.11	Summary.....	70
4	Functional characterisation of G20c large terminase protein.....	71
4.1	Non-hydrolysable ATP analogues.....	71
4.2	Effects of ATP analogues on protein stability.....	72
4.3	Nuclease activity.....	73
4.3.1	Effects of sodium chloride concentration.....	73
4.3.2	Effects of temperature and protein concentration.....	74
4.3.3	Effects of divalent metal ions and metal concentrations.....	75
4.3.4	Effects of ATPase Walker B mutation.....	76
4.3.5	Effects of ATP, ADP and non-hydrolysable ATP analogues.....	77
4.3.6	Effects of small terminase.....	78
4.3.7	Effects of Raltegravir and PicoGreen.....	79
4.4	Summary.....	80
5	Structural and functional studies of the G20c large terminase.....	82
5.1	Structure determination of the G20c large terminase nuclease domain.....	82
5.2	Structure of the G20c large terminase nuclease domain.....	82
5.3	Structure comparison with other large terminases.....	84
5.4	Structures of nuclease-metal complexes.....	84
5.5	Metal dependence of nuclease activity.....	86
5.6	Effect of active site residues on nuclease activity.....	87
5.7	Structural features.....	88
5.8	Role of active site residues.....	88
5.9	Structural basis for metal dependence of nuclease activity.....	89
5.10	RuvC is the closest structural homologue of large terminase nuclease.....	90
5.11	DNA binding surface plasticity.....	91
5.12	RuvC-like, canonical two-metal dependent catalysis.....	92

5.13	Insights into the headful DNA packaging mechanism.....	94
5.14	Summary.....	94
6	Structure determination for the D6E large terminase.....	96
6.1	Structure determination (Crystal form 1).....	96
6.2	Post-crystallisation manipulation and structure determination (crystal form2).....	99
6.3	Structure of the D6E large terminase protein.....	103
6.4	Effects of ATP analogues on protein thermostability.....	103
6.5	Crystallisation of full-length protein in complex with ADP and non-hydrolysable ATP analogues.....	104
6.6	C-terminus truncation.....	105
6.7	Crystallisation and structure determination of complexes with ADP and non-hydrolysable ATP analogues.....	106
6.8	Structures of complexes with ATP analogues.....	107
6.9	Summary.....	108
7	DNA binding properties of G20c and D6E large terminases.....	110
7.1	DNA binding of the G20c large terminase.....	110
7.1.1	Electrophoretic gel mobility shift assays.....	110
7.2	DNA binding of the D6E large terminase.....	112
7.2.1	Preparation and nuclease activity of the D6E large terminase nuclease deficient mutant.....	112
7.2.2	Electrophoretic gel mobility shift assays.....	113
7.2.3	Microscale thermophoresis.....	114
7.2.4	Analytical ultra-centrifugation.....	117
7.2.5	Size exclusion chromatography.....	120
7.3	Crystallisation and preliminary X-ray analysis of the D6E large terminase complex with DNA.....	121
7.3.1	Crystallisation of protein-DNA complex using poly (T) ssDNA.....	121
7.3.2	Crystallisation of protein-DNA complex using dsDNA.....	122
7.3.3	Analysis of X-ray diffraction data.....	123

7.4	Summary.....	125
8	ATP hydrolysis and oligomerisation of the D6E large terminase.....	126
8.1	Cloning and expression of the ATPase domain.....	126
8.2	Analysis of the oligomeric state.....	127
8.3	ATP hydrolysis by the ATPase domain and full-length terminase.....	128
8.4	Structure of the C-terminal arm.....	129
8.5	The role of the C-terminal arm in ATP hydrolysis.....	131
8.6	Model of the D6E ATPase pentamer.....	131
8.7	Cloning, expression and purification of R158A, R44A and E143A.....	132
8.8	Identification of the trans-acting arginine finger.....	133
8.9	The trans arginine finger.....	134
8.10	The cis arginine finger.....	136
8.11	Insights into ATP hydrolysis by large terminase.....	136
8.12	DNA translocating “tunnel” loops.....	137
8.13	Comparison of inter-domain conformations and interactions.....	138
8.14	Formation of oligomeric assembly of large terminase requires large conformational changes.....	139
8.15	Summary.....	141
9	Conclusions.....	142
Appendices.....		145
Appendix 1.....		145
Appendix 2.....		145
Appendix 3.....		146
Appendix 4.....		148
Appendix 5.....		150
Appendix 6.....		150
Abbreviations.....		152
References.....		155

## List of Figures

Figure 1- 1 Schematic illustration of phage-induced bacteriolysis.....	19
Figure 1- 2 Headful DNA packaging process.....	22
Figure 1- 3 Representative structures of small terminase and portal proteins.....	23
Figure 1- 4 Domain arrangement of T4 and Sf6 large terminase proteins.....	24
Figure 1- 5 Structure and sequence alignment of large terminase ATPases.....	25
Figure 1- 6 Structural comparison of the RNase H from <i>Bacillus halodurans</i> and large terminase nuclease domain from bacteriophage SPP1.....	28
Figure 1- 7 The proposed transition state of the two-metal catalysis mechanism based on RNase H proteins.....	30
Figure 2- 1 Scheme of the pET expression system.....	38
Figure 2- 2 The forces acting on a solute particle in a gravitational field.....	44
Figure 2- 3 A typical thermophoretic curve.....	47
Figure 2- 4 Reactions involved in coupled ATPase assays.....	49
Figure 2- 5 Calibration curve.....	50
Figure 2- 6 Crystallisation phase diagram.....	51
Figure 3- 1 PCR amplification.....	56
Figure 3- 2 Expression and solubility tests of large terminase proteins by SDS-PAGE.....	57
Figure 3- 3 Initial solubility screen of GVE2 large terminase protein.....	58
Figure 3- 4 Secondary solubility screen of the GVE2 large terminase using SDS-PAGE.....	59
Figure 3- 5 Purification of large terminases by nickel affinity and size exclusion chromatography.....	61
Figure 3- 6 Thermal shift assays for G20c large terminase D294N.....	62
Figure 3- 7 Thermal shift assays for D6E and GBSV1 large terminase.....	64
Figure 3- 8 Secondary structure prediction and SDS-PAGE.....	65
Figure 3- 9 Cloning, expression and purification of the C-terminally truncated G20c large terminase.....	66
Figure 3- 10 ATPase activities measured by malachite green phosphate assays.....	67
Figure 3- 11 In vitro nuclease activity assays.....	68

Figure 3- 12 Images of protein crystals.....	69
Figure 4- 1 Chemical structure of AMPPCP, ATP- $\gamma$ -S and AMPPNP.....	71
Figure 4- 2 The stabilisation of ATP, ADP and non-hydrolysable ATP analogues on the thermal stability of G20c large terminase $\Delta$ 443 are characterised using thermal shift assays .....	73
Figure 4- 3 The effect of sodium chloride concentration on nuclease activity.....	74
Figure 4- 4 The effects of temperature and protein concentration on nuclease activity.....	75
Figure 4- 5 The effects of the types of metal ions and metal ion concentrations on nuclease activity.....	76
Figure 4- 6 The effects of ATPase active site mutation on nuclease activity.....	77
Figure 4- 7 The effects of ATP analogues on the nuclease activity of the wild type $\Delta$ 443 and ATPase Walker B $\Delta$ 443 mutant.....	78
Figure 4- 8 The effects of small terminase on nuclease activity.....	79
Figure 4- 9 The effects of small molecules on the nuclease activity.....	80
Figure 5- 1 X-ray data collection.....	82
Figure 5- 2 Comparison of large terminase nuclease domains from different viruses.....	83
Figure 5- 3 Metal coordination in crystal structures.....	85
Figure 5- 4 Dependence of nuclease activity on metals and metal coordination in crystal structures.....	86
Figure 5- 5 Metal binding sites.....	87
Figure 5- 6 Comparison of large terminase and RuvC structures.....	89
Figure 5- 7 Comparison of metal location and DNA orientation.....	90
Figure 5- 8 Conformational flexibility in the active site.....	92
Figure 5- 9 Nuclease active site residues and metal-metal distances observed in crystal structures.....	93
Figure 5- 10 Proposed mechanism of DNA cleavage.....	93
Figure 6- 1 Crystals of the D6E large terminase protein.....	97
Figure 6- 2 Structure determination of the D6E large terminase crystal form 1 by MAD.....	98
Figure 6- 3 Crystals obtained using PGA screen.....	99

Figure 6- 4 X-ray diffraction patterns.....	101
Figure 6- 5 Structure and comparisons of packing in crystal forms 1 and 2 of the D6E large terminase.....	102
Figure 6- 6 Overall structure of the D6E large terminase.....	103
Figure 6- 7 Protein stabilisation by binding of ATP analogues.....	104
Figure 6- 8 Electron density of the ATPase active and the C-terminal arm.....	105
Figure 6- 9 PCR amplification, expression test and purification.....	105
Figure 6- 10 Structure of the D6E large terminase (space group P1, crystal 3) in complex with AMPPNP.....	106
Figure 6- 11 The ATPase active site.....	108
Figure 7- 1 DNA binding assays of the G20c large terminase protein and its C-terminal nuclelease domain.....	111
Figure 7- 2 Structure comparisons of the nuclease active site.....	112
Figure 7- 3 Comparison of in vitro nuclease activities of the wild type D6E large terminase protein and D326N mutant.....	113
Figure 7- 4 DNA binding assays for D6E D326N large terminase.....	114
Figure 7- 5 MST analysis of the DNA binding affinity and cooperativity of poly(T) ssDNA .....	116
Figure 7- 6 MST analysis of the stability of protein-poly(T) ssDNA complex.....	117
Figure 7- 7 Analysis of the stoichiometry of the D6E D326N large terminase complex with ssDNA.....	118
Figure 7- 8 Analysis of stoichiometry of the D6E D326N large terminase complex with dsDNA.....	119
Figure 7- 9 Analysis of the stability of protein-DNA complexe.....	121
Figure 7- 10 Crystals obtained from samples of protein-DNA mixture.....	123
Figure 7- 11 Structure and electron density maps.....	124
Figure 8- 1 PCR amplification and expression tests.....	126
Figure 8- 2 Characterisation of the oligomer state of the full-length D6E large terminase and the N-terminal ATPase domain.....	128
Figure 8- 3 Kinetics of ATP hydrolysis.....	129

Figure 8-4 Structure of the C-terminal arm and comparisons with equivalent structural elements in large terminases of other bacteriophages.....	130
Figure 8-5 Plot of ATP hydrolysis rate versus ATP concentration for the wild type D6E large terminase (squares), R421A mutant (circles) and C-terminal arm truncated protein (triangles).....	131
Figure 8-6 Model of the D6E ATPase domain pentamer.....	132
Figure 8-7 PCR amplification and SDS-PAGE.....	133
Figure 8-8 ATP hydrolysis by the wild type D6E large terminase and its mutants.....	134
Figure 8-9 ATPase comparison.....	135
Figure 8-10 Identification of DNA tunnel loops.....	138
Figure 8-11 Proposed mechanism for large terminase motor assembly.....	140

## List of Tables

Table 1- 1 Examples of isolated and sequenced thermophilic bacteriophages.....	20
Table 2- 1 Standard PCR reaction conditions for large terminase gene amplification.....	34
Table 2- 2 PCR program details for large terminase gene amplification.....	34
Table 2- 3 Standard in-fusion reaction 1.....	35
Table 2- 4 Standard in-fusion reaction 2.....	35
Table 2- 5 Standard colony PCR reaction.....	35
Table 2- 6 Standard colony PCR programme.....	35
Table 2- 7 Standard site directed mutagenesis reaction.....	36
Table 2- 8 Selenomethionine protein expression medium.....	39
Table 2- 9 Calculated Kav and logarithm of molecular weight for selected protein standards .....	50
Table 3- 1 Buffer, salt concentration and pH used in initial solubility screening.....	57
Table 3- 2 X-ray data collection statistics (D6E large terminase).....	70
Table 5- 1 X-ray data collection and refinement statistics (G20c nuclease crystal form 1, 2 and 3, table reproduced from (81)).....	85
Table 5- 2. Structure alignment of G20C large terminase nuclease with Bh-RNase H and Tth- RuvC (177) (table reproduced from (81)).....	91
Table 6- 1 X-ray data collection statistics (D6E large terminase crystal form 1, table reproduced from (186)).....	98
Table 6- 2 Crystal dehydration using ammonium sulfate at various concentrations in the presence of additives.....	100
Table 6- 3 Conditions that are tested as Cryo-protectants for crystals of the D6E large terminase.....	100
Table 6- 4 X-ray data collection and refinement statistics (D6E large terminase crystal form 2, table reproduced from (186)).....	102
Table 6- 5. X-ray data collection and refinement statistics (D6E large terminase crystal form 3, table reproduced from (186)).....	107
Table 7- 1 The length and sequence of the dsDNA constructs used in electrophoretic gel mobility shift assays.....	111

Table 7-2 Sequences of the DNA constructs used in protein-DNA complex crystallisation	122
Table 7-3 Resolution, space group and unit cell parameters of the three data sets collected using DNA 2 (data set 1 and 2) and DNA 5 (data set 3).....	124
Table 8-1 PDBe fold alignment of D6E large terminase ATPase with UvrD and E1 helicases (177).....	138
Table 8-2 Comparison of intra-domain electrostatic interactions (table reproduced from (186)).....	139
Table 8-3 Comparison of intra-domain VDW interaction areas (table reproduced from (186))	
.....	139

## **List of Equations**

Equation 2- 1 The balance of the three forces in a gravitational field.....	45
Equation 2- 2 Estimation of protein-DNA complex partial specific volume.....	46
Equation 2- 3 Thermophoresis depletion.....	46
Equation 2- 4 Normalised fluorescence intensity.....	47

## Acknowledgements

First and foremost, I would like to thank my supervisor, Prof. Fred Antson, for bringing me into the interesting research field of the molecular mechanisms of viral motors and more broadly, the field of structural biology. I appreciate that he has given me enough freedom to carry out research and develop skills within the main objective of the project. I am very grateful for his support, guidance and encouragement throughout my PhD.

Special appreciation and thanks to my co-supervisor, Dr. Sandra Greive, for her time and effort on reading my thesis and reports; for her continued support and encouragement on my research skill development. I would also like to thank her for the inspiring discussions on the project and for introducing me to the field of biophysics.

Thanks are extended to: Maria Chechik for her help on troubleshooting problems that I encountered in cloning, protein expression and purification. I appreciate her efforts on supporting the lab and reagents generation throughout the project; Dr. Huw Jenkins for his help and informative discussions on X-ray data collection strategies, processing and structure determination of G20c and D6E large terminases.

I would also like to thank Dr. Christoph Baumann for reading my TAP report and giving helpful suggestions on my research work; Haochen Zhu for contributing the work on characterising phiOH2 large terminase; Dr. Juan Loredo-Varela for getting me started laboratory work at the beginning of the project; Dr. John Darby and Oliver Bayfield for their help on the setting up thermal shift assays and data analysis for G20c large terminase; Dr. Andrew Leech for AUC data collection and analysis; Dr. Johan Turkenburg and Sam Hart for X-ray data collection. Katarina Meze, Herman Fung and Dr. Evgeny Klimuk for inspiring discussions on large terminase proteins. Simon Grist, Sally Lewis, Louise Haigh and Juliet Borgia for maintaining the wet lab.

Thanks to the following people who have kindly provided the genomic DNA of several thermophilic bacteriophages used in this work: Prof. Konstantin Severinov (G20c and 1-1), Prof. Xiaobo Zhang (GVE2) and Prof. Katsumi Doi (phiOH2).

Thanks to the rest of Antson group and the YSBL people for making my time at the laboratory enjoyable during these years.

Finally, I would like to thank all my family members who have been giving me moral support and encouragement continuously during my PhD.

## **Author's declaration**

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References. Work that has been performed by someone else is detailed below:

The G20c large terminase D294N gene construct cloned in the YSBL3CLIC+ vector was provided by Dr. Sandra Grieve at the beginning of the project. The PacL26 plasmid used as supercoiled DNA substrate for *in vitro* nuclease assays were cloned by Maria Chechik.

Some published or accepted works from this thesis are listed below:

1. Xu, R. G., Jenkins, H.T., Chechik, M., Blagova, E.V., Lopatina, A., Klimuk, E., Minakhin, L., Severinov, K., Greive, S.J. and Antson, A.A. (2017) Viral genome packaging terminase cleaves DNA using the canonical RuvC-like two-metal catalysis mechanism. *Nucleic Acids Research*, 45(6), 3580-3590.
2. Hilbert, B. J., Hayes, J. A., Stone, N. P., Xu, R. G., and Kelch, B. A. (2017). The large terminase DNA packaging motor grips DNA with its ATPase domain for cleavage by the flexible nuclease domain. *Nucleic acids research*, 45(6), 3591-3605.
3. Xu, R. G., Jenkins, H.T., Antson, A.A. and Greive, S.J. Structure of the large terminase from a hyperthermophilic virus reveals a unique mechanism for oligomerisation and ATP hydrolysis. *Nucleic Acids Research*, doi:10.1093/nar/gkx947.

# 1 Introduction

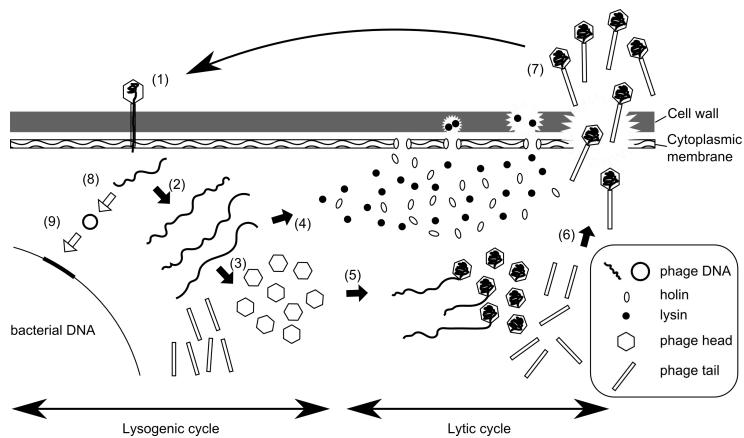
## 1.1 Bacteriophage

Bacteriophages are viruses that can specifically infect and lyse bacteria. Since their discovery by Twort (1) and d'Herelle (2) in 1915 and 1917, they have been playing a crucial role in our understanding of both viral biology and bacterial genetics, and more generally, in the establishment of modern molecular biology. With the development of the electron microscope, the bacteriophage particles were first visualized in the 1940s (3), a significant landmark in understanding their morphology and structure. Bacteriophages normally contain a DNA or RNA genome within the nucleoprotein particle that has an attached proteinaceous tail required for infection. Based on their morphology, nucleic acid properties and the presence or absence of lipid or envelop, bacteriophages can be classified into 13 families in which 96% of the bacteriophages contain dsDNA, with the remaining bacteriophages having ssDNA, ssRNA or dsRNA as their genome (4). Based on their shape, bacteriophages can be subdivided into a variety of morphological classes including tailed, polyhedral, filamentous or pleomorphic.

### 1.1.1 Life cycles

Bacteriophages cannot multiply through division, as cells do, because they lack the necessary molecular machines. Instead, they seek a host cell in which to replicate and assemble new virions using the metabolism and machinery of the host cell. Based on their distinct life cycle during infection, bacteriophages can be divided into two groups: the virulent phages with a lytic cycle in which self-assembly is accompanied by the lysis of bacteria and the temperate phages, which have a lysogenic life cycle in addition to lytic cycle.

In the lytic cycle, the multiplication of the genomic material of the bacteriophage in the host bacterial cell is accompanied by the synthesis of viral proteins. The viral proteins then assemble to form procapsids, that are then packaged with the genome, followed by addition of tail fibers and other maturation proteins that complete bacteriophage head assembly. Finally, the assembled bacteriophages are released after lysis of the bacterium. However, in the lysogenic cycle (5), the genetic material of the phage does not multiply. It inserts into the bacterial chromosome to form a prophage which is replicated along the host genome during the division of the bacterium. In this way, the phage genome is transmitted to the daughter bacteria. The lysogenic cycle can persist until induction occurs, allowing the prophage to be removed from the bacterial genome and starting the lytic cycle, leading to the lysis of bacterium and the release of mature bacteriophage (Figure 1-1).



*Figure 1-1 Schematic illustration of phage-induced bacteriolysis*

1) Adsorption and DNA injection; 2) DNA replication; 3) production of head and tail; 4) synthesis of holin and lysin; 5) DNA packaging; 6) completion of phage particle; 7) disruption of the cell wall and release of the progeny; 8) circularization of phage DNA; 9) integration of the phage DNA into the host genome (Figure adapted from (4)).

### 1.1.2 Bacteriophage therapy

The first bacteriophage therapy on humans was reported in 1921 by Richard Bruynoghe and Joseph Maisin (6), who utilized bacteriophages to treat staphylococcal skin disease. Early research institutes and companies were then set up, including the Eliava Institute, Hirschfeld Institute and Eli Lilly Company, which focused on the commercial production of bacteriophage for human use (7). With the advent of antibiotics in 1940s, the commercial production of bacteriophages was almost ceased in the Western world while in the Eastern Europe and former Soviet Union, the study of bacteriophage therapy was still carried on. Bacteriophage therapy was proved by experiments to be effective as an alternative therapy for patients in cases where antibiotics were ineffective (8). Compared to antibiotic therapy, bacteriophage therapy has several advantages: 1) Bacteriophages specifically infect the pathogen bacterium with less effect on the other natural microflora. Furthermore, their metabolites do not affect eukaryotic cells so adverse effects are minimised; 2) It is effective against multidrug-resistant pathogenic bacteria because the mechanisms by which it induces bacteriolysis differ completely from those of antibiotics; 3) It can respond rapidly to the appearance of phage-resistant mutants because the phages themselves are able to mutate; 4) Bacteriophage can be rapidly modified to combat the emergence of newly arising bacterial threats.

Nowadays, the widespread application of antibiotics in both medicine and agriculture has been causing the rapid growth of multiple drug resistant bacterial pathogen strains. Bacteriophage therapy is undoubtedly the most promising therapy to bring those diseases caused by multiple drug resistance of bacteria pathogens under control.

## 1.2 Thermophilic Bacteriophages

Thermophilic bacteriophages are bacterial viruses that infect bacteria living at extremely high temperature. They are particularly interesting due to the improved thermal stability, thereby being regarded as model systems to investigate the genetic transfer, diversity and evolution of life at high temperature. Although many bacteriophages have been isolated and characterized, only a few of them infect thermophiles. Therefore, our knowledge of thermophilic bacteriophages is limited compared to that of mesophilic bacteriophages (9). It has been reported that thermophilic bacteriophages were isolated from hot springs, mud pots, solfataric fields and deep-sea vents (10). To date, the majority of the isolated and sequenced thermophilic bacteriophages contain double-stranded DNA as their genome and most of their host strains belong to the *Thermus* genus. According to their morphology observed under the electron microscopy, most often they belong to the *Myoviridae* and *Siphoviridae* families, although thermophilic bacteriophages belonging to *Tectiviridae*, *Inoviridae* and *Sphaerolipoviridae* families were also recently reported (examples see Table 1-1).

The deep-sea hydrothermal vents have one of the most extreme environments on the earth, with high temperatures up to 300-400 degrees, high pressures up to 25 Mpa and toxic fluids of sulfur minerals. Thermophilic bacteriophages which infect *Geobacillus* including GVE2 (11) and D6E (12) from deep-sea hydrothermal vents were isolated and sequenced, providing insight into the genomic components and molecular mechanisms of bacteriophage-host interactions of extremophilic deep-sea viruses. For structural biology, thermophilic bacteriophages provide valuable genomic sources for understanding the protein-protein, protein-DNA interactions and DNA translocation of their viral DNA translocation motor, an essential component for their viral self-assembly in extreme environments.

*Table 1-1 Examples of isolated and sequenced thermophilic bacteriophages*

Thermophilic bacteriophage	Family	Host	Isolation	Genome size
BV1(13)	Myoviridae	<i>Bacillus</i> sp. 6k512	Inshore hot spring,	35,055 bp
D6E(12)	Myoviridae	<i>Geobacillus</i> sp. E263	Deep-sea hydrothermal field	49,335bp
GBSV1(14)	Myoviridae	<i>Geobacillus</i> sp. 6k51	Offshore hot spring.	34,683 bp
GVE2(11)	Siphoviridae	<i>Geobacillus</i> sp. E263	Deep-sea hydrothermal vents	40,863bp
RM378(15)	Myoviridae	<i>Rhodothermusmarinus</i>	Geothermal sites	129,908 bp
phiIN93(16)	Sphaerolipoviridae	<i>Thermus aquaticus</i> TZ2	Hot spring soil	19,604 bp
phiOH2(17)	Siphoviridae	<i>G.kaustophilus</i> NBRC	Hot spring	unknown
phiYS40(18)	Myoviridae	<i>Thermus thermophilus</i> HB8	Hot spring	152,372bp
P23-45(19)	Siphoviridae	<i>Thermus thermophilus</i>	Hot spring	84,201bp
P74-26(19)	Siphoviridae	<i>Thermus thermophilus</i>	Hot spring	83,319bp
phiTMA(20)	Myoviridae	<i>Thermus thermophilus</i> HB27	Hot spring	151,483 bp
TS2126(21)	Myoviridae	<i>Thermus scotoductus</i>	Hot tap water	Around 90kbp
MMP17(22)	Myoviridae	<i>Meiothermus</i>	Hot spring	33.5-39.5 kbp
TSP4(23)	Siphoviridae	<i>ThermusTC4</i>	Hot spring	Around 80 kbp.

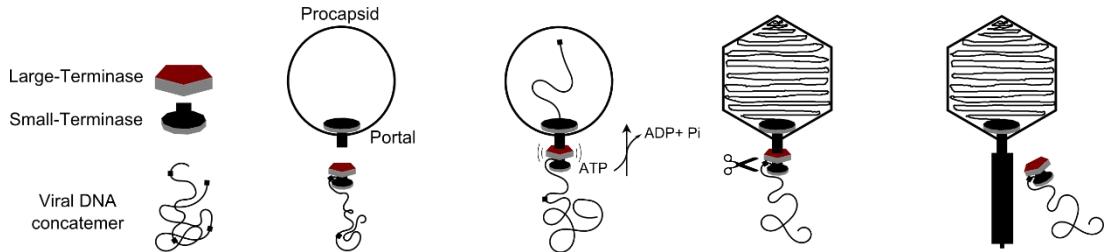
### **1.3 Viral DNA packaging systems**

Genome encapsulation plays an essential role in the life cycles of viruses. Based on strategies employed and the relatedness of components during genome packaging, encapsulation can be classified largely into 3 types: the energy-independent packaging systems, the portal-translocase/terminase system and the FtsK/HerA-type system (24). Type 1 genome encapsulation uses positively charged protein subunits to bind segments of the negatively charged genome. In this manner, the viral genome can be coated with capsid proteins that self-assemble into a capsid around the nucleic acid in an ATP independent process. This type of encapsulation is commonly seen among small DNA and RNA viruses. Giant viruses such as *Mimivirus* from the nucleo-cytoplasmic large DNA virus (NCLDV) family that have a larger genome size utilize the type 3 genome encapsulation mechanism in which the FtsK/HerA-like packaging ATPase plays a crucial role in genome encapsulation in a process that is powered by ATP. Tailed dsDNA bacteriophages and evolutionarily related eukaryotic herpes viruses use the type 2 portal-translocase/terminase genome encapsulation mechanism in which the powerful ATP-dependant packaging machinery, composed of the portal protein and terminases, encapsidates the viral genome into a procapsid, a preformed empty protein shell.

### **1.4 Bacteriophage DNA packaging motor**

The type 2 portal-translocase/terminase genome packaing systems normally involves three protein components: the portal protein which is embedded in the procapsid and contains a tunnel for DNA translocation, and also the small and large terminase proteins (25) (Figure 1-2). The small terminase is responsible for the recognition of viral DNA and positioning of large terminase onto the viral DNA concatemer for the cleavage reaction that generates the viral genome ends. In addition to the endonuclease activity, the large terminase also plays a significant role in converting chemical energy into mechanical movement of the molecular motor via ATP-driven conformation changes resulting in DNA tanslocation into the procapsid (26). The portal protein is located at a unique vertex of the procapsid, providing the channel for the passage of DNA. In the head assembly process, the dodecameric ring of the portal protein acts as a nucleation point for procapsid formation, becoming embedded in one of the twelve 5-fold symmetrical vertices of the icosahedral shell. During this process, the major capsid protein and the scaffolding protein form an empty capsid. The packaging terminases complexed with viral DNA is then docked to the portal ring to start headful packaging coupled with ATP hydrolysis. During this process, the packaging terminases work against high pressure that builds up inside the capsid, where negatively charged DNA becomes packaged to crystalline density (27). After the interior of the procapsid is filled with DNA, the large terminase cuts the DNA concatemer, terminating the DNA

translocation. The large terminase remains bound to the new end of the DNA concatemer while dissociating from the portal ring, and is poised to dock onto another procapsid to begin translocating the next copy of viral genomic DNA. Finally, a connector protein assembles on the portal ring to seal off the portal channel, preventing the packaged viral DNA from leaking and creating a platform for tail assembly (28) (Figure 1-2).



*Figure 1-2 Headful DNA packaging process*

#### 1.4.1 The small terminase protein

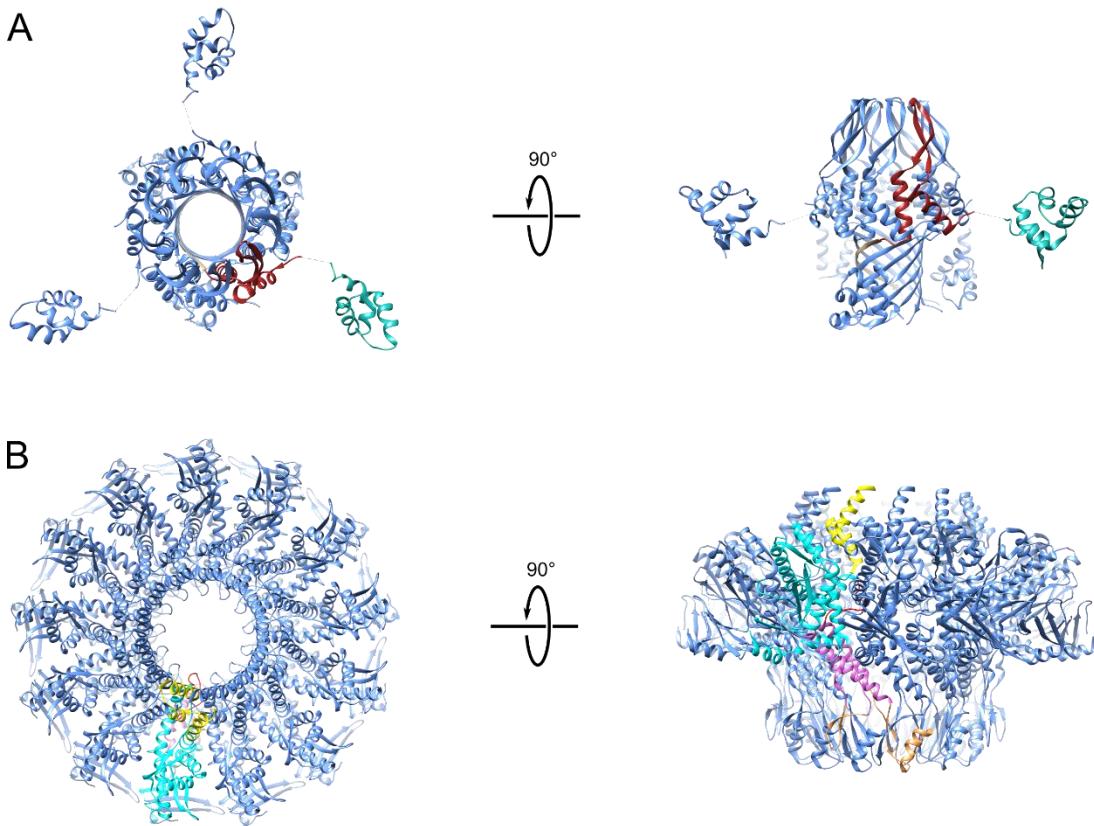
##### 1.4.1.1 Structure and function

During viral DNA translocation, the small terminase protein recognises specific sites on the viral DNA concatemer to position the large terminase for endonucleolytic cleavage. In phage  $\lambda$  and headful phages such as T4, SPP1 and P22, this recognition site is *cosB* (29) and *pac* (30-32), respectively. Structural studies show that the small terminases from SF6, P22 and Sf6 adopt vase-shaped hollow circular structures that are formed by multiple subunits (8 to 11) assembled using circular symmetry (33-35) (Figure 1-3). The central channel that has been proposed to be involved in DNA translocation (“threading model”), has diameters of around 10 to 20 Å. The small terminase monomer is composed of three domains: a globular N-terminal domain which is involved in DNA binding, an oligomerization domain and a flexible C-terminal domain which interacts with the large terminase (35). It has been recently proposed that in contrast to the threading mode (35), the small terminase is more likely to interact with DNA by using N-terminal DNA binding domains of the circular oligomer, via non-sequence specific recognition of curved DNA by its helix-turn-helix motifs (“wrapping model”) (36,37).

##### 1.4.1.2 Small-large terminase interaction

It has been observed for the bacteriophage P22 that the co-expressed small and large terminase proteins form a stable complex (38). The interaction was observed at an overlapped region of the small and large terminase genes which encode the C-terminus of the small terminase and the N-terminus of the large terminase. In the N-terminal ATPase domain of the large terminase, a 58-residue region was identified to be crucial for interaction with small terminase. This region was predicted to form an acidic helix-loop-helix structure to interact with the highly basic C-terminus of the small terminase. Moreover, the

stoichiometry of the complex has also been investigated, suggesting a composition of one small terminase nonomer interacting with two to three large terminase monomers (38).



*Figure 1-3 Representative structures of small terminase and portal proteins*

(A) Structure of the SF6 small terminase. The N-terminal DNA binding domain, oligomerisation domain and C-terminal domain are coloured in sea green, dark red and tan, respectively. (B) Structure of the SPP1 portal protein. A single subunit is coloured as follows, yellow: “Crown”, Cyan: “Wing”, Orchid: “Stem”, Brown: “Clip”. The tunnel loop is highlighted in red.

### 1.4.2 The portal protein

#### 1.4.2.1 Structure and function

Structures of the portal protein have been determined for bacteriophages SPP1, P22, phi29 and T4 by X-ray crystallography or electron microscopy. The portal protein is a cone shaped multi-subunit oligomer that is located at the pentameric vertex of the procapsid (Figure 1-3). The central channel has been thought to allow the DNA to be translocated into the prohead during viral packaging. The portal assembly has been proposed to serve three functions in the life cycles of most tailed bacteriophages, including initiating head assembly, providing a platform for the packaging motor during DNA packaging into the prohead and also during tail attachment after DNA packaging (39).

#### 1.4.2.2 Portal-large terminase interaction

Electron microscopy studies suggest that the large terminase subunits assemble to form pentameric donut shaped rings to power DNA translocation (40,41). Direct interaction between the portal and large terminase proteins has been observed in bacteriophage SPP1 (42). In a more recent study on bacteriophage T4, it has been proposed that the N-terminal ATPase subdomain II of large terminase gp17 interacts with the “stalk” (“clip”) subdomain of the portal protein by a HLH motif (helix-loop-helix) (Figure 1-3). In this study, interaction between the C-terminal nuclease domain and the portal protein has also been observed, but the affinity was low and appeared to be non-specific compared to the N-terminal domain (43). A molecular lever mechanism was proposed in which the HLH motif couples ATP hydrolysis and DNA movement. However, a later study based on FRET analysis argued that not only should the C-terminal nuclease domain not be eliminated as interacting with portal, but is likely to be even closer to the portal assembly than the ATPase domain and play more important roles in portal-large terminase interaction (44).

### 1.4.3 The large terminase protein

The large terminase protein is composed of an N-terminal ATPase domain powering the DNA translocation by utilizing energy from ATP hydrolysis and a C-terminal endonuclease domain cleaving the genomic DNA concatemer at the initiation and a later termination stages of DNA packaging (45). The physical linker that connects these two domains is thought to be essential for its *in vivo* packaging activity, based on the fact that while the individual domains maintained their activity, the packaging activity was lost (46). The domain arrangement is shown in Figure 1-4 using T4 and Sf6 large terminases as examples.

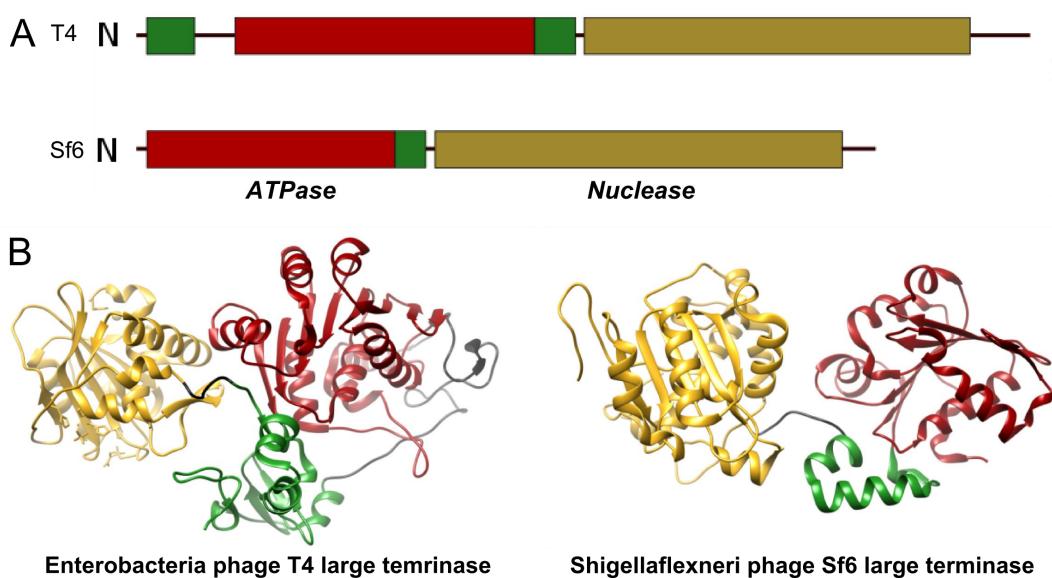


Figure 1-4 Domain arrangement of T4 and Sf6 large terminase proteins

(A) Schematic diagram of T4 and Sf6 large terminases comprising the N-terminal ATPase domain (red/green for subdomains I/II, respectively) connected to a C-terminal nuclease domain (yellow) by a flexible linker (black). Residue numbers at domain boundaries are indicated. (B) Ribbon diagram showing the structures of the T4 and Sf6 large terminases. The ATPase and nuclease domains are coloured in accordance with (A).

#### 1.4.3.1 The ATPase domain

The ATPase domain uses the energy from ATP hydrolysis to power DNA translocation (47,48). Due to its topological differences with other ASCE (Additional Strand Catalytic E (glutamate)) ATPases, the large terminase ATPase was classified into a unique division among the HerA/Ftsk, RecA, PilT and helicase superfamily (49,50).

#### Structure

The ATPase domain is located at the N terminus of the large terminase protein. Structural studies show that the ATPase domain consists of two subdomains: an N-terminal ASCE subdomain containing the ATPase active site (subdomain I in Figure 1-4) and a C-terminal lid subdomain (subdomain II in Figure 1-4, also referred to as linker or hinge) which sits on the top of the N-terminal ASCE subdomain and is formed by a bundle of  $\alpha$ -helices (Figure 1-5). Like other ASCE superfamily ATPases (50), the N-terminal ASCE subdomain contains five  $\beta$ -strands in the topological order of  $\beta_5\text{-}\beta_1\text{-}\beta_4\text{-}\beta_3\text{-}\beta_2$  that are sandwiched by two stacks of  $\alpha$ -helices. Additional secondary structural elements, including an  $\alpha$ -helix and three  $\beta$  strands were found inserted between  $\beta_2$  and  $\alpha_3$  (50) (Figure 1-5). Sequence alignment with several cellular helicases indicate that they share similar functional motifs such as the Walker A-, Walker B-, adenine binding- and C- (Sensor I) motifs (51).

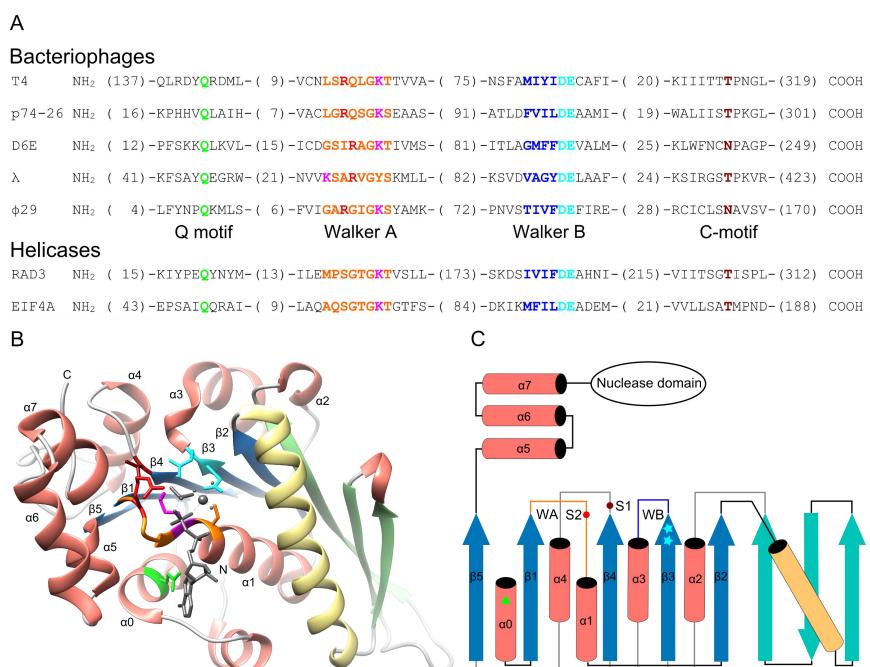


Figure 1-5 Structure and sequence alignment of large terminase ATPases

(A) Sequence alignment of large terminase ATPases with helicases. Amino acids belonging to key motifs are coloured as follows: Q motif (green), Walker A (orange/ magenta), Sensor I (red), Walker B (blue/cyan), sensor II (dark red). (B) Ribbon diagram showing the structure of the P74-26 large terminase ATPase domain. Amino acids of the key motifs are highlighted using the same color as in (A). (C) Topological diagram of large terminase ATPase domain. WA:Walker A (Red line), WB: Walker B (cyan stars), S1: Sensor I (dark red ball), S2: Sensor II (red ball), Q (green triangle): Q motif. Conserved secondary elements found in other AAA+ proteins are coloured in blue ( $\beta$ -strands) and coral ( $\alpha$ -helices); Additional secondary elements are coloured in spring green ( $\beta$ -strands) and yellow ( $\alpha$ -helix)

### **Walker A motif**

The Walker A motif is a common nucleotide binding motif that was first identified in the  $\alpha$ - and  $\beta$ -subunits of the F1-ATPase, myosin and other ATP-dependent enzymes (52). It is positioned at the beginning of helix  $\alpha$ 1 which links the  $\beta$ 1 and  $\beta$ 2 strands (Figure 1-5). The consensus sequence of the Walker A motif is G/A-XXXXGK(T/S) (where X is any amino acid) (53). Mutations in this motif inhibit ATP binding. Interestingly, large terminases display several deviations from the consensus Walker A motif sequence of ATPases. Firstly, in some large terminases, the lysine residue normally present at the C-terminus of the consensus sequence, has migrated to the N-terminus of the Walker A motif. It has been suggested that although the position of the lysine residue varies within the motif, it is likely to play the same functional role, i.e. to interact with the  $\beta$  and  $\gamma$  phosphates and stabilise the Mg<sup>2+</sup>-ADP product, during ATP hydrolysis (54). Secondly, an arginine residue, not found in the consensus sequence, is highly conserved among many large terminases. Mutagenesis of this arginine has been shown to be crucial for ATPase activity and/or DNA packaging in studies of bacteriophage T4, P74-26 and  $\lambda$  large terminases (40,55,56). Based on structural comparisons with other ATPase motors, it has been suggested that this conserved arginine serves as a *cis*-arginine finger which is likely to act as a trigger for ATP hydrolysis and/or mechano-chemical coupling during DNA translocation (40,57). Recent structural and mutagenesis studies on bacteriophage P74-26 large terminase ATPase suggests that, instead of acting as a *cis*-arginine finger, this arginine is more likely to function as sensor II, as seen in other ATPase motors, to facilitate ATP hydrolysis in *cis* by inducing conformational changes in the lid subdomain upon ATP hydrolysis (56).

### **Walker B motif**

For ASCE superfamily proteins, the Walker B motif (52) is located at the C-terminus of the  $\beta$ 3 strand (Figure 1-5). It contains a conserved hydrophobic amino acids rich sequence which can be described as hhhhDE (h represents a hydrophobic amino acid) (53). In common with other well studied ATPases such as RecA, F1-ATPase, DNA/RNA helicases and translocases, structural studies on the Sf6 large terminase showed that the aspartate residue coordinates the Mg<sup>2+</sup> ion, which positions the  $\beta$  and  $\gamma$  phosphates of ATP, facilitating

the nucleophilic attack on the  $\gamma$ -phosphate by a glutamatic acid activated water molecule (57). This motif has been reported to be crucial for ATP hydrolysis (53). Similarly, in bacteriophage large terminases, this motif is intolerant of any mutations, further emphasising its important roles in catalysis (56,58).

### ***Adenine binding motif***

The adenine binding motif (also called Q motif) was first discovered in DEAD box proteins, such as SF1 and SF2 RNA helicases, and lies about 17 amino acids upstream of the Walker A motif (59,60). It is composed of 9 amino acids including a highly-conserved glutamine that contacts N6 and N7 of the base heterocyclic ring via hydrogen bonds (61) (Figure 1-5). The hydrophobic residues within this motif were observed to form hydrophobic contacts with the adenine ring. Mutational studies of the Q motif in RNA helicases concluded that it directly regulates the affinity of the protein for the substrate through conformational changes associated with nucleotide binding (62). This motif in bacteriophage  $\lambda$  large terminase had been proposed to be important in determining the power and efficiency of the packaging motor (60).

### ***C-motif***

The C-motif (Sensor I) representing a conserved threonine/asparagine-serine, positioned 20–30 amino acids downstream of the Walker B motif was first discovered in helicases (63). It has been suggested that this motif interacts with the  $\gamma$  phosphate of ATP and participates in ATP hydrolysis (53) (Figure 1-5). Mutagenesis studies of this region in T4 large terminases generated a novel phenotype which only binds and hydrolyzes ATP once but has lost the ability to turnover (64). In the case of phage  $\lambda$  large terminase, the C-motif mutant protein can package only part of the phage genome (65), indicating that this region plays an important role not only in allowing product release and substrate rebinding, but in coupling ATP hydrolysis to DNA packaging.

### ***DNA binding regions***

The N-terminal domain of the bacteriophage T4 large terminase has been suggested to play a role in interacting with DNA, supported by the observation that DNA binding was only observed for the N-terminal ATPase but not for the C-terminal nuclease domain (66). In contrast, in the DNA translocation models proposed later, the C-terminal nuclease domain appears to form a major contact with DNA (40). However, consistent with the earlier observation that the ATPase domain of the T4 large terminase interacts with DNA (66), recent mutagenesis studies of the P74-26 ATPase also emphasized the role of ATPase domain in DNA binding and viral DNA packaging (56). In addition, arginine residues that

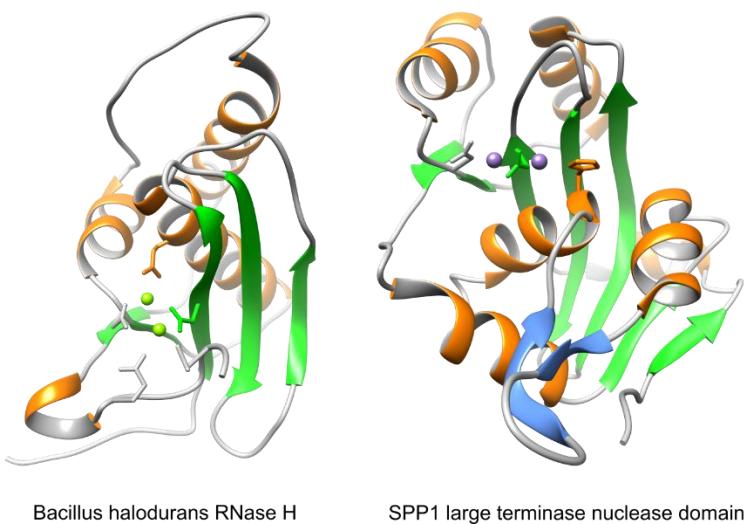
are critical for DNA binding were only found in the ATPase domain, supporting the ATPase domain as a primary DNA binding region.

#### 1.4.3.2 The nuclease domain

The large terminase nuclease domain is a member of the RNase H-like endonuclease superfamily (retro-viral integrase superfamily). Proteins in this superfamily are commonly found to require the presence of divalent metal ions, such as  $Mn^{2+}$  and  $Mg^{2+}$ , for cleaving their DNA substrates.

##### ***Structure of the nuclease domain***

The nuclease domain consists of a seven-stranded anti-parallel  $\beta$ -sheet flanked by  $\alpha$ -helices on both sides, closely resembling the RNase H fold (40,67-69). The RNase H fold, found in the RNase H superfamily proteins is found in several other cellular enzymes, such as RuvC resolvases, integrases and translocases. Despite their structural similarities with large terminase nuclease, these proteins typically have five-stranded central  $\beta$ -sheets instead of the seven-stranded sheet found in large terminases. In addition, an auxiliary  $\beta$ -hairpin, found only in large terminase nucleases and proposed to participate in nuclease activity regulation and DNA interaction (67), was not observed in other members of the RNase H like superfamily. Several negatively charged residues and bound divalent metal ions found in the nuclease active site have been reported to play crucial roles in phosphodiester bond cleavage (67,70)(Figure 1- 6).



*Figure 1-6 Structural comparison of the RNase H from *Bacillus halodurans* and large terminase nuclease domain from bacteriophage SPP1*

*The  $\beta$ -hairpin is coloured in blue. The side chains of active site carboxylates are shown in sticks. Catalytic metal ions are shown as sphere and coloured in green (magnesium) and grey (manganese).*

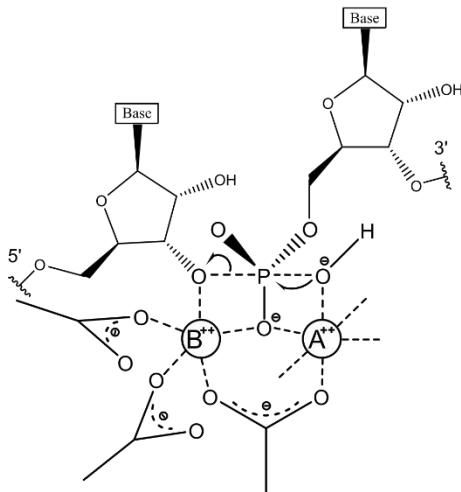
### **Nuclease activity regulation**

The packaging nuclease activity of the large terminase needs to be regulated, as the enzyme is inactive during DNA translocation. It has been proposed that the nucleotide binding state of the N-terminal ATPase domain regulates the nuclease activity of the T4 large terminase (71), as stimulation of the non-sequence specific nuclease activity has been observed in the presence of ATP analogues. It has also been proposed that an auxiliary  $\beta$ -hairpin that is not found in other RNase H family proteins, plays an important role in regulating nuclease activity (67). In the proposed model, the  $\beta$ -hairpin alters its conformation to allow DNA to access the nuclease active site. Observations that an isolated C-terminal nuclease domain is not as active as full-length large terminase (67) or even completely inactive (68) suggested the importance of cross-talk between the N-terminal ATPase and C-terminal nuclease domains, for regulating nuclease activity. Recently, structural studies on the Sf6 GP2 nuclease domain provided further insight into the regulation of the nuclease activity (72). Binding of divalent metals in the structure of Sf6 GP2 nuclease could only be observed following mutation of K428, which blocks the active site by forming a salt bridge with one of the catalytic aspartates. The location of K428 in the  $\beta$ -hairpin further emphasized the importance of this secondary structure element for nuclease activity regulation.

### **Catalytic mechanism**

It has been assumed that the large terminase nuclease utilizes the two-metal catalysis mechanism proposed for other members of the RNase H-like superfamily (73) such as transposases, retroviral integrases, and RuvC Holliday junction resolvases (74,75). This assumption is supported by three observations: firstly, the large terminase nuclease domain resembles the RNase-H fold (67-69,76). Secondly, similar to RNase H proteins, the large terminase nuclease generates 3'-hydroxyl and 5'-phosphate ends during phosphoryl transfer reaction (77,78). Thirdly, two catalytic metals ions, bound at positions A and B, have been observed in crystal structures of SPP1, HCMV and Sf6 nucleases (67,69,79). The catalytic mechanism involving two metal ions was previously proposed for phosphoryl-transfer reactions catalysed by DNA polymerase I 3', 5'-exonuclease, RNase P, alkaline phosphatase, group I and group II self-splicing introns and spliceosome (80). Specifically, the two metal ions form inner-sphere complexes with the scissile phosphate, the active site carboxylates and coordinated water molecules (Figure 1-7). Metal A activates a coordinated water or sugar hydroxyl for nucleophilic attack, while metal B stabilises the oxyanion leaving group in the transition state (80). A structural study on the *Bacillus halodurans* RNase H (Bh-RNase H) complex with an RNA/DNA hybrid suggested that during catalysis, the two metal ions, initially separated by  $\sim$ 4.0 Å, are likely to move closer together to the inter-ionic distance of  $\sim$ 3.5 Å, facilitating the formation of the pentavalent transition state (70).

Recently, two different metal binding modes were reported for the Sf6 GP2 nuclease (69). In the first metal binding mode, the two  $Mg^{2+}$  or  $Mn^{2+}$  ions were modelled at ultra-short metal-metal distances of 2.45 Å and 2.63 Å, respectively, whereas in the second mode the two  $Mn^{2+}$  ions are separated by 3.75 Å. It was argued that binding two metals at the ultra-short metal-metal distance generates a positive niche, facilitating formation of the transition state (69).



*Figure 1-7 The proposed transition state of the two-metal catalysis mechanism based on RNase H proteins*

*In this mechanism, the scissile phosphate is jointly coordinated by two catalytic metal ions. Metal A activates a water nucleophile while metal B stabilises the pentavalent phosphorane transition state (figure reproduced from (81)).*

## 1.5 DNA cleavage specificity

Based on the observations for several well-characterised bacteriophages ( $\lambda$ , HK97, P2, T4, SPP1, P22, Sf6), their DNA cleavage strategies can be classified into sequence specific and non-sequence specific (headful) (82). It is possible to predict the DNA cleavage specificity of an unknown bacteriophage, by performing amino acid sequence and structural comparison of the large terminase with other well characterised bacteriophage large terminases (82). However, due to sequence diversity, as well as the lack of structural information, it is necessary to determine the chromosome end sequence of an unknown bacteriophage to gain further understanding on its DNA cleavage specificity.

## 1.6 Proposed mechanisms for large terminase DNA translocation

To elucidate the role of the large terminase protein during viral DNA packaging, several DNA translocation mechanisms were proposed.

### **1.6.1 Inch-worm (electrostatic force driven) mechanism**

In this model, electrostatic interaction was claimed to be the force for the DNA translocation of the T4 bacteriophage (76). Specifically, the conformational change generated by ATP hydrolysis triggers the rotation of the transmission domain (subdomain II in Figure 1-4, green), also referred to as the linker/lid subdomain or the hinge domain, to align the charged residues on the surface of the N- and C-terminal domains. The electrostatic attraction caused by the opposite charges pull the C-terminal domain towards the N-terminal domain, thereby translocating DNA by two base pairs. In this model, two DNA binding sites at opposing surfaces of the C-terminal nuclease domain have been proposed, one for DNA translocation and the other for DNA cleavage. However, it is difficult to see how the nuclease activity can be switched on, as this would require a significant rotation of the C-terminal nuclease domain within the oligomerised state during DNA packaging.

### **1.6.2 Chemo-mechanical coupling mechanism**

Based on the Sf6 large terminase structural data it has been proposed that the large terminase utilizes a chemo-mechanical coupling and coordination mechanism for DNA translocation and cleavage (83). According to this model, the conformational changes of the linker domain induced by the ATP hydrolysis and product release, physically move the DNA. The large terminase switches between pre-hydrolysis and ATP hydrolysis states, leading to the continuous translocation of the DNA. At the final stage of DNA translocation, the nuclease active site becomes exposed to DNA cleavage owing to conformational changes, possibly induced by the fully filled capsid.

### **1.6.3 Torsional compression mechanism**

In the DNA packaging assays of the bacteriophage T4, normal DNA can be packaged whereas no translocation was observed in the presence of abnormal DNA. The inhibitory effects on T4 DNA packaging were likely caused by the abnormal DNA structures, suggesting a “torsional compression mechanism” (84). According to this mechanism, DNA binding by the portal protein is coupled with the large terminase making a power stroke, inducing a transient compression on the DNA through local conformational changes in the DNA duplex. When normal dsDNA was used in DNA packaging assays, sufficient compression at the entrance to portal’s tunnel would promote translocation through the tunnel and the release of the translocated DNA segment at tunnel’s exit, into the prohead. However, in the presence of abnormal DNA substrates (nicked, hairpin, mismatched, Y-DNA), the force generated by the large terminase could be dissipated before translocation, resulting in DNA movement in the reversed direction. This mechanism is supported by

molecular dynamics simulations which predicted transitions between the A-DNA and B-DNA forms (85).

#### 1.6.4 The “lever-like motion” mechanism

Recently, X-ray structures of the P74-26 bacteriophage ATPase domain, in the apo form and in complex with ADP·BeF<sub>3</sub><sup>-</sup> have been determined (56). The largest conformational change between the apo and ADP·BeF<sub>3</sub><sup>-</sup> bound structures occurred in the lid subdomain, induced by its interaction with the P-loop upon the binding of the ATP analogue, resulting in 13-degree rotation of the lid subdomain. This observation is consistent with the chemo-mechanical mechanism proposed previously. Apart from the structural findings, a conserved *trans*-acting arginine finger mediating ATP hydrolysis was identified by mutagenesis studies (56). A model for DNA translocation was then proposed by molecular docking based on the biochemical and crystallographic observations. In this model, 5 ATPase domains assemble to form a pentameric ring with a *trans*-arginine finger of each domain projecting into the active site of the adjacent subunit.

### 1.6 Aims of research

The aims of the research on large terminase proteins, presented in this thesis were:

- 1) To understand the structural basis of action, based on structural data on large terminase proteins from several thermophilic bacteriophages;
- 2) To analyse the DNA binding properties of the large terminase, including affinity and stoichiometry of binding;
- 3) To investigate the DNA cleavage mechanism of the large terminase protein;
- 4) To understand how large terminase assembles to form an active motor and how it hydrolyses ATP.

## 2 Material and methods

This section describes the basic principles and general protocols that were used to perform each experiment.

### 2.1 DNA Cloning

DNA cloning is a process that makes multiple and identical copies of a chosen DNA template. It is usually the starting point of many genetic engineering approaches in molecular biology. DNA cloning has important applications, for example, it has been commonly used in making large amount of gene products for biochemical studies or in characterising the function of specific genes by introducing mutations.

#### 2.1.1 PCR amplification

PCR amplification is one of the tools for DNA cloning. It starts with heat induced melting of the double-strand DNA template, followed by annealing of complementary PCR primers to each of the template strands when reducing the temperature. In the next step, the polymerase starts DNA synthesis and chain elongation using the provided nucleotides. The same process repeats per cycle and the products obtained from the previous cycle are used as templates for the subsequent amplification. This leads to a robust exponential increase of the target PCR products.

Primers (sequences see Appendix 1) were designed manually (86) and synthesised by Eurofins. The lyophilised oligoes were re-suspended in TE buffer (10 mM Tris pH 8.0 and 1 mM EDTA). Genes that encode the full-length large terminase proteins of bacteriophage D6E (1-427, 1-234), GBSV1 (1-573), GVE2 (1-567), phiOH2 (1-420), 1-1 (1-616) and G20c (1-443, 1-485, 256-443) were amplified using PCR. The full-length D6E and GBSV1 large terminase genes were synthesised and pre-cloned into the pUC57 plasmid by Genewiz. Their codons were optimised for *E.coli* protein expression. In the case of GVE2, phiOH2 and 1-1, genomic DNA (prophage) was used as the DNA template. Linearised plasmid vector (YSBL3CLIC+ (87) and Champion SUMO vector (88) (Thermo Fisher Scientific)) were amplified using a similar protocol, except different primers and longer extension times were used (see below). YSBL3CLIC+ vector was used to express recombinant proteins with HRV 3C protease cleavable N-terminal 6-histidine tag using pET expression system (89-91). Champion SUMO vector was used to express N-terminal his-tag and SUMO fusion proteins for increased expression and solubility and generation of native protein following cleavage by SUMO Protease. PCR was performed using a Life ECO thermocycler. Reaction set-up and the programme used to amplify the DNA template are listed in Table 2-1 and Table 2-2, respectively.

*Table 2- 1 Standard PCR reaction conditions for large terminase gene amplification*

<b>Components</b>	<b>Volume (µl)</b>	<b>Final concentration</b>
5× reaction buffer	10	
Forward primer	5	2 µM
Reverse primer	5	2 µM
Polymerase*	0.5	0.05 U/µl
Template	0.5	1-2 ng/µl
dNTP	0.4	0.2 mM
Water	28.6	
Total volume	50	

\* Phusion® or Q5® High-Fidelity DNA Polymerase (New England Biolabs) were used.

*Table 2- 2 PCR program details for large terminase gene amplification*

<b>Step</b>	<b>Temperature (°C)</b>	<b>Time (s)</b>	<b>Number of cycles</b>
Initial denaturation	98	30-120	1
Denaturation	98	15	35
Annealing	45-60	30	
Extension	72	90-180*	
Final extension	72	420	1
Hold	4-10	∞	

\* The selection of extension time was based on 15-30 seconds per kb

### 2.1.2 Ligation Independent Cloning

Ligation independent cloning is a gene recombination method that does not require the use of DNA restriction enzymes and DNA ligase. In this method, the vector and gene of interest were prepared by PCR amplification followed by treatment with T4 DNA polymerase. The T4 DNA polymerase has 3'-5' exonuclease activity which creates complementary overhangs on both the vector and gene of interest. The resulting circular recombinant molecules can be used for efficient bacterial transformation (92).

In this work, PCR amplified bacteriophage large terminase genes, including full-length D6E, GBSV1, GVE2, phiOH2, 1-1 and truncated G20c and D6E were cloned into linearised YSBL3CLIC+ (87) using ligation independent cloning. To improve solubility, GVE2, phiOH2 and 1-1 large terminase genes were also cloned into Champion SUMO vector (88) (Thermo Fisher Scientific). All in-fusion reactions were performed using two commercial kits from ClonTech or New England Biolabs are listed in Table 2-3 and Table 2-4, respectively. The reaction mixture was incubated at 50 °C for 1h before cell transformation (see section 2.2 below).

*Table 2- 3 Standard in-fusion reaction 1*

<b>Components</b>	<b>Volume (μl)</b>	<b>Mass (ng)</b>
5× In-Fusion HD Enzyme Premix	2	
Insert	1	50*
Vector	1	100*
Water	6	
Total volume	10	

\*Mass calculated based on the molar ratio of 2:1 (Insert: Vector).

*Table 2- 4 Standard in-fusion reaction 2*

<b>Components</b>	<b>Volume (μl)</b>	<b>Mass (ng)</b>
2× HiFi DNA Assembly Master Mix	5	
Insert	1	50*
Vector	1	100*
Water	3	
Total volume	10	

\*Mass calculated based on the molar ratio of 2:1 (Insert: Vector).

### 2.1.3 Colony PCR

Colony PCR is a convenient and efficient method to determine if the insert is successfully recombined into the vector. Colonies obtained from transformation were added to the PCR reaction. In the initial heating step, the bacterial cells will be lysed by the heat and release the plasmid as PCR template. In this work, colony PCR was used as a standard technique to screen for which colonies carry the target insert genes. The standard colony PCR reaction and programme are listed in Table 2- 5 and Table 2- 6, respectively. Primer sequences used in colony PCR are listed in Appendix 1.

*Table 2- 5 Standard colony PCR reaction*

<b>Components</b>	<b>Volume (μl)</b>	<b>Final concentration</b>
5× reaction buffer	1	
Forward primer (Insert)	1	10 μM
Reverse primer (T7 terminator)	1	10 μM
Polymerase*	0.04	0.02 U/μl
dNTP	0.08	0.2 mM
Water	6.88	
Total volume	10	

\* Dream Taq™ Hot Start DNA Polymerase (Thermo Fisher Scientific) were used.

*Table 2- 6 Standard colony PCR programme*

<b>Step</b>	<b>Temperature (°C)</b>	<b>Time (s)</b>	<b>Number of cycles</b>
Initial denaturation	94	60	1
Denaturation	94	15	25
Annealing	50	30	
Extension	72	90*	
Final extension	72	300	1
Hold	4-10	∞	

\*The selection of extension time was based on 15-30 seconds per kb

#### 2.1.4 Site directed mutagenesis

In molecular biology, site directed mutagenesis methods are used to introduce codon changes to the DNA sequence of a gene. This technique requires the use of DNA primers which contains the desired mutations to hybridise with the target gene of interest. The polymerase can then extend the primers to generate genes which carry the expected mutations

In this thesis, site directed mutagenesis was used as a routine method to introduce mutations or C-terminal truncations in bacteriophage G20c and D6E large terminase genes. The standard reaction is listed in Table 2-7. The PCR programme used is similar to that used in PCR amplification in Table 2-2, except that 16 cycles were used instead of 35. The primer sequences are listed in Appendix 2.

*Table 2-7 Standard site directed mutagenesis reaction*

Components	Volume (μl)	Final concentration
2× HiFi PCR premix*	12.5	
Forward primer	2.5	10 μM
Reverse primer	2.5	10 μM
Template	0.5	1-2 ng/μl
Water	7	
Total volume	25	

\*ClonAmp HiFi PCR Premix contains dNTPs and optimised buffer for rapid set-up of PCR reactions.

## 2.2 Transformation

In nature, transformation is a form of horizontal gene transfer which passes plasmid DNA of a certain function from one bacterium to another. For transformation to occur, the recipient bacteria must be in a state of competence which may take place as a time-limited response to environmental conditions, such as starvation and cell density (93). This competence can also be induced in a laboratory, which resulted in competent cells stocks that have been widely used in molecular biology for transferring the plasmid DNA of interest for protein or DNA production.

Typically, 1 μl (50-100 ng/μl) of recombinant plasmid DNA was added into 20 μl chemically competent cells. For genes encoding protein that may cause stress to the cell, Rosetta (DE3) pLysS (94,95) were used instead of BL21 (DE3) and B834 (DE3) strains (96-98) (Novagen EMD Millipore, USA). The mixture was left on ice for up to 30 mins before heat shock at 42 °C for 1 min using a water bath heating block. The mixture was cooled on ice for 2 mins, followed by adding 200 μl of lysogeny broth (LB) medium and incubating at 37 °C for up to 60 mins. Competent cells were then spread on to an agar plate containing

desired antibiotics and incubated at 37 °C for 18h. After incubation, colonies of bacterial which carry the target plasmid containing the antibiotics resistant gene will grow on the plate.

### **2.3 Plasmid DNA production and purification**

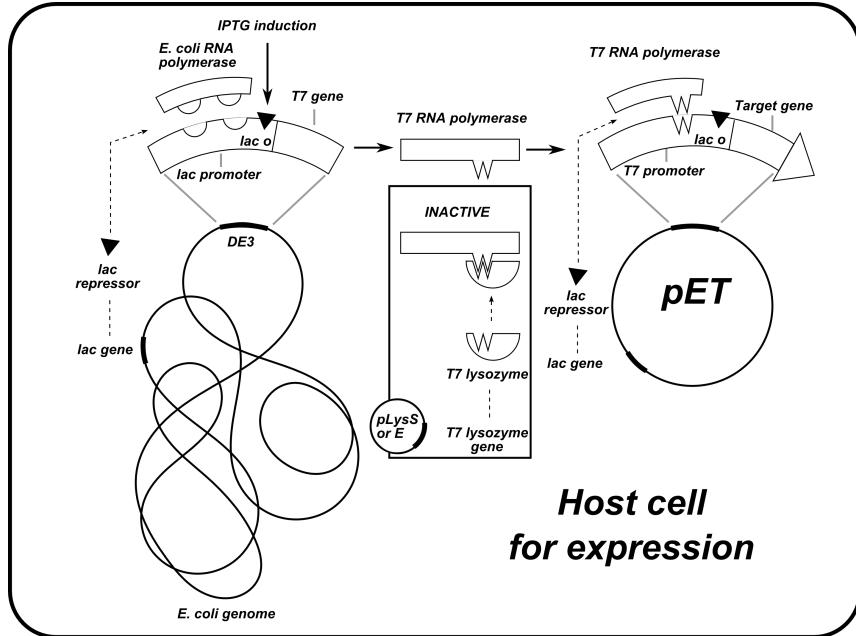
After transformation, plasmid DNA was produced by small scale cell cultures and purified using a commercial plasmid DNA purification kit. Specifically, a 10ml cell culture was incubated at 37 °C for 18h inoculated from a single colony obtained from cell transformation. Appropriate antibiotics was added to suppress the growth of bacterial that does not carry the plasmid with the resistance gene. The cell pellets were collected by centrifugation from which plasmid DNA was purified using NucleoSpin® Plasmid kit. A resuspension buffer containing RNase A was firstly added to resuspend the cell pellets, before the lysis buffer was added. The NaOH and SDS in the lysis buffer lyse the cell by destroying the cell wall and dissolving cell membrane. The RNase A added before cell lysis will degrade the undesired cellular RNA. Neutralisation buffer was added to turn the pH back to neutral. During this process, denatured large chromosome DNA and cellular proteins that are released by cell lysis will aggregate and precipitate. The denaturation of supercoiled plasmid DNA is reversible and stays in the supernatant. This solution was passed through a silicone gel membrane column and the plasmid DNA stays bound to the membrane. After an ethanol wash to remove other impurities bound to the membrane, the plasmid DNA was eluted and collected by centrifugation. The above-mentioned protocol was used as a standard step for plasmid production and purification in this thesis.

### **2.4 *E. coli* protein expression**

#### **2.4.1 The pET expression system (89-91)**

The λ (DE3) prophage containing the gene coding T7 RNA polymerase is engineered into *E. coli* strains under a modified lac operon system. Instead of a T7 promoter sequence in front of the *lac* operator sequence, there is a *lac* promoter sequence that the native *E. coli* RNA polymerase is able to bind. Upon IPTG induction, the lac repressor protein (LacI) falls off the *lac* operator sequence of DNA in the host chromosome so that T7 RNA polymerase can be transcribed and translated. For some strains like Rosetta(DE3)pLysS, an additional plasmid containing the gene coding for T7 lysozyme was introduced, which inactivates the T7 promoter. This allows more tightly controlled target protein expression and helps to reduce background expression levels of T7 RNA polymerase from expressing the protein of interest prior to induction. These strains are especially useful for the expression of genes that

are toxic to *E. coli* growth. These strains also carry tRNA genes encoding codons that are rarely used in *E. coli*, which makes it possible to express eukaryotic proteins.



*Figure 2-1 Scheme of the pET expression system*

(Figure adapted from <https://www.quora.com/How-does-IPTG-induced-gene-expression-work-at-a-molecular-level>).

#### 2.4.2 Protein expression protocol

The full-length large terminase genes together with all the mutants and truncated constructs were transformed and expressed in *E. coli* BL21-Gold (DE3) (Agilent Technologies USA, Inc) or *E. coli* Rosetta (DE3) pLysS (Novagen EMD Millipore, USA) in LB medium containing 34 µg/ml kanamycin (BL21-Gold) or 34 µg/ml kanamycin and 34 µg/ml of Chloramphenicol (Rosetta pLysS). Cell growth starts with a 10 ml LB medium overnight culture containing the above mentioned anti-biotics at 37°C. Cells were inoculated into 2 litre flasks containing 750 ml LB medium and grown at 37°C until OD<sub>600</sub> reached 0.6-0.8 followed by induction with 1mM isopropyl 1-thio-β-D-galactopyranoside (IPTG) and further growth for 2 h. Cells were harvested by centrifugation for 20 min at 5,000x g at 4°C and frozen at -80°C before purification.

#### 2.4.3 Selenomethionine labelled protein expression protocol

In selenomethionine labelled proteins, the methionines were replaced by Selenomethionines. Such proteins are particularly useful for structure determination by anomalous dispersion approaches, such as SAD (single wave-length anomalous dispersion) or MAD, where phase information is gained due to X-ray adsorption at the specific X-ray wavelength, corresponding to the Se adsorption edge. The Selenomethionine labelled protein was

produced using the *E. coli* B834 strain, a methionine auxotroph that is unable to synthesize methionine. Selenomethionine was provided in the minimum methionine-free cell culture medium which can be used for the cell to synthesize Selenomethionine protein. A protocol for producing Selenomethionine labelled protein is described in this section (Table 2-8):

*Table 2-8 Selenomethionine protein expression medium*

*Components	Concentrations	Volume for 1 litre
MgSO <sub>4</sub>	1 M	2 ml
20X M9 medium	1 X	100 ml
FeSO <sub>4</sub> 7H <sub>2</sub> O	12.5 mg/ml	2 ml
Glucose	40 % (w/v)	10 ml
Amino acids mix I	4 mg/ml	10 ml
Amino acids mix II	4 mg/ml	10 ml
Vitamins	1 mg/ml	1 ml
SelenoMethionine	10 mg/ml	4 ml
Sterilised water	-	860 ml
Antibiotics (Kanamycin)	34 µg/ml	1 ml

\*20X M9 medium is made by dissolving 10g NH<sub>4</sub>Cl, 30g KH<sub>2</sub>PO<sub>4</sub>, 68g Na<sub>2</sub>HPO<sub>4</sub> in 500ml water. Amino acids mix I is a mixture of 15 non-essential amino acid except phenylalanine, tryptophan and tyrosine in water. Amino acid mix II is a mixture of phenylalanine, tryptophan and tyrosine in water. pH was adjusted to 8 by NaOH to improve solubility. Vitamins are consisting of riboflavin, niacinamide, pyridoxine monohydrochloride and thiamine. All components were filtered and sterilised.

The recombined YSBL3CLIC+ (87) plasmid containing the D6E large terminase gene was transformed into B834 cells. An overnight culture was started in LB medium with 34 µg/ml Kanamycin. The cell pellets were collected after centrifugation and resuspended using the medium described above for selenomethionine protein expression (Table 2-8). The cell mixture was then inoculated into 1 litre selenomethionine protein expression medium and cultured at 37 °C. 1 mM IPTG was induced once the OD600 is around 0.6 to 0.8. The cell culture was left at 37 °C for 12h before harvest.

## 2.5 SDS-PAGE

SDS-PAGE (sodium dodecyl sulphate polyacrylamide gel electrophoresis) is a method used to separate proteins by electrophoresis using a discontinuous polyacrylamide gel as a support medium and sodium dodecyl sulphate (SDS) to denature the proteins. In this work, SDS-PAGE was used as a routine technique in checking the size and purity of the target protein after expression and during purification. Typically, purified protein or cell lysate 15 µl was mixed with 5 µl gel loading buffer containing 40% Glycerol, 250 mM Tris-HCl pH 6.8, 8% SDS, 0.04% bromophenol blue and 5% beta-mercaptoethanol to a final volume of 20 µl and heated in a PCR thermal cycler for 5 mins at 98°C. The heated samples were separated by

polyacrylamide gel electrophoresis at a voltage of 200V. The gel running buffer consists of 25 mM Tris pH 8.3, 0.1% SDS and 192 mM Glycine. The gel was stained by Coomassie Brilliant Blue G-250 followed by destaining by water overnight.

## 2.6 Agarose gel electrophoresis

Agarose gel electrophoresis is a technique to separate a mixture of macro molecules, including DNA, RNA or protein. The negative charges of the phosphate backbone move the DNA towards the positively charged anode during electrophoresis. In this work, agarose gel electrophoresis was used as a standard method to analyse the DNA product obtained in PCR. It was also used in *in vitro* nuclease assays to quantify the nuclease activity of the wild type G20c large terminase protein and various mutant proteins, by separating the reaction mixture containing supercoiled, linear, nicked and fragmented DNA.

## 2.7 Expression and solubility tests

An assay based on SDS-PAGE was used for the evaluation of both protein expression level and solubility before proceeding to large scale expression. Pellets obtained from 1 ml cell cultures were resuspended into 120 µl of standard buffer containing 1 M NaCl and 50 mM Tris-HCl 8.0 or other buffers for protein solubility optimization. The cell lysate was obtained by sonication and 12 µl samples were taken before and after centrifugation, followed by loading onto a 12% polyacrylamide gel for electrophoresis separation. The sample taken before sonication represents the level of total target protein expression while solubility level under the chosen buffer can be quantified by the band intensity of the target protein in the supernatant sample.

## 2.8 Protein purification

Most approaches aimed at protein characterisation at the molecular level require the protein to be as pure as possible to avoid the interference of other contaminants from the expression system. Thus, it is important to obtain highly purified protein material after expression, by performing protein purification. Depending on the differences in size, physical-chemical properties or binding affinity between the target protein and host proteins/nucleic acids, the target proteins can be separated by several purification methods, such as nickel affinity chromatography, ion exchange chromatography and size exclusion chromatography. In the following section, a general protein purification protocol used to purify large terminases will be described based on the three-purification chromatography methods mentioned above.

### 2.8.1 Nickel affinity chromatography

Before purification, cell pellets were resuspended in buffer A (20 mM Tris pH 7.5, 1 M NaCl) containing 1 mM AEBSF, 0.5 µg/ml leupeptin, 0.7 µg/ml pepstatin and 0.1 mg/ml lysozyme and sonicated extensively on ice. The lysate was clarified by centrifugation at 19,000x g for 1 h and filtration using a 0.45 µm filter. Proteins were first purified by nickel affinity chromatography with a His-Trap column (GE Healthcare) equilibrated with buffer A containing 10 mM imidazole, and eluted with a 10 to 500 mM imidazole linear gradient in buffer A. All proteins that were purified by this method in this work have a cleavable N-terminal 6-histidine tag which chelates the nickel ion on the column and remains bound to it at low concentrations of imidazole. When increasing the concentrations of imidazole, the chelating of nickel ions by the 6-histidine tag is competed with- and weaken by- imidazole, leading to dissociation of the protein molecules from the column. Due to lack of the histidine tag, most *E. coli* cellular contaminants cannot bind to the nickel ions specifically in the column and were eluted as flow-through. After the imidazole gradient, the eluted target protein fractions were collected and dialyzed into 20 mM Tris pH 7.5, 250 mM NaCl, 0.5 mM DTT at 4°C overnight. During the dialysis, HRV 3C protease was added to the protein in a 1:50 (w/w) ratio to remove the N-terminal 6 histidine tag. To remove the HRV 3C protease which also contains a N-terminal 6 histidine tag from the target protein, samples after digestion were analysed by SDS-PAGE and applied to the His-Trap column as before. In this process, the flow-through which is mainly the target protein, was collected whereas the HRV 3C protease remains bound to the column.

### 2.8.2 Size exclusion chromatography

The size exclusion column is made of agarose gel beads with pores. The molecules that pass through the gel beads are separated by size (molecular weight). Larger molecules that are too big to enter the pores are eluted from the column quickly while the elution of smaller molecules is delayed by passage through the bead pores. In this work, size exclusion chromatography was used as a subsequent purification step after nickel affinity chromatography. The flow-through containing target protein obtained from the second nickel affinity chromatography was concentrated and applied to a Superdex 200 Hiload 16/60 column pre-equilibrated in 20 mM Tris-HCl, pH 7.5, and 250 mM NaCl (buffer B). The final protein samples were concentrated to 10-100 mg/ml.

### 2.8.3 Ion exchange chromatography

The stationary phase of protein ion exchange chromatography consists of an immobile matrix that contains charged ionisable functional groups or ligands. The mobile phase is usually buffer and salt. Initially, a low salt mobile phase is used to allow the proteins to bind to the column. These proteins have opposite surface charges to those on the column matrix

and undergo electrostatic interaction with the immobile matrix. By calculating the theoretical protein isoelectric point (pI) and altering the pH of the mobile phase, one can predict the surface charge properties of a certain protein molecule. This can be used in selecting the type of immobile matrix prior to the experiment. Once the target protein molecule is bound, a salt gradient can be applied to compete the immobilised protein from the column matrix. The protein is eluted from the column due to the weakened interaction by the counterions from the salt in the mobile phase.

The large terminase proteins used in this work were firstly purified by nickel affinity chromatography and size exclusion chromatography. To avoid the contamination of any DNA or RNA that may form strong interactions with the protein, ion exchange chromatography was used as an extra polishing method for D6E large terminase protein. This approach was used to produce ultrapure protein used for preparation of DNA-protein complexes for crystallisation.

## 2.9 Thermal shift assay

Thermal shift assay measures the fluorescence emission upon binding of a probe to an exposed hydrophobic region when protein is denatured by heat. It has been used in optimisation of conditions for stabilisation of proteins for structural studies (99,100) and high throughput studies of protein-ligand interactions (101,102). In the following section, a protocol was developed based on thermal shift assays to determine the optimal buffer and pH conditions for the stability of G20c, D6E and GBSV1 large terminase proteins under increased temperature. Furthermore, the experiment set-up of thermal shift assays to study the effects of ATP, ADP and non-hydrolysable ATP analogues on the protein stability of G20c and D6E large terminase was also described. All assays are performed using 5 µM purified protein and 5×SYPRO® Orange (diluted from 5000×stock) in the presence of various types buffer or ATP analogues in a 20 µl mixture. Prior to the measurement, samples were centrifugated at 568 × g for 1 min to remove precipitates and bubbles. Melting curves were measured in the temperature range of 25 °C to 95 °C at 1 °C /min on a real-time q-PCR machine. Tm values were obtained by fitting the melting curves with a sigmoid equation.

### 2.9.1 Buffer type and pH optimisation

These experiments were performed for the wild type full-length D6E and GBSV1 terminases and also for the D294N mutant of the G20c large terminase protein. For D6E and GBSV1 large terminases, all experiments were performed with a 250 mM NaCl but with 100 mM of different buffers types (with pH ranging from 5 to 10). The same set-up was used for the D294N mutant of the G20c large terminase, except that the NaCl concentration was increased to 500 mM and 1M.

### **2.9.2 Additives screen**

The experiment set-up for additives screen of the G20c D924N large terminase protein was the same as the buffer type and pH screen of D6E and GBSV1 terminases, except that NaCl concentration was increased to 500 mM.

### **2.9.3 Protein-ligand interaction**

The assays were performed using wild type full-length G20c and D6E large terminase proteins. For D6E large terminase, 5 mM ADP or non-hydrolysable ATP analogues, 5 mM MgCl<sub>2</sub>, 5 mM HEPES pH 7.5 and 5 mM potassium glutamate were added. As magnesium chloride was added in the assay buffer to facilitate nucleotide binding, its effect on protein stability was evaluated prior to the addition of ADP or non-hydrolysable ATP analogues. For G20c large terminase, the assays were performed with 10 mM ATP, ADP or non-hydrolysable ATP analogues, in a solution containing 10 mM MgCl<sub>2</sub>, 10 mM HEPES pH 7.5 and 10 mM NaCl.

## **2.10 Malachite green phosphate assay**

Malachite green, ammonium molybdate and phosphate form a complex under acidic conditions, which absorbs light at 660 nm. The amount of phosphate can then be determined by measuring the absorption of the coloured complex. A phosphate standard curve is required to determinate the linear relationship between light absorption by the complex and the amount of phosphate used (103). In this work, a general protocol that was used to measure the ATPase activity of the wild type D6E, phiOH2 and G20c large terminases is described below.

The colour reagent was prepared by mixing 0.045% (w/w) malachite green solution with 4.2% (w/w) ammonium molybdate-4M hydrochloride solution in a 1:3 (v/v) ratio. This colour reagent was then filtered and stored at room temperature before use. An assay buffer which contains 20 mM Tris pH 8.0 and 0.1 mg/ml BSA was used as a standard buffer for the phosphate standard and ATP hydrolysis reaction preparation. To prepare the standard curve, 2 nM, 4 nM, 8 nM, 12 nM and 16 nM potassium phosphate solutions in the assay buffer was used as phosphate standards solutions. To a 100 µl phosphate standard solution, 800 µl of colour reagent was added and incubated at room temperature for 2 mins. 100 µl 34% sodium citrate solution was then added as a neutralization reagent and incubated at room temperature for 20 mins. Absorption at 660 nm was measured using a UV-Visual light spectrometer, which was then recorded and plotted against the corresponding phosphate concentration to generate the phosphate standard curve. For the ATP hydrolysis reaction, 1 µM of protein, 1 mM ATP, 2 mM MgCl<sub>2</sub> and 20 µl 5×assay buffer was mixed in a 100 µl reaction mixture before incubation at 37°C for 30 mins. The reaction was stopped by the

addition of EDTA to a final concentration of 10 mM. The amount of phosphate release was measured using the same protocol as described above for the standard phosphate solutions.

## 2.11 *In vitro* nuclease assay

*In vitro* nuclease assays were developed to test the DNA cleavage properties of G20c, D6E and phiOH2 large terminase proteins and their mutant forms. Supercoiled DNA pacL26 (cloned by inserting the 100 bp phage SPP1 pacL site into pUC 18 vector (104)) was incubated with the purified G20c or D6E large terminase protein (1  $\mu$ M) in a 20  $\mu$ l reaction mixture containing 7 mM Tris pH 7.5, 7 mM potassium glutamate and 1 mM or 10 mM MnCl<sub>2</sub> at 37 °C for 30 mins. The reaction was stopped by the addition of EDTA (50 mM), SDS (0.5%) and proteinase K (50  $\mu$ g/ml) with a further incubation at 37°C for 30 mins. The resultant cleavage products were then separated on a 0.8% agarose gel followed by ethidium bromide staining. 1×TAE buffer was used as running buffer.

## 2.12 Analytical ultra-centrifugation

Analytical ultracentrifugation (AUC) is a powerful technique for the quantitative analysis of the molecular weight and shape of macromolecules in solution. For a given gravitational field, the mass of a macromolecule will redistribute when the gravitational potential energy exactly balances the chemical potential energy at each radial position. A sedimentation velocity experiment monitors the rate at which boundaries of molecules move during this redistribution, while a sedimentation equilibrium experiment determines the concentration distribution after equilibrium is reached (105). The basic theories of a sedimentation velocity experiment will be described in the following section, together with the experiment set-up in characterising the oligomeric state of the wild type D6E large terminase and the protein-DNA complex formed by the D326N mutant with DNA molecules of different lengths.

### 2.12.1 Sedimentation velocity theory

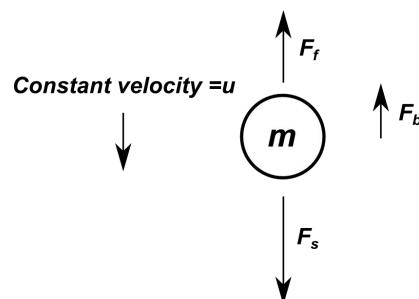


Figure 2-2 The forces acting on a solute particle in a gravitational field  
(Figure adapted from (106)).

In a sedimentation velocity experiment, three forces are applied to a macromolecule in a gravitational field: gravitational force ( $F_s$ ), buoyant force ( $F_b$ ) and frictional force ( $F_f$ )

(Figure 2-2).  $F_s$  is generated by the spin of the motor. It is proportional to molecular weight of the macromolecule and the rotor speed and the radius of the rotor and drives the sedimentation forward.  $F_b$  is equal to the weight of fluid displaced;  $F_f$  is proportional to the velocity when particles move through a viscous fluid.

Within a short time, the three forces come into balance:

$$F_s + F_b + F_f = 0$$

*Equation 2-1 The balance of the three forces in a gravitational field*

Based on the above equation, sedimentation coefficient ( $s$ ), the velocity of the particle per unit gravitational acceleration, can be obtained. It only depends on the properties of the particle and is independent of the operating conditions. Molecules with different molecular weights, different shapes or sizes, will move with different velocities in a given gravitational field and will have different sedimentation coefficients.

### 2.12.2 Protein oligomeric state analysis

Purified wild type D6E large terminase at a concentration of 4  $\mu\text{M}$  was extensively dialyzed into a buffer containing 10 mM HEPES pH 7.5 and 50 mM KOAc using a 3500 Da cut-off ultrafiltration dialysis device, prior to the experiment. The sample, together with the dialysis buffer which was used as the reference, were pipetted into sector-shaped cells with quartz windows and analysed by a Beckman Optima XL-A analytical ultracentrifuge. The sample was centrifuged at 35000 rpm at 20°C and radical scans at a wavelength of 280 nm were obtained continuously. Sedimentation profiles were analysed using Sedfit (107), using a partial specific volume of 0.74 ml/g and density and viscosity 1.0014 g/ml and 0.0102 Poise for the buffer were estimated for the potassium acetate buffer using Sednterp tabulated values for sodium acetate (108). The fractional ratio of the observed oligomers is fitted as 1.28. Fits were considered to be satisfactory if the rmsd was less than 0.008 and the residuals were random and less than 10% of the original signal.

### 2.12.3 Analysis of the Protein-DNA complex with Poly (T) ssDNA

PolyT<sub>20</sub> ssDNA which has either a 5'-6-fluorescein or 3'-fluorescein label was used in this experiment. Stocks of fluorescent DNAs and D6E large terminase D326N protein were dialyzed into a buffer containing 20mM Tris pH 7.5, 250mM NaCl and 1mM MgCl<sub>2</sub> using a 3500 Da cut-off dialysis device prior to the experiment, which was followed by sample preparation by diluting the DNA and protein stock using dialysis buffer into a final concentration of 5  $\mu\text{M}$  fluorescent ssDNA with either 10  $\mu\text{M}$  or 100  $\mu\text{M}$  protein. 5  $\mu\text{M}$  fluorescent ssDNA with no protein added were used as control samples. Data were collected and analysed as described in the above section, except that samples were centrifuged at

45000 rpm and a wavelength of 495 nm were used for radical scan. The partial specific volume of the protein-DNA complexes used for data fitting was estimated using the following formula:

$$vB_{complex} = \frac{vB_{protein} * M_{protein} * n1 + vB_{DNA} * M_{DNA} * n2}{M_{protein} * n1 + M_{DNA} * n2}$$

*Equation 2- 2 Estimation of protein-DNA complex partial specific volume*

*n1 and n2 represent the ratio of protein and DNA that form the protein-DNA complex. Partial specific volume values of 0.74 ml/g (108), 0.44 ml/g (Steve Weitzel, personal communication) (109) and 0.54 ml/g (110) were used for D6E large terminase Poly (T) ssDNA and dsDNA, respectively.*

#### 2.12.4 Protein dsDNA-complex analysis

In this experiment, 5'-6-fluorescein labelled dsDNA with lengths of 10, 20 or 30 bp were used instead of PolyT20 ssDNA. The experiment follows the same procedure as mentioned in the above section, except that 500 µM amounts of the D326N D6E large terminase protein were used. Estimation of partial specific volumes of the complex and data fitting were performed as described above.

### 2.13 Micro-scale thermophoresis

Micro-scale thermophoresis (MST) is a powerful technique used to quantify interactions between molecules at the cost of low sample consumption in free solution. It allows high throughput determination of binding affinity as well as analysis of the component stoichiometry (111). In the following section, a brief introduction on the basic theory of thermophoresis and MST signal analysis will be described. This is followed by the description of a protocol for MST experiment set-up in studying the D6E large terminase D326N interaction with DNA.

#### 2.13.1 Thermophoresis theory

MST is based on thermophoresis in which molecules migrate in an applied temperature gradient. This initial migration leads to a redistribution of molecules along the temperature gradient. For a given spatial temperature difference  $\Delta T$ , the concentration changes at steady-state can be described as thermophoresis depletion:

$$c_{hot}/c_{cold} = e^{(-S_T\Delta T)}$$

*Equation 2- 3 Thermophoresis depletion*

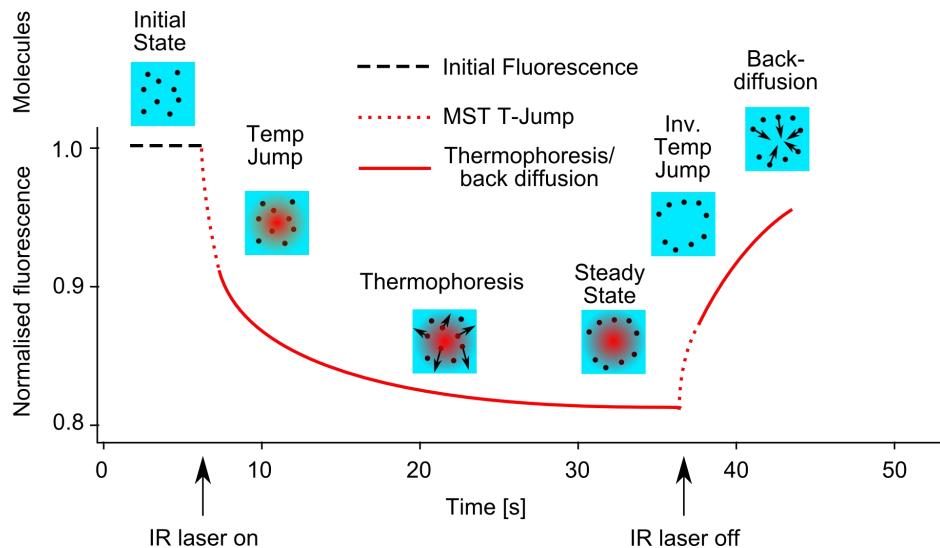
*c<sub>hot</sub>: molecule concentration in the hot area; c<sub>cold</sub>: molecule concentration in the cold area; S<sub>T</sub>: Soret coefficient.*

The Soret coefficient of a given molecule at a certain temperature is determined by its size, charge, hydration shell or conformation. Binding of partner molecules leads to changes of at

least one of these properties, from which a distinct thermophoresis depletion profile will be observed. Analysis of the binding induced changes in thermophoresis depletion can be used in quantifying binding affinity between biomolecules (112).

### 2.13.2 Thermophoresis signal analysis

In the MST experiment, samples prepared by serial dilutions are loaded into capillaries by capillary action. Before heating, the initial fluorescence is recorded. Due to pipetting errors or binding induced quenching of fluorescence from the fluorophore, the initial fluorescence may vary across capillaries. During heating, an infrared (IR) laser is focused onto the sample to exactly the spot where fluorescence intensity is measured on the capillary, generating a temperature gradient towards the solution around the spot. A dramatic change of fluorescence intensity will be first observed, called the *T-jump*. This is followed by a relatively slow thermophoresis phase which reaches plateau after several seconds and represents a steady state where thermal diffusion is counter balanced by mass diffusion. Upon turning off the heating, molecules start to migrate backwards which leads to a dramatic increase of the fluorescence intensity at the measurement spot, called *inversed T-jump*. Finally, the mass diffusion drives backward migration of molecules which compensates for the concentration gradient created by heating (Figure 2-3).



*Figure 2-3 A typical thermophoretic curve*

(Figure adapted from <http://lifeserv.bgu.ac.il/wb/azilha/pages/instruments/monolith-nt.php>).

The changes in thermophoresis at both bound and unbound state can be obtained as normalised fluorescence intensity (112):

$$F_{norm} = \frac{F_1}{F_0}$$

*Equation 2-4 Normalised fluorescence intensity*

$F_{norm}$ : normalized fluorescence;  $F_I$ : fluorescence after thermal diffusion;  $F_0$ : initial fluorescence or fluorescence after T-jump.

### 2.13.3 Determination of EC50 and binding cooperativity

To characterise the binding properties of the D6E large terminase D326N to single stranded poly(T) DNA at various length (10, 15, 20 and 50), 5'-FAM (6-Carboxyfluorescein) labelled DNA at 25 nM was titrated with protein at a wide concentration range. The binding buffer contained 10 mM HEPES, 200 mM NaCl and 1 mM MgCl<sub>2</sub>. After a serial dilution of the protein in the presence of fluorescent DNA at a constant concentration, the mixture was centrifuged at 11000x g for 5 min to remove precipitates that may affect thermophoresis. The mixture was then directly loaded into 16 Monolith NT.115 standard treated capillaries for analysis without further incubation. Data were collected on a NanoTemper instrument (115 series). The normalised fluorescence intensity data were plotted and fitted by Hill equation, from which EC50 and n number were obtained. Each measurement was repeated at least three times.

The competition experiment follows the same procedure as mentioned above, except that mixtures of 5'-FAM (6-Carboxyfluorescein) labelled poly(T)20 DNA at 25 nM and 2740 nM protein were titrated with increasing concentrations of unlabelled poly(T)20 DNA.

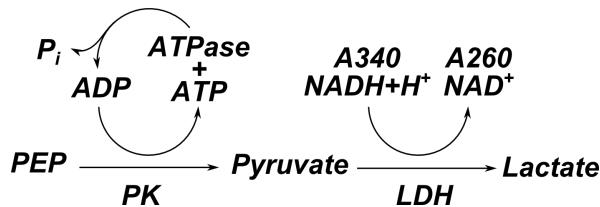
### 2.14 Electrophoretic motility shift assay

Electrophoretic mobility shift assay (EMSA) is a classic method used to characterise protein-nucleic acid complexes under native conditions (using non-denaturing gel electrophoresis in the absence of SDS). Due to changes in size, charge and shape upon protein-nucleic acid interaction, the resulting complex usually migrates slower than the nucleic acid but faster than the protein alone. To perform EMSA, protein and nucleic acids were mixed and subjected to a polyacrylamide or agarose gel under native conditions. After electrophoresis, the distribution of species containing nucleic acid is determined, usually by <sup>32</sup>P or fluorescent tag labelled nucleic acid.

To characterise the interaction of D6E large terminase D326N mutant with DNA, 25 nM poly(T) ssDNA at various length were used to mix with the mutant protein and load on to a native polyacrylamide gel. Both native and 5'-FAM labelled PolyT ssDNA were used. The polyacrylamide gel is made in TBE buffer (1 litre of 5X stock solution is made from 54 g of Tris base, 27.5 g of boric acid and 20 ml of 0.5 M EDTA pH 8.0). 1×TBE buffer was also used as running buffer. For gels with fluorescent DNA, gel images were collected on a Bio-Rad FX laser scanner (excitation wave-length 495 nm).

### 2.15 Coupled ATPase assay

Coupled ATPase assays use two well-studied enzymes including pyruvate kinase and lactate dehydrogenase, with the substrates Mg<sup>2+</sup>-ADP, phosphoenolpyruvic acid (PEP) and NADH. It allows the continuous measurement of the ATP hydrolysis rate of the protein of interest in real time. The theoretical basis of this assay is described briefly here. When ATP is hydrolysed to ADP and phosphate by the protein of interest, pyruvate kinase rapidly converts the ADP back to ATP and PEP to pyruvate. The lactate dehydrogenase can then reduce the pyruvate to lactate while oxidizing NADH to NAD<sup>+</sup>. In this way, for each ATP hydrolysed, one NADH is oxidized and the absorbance decreases. In this assay, the concentration of ATP is rapidly regenerated so that it remains constant throughout the life time of the reaction. The rate of absorption decrease of NADH can be traced at 340 nm which represents the hydrolysis rate of ATP by the protein of interest (Figure 2-4).



*Figure 2-4 Reactions involved in coupled ATPase assays*

(figure adapted from [https://labs.cellbio.duke.edu/Kinesin/Methods/ATPase\\_assay.html](https://labs.cellbio.duke.edu/Kinesin/Methods/ATPase_assay.html))

Coupled enzyme assays for the D6E large terminase and mutant proteins were performed at room temperature using: 6U/mL pyruvate kinase, 6U/mL lactate dehydrogenase, 1 mM phosphoenolpyruvate, 340 µM NADH, 10 mM HEPES pH 7.5, 50 mM potassium glutamate, 10 mM magnesium chloride, 0.5 mM ATP, and 4.5 µM TerL, unless otherwise noted. Absorbance was measured using sub-micro cell quartz cuvettes in a Cary 100 UV-visible spectrophotometer at a wave-length of 340 nm. In this assay, every NADH oxidized to NAD<sup>+</sup> corresponds to one ATP hydrolysed. To convert the measured absorbance directly to ATP concentration, we use the NADH extinction coefficient 6077 M<sup>-1</sup>cm<sup>-1</sup> as measure under our experimental conditions. The ATP hydrolysis rates obtained at various ATP concentrations were fitted by the Hill equation (18).

## 2.16 Calibration of Superdex 200 size exclusion column

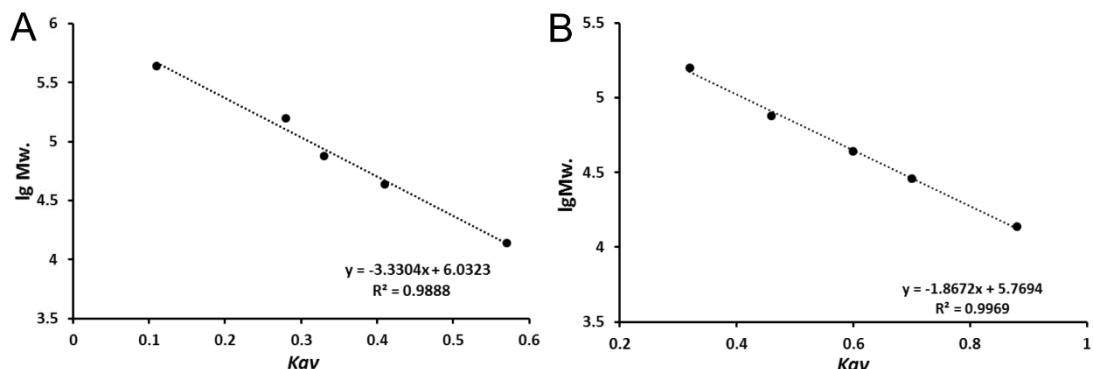
For estimation of the molecular weight of the full-length D6E large terminase protein and the ATPase domain, Superdex 200 increase 10/300 and Superdex 200 16/60 size exclusion columns were calibrated by running standard proteins. The standard proteins obtained from the calibration kit (GE healthcare) are prepared individually and mixed before loading onto the column and eluting with the desired buffer used for the full-length D6E large terminase or ATPase domain. The elution volumes ( $V_e$ ) were determined by running a mixture of Ferritin (0.3 mg/ml), Aldolase (4mg/ml), Conalbumin (3 mg/ml), Ovalbumin (4 mg/ml) and

Ribonuclease A (3 mg/ml) for a Superdex 200 10/300 increase or Aldolase (4mg/ml), Conalbumin (3 mg/ml), Ovalbumin (4 mg/ml), Carbonic Anhydrase (3 mg/ml) and Ribonuclease A (3 mg/ml) on Superdex 200 16/60 size exclusion column. The void volume  $V_o$  for the two columns are determined by running Blue Dextran 2000 (1 mg/ml) and eluting using the desired buffers. The elution volume for all the protein standards observed using the two columns are listed in Table 2-9. The void volume is determined as 8.03 ml and 43.24 ml, respectively for Superdex 200 10/300 increase and S200 16/60. The geometric volume was calculated by measuring the width and length of the packing inside the column. The geometric volume was determined as 24 ml and 114 ml, respectively for Superdex 200 10/300 increase and S200 16/60.

*Table 2-9 Calculated Kav and logarithm of molecular weight for selected protein standards*

S200 increase 10/300			S200 16/60		
Protein standards	Kav	lgMw.	Protein standards	Kav	lgMw.
Ferritin	0.11	5.6	Aldolase	0.32	5.2
Aldolase	0.28	5.2	Conalbumin	0.46	4.9
Conalbumin	0.33	4.9	Ovalbumin	0.6	4.6
Ovalbumin	0.41	4.6	Carbonic Anhydrase	0.7	4.5
Ribonuclease A	0.57	4.1	Ribonuclease A	0.88	4.1

The calibration curve is prepared by plotting  $Kav$  (ratio between the elution volume of a given molecule and the total available volume of the column, calculated by  $(V_e-V_o)/(V_c-V_o)$ ) and the logarithm of molecular weight for selected protein standards (Table 2-9, Figure 2-5).



*Figure 2-5 Calibration curve*

The plot of  $Kav$  versus logarithm of molecular weight of the standard proteins were shown for A. S200 increase 10/300 and B. S200 16/60. The functions explain the relation between  $Kav$  and logarithm of molecular weight obtained by linear fits.

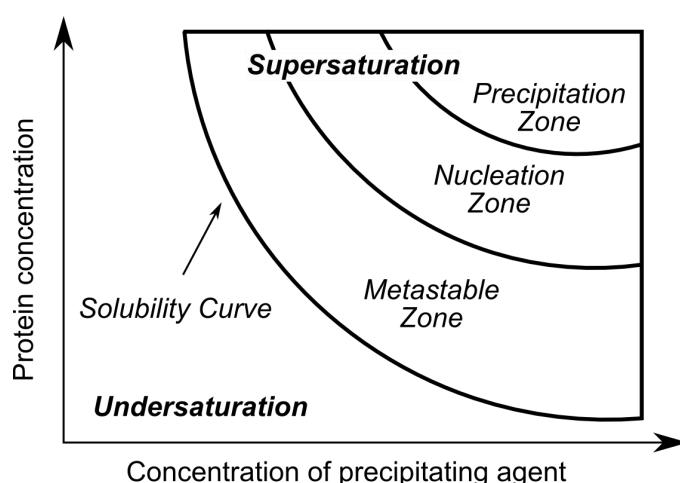
## 2.17 Protein Crystallisation

Like small molecules, proteins can be crystallised, for structure determination by X-ray crystallography. The protein molecule packing in crystalline arrays is usually driven by non-covalent interactions between protein molecules. Protein crystallisation can be performed by

vapor diffusion, micro batch (under oil), micro dialysis and several other techniques (113). In the following section, the principles of protein crystallisation and a general crystallisation protocol used in this work will be presented.

### 2.17.1 The crystallisation phase diagram

Crystallisation requires the macromolecule in solution to be brought to supersaturation. This can be achieved by mixing the protein at high concentration with precipitants. The addition of precipitants can also promote the nucleation of protein crystals in the solution and facilitates the formation of three-dimensional protein crystals. The process of crystallisation can be explained by a phase diagram of crystallisation (Figure 2-6). In this diagram, different areas correspond to different states: 1) Under saturation - Protein stays under saturated so that no precipitation occurs and crystals can be formed. 2) Meta stable zone - Protein reaches to super saturation, but no nucleation will occur unless the solution is mechanically shocked or a seed crystal introduced. 3) Nucleation zone - Protein nucleation occurs, crystals began to form and the concentration of protein in solution drops to saturation. 4) Precipitation zone - Protein precipitates, but crystals may still grow (114) (Figure 2-6).



*Figure 2- 6 Crystallisation phase diagram*

*Schematic representation of a two-dimensional phase diagram, illustrating the change of protein molecules concentration against precipitating agent concentration. The concentration space is divided by the solubility curve into two areas corresponding to under saturated and supersaturated state of a protein solution. The supersaturated area comprises of the metastable, nucleation and precipitation zones (figure adapted from [http://www.xray.bioc.cam.ac.uk/xray\\_resources/whitepapers/xtal-in-action/node3.html](http://www.xray.bioc.cam.ac.uk/xray_resources/whitepapers/xtal-in-action/node3.html)).*

In a vapor diffusion experiment, usually equal volumes of precipitant and protein are initially mixed in the drop. During the experiment, water will diffuse out of the drop and both the precipitant and protein concentrations will increase until equilibrium is achieved between the drop and the reservoir containing the precipitant solution. In a batch

crystallisation experiment, the precipitant and protein concentration do not change during the experiment.

## 2.18 Protocol of protein crystallisation using vapor diffusion

Initial crystallisation screens were conducted using commercial protein crystallisation suites, including PACT, CSS, Index, MPD, Hampton1/2, PEG Ion, ammonium sulfate and Salt Rx screens. 54 µl of well solutions from the commercial screens were either dispensed automatically using a Hydra robot (Art Robbins Instruments) or manually using multi-channel pipette into a 96-well MAXI MRC crystallisation plate. 150 nl of protein was mixed with 150 nl of crystallisation using a Mosquito robot (TTP LabTech). Crystals were tested using an in-house X-ray generator with rotating anode (Rigaku). Promising well-diffracting crystals were sent for data collection to synchrotron. Conditions where hits or small crystals were observed were recorded for future optimisation or seeding. Optimisations were performed by dispensing 100 µl of well solutions into a 48-well MRC crystallisation plate with an oryx 8 robot (Douglas Instruments). 0.5-1 µl of protein was mixed with equal volume of well solutions using the Mosquito robot, as described above. Plates were sealed and loaded into a Minstrel III plate-handling robot/hotel (Rigaku) for automated crystal imaging.

## 2.19 Seeding

For proteins that are hard to crystallise or from which only small crystals can be obtained, seeding is particularly helpful to facilitate crystallisation and reduce the number of nucleation events. Several different seeding techniques have proven successful in facilitating protein crystallisation, such as micro seeding and macro seeding (115). Micro seeding involves transferring sub microscopic seeds into new drops to control the nucleation. It has been reported that micro seeding can reduce crystal twining (116) and allows faster growth of protein crystals to avoid substrate degradation (117). Macro seeding instead transfers single crystals as seeds into a series of echo solutions to generate a fresh surface. It is particularly useful for attempts to increase the size of crystals.

Micro seeding was used for producing large crystals of the wild type D6E large terminase protein for phasing by multiple-wavelength anomalous dispersion (MAD). Seeds were prepared by crushing fresh crystals of the D6E large terminase protein with a seeding bead and vigorous mixing using a vortex. Different dilutions of the seeding stock were performed and added to the crystallisation to control nucleation. For crystallisation, 150 nl of protein solution was mixed with 100 nl of well solution and with 50 nl of seeding stock dilution, using the oryx 8 robot.

## **2.20 Post-crystallisation manipulation**

Determination of the three-dimensional structure of biological macromolecules by X-ray crystallography requires getting high-quality crystals. However, crystals are not always produced in the quality suitable for high resolution X-ray studies. Post-crystallisation manipulation can sometimes convert poorly diffracting crystals into diffraction quality crystals. Methods which may improve the crystal quality include annealing, dehydration, soaking and cross-linking (118,119). In this section, a novel post-crystallisation method was introduced by transferring the crystal into high concentration salt solutions. By screening different salt conditions at various pH, conditions that can consistently improve crystal diffraction were identified.

The crystals obtained from wild type D6E large terminase protein suffered from poor diffraction, of around 3.0 Å. The crystal structure was partially determined with one domain remaining disordered. These crystals were soaked in various salt solutions for up to 3 mins and flash frozen in liquid nitrogen. The diffraction images collected using the in-house X-ray generator were used to monitor diffraction quality.

## **2.21 X-ray data collection and processing**

X-ray diffraction images are collected by the rotation method (120). During data collection crystal is slowly rotated, while being exposed by an incident X-ray beam, to bring different sets of planes into diffracting position. Data processing could be divided into several steps (121): 1) Indexing: to work out cell dimensions and crystallographic symmetry. 2) Parameter refinement: it optimises the fit of calculated to observed spots on the image while varying a number of geometric parameters including the crystal to detector distance. 3) Integration: to sum up intensity from pixels of the spot and subtract intensity of the background (averaged inside a defined box, typically 2x2mm, centred at the spot, with exception of the area occupied by the spot) 4) Scaling and merging: due to reasons such as beam intensity change, cell volume changes induced by radiation damage, potential beam and wavelength drift, the intensity data measured are not on a common scale. Scaling puts the intensities measured for each image on a common scale, by minimising differences between intensities of equivalent reflections (related by symmetry) recorded on different frames during the data collection. Merging combines multiple measurements together. In this work, the X-ray diffraction data were collected remotely from cryo-cooled crystals using a Pilatus detector at Diamond Light Source beamlines I02, I03 and I04, at various wavelengths. The data were processed by XDS (X-ray Detector Software) (122), followed by merging symmetry related reflections using Aimless.

## **2.22 SAD, MAD phasing and molecular replacement**

During X-ray data collection, the intensities of the diffracted waves scattered from a series of Bragg planes in the crystal in all directions are recorded on the detector. Using these intensities, the amplitudes of the scattered waves can be derived. However, the phase information cannot be recorded, cause the so called ‘phase problem’. The phases can be derived by either using the atomic coordinates of a structurally similar protein (Molecular Replacement method) or by introducing a heavy atom or an “anomalous” atom which would absorb X-rays at certain wavelength which could be used for data collection (SAD, MAD and MIR). In some proteins “anomalous” atoms, such as metals (Zn, Cu, Fe), are naturally present, facilitating structure determination. In this thesis, SAD, MAD and molecular replacement were used as phasing methods (123). Crystallographic calculations were performed using the CCP4 suite of programs (124). Experimental phasing was performed using SHELXC/SHELXD (125,126). For MAD phasing, multiple data sets were scaled using XSCALE and merged using mts2sca, followed by phasing using SOLVE (127). Molecular replacement was performed by Phaser (128).

## **2.23 Density modification, model building, refinement and validation**

Density modification is a powerful method for improving the quality of electron density maps using improved phase estimates (129,130). Methods such as solvent flattening, NCS averaging and histogram matching are commonly used in density modification programs. These methods are often used in combination with automated model building and refinement, in particular in several automated protein structure building pipelines (130,131). In this work, density modification was conducted with either SHELXE (126) or RESOLVE (130,132,133), followed by automatic model building using ARP/wARP (134) or buccaneer (135). The resulting model was inspected by COOT (136), where iterative model building and refinement cycles were performed to remove errors from the automatically built model and to generate a more complete atomic model. Refinement was performed using REFMAC5 (137), followed by manual structure inspection and rebuilding. The refined atomic models were validated using the wwPDB validation server (138) and deposited with the Protein Data Bank (139).

### **3 Isolation and characterisation of large terminases from several thermophiles**

In this chapter, large terminase proteins from several thermophilic bacteriophages, namely G20c, D6E, GBSV1, phiOH2, GVE2 and 1-1, were cloned, purified and characterised for further structural and biochemical studies.

#### **3.1 Structure based sequence alignment**

The protein sequences of the G20c, D6E, GBSV1, phiOH2, GVE2 and 1-1 large terminases were aligned with the Sf6 large terminase using the structure based protein sequence analysis tool (140) (Appendix 3). The alignment revealed similar secondary structures among all the large terminases analysed and identified conserved key functional motifs in the ATPase domains such as the Walker A, Walker B and phosphate sensor motifs. In addition, several conserved carboxylates were also observed in the nuclease domain. Despite the predicted similarities in secondary structure, large differences in amino acid sequence were observed (average sequence identities (141) are between 11-25%).

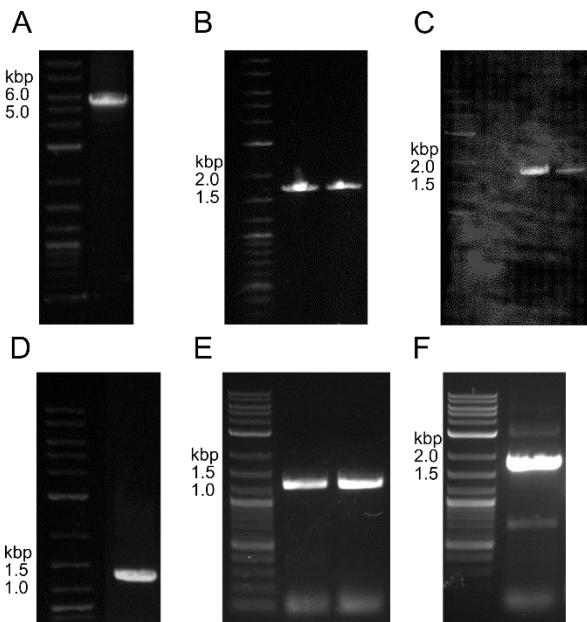
#### **3.2 Protein disorder prediction**

Successful structure determination of proteins by X-ray crystallography requires the production of protein crystals in which protein molecules are orderly packed in the unit cell. However, crystallisation of proteins containing conformationally flexible or intrinsically disordered regions can be problematic, as such regions may prevent crystal packing. To predict the disordered regions of several large terminase proteins, their protein sequences were submitted to a protein disorder prediction server IUPred (142). A plot of disorder tendency versus residue number were obtained for all large terminase proteins. According to this prediction, D6E, GBSV1, GVE2 and 1-1 are predicted to be more ordered than the G20c and phiOH2 large terminases (Appendix 4).

#### **3.3 Cloning, expression and solubility tests**

The expression vectors (YSBL3CLIC+ (87) and SUMO tagged) and the target large terminase genes including D6E, GBSV1, phiOH2, GVE2 and 1-1 (Appendix 5) were amplified by PCR and combined by infusion to produce the expression constructs (Figure 3-1) as described in Chapter 2. Final constructs were confirmed by Sanger sequencing (143). Initial crystallisation trials of the wild type and ATPase Walker B mutants of G20c large terminase resulted in non-diffracting crystals and degraded protein during purification and crystallisation (Maria Chechick and Dr. Elena Blagova). To solve this problem, a new construct with a single mutation at the predicted nuclease domain active site (D294N) was

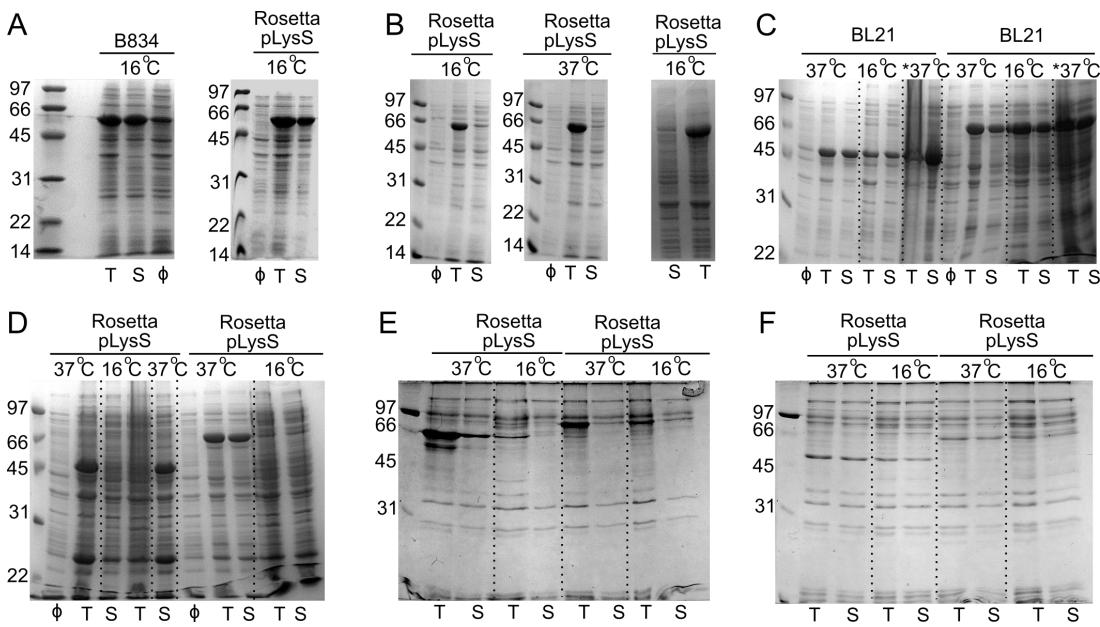
designed and cloned (Dr. Sandra Greive). The sequenced recombinant plasmids of the G20c D294N mutant, along with the D6E, GBSV1, phiOH2, GVE2 and 1-1 were transformed and tested for expression and solubility prior to large scale protein expression.



*Figure 3-1 PCR amplification*

*Agarose gel electrophoresis showing the amplified PCR products of (A) YSBL3CLIC<sup>+</sup> vector, (B) GVE2, (C) GBSV1, (D) D6E, (E) phiOH2 and (F) 1-1 large terminase gene.*

These tests showed that all the tested large terminase constructs were successfully expressed. G20c D294N large terminase was expressed at milligram levels and was soluble when expressed at 16 °C using either B834 or Rosetta pLysS *E. coli* strains. The expression levels for wild type D6E and GBSV1 was much lower at 16 °C than at 37 °C when using Rosetta pLysS, while similar expression levels at both temperatures were observed using the BL21 strain. The phiOH2 large terminase was also soluble with a modest expression level under the tested experimental conditions. GVE2 and 1-1 were expressed but the proteins were not very soluble. No apparent improvement of solubility was observed for these two proteins when expressed with an N-terminal SUMO tag (Figure 3-2).



*Figure 3-2 Expression and solubility tests of large terminase proteins by SDS-PAGE*

The expression temperature, bacterial strains are indicated at the top of each gel.  $\Phi$ : before IPTG induction, T: total cell lysate, S: Supernatant after cell lysis. The expression level of large terminase protein is shown for A. G20c (YSBL3CLIC+), B. left and middle: GVE2 (YSBL3CLIC+); right: GVE2 (Champion SUMO); C. and D. left: D6E (YSBL3CLIC+); right: GBSV1 (YSBL3CLIC+), E. left: I-1 (YSBL3CLIC+); right: I-1 (Champion SUMO); F. left: phiOH2 (YSBL3CLIC+); right: phiOH2 (Champion SUMO)

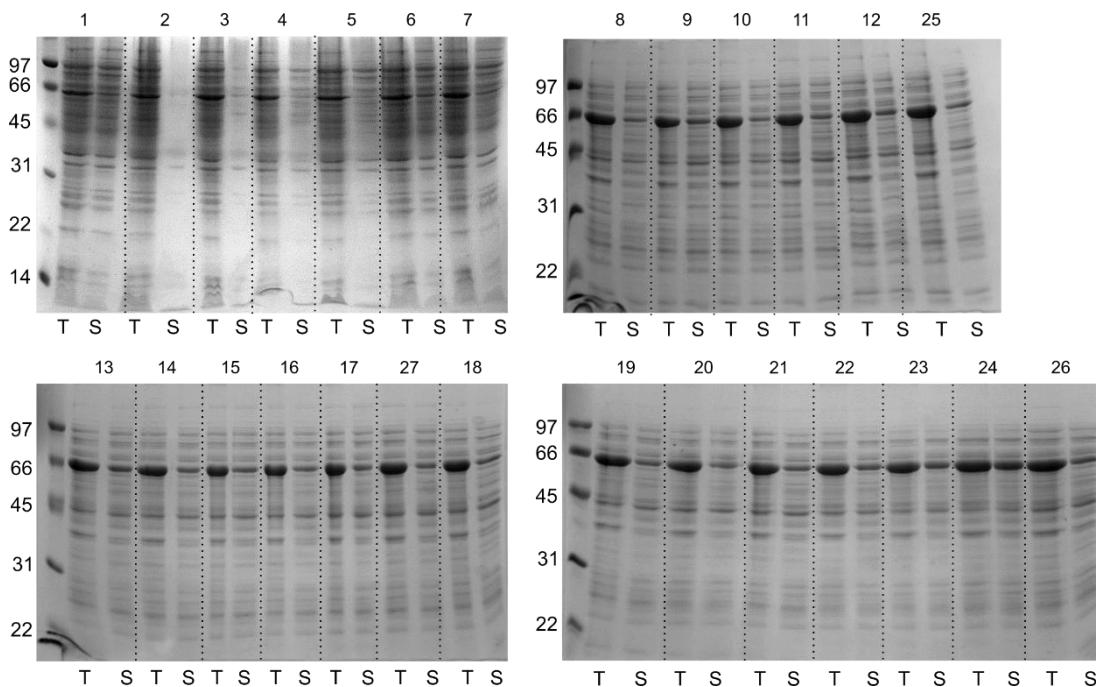
### 3.4 Solubility screen

While the GVE2 protein was expressed, the poor solubility limited the amount of protein that could be obtained from purification. To overcome this limitation for future structural and biochemical studies, solubility screens were used to search for alternative conditions that could solubilise this protein. All conditions used in the solubility screen are listed in Table 3-1. Table 3-1 Buffer, salt concentration and pH used in initial solubility screening

*Table 3-1 Buffer, salt concentration and pH used in initial solubility screening*

Condition	Buffer (100 mM)	Buffer pH	NaCl Conc.
1	Water	7	1 M
2	sodium citrate	4	1 M
3	sodium acetate	4.5	1 M
4	sodium citrate	5	1 M
5	sodium citrate	5.5	1 M
6	bis-tris	6	1 M
7	sodium phosphate	6	1 M
8	MES	6.2	1 M
9	bis-tris propane	6.5	1 M
10	ADA	6.5	1 M
11	MES	6.7	1 M

<b>12</b>	PIPES	6.7	1 M
<b>13</b>	sodium phosphate	7	1 M
<b>14</b>	MOPS	7	1 M
<b>15</b>	HEPES	7	1 M
<b>16</b>	HEPES	7.5	1 M
<b>17</b>	Tris-HCl	7.5	1 M
<b>18</b>	HEPES	8	1 M
<b>19</b>	Tris-HCl	8	1 M
<b>20</b>	Bicine	8	1 M
<b>21</b>	Tris-HCl	8.5	1 M
<b>22</b>	CHES	9	1 M
<b>23</b>	CHES	9.5	1 M
<b>24</b>	CHES	10	1 M
<b>25</b>	sodium phosphate	6.5	1 M
<b>26</b>	sodium phosphate	8.0	1 M
<b>27</b>	sodium phosphate	7.5	1 M



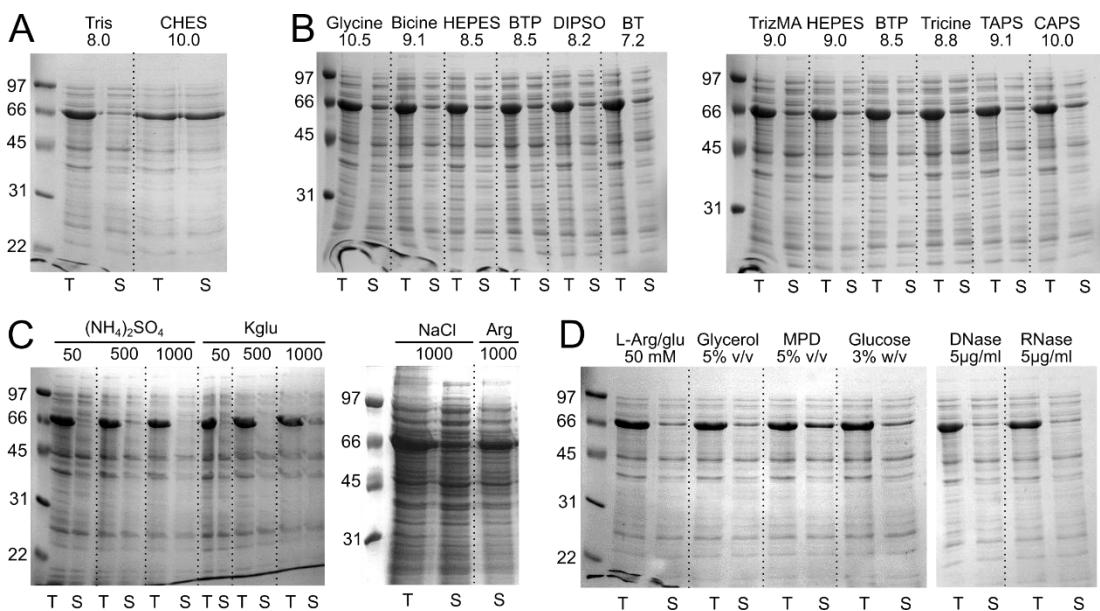
*Figure 3-3 Initial solubility screen of GVE2 large terminase protein*

The numbers at the top of each lane correspond to condition numbers used in Table 3-1. T: total cell lysate, S: Supernatant after cell lysis.

The initial solubility screen revealed that the GVE2 large terminase protein had poor solubility in most of the conditions tested. An improvement in the solubility level of this protein was only observed under conditions containing 100 mM CHES pH 10.0 and 1M NaCl (Condition 24; Table 3-1; Figure 3-3). When the concentration of CHES buffer at pH

10.0 was reduced to 50 mM, the level of solubility remained similar to that at 100 mM CHES, in the presence of 1 mM DTT (Figure 3-4).

Results of the initial screen indicated that the high pH was important to solubilise the protein, as an increase of pH by 0.5-1 unit improved the solubility (Table 3-1; Figure 3-3, 22, 23 and 24). In the subsequent screening step, buffers effective at higher pH were tested at a concentration of 50 mM (Figure 3-4). However, no significant improvement in solubility was observed. In these experiments, the effect of alternative salts (to NaCl) and additives was also probed (Figure 3-4). It was observed that in the presence of 1M arginine hydrochloride pH 8.0, the amount of soluble protein in the total lysate was increased while in the presence of other additives, no improvement in solubility was observed (Figure 3-4).



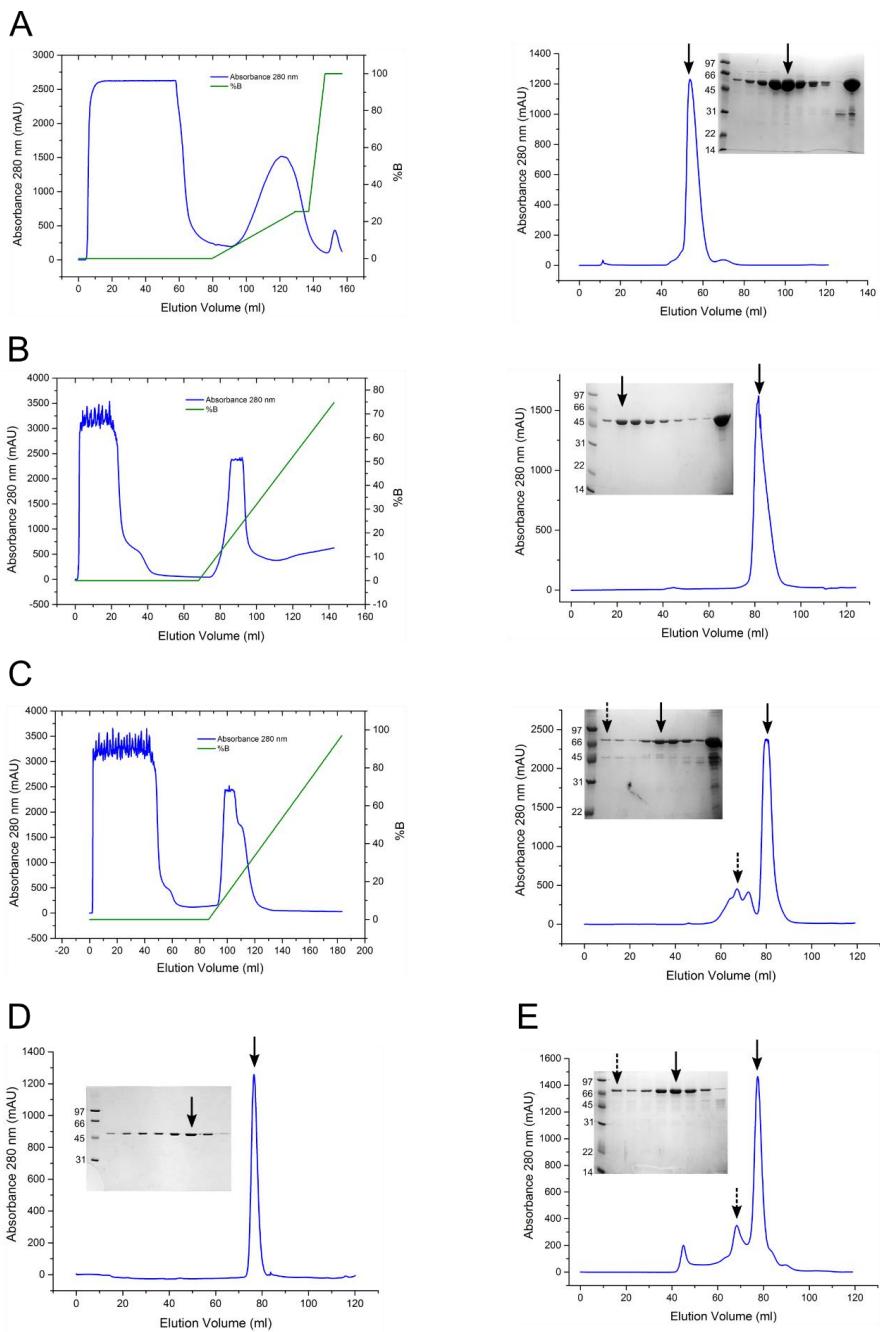
*Figure 3-4 Secondary solubility screen of the GVE2 large terminase using SDS-PAGE*

(A) Comparison of the solubility between Tris pH 8.0 and CHES pH 10.0 at 50 mM concentration. 1 M NaCl and 1 mM DTT were used. (B) Solubility screen using buffers at high pH. 1 M NaCl was included. (C) Effects of salt on solubility. Tris pH 8.0 was used as the buffer. Kglu: potassium glutamate, Arg: Arginine hydrochloride. (D) Effects of additives on solubility in the presence of Tris pH 8.0 and 1 M NaCl.

### 3.5 Purification

From the results of the expression tests and solubility screens, G20c, D6E, GBSV1, GVE2 and phiOH2 large terminases are expressed as soluble proteins at modest to high levels. Large scale cell culture was performed using the cell strains used in the expression tests. The cell pellets obtained from G20c, D6E, GBSV1 and phiOH2 large terminase expressions were lysed and purified using the general protocol described in the material and methods chapter, whereas a different purification method was used for the GVE2 protein (Figure 3-5). It has been observed that the GVE2 protein has excellent solubility in 1 M Arginine

hydrochloride pH 8.0 as mentioned above (Figure 3-4). However, 1 M Arginine hydrochloride is not compatible for nickel affinity chromatography. Combining the previously observed improved solubility in the presence of CHES buffer, a protocol was developed for effective purification of this protein: 1) 1 M Arginine hydrochloride pH 8.0 was used with 1 M NaCl as the lysis buffer to solubilize this protein. 2) the supernatant was pooled into a dialysis bag with a 50 kDa molecular weight cut off and dialysed against 50 mM Tris-HCl pH 8.0, 250 mM NaCl to remove the Arginine hydrochloride. In this process, the amount of contaminating proteins from *E. coli*. in the dialysis bag with molecular weights below 50 kDa should be reduced. As the solubility of the GVE2 protein is highly dependent on 1 M Arginine hydrochloride, this protein will precipitate during dialysis. 3) Precipitate was collected by centrifugation from the mixture in the dialysis bag and then resuspended in 50 mM CHES pH 10.0, 250 mM NaCl. 4) The solubilised protein was concentrated and loaded onto the S200 Size Exclusion Chromatography column, equilibrated with 50 mM CHES pH 10.0, 250 mM NaCl.



*Figure 3-5 Purification of large terminases by nickel affinity and size exclusion chromatography*

The chromatography curves were monitored at 280 nm wavelength and the purity of the proteins after purification are shown for several large terminases including (A) G20c D294N, (B) D6E, (C) GBSV1, (D) phiOH2 and (E) GVE2. The protein bands in SDS-PAGE corresponding to the peak of the size exclusion chromatography are indicated by solid arrows (monomer) and dashed arrows (oligomer). A superdex 200 size exclusion column was used for the purification of D6E, GBSV1, phiOH2 and GVE2 large terminases while the G20c D294N protein was purified on a Superdex 75 size exclusion column.

### 3.6 Protein thermal stability

Large terminase proteins from thermophiles are expected to have increased thermal stability compared to their mesophilic counterparts (20°C to 45°C). To investigate their thermal

stability, thermal shift assays were conducted for the G20c D294N, D6E and GBSV1 large terminases. In this experiment, the effect, on protein thermal stability, of buffer, pH and additives were also investigated (144).

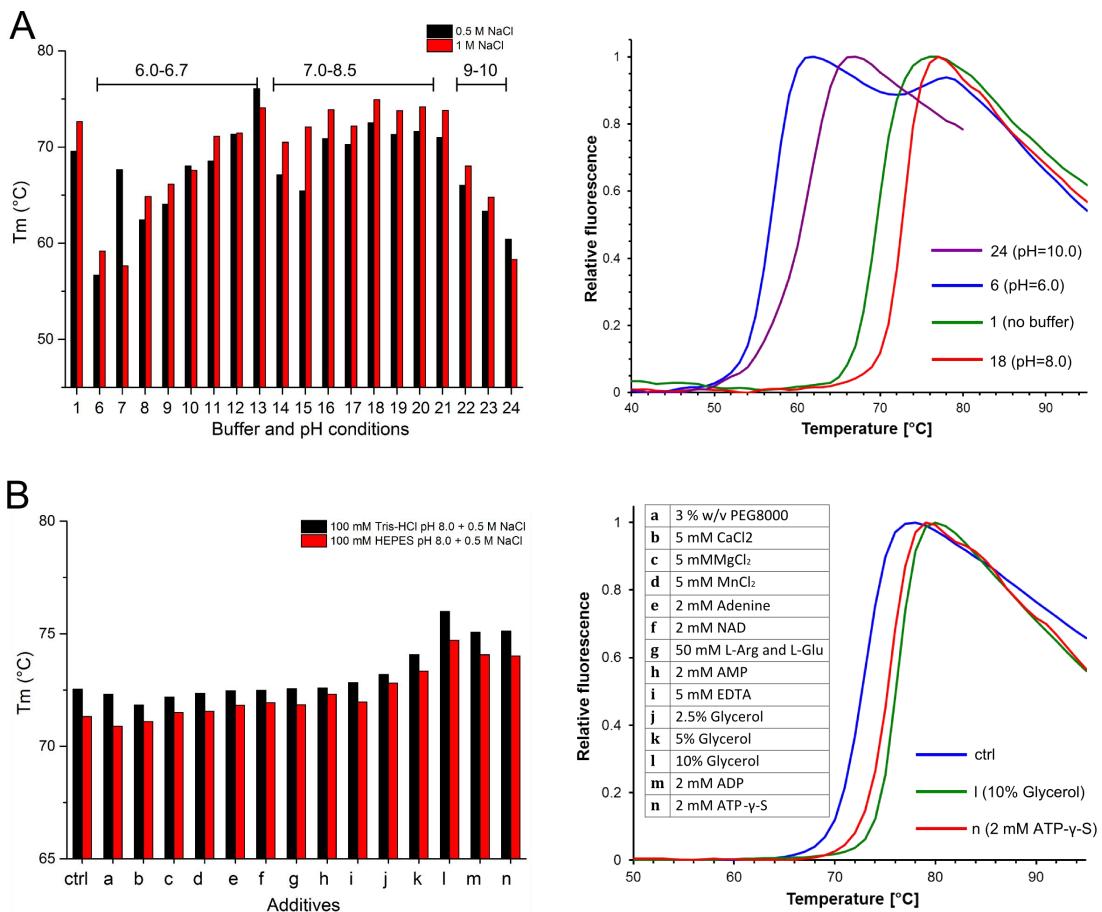


Figure 3-6 Thermal shift assays for G20c large terminase D294N

(A) buffer and pH screen. left: plot of melting temperatures under various buffer and pH conditions (Table 3- 1). pH range of buffers is indicated on the top, right: melting curves observed under different conditions are displayed as normalised fluorescent intensities. (B) additive screen. left: Plot of melting temperatures in the presence of various additives, right: Additives used in the assay and their concentrations.

Various buffer and pH conditions from Table 3- 1 that were used for solubility screens of the GVE2 large terminase were screened for improved stability of the G20c large terminase D294N protein by thermal shift assays. Two different NaCl concentrations were used in this experiment. It was observed that when increasing the NaCl concentration from 0.5 M to 1 M, the melting temperature increased by 1-5 °C under most of the conditions (Figure 3-6). When conditions 2-5 were used (pH 4-5.5), the melting curves lost the expected sigmoidal shape (data not shown). It is likely that under low pH conditions the protein is significantly destabilised even before the temperature gradient is applied. At pH 6.0 to 6.5, the curve exhibits a double melting behaviour (Figure 3-6). This may be caused by the unfolding of one domain prior to the other under low pH conditions. At pH 7.0-8.5, the highest melting

temperatures of the protein were obtained. Further increases in pH had a negative effect on protein stability, as a significant drop in the melting temperature was observed at this pH range. In the additive screen, three buffer conditions that were among the optimised pH range (7-8.5) were selected. Among all the additives tested, ADP, ATP- $\gamma$ -S and glycerol gave the highest Tm values compared to the rest (Figure 3-6). These conditions were expected to stabilise the protein and likely to facilitate successful crystallisation.

Buffer and pH screens were also performed for D6E and GBSV1 large terminase proteins. As observed in the previous characterisation of G20c large terminase D294N, increased salt concentration has a substantial impact on the protein stability. This suggests that to detect the small changes in melting temperature confidently, under various buffer and pH conditions, the salt concentration should be kept to the minimum required for maintaining protein solubility at the concentration used in the assays. In the subsequent experiment, the concentration of NaCl was reduced to 250 mM where no protein precipitation was observed for either protein. It was found that the melting temperature of the GBSV1 (68 °C) was around 4 °C higher compared to that of the D6E (64 °C) large terminase under the same reference conditions, indicating differences in protein thermostability (Figure 3-7). The melting temperature of the D6E large terminase protein exhibited an apparent increase at a wide pH range between 6 to 9. In contrast, the optimal pH range for the GBSV1 large terminase was determined to be 6 to 7.5, narrower than that observed for the D6E large terminase (Figure 3-7). At pH 5 and 10, a decrease of the melting temperatures for both proteins was observed. This decrease was more significant for GBSV1 than that for the D6E large terminase, consistent with the difference observed in pH range that stabilises the protein.

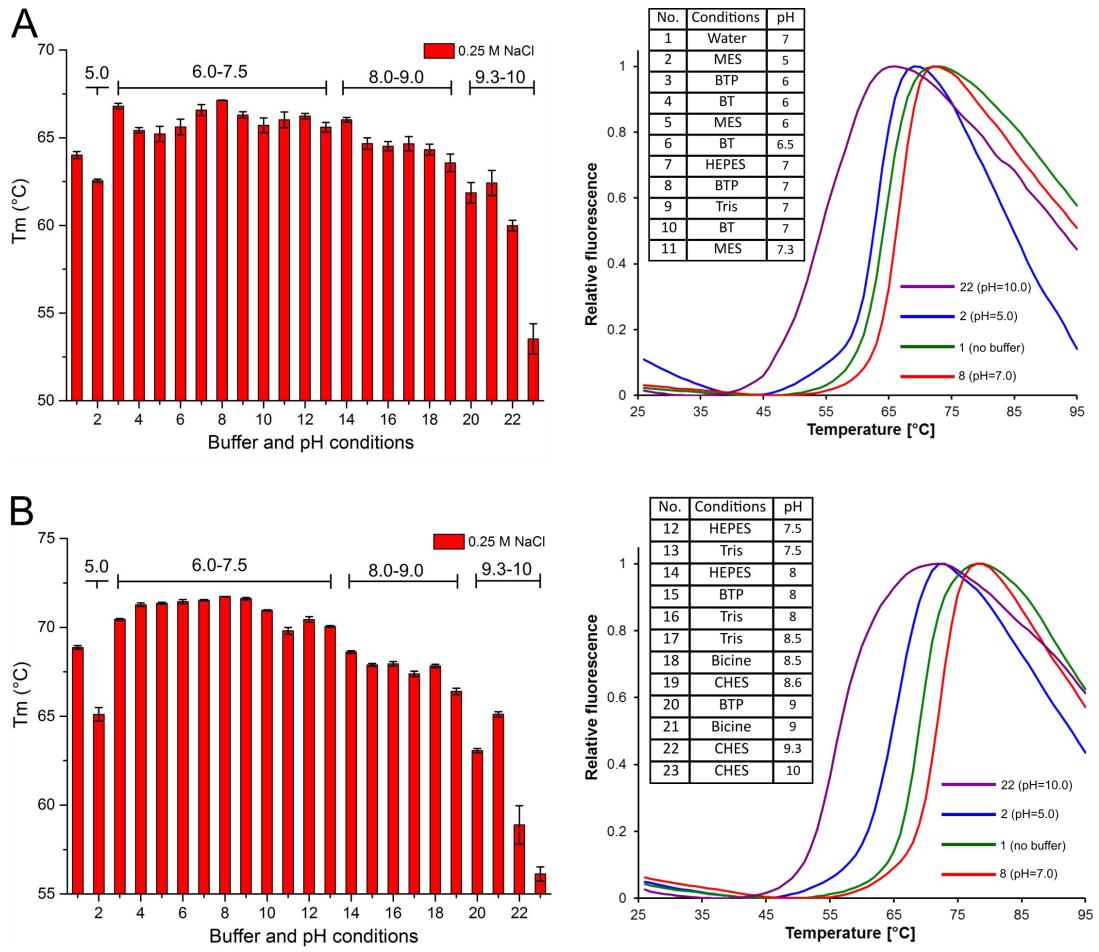


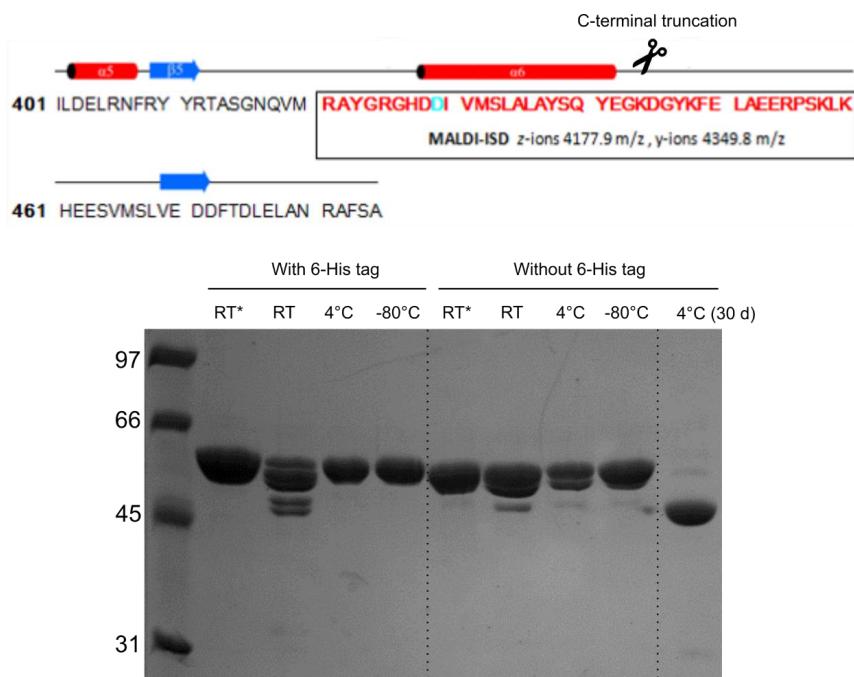
Figure 3-7 Thermal shift assays for D6E and GBSV1 large terminase

(A) buffer and pH screen for D6E large terminase. (B) buffer and pH screen for D6E large terminase. left: plot of melting temperatures under various buffer and pH conditions. pH range of buffers is indicated at the top, right: melting curves observed under different conditions are displayed at normalised fluorescent intensities. The buffer and pH conditions screened in this experiment are listed.

### 3.7 C-terminal truncation of G20c large terminase

Similar to the wild type full length G20c large terminase (Maria Chechick data), degradation was also observed for the D294N G20c large terminase. This degradation occurred during purification and proceeded at both 4 °C and room temperature (Figure 3-8). The degraded samples were sent for MALDI-MSD analysis that identified a fragment from the C-terminal region in the protein sample (Figure 3-8). Previous secondary structure analysis and protein disorder prediction suggested that the region downstream of this C-terminal fragment is likely to be unstructured and disordered (443-485). To facilitate crystallisation and improve protein stability, a C-terminal truncation of the wild type G20c large terminase was made ( $\Delta 443$ ). This truncation should not affect the nuclease active site, as the predicted residues important for nuclease catalytic activity are not present in this C-terminal segment. Previously reported crystallisation work for large terminase from bacteriophage T4 indicated

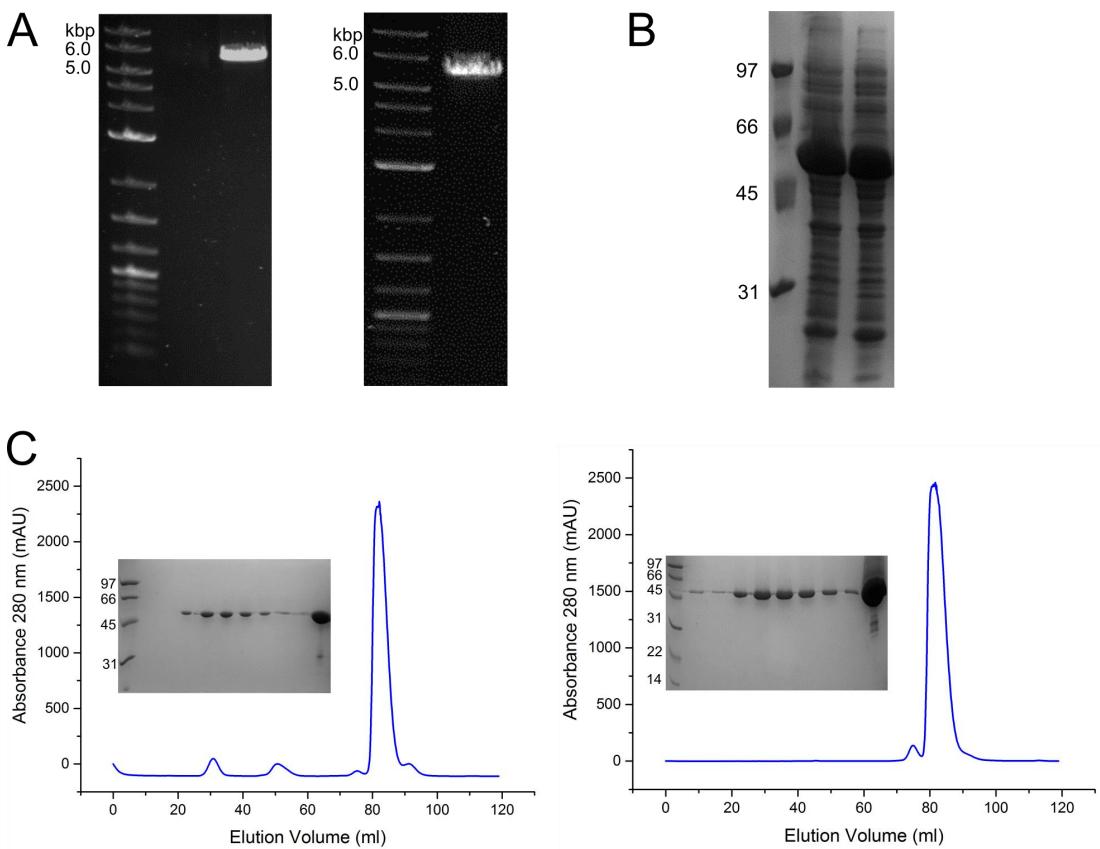
that an ATPase Walker B mutation facilitates successful crystallisation of this protein (40). Given the functional similarity of these two proteins, an equivalent mutation (D148E/E149D) for the C-terminal truncated G20c large terminase protein ( $\Delta$ 443) was also constructed. Since it was found that protease inhibitors reduced this degradation, a low concentration of AEBSF at 0.1 mM was therefore added to all buffers used in purification.



*Figure 3-8 Secondary structure prediction and SDS-PAGE*

*Top:* The sequence and predicted secondary structure of the C-terminus of G20c large terminase. The sequence of the peptide found in MALDI-MSD is highlighted in red. Aspartate residue predicted to be important for nuclease activity is highlighted in cyan. The position of the C-terminal truncation is indicated. *Bottom:* degradation of the G20c large terminase protein over time examined at different temperatures. \*: Samples that have been treated with protease inhibitors prior to incubation.

The C-terminally truncated wild type G20c large terminase  $\Delta$ 443 construct was cloned by replacing the codon for K444 with a stop codon in the wild type full length G20c large terminase construct cloned in YSBL3CLIC+ vector (87). The  $\Delta$ 443 D148E/E149D construct was made using the same mutagenesis method based on D148E/E149D full length G20c large terminase gene cloned in YSBL3CLIC+ vector. The PCR amplified recombinant vectors containing the desired truncation and mutation were analysed by agarose electrophoresis and are shown in Figure 3-9. The two constructs were expressed as soluble proteins in *E. coli*. (Figure 3-9). The proteins were purified using protocols described in Chapter 2, except that AEBSF was added to prevent protease degradation. No degradation bands were observed during purification (Figure 3-9).

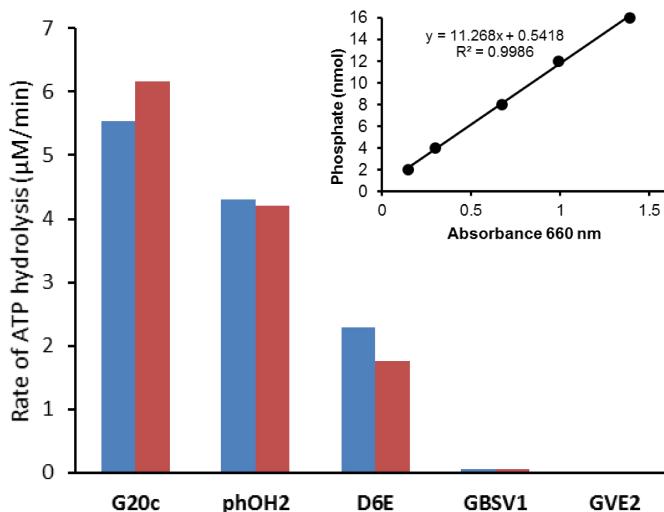


*Figure 3-9 Cloning, expression and purification of the C-terminally truncated G20c large terminase*

(A) the PCR amplified recombinant plasmid containing the G20c large terminases and a stop codon at K444, (B) The expression test of the G20c large terminase Δ443. (C) Size exclusion profiles of the G20c large terminase WT Δ443 (left) and D148E/E149D mutant Δ443 (right). The purity of proteins is demonstrated by SDS-PAGE.

### 3.8 ATPase activity

Earlier sections describe purification of the D6E, GBSV1, G20c, GVE2 and phiOH2 large terminase proteins. To check ATPase activity, each purified protein was subjected to a malachite green phosphate assay, performed using the protocol (103) described in Chapter 2. The amounts of phosphate released by the proteins over time were plotted and compared in Figure 3-10. It was observed that G20c Δ443, D6E and phiOH2 large terminases were competent in ATP hydrolysis under the experimental conditions. However, in the case of GVE2 and GBSV1 large terminase, no activity was measured using this assay. It is possible that the ATP hydrolysis by these two terminases was not triggered under the experimental conditions tested, and that these proteins would be active at alternative conditions. It is also possible that the ATPase activity of GVE2 and GBSV1 requires the participation of other motor components, such as the portal protein, small terminase and/or other DNA binding partners.



*Figure 3-10 ATPase activities measured by malachite green phosphate assays*

The rate of phosphate release by several large terminase proteins, including G20c Δ443, phiOH2, D6E, GBSV1 and GVE2 is shown by coloured bars. Two independent measurements were performed for each of the five large terminase proteins. The phosphate standard curve is displayed above the bar chart.

### 3.9 Nuclease activity

In addition to the ATPase activity assays, the *in vitro* nuclease activity was tested for the G20c (see Chapter 4), D6E, GVE2 and phiOH2 large terminases. The degradation products of the supercoiled plasmid in the presence of protein and metal ions at various concentrations were analysed by agarose gel electrophoresis. The assays were performed following the protocol described in Chapter 2. As observed for other members of the RNase H endonuclease superfamily, the nuclease activity requires the participation of divalent metal ions (74).

The effect of divalent metal ions on the nuclease activity of phiOH2 large terminase is shown in Figure 3-11. The nuclease is activated by manganese, zinc, nickel, copper and calcium while no nuclease activity was observed upon addition of magnesium and cobalt under the tested conditions. In the presence of manganese, zinc, nickel, copper and calcium, the protein digested the DNA substrate into shorter DNA fragments as indicated by smearable bands on the gel. To compare the rate of DNA digestion in the presence of different divalent metal ions, the reaction was stopped at various time points from 5 to 50 min, using zinc and manganese as co-factors. In the presence of manganese, the protein degraded all the DNA substrate within 5 min while when zinc was used, the rate of DNA digestion was much slower. These results suggest that the phiOH2 large terminase protein can be activated by several divalent metal ions. The nuclease reaction rate was faster in the presence of manganese than when zinc is added.

The effects of divalent metal ions on the nuclease activity of D6E large terminase was also investigated (Figure 3-11). It was found that the nuclease activity was supported by manganese, magnesium, zinc, cobalt and calcium. Among these divalent metal ions, manganese and cobalt strongly activated the protein, producing smeared DNA fragments while the activity under magnesium, zinc and calcium were modest. When copper was used at the same concentration, no nuclease activity as observed under the tested conditions. The nuclease activity was also investigated for GVE2 large terminase protein in the presence of magnesium as a co-factor. In this experiment, a protein concentration dependent increase of nuclease activity was observed.

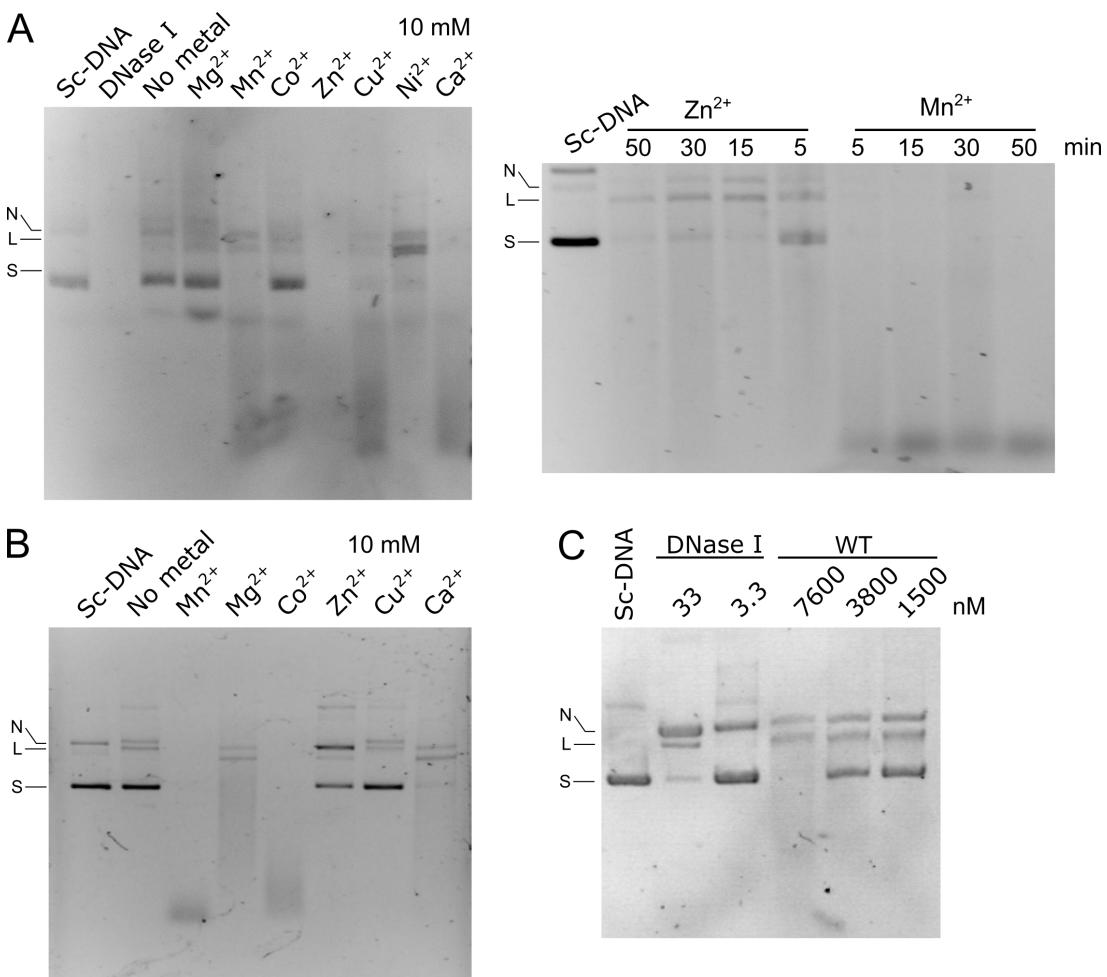


Figure 3-11 *In vitro* nuclease activity assays

(A) Nuclease activity of the phiOH2 large terminase. Left - comparisons of *in vitro* nuclease activities in the presence of various metal ions. Right - comparisons of *in vitro* nuclease activities in the presence of Zn<sup>2+</sup> and Mn<sup>2+</sup> over time. (B) Effects of various divalent metal ions (10 mM) on the nuclease activity of D6E large terminase. (C) Effect of protein concentration on the nuclease activity of GVE2 large terminase. DNase I was used as a positive control. Sc-DNA: for this condition, only supercoiled DNA was used as a control. No metal: equivalent volume of water was added instead of divalent metal ions. N: nicked, L: linear.

### 3.10 Crystallisation

Crystallisation experiments were performed for G20c D294N, G20c  $\Delta$ 443, GVE2, GBSV1, D6E and phiOH2 large terminase proteins. Initial screening of crystallisation conditions using sitting drop vapor diffusion at room temperature only yielded non-diffracting crystals for G20c D294N, GBSV1 and phiOH2. No hits were observed for the GVE2 large terminase, whereas diffracting crystals were found for D6E and G20c  $\Delta$ 443.

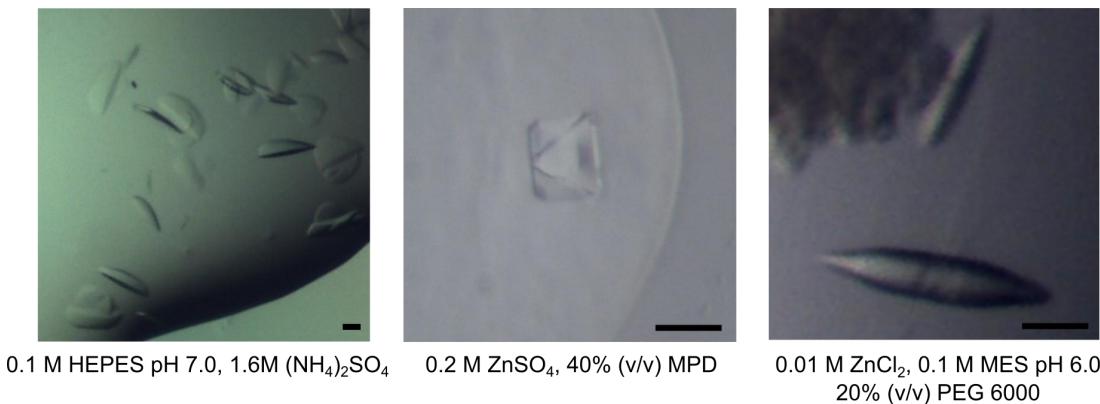


Figure 3-12 Images of protein crystals

(left): D6E; middle: G20c D294N; (right): G20c  $\Delta$ 443. Scale bars (50  $\mu$ m) are indicated on each image.

The crystals of D6E large terminase appeared around 3 days after setting up the crystallisation tray. The size of the crystals reached the maximum within around 10 days. Crystals found in the ammonium sulphate Screen grew with reservoir conditions 0.1 M HEPES pH 7.0, 1.6 M ammonium sulfate and 0.1 M HEPES pH 7.5, 1.6 M ammonium sulfate, 2% PEG1000. These crystals were cryo-protected by addition of 25% glycerol to the corresponding buffer and salt present in the reservoir, before vitrification and X-ray diffraction tests using in-house X-ray facilities. Most of the crystals tested only diffracted to 7-8 Å while the best crystals diffracted to 4-5 Å and were sent to the Synchrotron for data collection. The best data set was collected at the synchrotron (Table 3-2) to a resolution of ~3.2 Å. Due to the lack of phase information, the structure could not be solved. For successful structure determination, optimisation was needed to improve the crystal size and quality, as well as produce a heavy atom derivative or a selenomethionine labelled protein. The structure determination of the D6E large terminase will be described in detail in Chapter 6.

*Table 3- 2 X-ray data collection statistics (D6E large terminase)*

Wavelength (Å)	0.9795
Space group	R32
Unit-cell a, b, c (Å)	104, 104, 268
Unit-cell $\alpha$ , $\beta$ , $\gamma$ (°)	90.0, 90.0, 120.0
Resolution(Å)	46.06-3.21 (3.43-3.21)
Rmerge (%)	10.1 (110.8)
I/ $\sigma$ <I>	14.4(2.2)
Completeness (%)	99.9(99.6)
Multiplicity	9.9(9.9)
CC <sub>1/2</sub>	0.999(0.786)

Values in the parentheses are for the outermost resolution shell.

Crystals of the Δ443 G20c large terminase were found over 30 days after setting up crystallisation using the PACT Screen (Figure 3-12). These crystals grew in 0.01 M ZnCl<sub>2</sub>, 0.1 M MES pH 6.0 and 20% PEG 6000 and were flash frozen directly from the drop, without adding any cryo-protectant. The best crystal diffracted to ~2 Å in-house while a 1.6 Å Zn-SAD data-set was collected using the same crystal at the Synchrotron. Calculation of the Matthew's coefficient and the specific volume indicated that only a fragment of the large terminase protein could fit into the asymmetric unit of this crystal. The structure determination of this fragment is described in detail in Chapter 5.

### 3.11 Summary

This chapter describes the cloning, expression, purification as well as preliminary characterisation of the thermal stabilities and enzyme function of large terminases from several bacteriophages, including G20c D294N, D6E, GBSV1, phiOH2, GVE2 and 1-1. The G20c D294N, D6E, GBSV1 and phiOH2 large terminases were expressed and purified in soluble form while the GVE2 and 1-1 had limited solubility. Based on the results from the solubility screen, a protocol has been developed for solubilisation and purification of the GVE2 large terminase. The optimal pH range and conditions that can stabilise the proteins were found for the G20c D294N, D6E and GBSV1 large terminase proteins. ATPase assays showed that only G20c D294N, D6E and phiOH2 large terminases are active ATPases, while no apparent activity was detected for GVE2 and GBSV1 large terminases. In addition, characterisation of the nuclease activities of the D6E, GVE2 and phiOH2 large terminases showed that their nuclease activity is dependent on metal co-factors. Due to a proteolytic degradation, a C-terminally truncated construct of the G20c large terminase was designed and produced. Initial crystallisation screens using the purified and functionally characterised proteins yielded diffracting crystals for the D6E and G20c large terminases. These crystals were used for subsequent structure determination, paving the way towards understanding the structure-function relationships.

## 4 Functional characterisation of G20c large terminase protein

As described in Chapter 3 above, the G20c large terminase Δ443 was expressed, purified, and characterised by the malachite green ATPase assay. This chapter describes further characterisation of this large terminase, including the nuclease activity and ATP binding properties.

### 4.1 Non-hydrolysable ATP analogues

As non-hydrolysable ATP analogues, ATP- $\gamma$ -S and AMPPNP are chemically modified at oxygen atoms in distinct positions of the ATP molecule. ATP- $\gamma$ -S is modified by replacing a non-bridging oxygen on the  $\gamma$ -phosphate using a sulfur atom, while the modification of AMPPNP and AMPPCP are on the bridging oxygen between the  $\beta$  and  $\gamma$  phosphates. This bridging oxygen is replaced by a nitrogen or carbon, respectively for AMPPNP or AMPPCP (Figure 4-1). When ATP- $\gamma$ -S is bound to the ATPase active site, the sulfur atom from the  $\gamma$  phosphate may disturb the hydrogen bond formed between the non-bridging phosphate of ATP with the P-loop Walker A residues (145). The introduction of a sulfur atom to the  $\gamma$  phosphate is also likely to alter the charge distribution of the  $\gamma$ -S phosphate which may destabilise the formation of the catalytic transition state. In the cases of AMPPNP and AMPPCP, the modification on the bridging oxygen would affect the P-O-P bond angles and conformations between the  $\beta$  and  $\gamma$  phosphates. This is likely to negatively affect their interaction with P-loop Walker A residues and therefore inhibit catalysis. It is also believed that during phosphoryl transfer reactions, protonation of the leaving group should happen during bound breakage (146-148). Nitrogen and carbon atoms are less electron negative than oxygen which makes them less likely to be protonated as a leaving group.

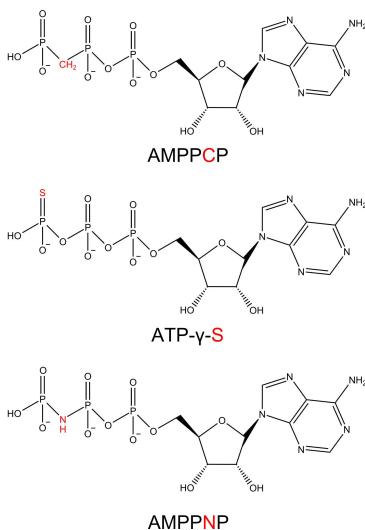


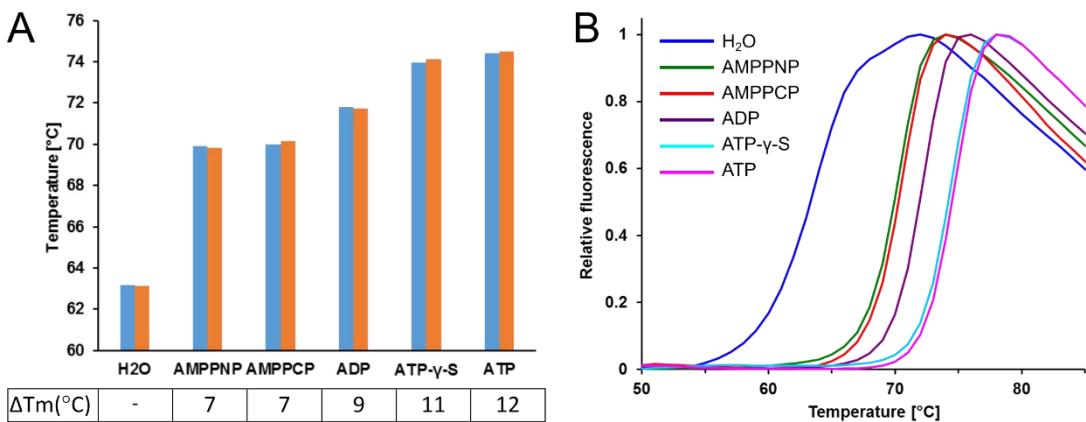
Figure 4-1 Chemical structure of AMPPCP, ATP- $\gamma$ -S and AMPPNP

## 4.2 Effects of ATP analogues on protein stability

As characterised by malachite green ATPase assays in Chapter 3, the G20c large terminase  $\Delta 443$  can turn over ATP, suggesting that the ATPase domain is functional. Thermal shift assays (144) were then performed to investigate the effects of ATP, ADP and non-hydrolysable ATP analogues on the thermal stability of the protein. The aim of this experiment is to determine the melting temperature of G20c large terminase  $\Delta 443$  in the presence and absence of various ATP analogues. By comparing the differences in melting temperature, the stabilisation effect of various ATP analogues on the protein can be quantified. The identified ATP analogues that stabilise the protein can be further used to facilitate protein crystallisation.

This experiment uses the general set-up and data analysis method described in Chapter 2. In the conditions containing ATP analogues, a concentration of 10 mM ATP analogue and 10 mM magnesium chloride were used. In the control experiment with no ATP analogue, an equal volume of de-ionised water was used instead of ATP analogue and magnesium chloride. The melting temperature (two repeats) were plotted and compared in Figure 4-2. Overall, the melting temperatures were significantly increased when adding ATP, ADP or non-hydrolysable ATP analogues in the presence of  $Mg^{2+}$ . Specifically, compared to the control condition, addition of ATP or ATP- $\gamma$ -S give the largest melting temperature increase among all the ATP analogues tested (12 °C and 11 °C, respectively). In the presence of ADP, a modest increase of 9 °C was observed. The stabilisation effects by AMPPNP and AMPPCP are the smallest among all ATP analogues, of ~7 °C (Figure 4-2).

As mentioned above, it was observed that ATP and ATP- $\gamma$ -S have a significant stabilisation effect on the thermal stability of the protein while the effect of other non-hydrolysable analogues such as AMPPNP, AMPPCP and ADP are less prominent (Figure 4-2). It is therefore possible that the conformation and/or nature of the  $\beta/\gamma$  phosphate moiety plays a role in binding. This would explain the observed difference in binding affinity reflected by changes in protein thermal stability upon the addition of various non-hydrolysable ATP analogues.



*Figure 4-2 The stabilisation of ATP, ADP and non-hydrolysable ATP analogues on the thermal stability of G20c large terminase Δ 443 are characterised using thermal shift assays*

(A) Bar chart showing the average melting temperatures with and without the addition of ATP analogues (two repeats). (B) Representative melting curve plotted using normalised fluorescent intensity versus temperature.

### 4.3 Nuclease activity

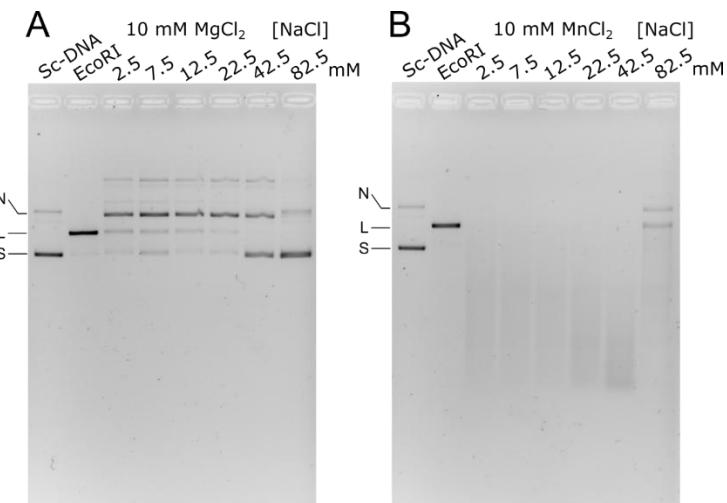
In this section, the nuclease activity of the G20c large terminase Δ443 was tested by *in vitro* nuclease assays. This assay allows a comprehensive and fast characterisation of the nuclease activity under various conditions. The aims of these experiments were to investigate the factors that affect the nuclease activity.

#### 4.3.1 Effects of sodium chloride concentration

The effects of sodium chloride concentration on nuclease activity was characterised by titrating the reactions with sodium chloride at an increasing concentration gradient. As the protein belongs to the RNase H endonuclease superfamily, the nuclease activity is expected to require participation of divalent metal ions, such as magnesium and manganese. In this experiment, 1 mM magnesium or manganese were provided as co-factors for nuclease activity. In both cases, the increased sodium chloride concentrations inhibited the nuclease activity. The inhibitions were observed at sodium chloride concentrations of 42.5 mM and 82.5 mM for the reactions activated by 1 mM magnesium and manganese, respectively (Figure 4-3).

Salt concentration has been reported to be important for DNA binding (149) and catalytic activities (150). As suggested for T4 (31) and SPP1 (32) large terminases, DNA is cleaved in a non-sequence specific manner in headful DNA packaging. It is possible the reduction in the nuclease activity at high sodium chloride concentration for the G20c large terminase is caused by a decrease in protein-DNA interaction. As an RNase H endonuclease superfamily protein (73), the DNA cleavage by the large terminase protein requires the participation of

divalent metal ions. It is also possible that at high concentrations of sodium chloride, the sodium ion can inhibit the binding of active metal co-factors, such as magnesium and manganese, as has been observed (151). This suggestion is also supported by the similar van der waals (VDW) radii of sodium, magnesium and manganese ions. It was also observed that sodium can be accommodated into the active site of enzymes which use magnesium and manganese as co-factors for catalysis (152).



*Figure 4- 3 The effect of sodium chloride concentration on nuclease activity*

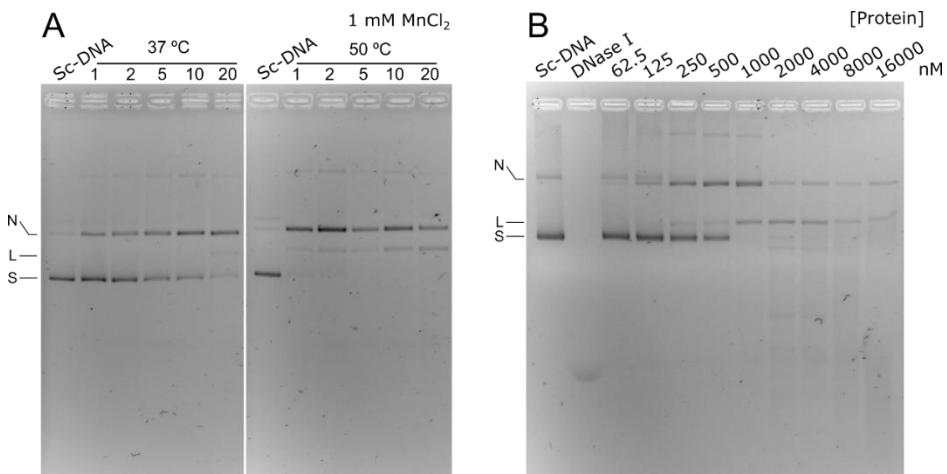
The assays were performed under (A) 10 mM magnesium chloride and (B) 10 mM manganese chloride. N: Nicked, L: linearised, S: Supercoiled. The linearised DNA control was obtained by treating the supercoiled DNA substrate with EcoRI. N: nicked, L: linearised, S: supercoiled.

#### 4.3.2 Effects of temperature and protein concentration

The effects of temperature on nuclease activity were characterised in the presence of 1 mM magnesium chloride. The digested DNA products over time was analysed using agarose gel electrophoresis. The results show an apparent increase of the nuclease activity when the temperature is increased from 37 °C to 50 °C (Figure 4-4). The close relatives of bacteriophage G20c, including P23-45 and P74-26 (19), have an optimum living temperature of around 65 °C. This experiment suggests that as a large terminase protein from a thermophilic bacteriophage, the protein remains stable and can cleave DNA faster at increased temperatures which are closer to the physiological temperature than 37 °C.

In the following experiment, the effect of protein concentration on nuclease activity was investigated by titrating the nuclease reactions with increasing protein concentrations. The results show that the nuclease activity undergoes a protein concentration dependent increase. At 62.5 nM protein, no apparent nuclease activity was observed. Nicked product was first observed when the protein concentration was increased to 125 nM. Further increase of protein concentration to 500 nM produces a mixture of nicked and linearised DNA products. When

the protein concentration is above 500 nM, full conversion of the supercoiled DNA to cleaved DNA products (nicked, linearised and small DNA fragments) was observed (Figure 4-4). These data suggest that at 1 mM manganese chloride concentration, the minimum protein concentration needed to digest all supercoiled DNA substrates is above 500 nM. This observation is important for estimating the nuclease activity of the G20c large terminase and designing subsequent experiments.



*Figure 4-4 The effects of temperature and protein concentration on nuclease activity*

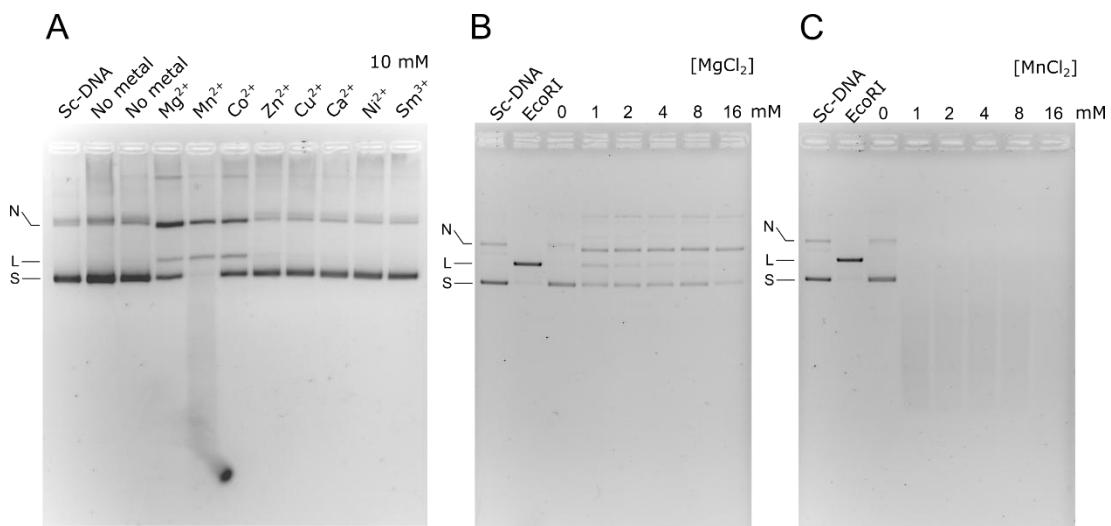
(A) Comparison of nuclease activity at  $37^{\circ}\text{C}$  and  $50^{\circ}\text{C}$  in the presence of  $1\text{ mM manganese chloride}$ . (B) Effect of protein concentration on nuclease activity. N: nicked, L: linearised, S: supercoiled.

#### 4.3.3 Effects of divalent metal ions and metal concentrations

To investigate the effects of divalent metal ions on nuclease activity, various metal ions including manganese, magnesium, cobalt, zinc, calcium, copper and nickel were used at a concentration of  $10\text{ mM}$ . Samarium, a trivalent metal ion, was also tested in parallel with other divalent metal ions for its effect on nuclease activity. It was observed that manganese, magnesium, and cobalt activated the nuclease activity, whereas no apparent activity is observed in the presence of the other tested metal ions (Figure 4-5).

Among the active co-factors, the effects of metal ion concentrations on nuclease activity were further investigated for magnesium and manganese. Metal ions were added to the reaction at concentrations ranging between  $1$  and  $16\text{ mM}$ . When magnesium was used at a  $1\text{ mM}$  concentration, the reaction produced nicked and linearised DNA products. Further increases in metal ion concentration to  $16\text{ mM}$  leads to a mild increase of the nuclease activity, which is reflected by the observed disappearance of linearised DNA products. In the presence of manganese, the supercoiled DNA substrates were fully digested into shorter DNA fragments even at  $1\text{ mM}$  metal ion concentration, generating smears on the gel (Figure 4-5).

It has been reported that for some RNase H proteins of the RNase H-like endonuclease superfamily, a significant decrease of the nuclease activity was observed when the concentration of magnesium was increased to 10 mM, an effect that is termed ‘attenuation’ (153). This is likely to be caused by the occupation of a magnesium ion at a third metal ion binding site next to the active site, which would perturb the coordinating geometry of the catalytic metal ions and thereby affecting catalysis. It has also been proposed that the decrease in activity is possibly because of the deficiency of product release due to inhibition of metal ion dissociation at high magnesium concentrations after catalysis. In the case of the G20c large terminase protein, no apparent attenuation was observed at similar metal ion concentrations (Figure 4-5), suggesting some differences in the metal binding site and catalytic mechanism. As observed in the structures of the nuclease domain of the SPP1 large terminase where only two metal ions were located in the absence of DNA, it is possible that the third binding site does not exist and therefore no attenuation can be found for large terminase protein at similar metal ion concentrations.



*Figure 4-5 The effects of the types of metal ions and metal ion concentrations on nuclease activity*

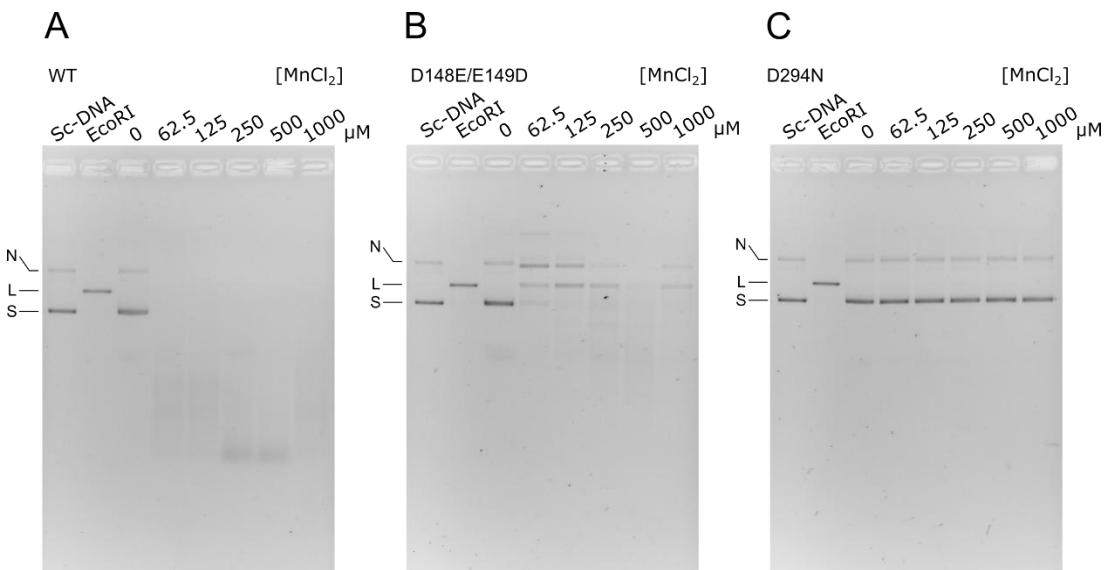
(A) The nuclease activity measured using various metal ions at 10 mM concentration. (B) and (C) The effect of magnesium and manganese concentration on nuclease activity. N: nicked, L: linearised, S: supercoiled.

#### 4.3.4 Effects of ATPase Walker B mutation

To investigate the role of the ATPase domain in DNA cleavage, the G20c  $\Delta$ 443 large terminase Walker B (ATPase activity deficient) mutant was used in the nuclease assays. The nuclease domain active site D294N mutant was also used as a negative control. The nuclease activities at various manganese and magnesium ion concentrations determined for these two mutants were compared with the wild type protein (Figure 4-6). The results show that compared to the wild type  $\Delta$ 443 protein, the Walker B mutant has a decreased nuclease

activity under experimental conditions. The wild type  $\Delta 443$  protein digested all DNA substrate into small DNA fragments even at 62.5  $\mu\text{M}$  manganese concentration whereas for the Walker B mutant, nicked and linearised DNA products were retained when 1000  $\mu\text{M}$  manganese concentration was used. When the nuclease active site mutation was used, the nuclease activity is completely abolished (Figure 4-6).

The data presented here suggest that D294N mutation in the active site of the nuclease domain has lethal effects on the nuclease activity. This is likely to be caused by a perturbation of catalytic metal ion binding to the active site. Compared to the wild type protein, the Walker B mutant has reduced nuclease activity at the same concentration range of manganese chloride used. As reported for the T4 (66) and P74-26 (56) large terminase proteins, the ATPase domain is the primary DNA binding region of the large terminase. Mutation in the ATPase domain is likely to affect DNA binding during DNA cleavage, which can cause a reduction in nuclease activity.



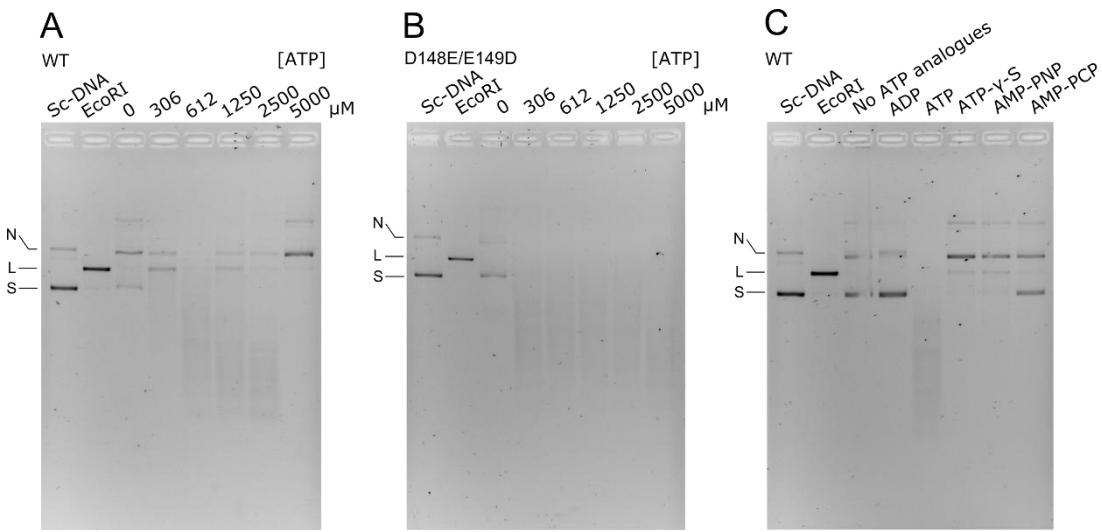
*Figure 4-6 The effects of ATPase active site mutation on nuclease activity*

*The nuclease activity under various manganese concentrations are shown using (A) wild type  $\Delta 443$ , (B) D148E/E149D  $\Delta 443$  mutant and (C) D294N mutant. N: nicked, L: linearised, S: supercoiled.*

#### 4.3.5 Effects of ATP, ADP and non-hydrolysable ATP analogues

The effects of ATP, ADP and non-hydrolysable ATP analogues on nuclease activity were characterised using wild type  $\Delta 443$  and ATPase Walker B mutant in the presence of 5 mM magnesium chloride. In this experiment, the results show that addition of ATP stimulates the nuclease activity of both the wild type and ATPase Walker B mutant under various ATP concentrations (Figure 4-7). In the cases of wild type protein, an inhibition of the nuclease activity of the nuclease domain was observed at higher magnesium concentrations while similar inhibition was not seen for the ATPase Walker B mutant. A similar inhibitory effect

of ATP on the nuclease activity at high concentration has also been reported for the T4 large terminase (78). The effects of ADP and non-hydrolysable ATP analogues were also tested on wild type  $\Delta$ 443 protein. It was observed that addition of ADP does not affect the nuclease activity. Addition of ATP- $\gamma$ -S, AMPPNP and AMPPCP leads to an increase in the nuclease activity, producing nicked and linearised DNA fragments. Compared to the strong stimulatory effect of ATP, the enhancement of the nuclease activity by non-hydrolysable ATP analogues are much weaker. Similar observations have also been seen for T4 large terminase (78).



*Figure 4-7 The effects of ATP analogues on the nuclease activity of the wild type  $\Delta$  443 and ATPase Walker B  $\Delta$  443 mutant*

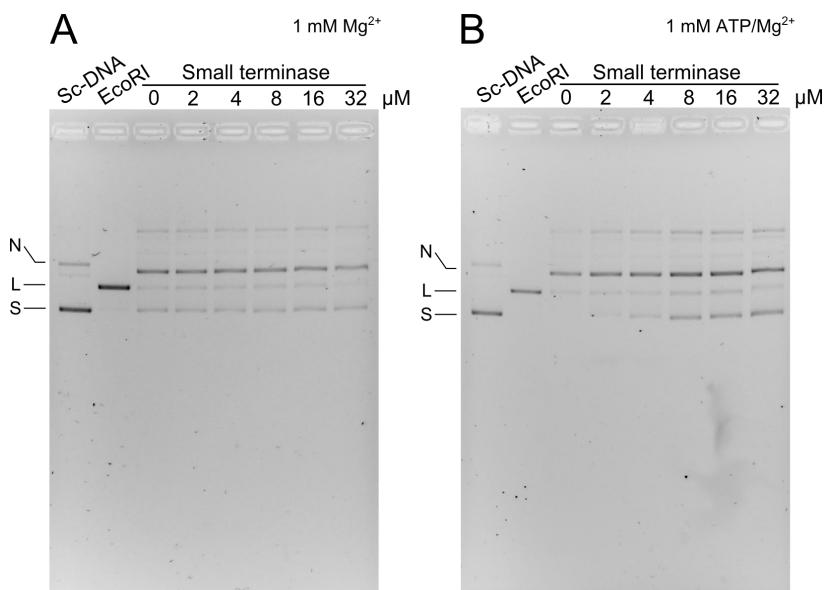
*The DNA product after the nuclease reactions in the presence of 0 to 5 mM concentration of ATP were analysed by agarose gel electrophoresis using (A) wild type and (B) ATPase Walker B mutant. The effects of other non-hydrolysable ATP analogues and ADP were also analysed as shown in (C). N: nicked, L: linearised, S: supercoiled.*

#### 4.3.6 Effects of small terminase

In bacteriophage DNA packaging, small terminase plays a role in DNA recognition. This section investigates the effects of the G20c small terminase on the nuclease activity of the G20c  $\Delta$ 443 large terminase. It was found that small terminase has no apparent effect on the nuclease activity of the large terminase protein in the presence of magnesium chloride. When ATP is added, a decrease of the nuclease activity is observed (Figure 4-8). It is likely that in the presence of ATP, addition of small terminase inhibits the nuclease activity of the large terminase.

In nuclease assays performed using T4 large terminase, it was observed that addition of small terminase inhibited the nuclease activity of the large terminase in the absence of ATP. However, in the presence of ATP, the nuclease activity was stimulated as the concentration of small terminase increased (78). This observation was different to that was observed for

G20c large terminase protein in this work, where a decrease in the nuclease activity was observed (Figure 4-8). A possible explanation for the observed inhibition of the G20c large terminase nuclease by the small terminase is that in the presence of ATP, the large terminase is likely to interact with the small terminase. This would have some effect on the interaction between the large terminase and DNA, which would affect the nuclease activity. It should not be ignored that a decrease in nuclease activity can also be caused by binding of small terminase to the DNA regions where large terminase preferred to make the cleavage. Binding of the small terminase to DNA may also alter the DNA conformation which would have effect on the DNA cleavage and binding by large terminase.



*Figure 4-8 The effects of small terminase on nuclease activity*

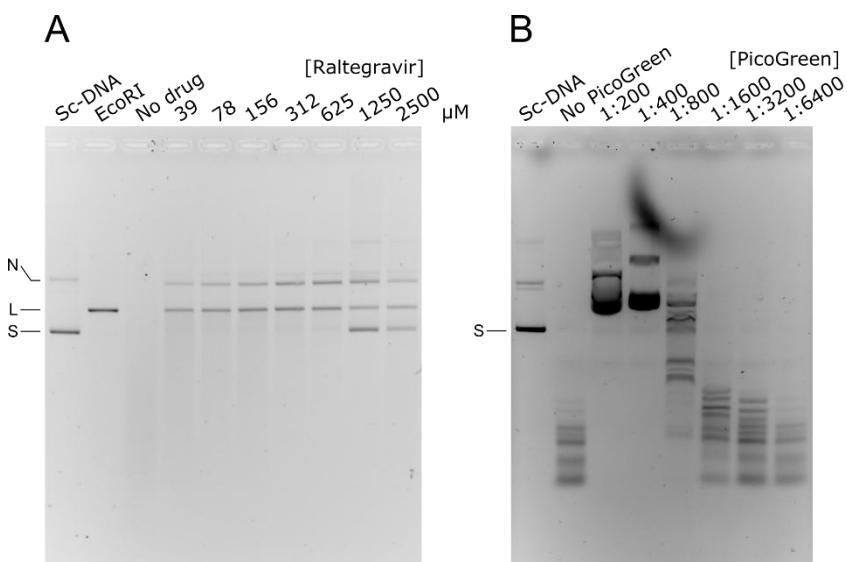
The experiment is performed in the addition of (A) 1mM magnesium, (B) 1 mM magnesium and ATP. The wild type G20c small terminase protein was titrated to the reaction using a concentration range of 2 to 32  $\mu$ M. N: nicked, L: linearised, S: supercoiled.

#### 4.3.7 Effects of Raltegravir and PicoGreen

Raltegravir (154) is a FAD approved anti-HIV drug, targeting the HIV reverse transcriptase RNase H domain active site. Like viral large terminase nucleases, RNase H proteins also belong to the RNase H-like endonuclease superfamily (73). Proteins from this superfamily have been proposed to use a similar two-metal catalytic mechanism in phosphoryl transfer reactions. It is likely that given the similarity in active site architecture and catalytic mechanism of RNase H proteins and viral large terminases (see Chapter 1, Figure 1-5, 1-6), Raltegravir may also interact with viral large terminase nucleases and inhibit its metal dependent nuclease activity. To test this, *in vitro* nuclease assays were set up to analyse the effects of Raltegravir on the nuclease activity of G20c large terminase  $\Delta$ 443. The raltegravir stock solution was prepared by dissolving the drug powder in 100% DMSO and titrated by dilution (from a 50 mM stock) into the nuclease reactions. It was observed that the nuclease

activity is reduced upon addition of Raltegravir at increasing concentrations (Figure 4-9). This is likely due to binding of Raltegravir as an RNase H inhibitor to the active site of the RNase H-like large terminase nuclease domain, which prevented the binding of the incoming DNA to the nuclease active site.

PicoGreen (155) is a fluorescent DNA minor groove binder. Its effect on the nuclease activity of G20c large terminase  $\Delta 443$  was probed using *in vitro* nuclease assays. A significant reduction of the nuclease activity was observed at a concentration range between 1:200 and 1:800 (Figure 4-9). As a minor groove binder, binding of this molecule to the DNA may alter the conformation of the DNA substrate, making it unfavourable for nuclease binding and cleavage. Alternatively, if DNA cleavage requires interaction with the minor groove, the binding of PicoGreen would inhibit DNA cleavage.



*Figure 4-9 The effects of small molecules on the nuclease activity*

*The reduction of the nuclease activity is shown in the presence of (A) Raltegravir and (B) PicoGreen titrated at various concentrations. N: nicked, L: linearised, S: supercoiled.*

#### 4.4 Summary

In this chapter, the factors that affect G20c  $\Delta 443$  large terminase ATP binding and DNA cleavage were analysed by thermal shift assays and *in vitro* nuclease activity assays, respectively. It was proposed that the conformation of the  $\beta/\gamma$  phosphate moiety plays a role in ATP binding, supported by the observed differences in protein thermal stability when non-hydrolysable ATP analogues of different chemical modifications were added. As seen in the nuclease assays, the nuclease activity is inhibited by sodium chloride, indicating non-sequence specific protein-DNA binding. This is further supported by the inhibition effects of Picogreen, a DNA minor groove binder. These results suggest that the G20c large terminase is likely to recognise DNA based on its conformation rather than specific sequence. As

observed for other RNase H-like proteins, the G20c  $\Delta$ 443 large terminase is activated by divalent metal ions, such as magnesium and manganese. However, in contrast to several other RNase H family members, no attenuation was observed for this protein when high concentration of magnesium and manganese were used. This indicated that the architecture and location of metal binding sites are very likely to be different to those reported for other RNase H family proteins. As observed in T4 large terminase, the nuclease activity is stimulated in the presence of  $Mg^{2+}$ -ATP. ADP or non-hydrolysable ATP analogues show either nil- or modest- stimulation of the nuclease activity. It is possible that the apparent stimulation of the nuclease activity needs the bound ATP to be hydrolysed. Distinct to the T4 large terminase protein, addition of the small terminase protein inhibited the nuclease activity which was likely due to either small terminase-DNA or small terminase-large terminase interactions. Addition of an anti-HIV reverse transcription metal ion chelator inhibited the nuclease activity, suggesting interaction of the drug molecule with the protein. Mutagenesis on the ATPase Walker B residues reduces the nuclease activity. It is likely that these mutations interrupted the interaction between ATPase domain and DNA, supporting the proposal that ATPase domain serves as a primary DNA binding region (56,66).

## 5 Structural and functional studies of the G20c large terminase

This chapter describes structural studies investigating how large terminase cleaves DNA. Structures of the G20c large terminase nuclease domain were determined in complex with various divalent metal ions at the nuclease active site. In addition, mutagenesis studies of the residues at the nuclease active site, along with *in vitro* nuclease assays were performed. These data allow to propose a model for DNA cleavage by the G20c large terminase nuclease, as well as understand the role of metal ions and several catalytic residues.

### 5.1 Structure determination of the G20c large terminase nuclease domain

As described in chapter 3, diffracting crystals were obtained using G20c large terminase protein construct ( $\Delta 443$ ) with a truncated C-terminal region. The best crystal diffracted to 1.6 Å at the DLS-I03. As  $Zn^{2+}$  was present in the crystallisation conditions, the data were collected at the wavelength close to the absorption edge of the  $Zn^{2+}$  atom ( $\sim 1.28$  Å). The optimum wavelength for the SAD data collection was chosen based on the X-ray fluorescent spectrum scan of the crystal. The Matthew's coefficient calculation showed that it is unlikely that the full-length protein containing both ATPase and nuclease domain is present in the unit cell. It was thus possible that only part of the full-length protein was crystallised due to proteolysis, given the time needed for crystal formation. The structure of this crystal form (crystal form 1) was determined using the protocol described in chapter 2 for Single wavelength anomalous dispersion.

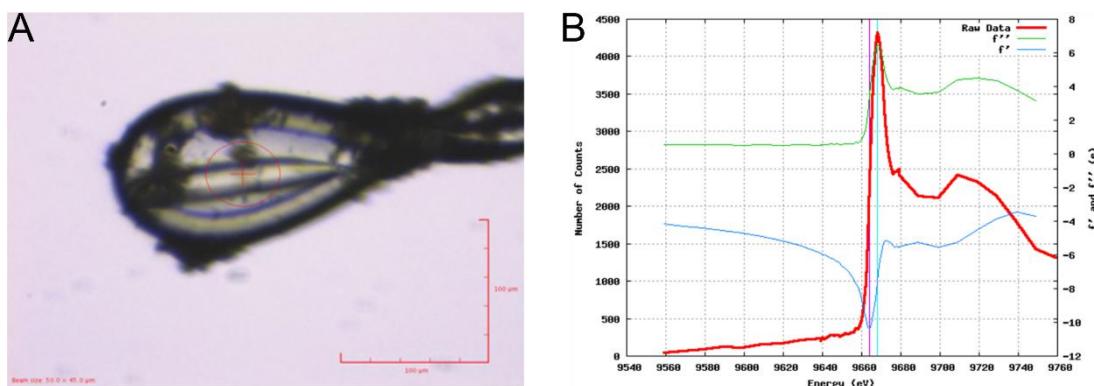


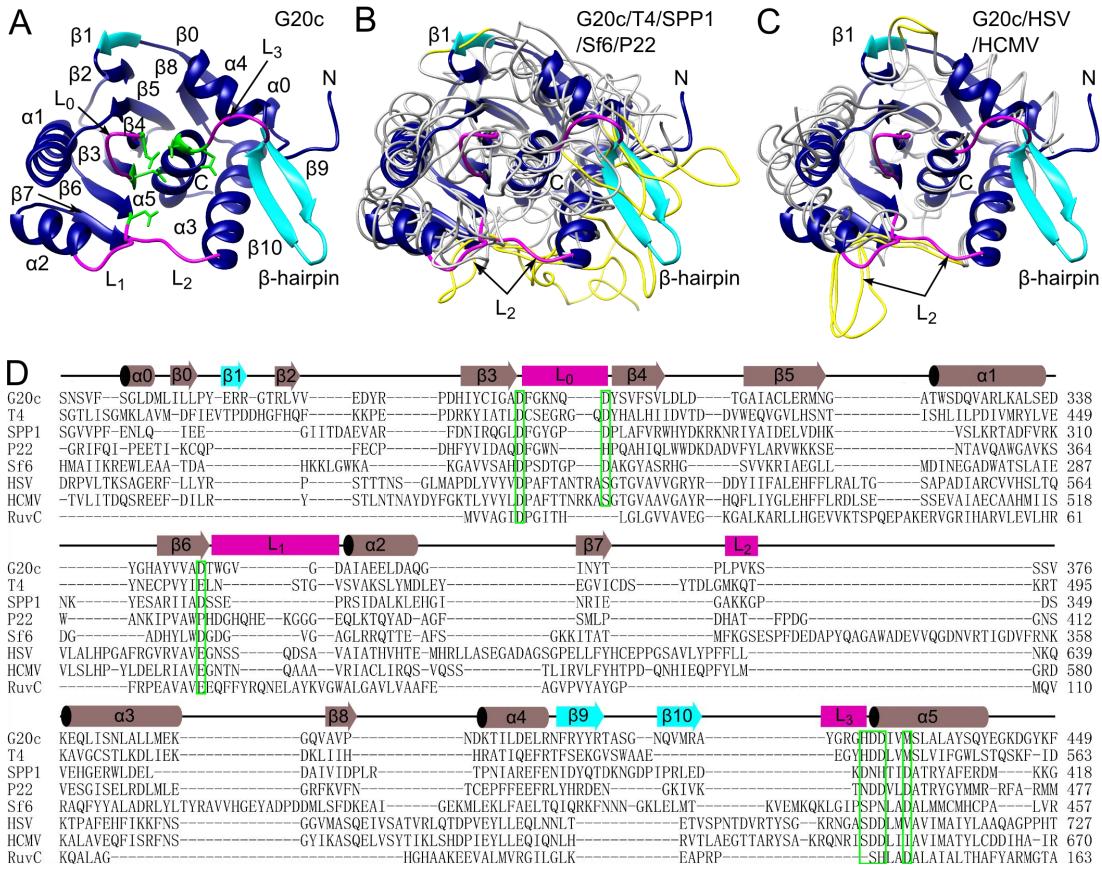
Figure 5-1 X-ray data collection

(A) The crystal used for SAD data collection. (B) X-ray fluorescent spectrum.

### 5.2 Structure of the G20c large terminase nuclease domain

As mentioned above, the structure of the proteolytically truncated G20c large terminase was determined by  $Zn^{2+}$  SAD. This structure corresponded to residues 257 to 443 of the nuclease domain. Subsequently, the same protein construct as observed in the crystal was designed

and crystallized in two different crystal forms, 2 and 3 (Table 5-1). These structures were determined by molecular replacement, following the protocol described in chapter 2. Crystal forms 1 and 2 contain a single molecule in the asymmetric unit, whereas two protein molecules were found per asymmetric unit for crystal form 3. As reported for other members of the RNase H-like endonucleases (74), the overall structure adopts an RNase-H fold where a cluster of carboxylic acids is contributed to the active site by strands  $\beta$ 3,  $\beta$ 4 and  $\beta$ 6, helix  $\alpha$ 5, and loops L<sub>0</sub>-L<sub>3</sub>. These residues were shown to be critical for bacteriophage function, DNA packaging or nuclease activities in bacteriophages T4 (76,156) and SPP1 (67,157) (Figure 5-2, panel D). Loops L<sub>1</sub> and L<sub>2</sub> were defined earlier for the P22 large terminase nuclease (68) and correspond to residues 347-352 and 369-372, respectively, in the G20c large terminase. The two other loops that also form part of the active site are L<sub>0</sub> (residues 295-301) and L<sub>3</sub> (residue 423-427), respectively. The  $\beta$ -hairpin ( $\beta$ 9 and  $\beta$ 10 strands on Figure 5-2), a unique feature of viral large terminases not observed in other members of the RNase H-like endonucleases (67), is well ordered in crystal forms 2 and 3, but is invisible in the structure of the proteolytic fragment (crystal form 1).



*Figure 5-2 Comparison of large terminase nuclease domains from different viruses*

(A) Ribbon diagram of the G20c large terminase nuclease.  $\beta$ 1 and  $\beta$ -hairpin ( $\beta$ 9 &  $\beta$ 10) are highlighted in cyan. Loops L<sub>0</sub>, L<sub>1</sub>, L<sub>2</sub> and L<sub>3</sub> are coloured in magenta. Carboxylates located in the active site cleft are coloured in green (B) Superposition of the G20c nuclease (blue/cyan ribbon, as in panel A) with large terminase nucleases from T4, SPP1, Sf6 and

*P22 (grey chain traces). Segments equivalent to L2,  $\beta$ 1 and the  $\beta$ -hairpin of the G20c nuclease are in yellow. (C) The same superposition as in (B) but with the Herpesviridae HSV and HCMV nucleases. (D) Structure based sequence alignment of large terminase nucleases from bacteriophages G20c, T4, SPP1, P22, Sf6; herpesviruses HSV, HCMV and *T. thermophilus* RuvC resolvase. Secondary structure elements of the G20c nuclease are shown above the alignment and are coloured as shown in (A). Conserved or semi-conserved active site residues are in green (figure reproduced from (81)).*

### 5.3 Structure comparison with other large terminases

In this section, the structure and sequence of the G20c large terminase nuclease were compared to nucleases from other viruses. Superposition of the G20c large terminase nuclease with bacteriophage (40,67-69) and herpes virus (83,158) nucleases reveals highly similar three-dimensional structures, despite low sequence identity (Figure 5-2). The G20c nuclease shares the highest similarity with T4 gp17 nuclease (40), with  $\text{Ca}$  rmsd of 1.9 Å for 176 residues and 25.4% sequence identity, while the lowest similarity is with the SPP1 G2P (67), with  $\text{Ca}$  rmsd of 2.9 Å for 143 residues and 12.6% sequence identity).  $\text{Ca}$  rmsd with the HSV pUL15 (83) (135 residues, 20.5% sequence identity) and HCMV UL89 (158) (141 residues, 19.2% sequence identity) nucleases is 2.4 Å and 2.5 Å, respectively.

The G20c nuclease exhibits three apparent structural differences to nucleases from other viruses. Firstly, the central  $\beta$ -sheet contains an additional  $\beta$ -strand ( $\beta$ 1 on Figure 5-2). Secondly, a  $\beta$ -hairpin that is unique to viral terminases (strands  $\beta$ 9 &  $\beta$ 10), is more extended and better ordered (Figure 5-1). Thirdly, loop L<sub>2</sub>, defined as the “hairpin loop” and implicated in interaction with DNA in the HSV pUL15 nuclease (83) is much shorter (Figure 5-2).

### 5.4 Structures of nuclease-metal complexes

To investigate how the G20c nuclease binds to catalytic metal ions, structures for complexes with  $\text{Mn}^{2+}$ ,  $\text{Mg}^{2+}$ ,  $\text{Co}^{2+}$  and  $\text{Ca}^{2+}$  were determined by soaking crystals of the apo protein, whereas structure of the  $\text{Zn}^{2+}$  complex was obtained by co-crystallisation (Figure 5-3, Table 5-1). It was observed that when  $\text{Mn}^{2+}$ ,  $\text{Mg}^{2+}$ ,  $\text{Co}^{2+}$  was used in soaking, only one metal ion was found accommodated in metal A site, coordinated in the canonical octahedral geometry by the side chains of D294 and D429 and four water molecules (159). A  $\text{Zn}^{2+}$  ion was also bound at the site A as well as at an additional satellite site. Finally, a single  $\text{Ca}^{2+}$  ion was found bound to the active site, but in two completely different binding modes in each of the two protein molecules present in the asymmetric unit. The  $\text{Ca}^{2+}$  ion adopts an octahedral geometry in molecule A, coordinated by D294, D429 and D300 side chain oxygens, and by F295, Y301 main chain oxygens, as well as a water molecule. In molecule B, the  $\text{Ca}^{2+}$  ion appears to be bound non-specifically, with only a single direct contact with the protein, mediated by the main chain oxygen of F295 (Figure 5-3).

Table 5-1 X-ray data collection and refinement statistics (G20c nuclease crystal form 1, 2 and 3, table reproduced from (81))

Crystal form	1 Zn <sup>2+</sup> -SAD	1 Zn <sup>2+</sup>	2 Apo	3 Mg <sup>2+</sup>	3 Mn <sup>2+</sup>	3 Co <sup>2+</sup>	3 Ca <sup>2+</sup>
Wavelength (Å)	1.2841	0.9763	0.9795	0.9796	0.9795	0.9796	0.9795
Space group	P6 <sub>5</sub>	P6 <sub>5</sub>	P3 <sub>2</sub> 1	P2 <sub>1</sub>	P2 <sub>1</sub>	P2 <sub>1</sub>	P2 <sub>1</sub>
Unit-cell a, c (Å)	60.8, 90.4	60.9, 90.4	55.1, 95.7	43.8, 69.9	43.8, 70.0	42.9, 69.7	43.4, 69.8
Unit-cell b (Å)				54.0	53.9,	53.9,	53.6
Unit-cell β (°)				91.94	92.06	91.62	91.8
Resolution(Å)	45.53-1.6 (1.63-1.60)	45.60-1.45 (1.47-1.45)	47.72-2.15 (2.21-2.15)	43.78-1.2 (1.22-1.2)	43.78-1.2 (1.22-1.2)	42.92-1.60 (1.63-1.6)	43.43-1.10 (1.11-1.1)
Rmerge (%)	7.9(147)	8.6(129)	4.3(80.4)	7.3(76.5)	4.2(34.6)	7.2(71.9)	7.8(97.7)
I/σ<I>	13.4(1.7)	8.2(1.0)	10.3(1.0)	7.4 (1.6)	17.6(4.0)	6.7(1.1)	6.0(1.0)
Completeness (%)	100(100)	99.6(99.5)	95.6(90.6)	97.0(97.6)	95.0(89.1)	95.5 (84.6)	98.8(87.3)
Multiplicity	8.8(8.8)	4.5(3.9)	2.2(2.2)	2.5(2.5)	4.3(4.3)	2.1(2.0)	3.2(2.8)
CC <sub>1/2</sub>	0.998(0.746)	0.997(0.420)	0.999(0.365)	0.996(0.522)	0.999(0.906)	0.997(0.611)	0.996 (0.385)
Refinement							
Rwork/Rfree		0.201/0.211	0.200/0.280	0.183/0.209	0.169/0.182	0.185/0.224	0.222/0.235
Mean B factor(A <sup>2</sup> )		23	56	14	11	22	12
R.M.S.D.							
Bond lengths (Å)		0.008	0.009	0.007	0.006	0.007	0.006
Bond angles (°)		1.2	1.2	1.2	1.2	1.2	1.2
Ramachandran (%)							
Favored	98	97	99	98	99	99	98
Allowed	2	3	1	2	1	1	2
Outliers	0	0	0	0	0	0	0

Values in the parentheses are for the outermost resolution shell.

SAD, single-wavelength anomalous diffraction.

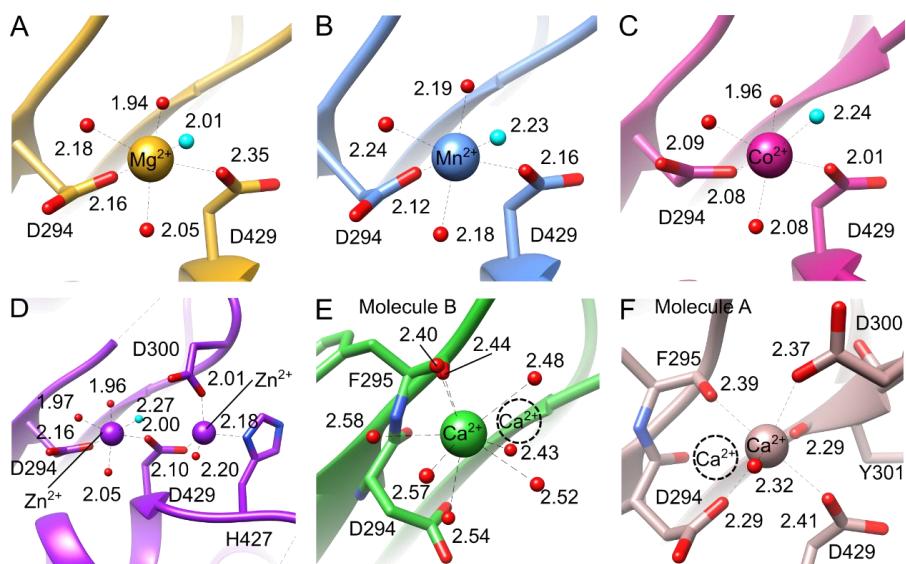
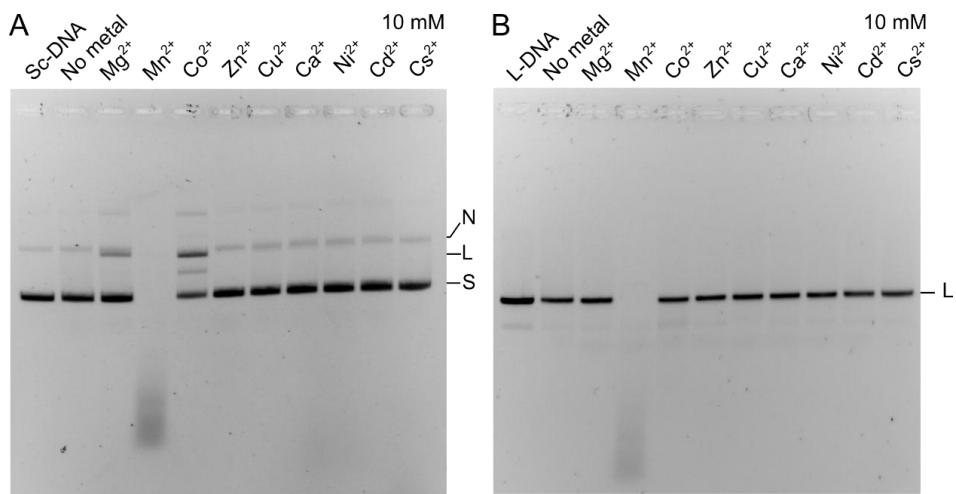


Figure 5-3 Metal coordination in crystal structures

Metal coordination in crystal structures of the G20c nuclease in complex with (A) Mg<sup>2+</sup>, (B) Mn<sup>2+</sup>, (C) Co<sup>2+</sup>, (D) Zn<sup>2+</sup>, (E, F) Ca<sup>2+</sup>. The nucleophilic water molecule is highlighted in cyan. The Ca<sup>2+</sup> ion from the second molecule in the asymmetric unit is shown as dashed circle. The nucleophilic water molecule is highlighted in cyan (figure reproduced from (81)).

## 5.5 Metal dependence of nuclease activity

In chapter 3, the C-terminally truncated  $\Delta 443$  construct of the G20c large terminase was designed and used for crystallisation trials and nuclease activity assays. Although this construct is less vulnerable to protease degradation, it also has reduced solubility compared to the wild type full-length protein, which makes it more likely to precipitate during purification or biochemical assays performed at low salt concentrations. To prevent protein precipitation and easily detect the catalytic activity, assays were performed using the full-length wild type G20c large terminase protein under low salt concentrations which are likely to facilitate binding of DNA and divalent metal ions. In common with observations for the  $\Delta 443$  construct discussed in chapter 4, the nuclease activity of the full-length wild type G20c large terminase was activated by divalent metals  $Mn^{2+}$ ,  $Mg^{2+}$  and  $Co^{2+}$  but inhibited by  $Ni^{2+}$ ,  $Zn^{2+}$  or  $Ca^{2+}$  (Figure 5-4), consistent with observations for T4 GP17 (78), SPP1 GP2 (157) and HCMV UL89 (158). Addition of  $Cu^{2+}$ ,  $Ni^{2+}$ ,  $Cd^{2+}$  and  $Cs^{2+}$  were also inhibitory. Similar to GP2 and UL89, addition of  $Mg^{2+}$  resulted only in a limited activity (157), leading to production of nicked or linearized DNA when supercoiled DNA was used as substrate. However, G20c nuclease had limited activity with  $Co^{2+}$ , unlike the high levels of non-specific *in vitro* nuclease activity observed for the SPP1 GP2 nuclease (comparable with  $Mn^{2+}$  (157)), or the inhibitory effect on the nuclease activity of T4 GP17 (78). Finally, when linear DNA was used as DNA substrate, the metal dependent properties were similar to the properties observed in the presence of supercoiled DNA (Figure 5-4).



*Figure 5-4 Dependence of nuclease activity on metals and metal coordination in crystal structures*

(A) Effect of divalent metal ions on the nuclease activity. N, nicked; L, linear (L-DNA); S, supercoiled DNA (Sc-DNA). Effect of active site residues on nuclease activity (figure reproduced from (81)).

## 5.6 Effect of active site residues on nuclease activity

Functional importance of the metal binding sites A, B and the second Zn<sup>2+</sup> binding site were investigated using the full-length protein containing both ATPase and nuclease domains. Aspartic acids coordinated by metals A and B were replaced by asparagines: D294N, D429N, D347N. Aspartate 428 which coordinates an inner shell water of the metal A was mutated to alanine or asparagine (D428A and D428N). Aspartate 300 and histidine 427 that co-ordinate the second Zn<sup>2+</sup> were mutated to alanine or asparagine (D300A, H427A and D300N and H427N). Nuclease assays for all mutant proteins were performed *in vitro* in the presence of 0.1mM MnCl<sub>2</sub>. Wild-type large terminase converted the entire supercoiled DNA substrate into DNA fragments with defined length. The lowest nuclease activity was observed for the D294N, D347N, D429N mutant proteins. In contrast, D300N shows nuclease activity comparable to that of the wild-type protein whereas a modest decrease in nuclease activity was observed for D300A. A reduction in the nuclease activity was also observed for the H427N mutation. However, this mutant protein retained the ability to process longer DNA into smaller fragments. Replacing this residue with Alanine (H427A) resulted in deficiency in digestion of the supercoiled DNA, as for the D347N mutation. Both D428N and D428A mutant proteins showed a significant drop in nuclease activity and were deficient for production of shorter DNA fragments (Figure 5-5).

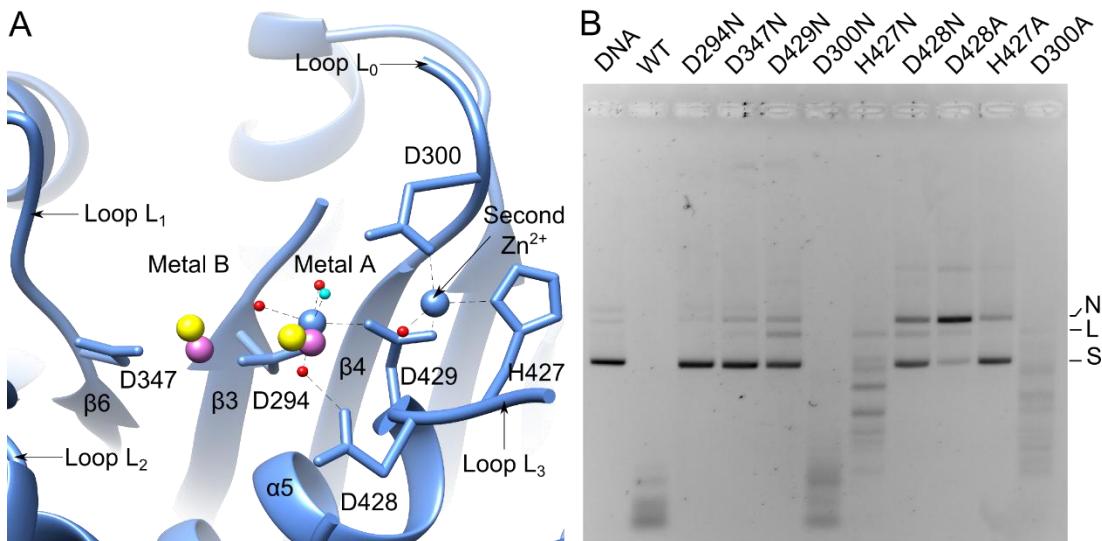


Figure 5-5 Metal binding sites

(A) Active site of the G20c nuclease, complex with Zn<sup>2+</sup>. The two Mn<sup>2+</sup> ions, taken from structure superposition with SPP1 and HCMV nuclease-Mn<sup>2+</sup> complexes are in magenta and yellow, respectively. The water nucleophile is shown in cyan. (B) In vitro nuclease assays. Activity is shown for the wild type and active site mutant proteins. N, nicked; L, linear; S, supercoiled DNA (figure reproduced from (81)).

## **5.7 Structural features**

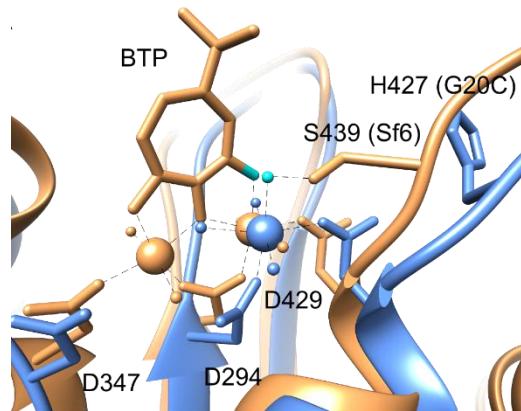
The structure of the nuclease from the thermophilic G20c bacteriophage differs from its mesophilic counterparts by shortened surface loops, notably a shorter loop L<sub>2</sub> than that present in other bacteriophage and herpes virus nucleases (Figure 5-2). There is also an increased number of salt bridges (160): 9 versus 5 to 7 found in nucleases of mesophilic viruses (40,67-69,83,158). In addition, the β-hairpin, present only in viral nucleases, is more extended and ordered. These differences are expected to increase the protein stability at the higher environmental temperatures encountered by the G20c bacteriophage (161).

## **5.8 Role of active site residues**

Of the residues that coordinate the second Zn<sup>2+</sup> ion, D300 is conserved among large terminases of bacteriophages T4, RB49, SPP1, Sf6 and P22. D409, the equivalent residue in T4 GP17, has been reported to be crucial for bacteriophage function (156). Comparison of the catalytic activity for mutants D300N and D300A indicated that the negative charge of D300 is not essential for catalysis (Figure 5-5). The slight reduction in activity observed for D300A is likely caused by disturbance of the water mediated hydrogen bonding network at D429 and/or the DNA backbone.

Histidine and glutamatic acid residues adjacent to the metal A site in RNase H proteins have been suggested to play important roles in catalysis by: affecting product release (162); participating in proton transfer pathway (163); or binding to a third Mg<sup>2+</sup> during catalysis (164). The significant reduction in nuclease activity observed for the H427A mutant (Figure 5-5) suggests that this residue is involved in catalysis. Interestingly, an equivalent serine residue found in the Sf6 nuclease forms a hydrogen bond with an oxygen atom of the bound metal chelator, occupying the position where the water nucleophile is normally coordinated by metal A (Figure 5-6). As H427 is located on the flexible loop L<sub>3</sub> downstream of the β-hairpin and is solvent exposed, it likely plays a role in orientating the water nucleophile and is facilitated by the conformational flexibility of loop L<sub>3</sub> and the β-hairpin to participate in the proton transfer pathway during catalysis.

The deficiency in catalysis observed for mutants of D428 (Figure 5-5) suggests that this residue may be responsible for stabilization of metal A binding, since it is proximal to metal A and forms a hydrogen bond with an inner shell water coordinating metal A.



*Figure 5-6 Comparison of large terminase and RuvC structures*

(A) Superposition of the Sf6 (brown) and G20c (blue) nucleases. The oxygen of  $\beta$ -thujaplicinol (BTP; cyan) occupies the same position as the nucleophile water. This oxygen forms a hydrogen bond with S439, equivalent to H427 in the G20c nuclease (figure reproduced from (81)).

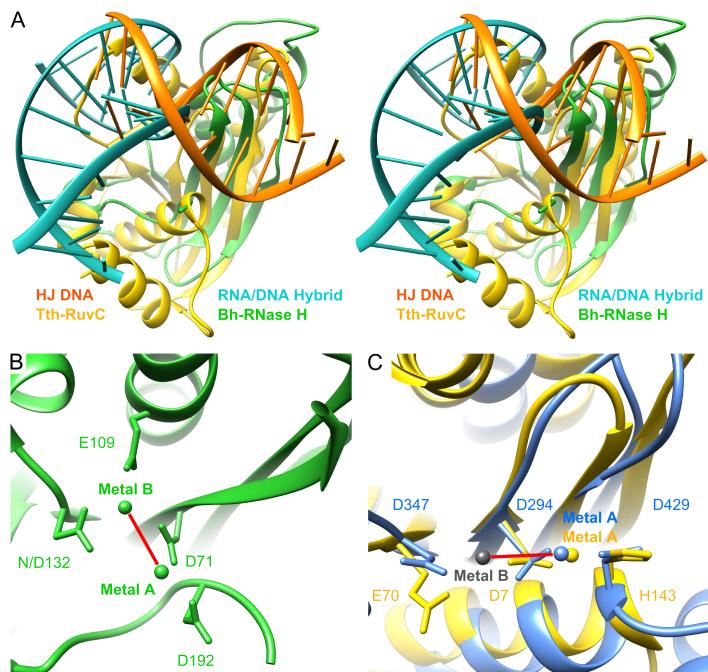
### 5.9 Structural basis for metal dependence of nuclease activity

RNase H-like superfamily proteins require divalent metals such as  $Mg^{2+}$  or  $Mn^{2+}$  for catalysis (153,165-167). The mechanism for  $Ca^{2+}$  inhibition of nuclease activity (153,168) can be explained by the different  $Ca^{2+}$  coordination observed in our structure (Figure 5-3), where binding to site A is non-specific (169) or is only coordinated by main chain and side chain oxygens, potentially induced by the larger atomic radius and longer coordination distances of  $Ca^{2+}$ , compared with  $Mg^{2+}$  or  $Mn^{2+}$ . Furthermore, the metal coordinated nucleophile water was replaced by the side chain of D300, likely to disturb catalysis. These observations provide further insight into the inhibition mechanism of  $Ca^{2+}$  proposed for other RNase H-like enzymes (70,168).

Due to the similarity in atomic radius, it was suggested that  $Zn^{2+}$  can substitute  $Mg^{2+}$  for catalysis (168), but this is not the case for RNase H, shown to be inhibited (162,170-173), or having significantly reduced activities in the presence of  $Zn^{2+}$  and  $Ni^{2+}$  (162). However, the structural basis for this inhibition has remained unclear. In our structure complexed with two  $Zn^{2+}$  ions, the  $Zn^{2+}$  ion bound to the metal A site adopts an octahedral geometry, resembling the canonical coordination of active cofactors such as  $Mg^{2+}$  (Figure 5-3), that would support catalysis. However, a second  $Zn^{2+}$  found bound to a site adjacent to the metal A binding site, not previously reported for the RNase H-like superfamily, is coordinated by catalytic important residues D429 and H427 (Figure 5-3). This geometry is likely to perturb catalysis by altering the ability of these residues to stabilise metal A binding (D429), orient the water nucleophile or participate in catalytic proton transfer (H427).

## 5.10 RuvC is the closest structural homologue of large terminase nuclease

A structure similarity search of the protein structure database (PDB) (174) for G20c large terminase nuclease allowed the identification of its closest structural homologues, RuvC resolvases (175). As members of the RNase H-like superfamily, RuvC proteins show similar folding topology and active-site geometry to those of the RNase H proteins (176). Interestingly, structural comparison of the Bh-RNase H and *T. thermophilus* RuvC (Tth-RuvC) with bound RNA/DNA hybrid or dsDNA respectively revealed an unexpected difference (Figure 5-7). Notably, an  $\alpha$ -helix which forms part of the active site is replaced by a loop in Bh-RNase H, which changes the position of its metal A coordinating residue, D192 vs H143 in Tth-RuvC (Figure 5-7). Moreover, an additional catalytic residue at the metal B binding site, E109, is also absent in RuvC (Figure 5-7). These differences result in a distinctly different orientation of the active site metals and the bound nucleic acid duplex. It is possible that these RNase H-like superfamily proteins evolutionally adjusted the position of their metal binding sites and metal coordinating residues to adapt to different nucleic acid substrates while maintaining the classic RNase H fold. Structural superposition of the G20c large terminase nuclease with Tth-RuvC, unlike for Bh-RNase H, results in good alignment of the three catalytic important residues, consistent with the sequence alignment (Figure 5-2). Therefore, given the similarity between the RuvC and large terminase nucleases, a similar catalytic mechanism can be deduced.



*Figure 5-7 Comparison of metal location and DNA orientation*

(A) Stereo view showing the superposition of Bh-RNase H (green, PDB accession code: 1ZBI) and *T. thermophilus* RuvC (Tth-RuvC, yellow, 4LD0) with bound nucleic acid duplexes (cyan for RNA/DNA hybrid, orange for Holliday junction DNA). (B) Active site of

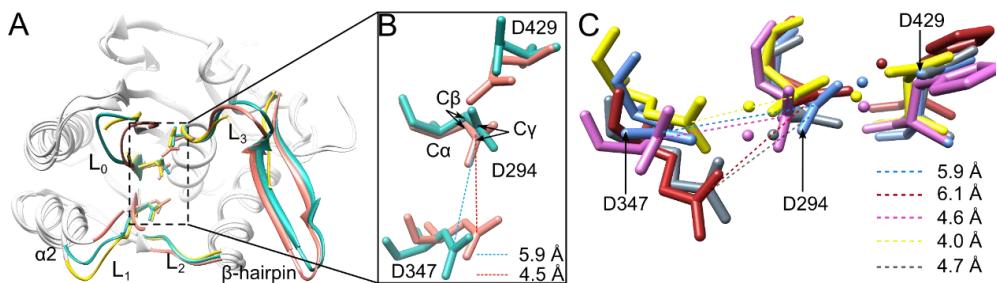
*Bh-RNase H nuclease with bound metal ions. (C) Superposition of the G20c (blue, 5MIN) and Tth-RuvC (yellow, 4EP4) nuclease active sites, with bound metal ions, shown in the same orientation as (B). The metal ion B from bacteriophage bIL47 RuvC (grey, 4KTZ) (figure reproduced from (81)).*

*Table 5-2. Structure alignment of G20C large terminase nuclease with Bh-RNase H and Tth-RuvC (177) (table reproduced from (81))*

	Q-score	P-score	Z-score	RMSD (Å)	N <sub>align</sub>	N <sub>sse</sub>	Seq-%	N <sub>md</sub>	N <sub>res-Q</sub>
RNase H	0.117	-	1.86	4.1	93	4	0.043	4	135
RuvC	0.226	3.84	6.96	2.4	106	8	0.132	2	158

## 5.11 DNA binding surface plasticity

Superposition of structures determined from crystal forms 1, 2 and 3 (Table 5-1) reveal that loops L<sub>0</sub>, L<sub>1</sub>, L<sub>2</sub> and L<sub>3</sub> that contribute to the active site and the β-hairpin adopt different conformations. Additionally, differences in the position of helix α2 at the C-terminus of L<sub>1</sub> is also observed (Figure 5-8). Notably, the above mentioned flexible structural segments, including the loops, β-hairpin and α2, are well conserved in structures of all large terminase proteins, including bacteriophage (40,67,68,178) and herpesvirus (79,179). It was observed that D347 predicted in stabilising the transition state during the catalysis (see below), from RNase H family proteins, is located at the N-terminus of L<sub>1</sub>. Importantly, the position of this residue differs significantly between crystal forms 2 and 3 (Figure 5-8), bringing the carboxyl group of D347 1.5 Å closer to the catalytic D294, which is coordinated to both metals A and B. It was also observed that a shorter distance between these two residues was observed earlier in Mn<sup>2+</sup> complexes of SPP1(67) and HCMV(79) nucleases and for Mg<sup>2+</sup> complexes of RuvC proteins from poxvirus and phage bIL67 (Figure 5-8). A model of the G20c nuclease-DNA complex was constructed by superposing the structures for G20c nuclease and Tth-RuvC resolvase in complex with a Holliday junction DNA (175) (Figure 5-10). In this model, conserved loops L<sub>0</sub>, L<sub>1</sub>, L<sub>2</sub> form direct contacts with DNA in the Tth-RuvC resolvase are also in proximity to the DNA in the modelled G20c nuclease-DNA complex. Loop L<sub>3</sub> and β-hairpin, absent in Tth-RuvC resolvase also contact the modelled DNA (Figure 5-10). This predicted DNA binding region is consistent with the residues implicated in DNA binding or cleavage by mutagenesis of the large terminase protein from bacteriophage P74-24 (180), a close homologue of the bacteriophage G20c.



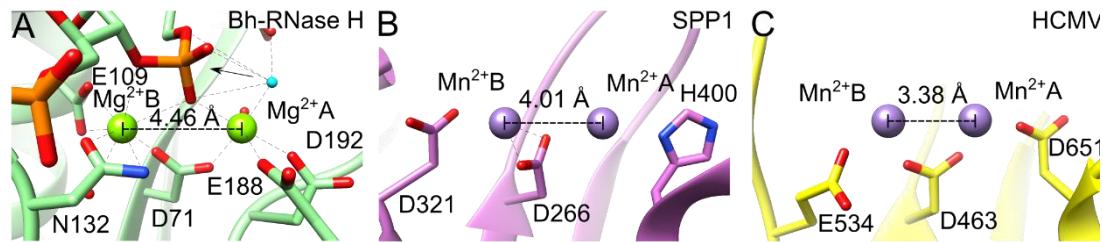
*Figure 5-8 Conformational flexibility in the active site*

(A) Superposition of G20c structures with the most flexible regions highlighted (yellow, crystal form 1; salmon, crystal form 2; cyan, crystal form 3). (B) Close-up view of (A) with residues from crystal form 1 omitted. The C $\gamma$ -C $\gamma$  distances between D347 and D294 are indicated using dashed lines. (C) Superposition of the G20c (blue), SPP1 (purple) and HCMV (yellow) large terminase nucleases with RuvC resolvases from Canarypox virus (dark red) and Lactococcus phage bIL67 (grey). Distances between the two carboxyl groups coordinating metal B are shown (figure reproduced from (81)).

## 5.12 RuvC-like, canonical two-metal dependent catalysis

Modelling of the G20c nuclease-DNA complex results in the B site bound manganese ion (Figure 5-10) corresponding to its position in the crystal structure of SPP1 GP2 (67), with the scissile phosphate of the Holliday junction DNA placed between this and Mn<sup>2+</sup>. The water nucleophile (Figure 5-10, cyan) coordinated by manganese A is in proximity to the scissile phosphate, creating a near attack conformation (181). (Figure 5-10).

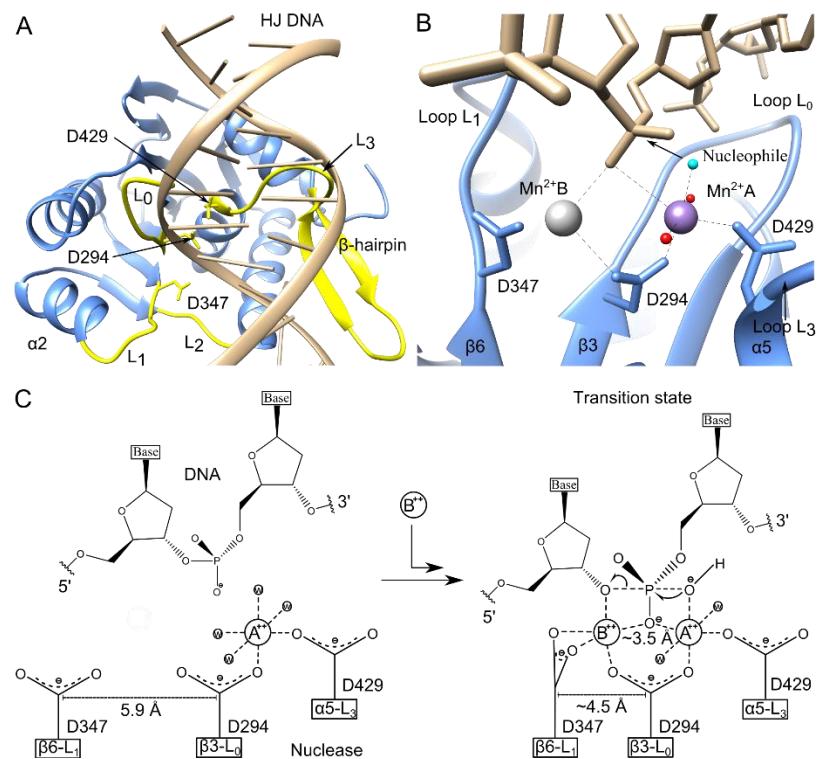
Unlike RNase H family proteins in which the metal site B is surrounded by three carboxylate side chains, only two carboxylates are present in the nucleases of bacteriophages T4, SPP1, Sf6, G20c and herpes virus HCMV and HSV. It has been suggested that during catalysis metal B adopts a fully dehydrated coordination favouring formation of a high-energy transition state (182). Therefore, we suggest that D347 should coordinate metal B in a bidentate conformation to provide two oxygen ligands from its side chain carboxylate in the pentavalent transition state. Equivalent residues coordinating metal B have been observed in both the SPP1, HCMV and Sf6-drug complexed nuclease structures (Figure 5-9, Figure 5-6). The flexibility in the position of D347, observed in the crystal structures presented here, would allow D347 to move closer to D294, co-ordinating manganese B and forming the pentavalent transition state. In accordance, structure superposition of the large terminase nucleases from G20c, SPP1, HCMV and other RuvC proteins, show that in the absence of metal B the two aspartates are more separated than in its presence (Figure 5-8).



*Figure 5-9 Nuclease active site residues and metal-metal distances observed in crystal structures.*

(A) Active site of Bh-RNase H complex with RNA/DNA hybrid and  $Mg^{2+}$  ions. RNA phosphorus atoms are in gold and nucleophile water molecule is in cyan. (B-C) Large terminase nucleases from (B) SPP1 and (C) HCMV (figure reproduced from (81)).

Based on this structural evidence, we propose the following mechanism for the large terminase nuclease. In the absence of DNA, site A is occupied by a divalent metal ion, as in structures of Poxvirus (183) and Tth-RuvC (184) resolvases from the RNase H-like superfamily. However, binding of a metal ion at site B depends on the conformation of loop L<sub>1</sub> which defines the distance between the carboxyl side chains coordinated to this metal. Upon DNA binding, the negative charge provided by the scissile phosphate facilitates the recruitment of the second metal ion to position B. Thus, the observed flexibility in loop L<sub>1</sub> (and position of D347) stabilises the formation of the transition state.



*Figure 5-10 Proposed mechanism of DNA cleavage*

(A) Model of G20c nuclease complex with DNA. DNA model was obtained from nuclease superposition with the structure of *T. thermophilus* RuvC-DNA complex. Structural segments that are in proximity to the modelled DNA are in yellow. (B) Close-up view at the active site of the model (shown in panel A), with  $Mn^{2+}$  position at site B (grey sphere) corresponding to

*its observed position in the crystal structure of SPP1 nuclease. The nucleophile water molecule which is expected to perform the SN2 nucleophilic attack on the scissile phosphate is coloured in cyan. (C) Schematic of the catalytic mechanism showing the proposed movement of D347 towards D294, concomitant with metal B and DNA binding, leading to transition state formation (figure reproduced from (81)).*

### 5.13 Insights into the headful DNA packaging mechanism

The observed flexibility of loop L<sub>1</sub> and the presumed dependence of metal binding on its conformation suggests that nuclease activity can be regulated via this loop. During initiation of DNA packaging, when viral DNA is recognised by the small terminase protein, DNA is expected to adopt a bent conformation, which can favour insertion into the active site of large terminase leading to DNA cleavage (36,185). When the capsid is filled with DNA, DNA bending (and cleavage) can be induced either by the internal pressure inside the capsid (headful cleavage) or other factors. Intriguingly, the closest structural homologue of the large terminase nuclease, RuvC, binds to Holliday junction DNA which is also known to be distorted and bent. Upon DNA interaction, the β-hairpin exposes the nuclease active site and allows the DNA scissile phosphate to be correctly placed. Concerted conformational changes in loop L<sub>1</sub> and DNA are likely to mediate DNA cleavage by bringing D347 closer to the active site, facilitating binding of the second metal and formation of the transition state.

While this model only describes the cleavage of one strand of the dsDNA substrate, producing a nicked DNA product, it can be postulated that cleavage of the second strand may be facilitated by rotation of the nuclease active site around this flexible DNA substrate to position the scissile phosphate of the second DNA strand within the active site. Alternatively, another binding event results in cleavage by a second large terminase molecule, either recruited to the initiation complex, or present as a subunit within the pentameric motor for the headful event. Further work will ascertain the validity of either of these models.

### 5.14 Summary

In summary, the crystal structures of the large terminase nuclease from the thermophilic bacteriophage G20c show that it is most similar to the RuvC subfamily of RNase H-like endonucleases. Like RuvC proteins, the nuclease requires either Mn<sup>2+</sup>, Mg<sup>2+</sup> or Co<sup>2+</sup> ions for activity, but is inactive with Zn<sup>2+</sup> and Ca<sup>2+</sup>. High resolution crystal structures of complexes with different metals reveal that in the absence of DNA, only one metal ion is accommodated in the active site. Binding of the second catalytic metal may be facilitated by the observed conformational variability, which enables the two catalytic aspartic acids to be brought 1.5 Å closer to each other. Structural comparison and nuclease activity data indicate that in common with the RuvC family, the location of the two catalytic metals differs from

other members of the RNase H family. Taken together, these data suggest that viral terminases cleave DNA by the canonical RuvC-like mechanism.

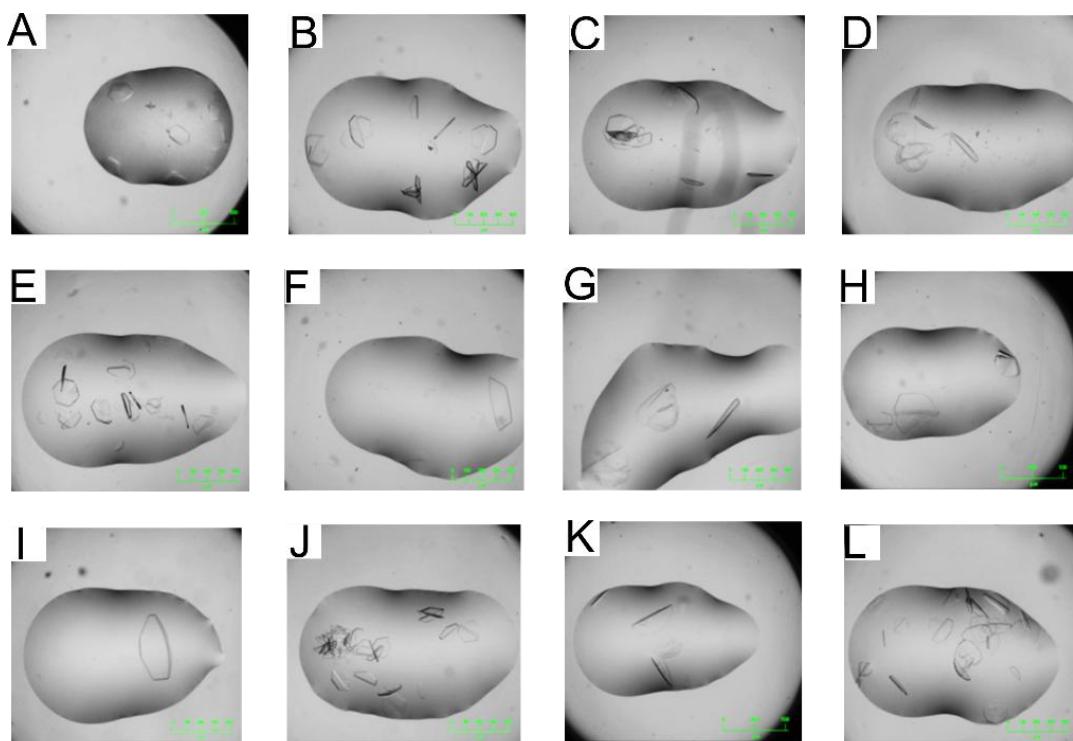
## 6 Structure determination for the D6E large terminase

This chapter describes the structure determination of the D6E large terminase. In the initial structure, the electron density corresponding to the ATPase domain was poorly defined compared to the rest of the structure, indicating disorder of the complete domain. Subsequent soaking with high NaCl concentration led to the successful determination of the protein structure containing ordered ATPase domain. To gain understanding on how the large terminase motor powers DNA translocation, structures in complex with ADP and non-hydrolysable ATP analogues were also determined after the removal of a C-terminal arm from the nuclease domain.

### 6.1 Structure determination (Crystal form 1)

As described in Chapter 3, diffracting crystals (space group R32, crystal form 1) were obtained for the D6E large terminase protein. Due to the poor sequence identity of the D6E large terminase compared to the available large terminases structures in the Protein Data Bank, attempts to solve the structure by molecular replacement failed. Sequence alignment and secondary structure prediction of the D6E large terminase suggested that it contains a nuclease domain that is a member of the RNase H-like endonuclease superfamily (73). Proteins of this superfamily catalyse phosphoryl transfer reactions using divalent metal ions. It was observed in Chapter 3 that D6E large terminase exhibits excellent nuclease activity in the presence of manganese ions. This indicated that the protein should be able to bind catalytic metal ions for DNA cleavage. Therefore, manganese was selected as the metal ion for producing protein-metal ion derivatives to facilitate phase determination and structure solution.

To obtain phase information, a metal derivative was produced by co-crystallisation of the protein with 1 mM manganese. During this process, crystallisation conditions were optimised (as described in Chapter 3) by varying the concentration of the precipitants (ammonium sulfate), pH, and buffer type, as well as by screening with additives (Figure 6-1). Micro-seeding (117) was also used to improve the size of the crystals for successful multiple wave-length dispersion experiment (MAD). The freshly made seeding stock was diluted using the mother liquor where the crystals grew at a series of concentrations before mixing with the protein and reservoir solutions. The largest crystal was selected for the MAD experiment.



A	0.01 M Strontium chloride hexahydrate	G	2 % w/v Benzamidinehydrochloride
B	3 % w/v Sucrose	H	3 % v/v Dimethyl sulfoxide
C	3 % w/v D-(+)-Trehalose dihydrate	I	1.2 % w/v myo-Inositol
D	30% w/v D-(+)-Galactose	J	0.2 M NDSB-201
E	0.2M NDSB-221	K	0.1 M Guanidine hydrochloride
F	0.1 M NDSB-256	L	3 % w/v D-(+)-Glucose monohydrate

Figure 6-1 Crystals of the D6E large terminase protein

Crystals obtained using Hampton additive screen are shown in (A-L). The concentration and chemical compositions of the additives are shown in the table below.

Three data sets were collected at various wavelength on the crystal at the peak (1.8926 Å), inflection (1.8942 Å) and high energy remote (1.7101 Å) of the theoretical X-ray anomalous scattering curve for manganese (Table 6-1, Figure 6-2). Matthew's coefficient calculation of the cell content suggested that only one molecule was found in the unit cell. The structure was determined following the protocol described in Chapter 2. The structure determined for this crystal form contained both an ordered nuclease domain and ATPase lid subdomain of the ATPase domain (residues 200-417). However, only weak electron density was observed for the ASCE subdomain of the ATPase domain. The central β-sheet region and most of the α-helices were disordered (Figure 6-2).

Table 6-1 X-ray data collection statistics (D6E large terminase crystal form 1, table reproduced from (186))

Crystal form	1 Peak	1 Inflection	1 High energy remote
Wavelength (Å)	1.8926	1.8942	1.7101
Space group	R32	R32	R32
Unit-cell a, b, c (Å)	103.3, 103.3, 268.9	103.6, 103.6, 269.3	103.8, 103.8, 268.8
Unit-cell $\alpha$ , $\beta$ , $\gamma$ (°)	90.0, 90.0, 120.0	90.0, 90.0, 120.0	90.0, 90.0, 120.0
Resolution(Å)	46.09-2.90 (3.08- 2.90)	46.18-2.90 (3.08- 2.90)	46.13-2.90 (3.08- 2.90)
Rmerge (%)	9.7 (114.3)	8.2(123.4)	7.2(109.5)
I/ $\sigma$ <I>	12.3(1.8)	13.8(1.7)	15.9(2.0)
Completeness (%)	100(99.9)	99.7(98.4)	99.8(99.1)
Multiplicity	9.4(9.0)	9.4(8.8)	9.0(8.5)
CC <sub>1/2</sub>	0.996(0.695)	0.998(0.660)	0.999(0.664)

Values in the parentheses are for the outermost resolution shell.

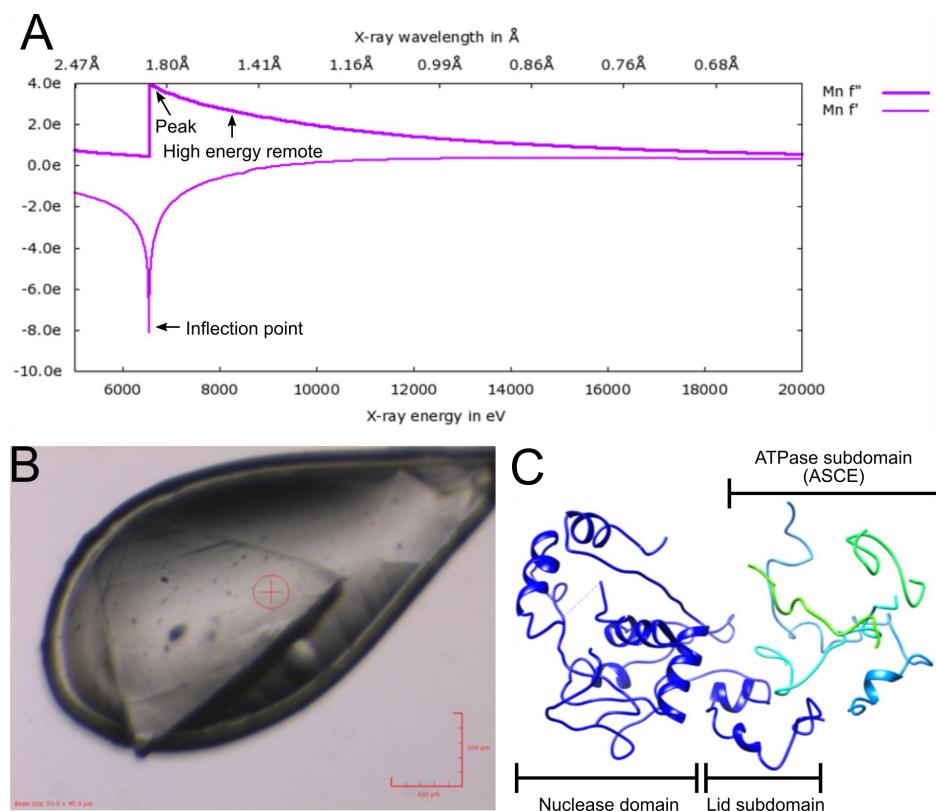


Figure 6-2 Structure determination of the D6E large terminase crystal form 1 by MAD

(A) Theoretical X-ray anomalous absorption edge of manganese. (B) The crystal used for MAD data collection. (C) The structure containing a disordered ATPase domain obtained after model building.

In order to determine the structure for both ATPase and nuclease domains, the protein was screened under broader crystallisation conditions. It was expected that the protein may be able to pack in a different manner to that observed in crystal form 1 under other crystallisation conditions. In this way, an alternative crystal form is likely to be used for high resolution structure determination. However, after screening of a broader range of crystallisation conditions, such as PGA (187), MPD (188) and Morpheus (189), no diffracting crystals were obtained (Figure 6-3).



0.1mM Sodium cacodylate,  
5% PGA-LM, 20% PEG 3350

0.1mM Sodium cacodylate,  
5% PGA-LM, 20% PEG 2K MME

0.1mM Sodium cacodylate, 5% PGA-LM, 12% PEG 8K

*Figure 6- 3 Crystals obtained using PGA screen*

*The crystallisation conditions are shown below each of the three (A-C) picture. None of these crystals exhibited promising X-ray diffraction. Scale bars (50 µm) are indicated on each image.*

## 6.2 Post-crystallisation manipulation and structure determination (crystal form2)

As described above, the structure of the D6E large terminase determined from crystal form 1 contains an order nuclease domain and a disordered ATPase domain. As reported previously for other large terminase proteins from bacteriophage T4 (40), P22 (68) and herpes virus (158), the crystallisation is difficult and is possibly hindered by the flexible inter-domain conformations. Therefore, it is likely that the disorder of the D6E large terminase protein ATPase domain is caused by heterogenous inter-domain conformations during molecule packing and crystal growth. Post-crystallisation treatments, including annealing, dehydration, soaking and cross-linking have been successfully utilised in many cases to improve the quality of the crystal (190,191). Annealing was firstly tried on the crystals containing the disordered ATPase domain. After the frozen crystal was mounted, the cryo stream was blocked for around 2 seconds, followed by re-cooling the crystal again using the cryo stream for 10 seconds. This process was repeated 3 times. The diffraction of the crystal had no apparent improvement after this treatment. Slow dehydration by equilibrating the drops containing the crystal or directly soaking of crystals with increased concentrations of precipitants has been reported as a promising method to improve diffraction. This is especially suitable for crystals with high solvent content. Following this method, the crystals of the D6E large terminase obtained in 1.2 M of ammonium sulfate were equilibrated or soaked with 1.4 to 2.0 M of ammonium sulfate with or without additives that have been used previously (192), for 2 hours or overnight (Table 6-2). It was observed that the crystals remain intact and diffract to similar resolutions when being dehydrated using this method.

*Table 6-2 Crystal dehydration using ammonium sulfate at various concentrations in the presence of additives*

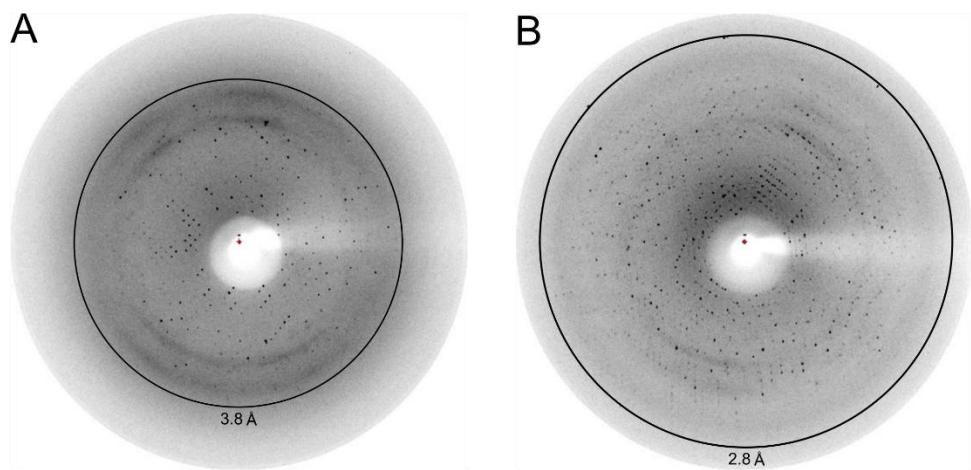
No additive	5% PEG3350	5% PEG8000	5% MPD	5% Glycerol	5% Ethylene glycol
1.2M	1.2M	1.2M	1.2M	1.2M	1.2M
1.6M	1.6M	1.6M	1.6M	1.6M	1.6M
2.0M	2.0M	2.0M	2.0M	2.0M	2.0M
2.4M	2.4M	2.4M	2.4M	2.4M	2.4M

For improvement of resolution, crystals were sometimes soaked directly into a cryo-protectant containing solutions with different chemical and physical properties compared to the precipitant used in crystal growth. Inspired by the success in achieving the desired improvement of the diffraction (190,191), a screen of a cryo-protectant conditions was conducted for the crystals containing the disordered ATPase domain (Table 6-3). Some conditions used in this experiment were based on the Hampton Cryoprotectant Screen (193), such as sodium chloride, ethylene glycol and glycerol.

*Table 6-3 Conditions that are tested as Cryo-protectants for crystals of the D6E large terminase*

1	25% Glucose	25% Sucrose	30% Ethylene glycol
2	4 M $(\text{NH}_4)_2\text{SO}_4$	4 M $(\text{NH}_4)_3\text{PO}_4$	4 M $\text{NH}_4\text{OAc}$
3	2.6 M $\text{Li}_2\text{SO}_4$	3.4 M NaMalonate	2.5 M NaCl
4	6 M LiCl	<b>5 M NaCl</b>	30% Glycerol

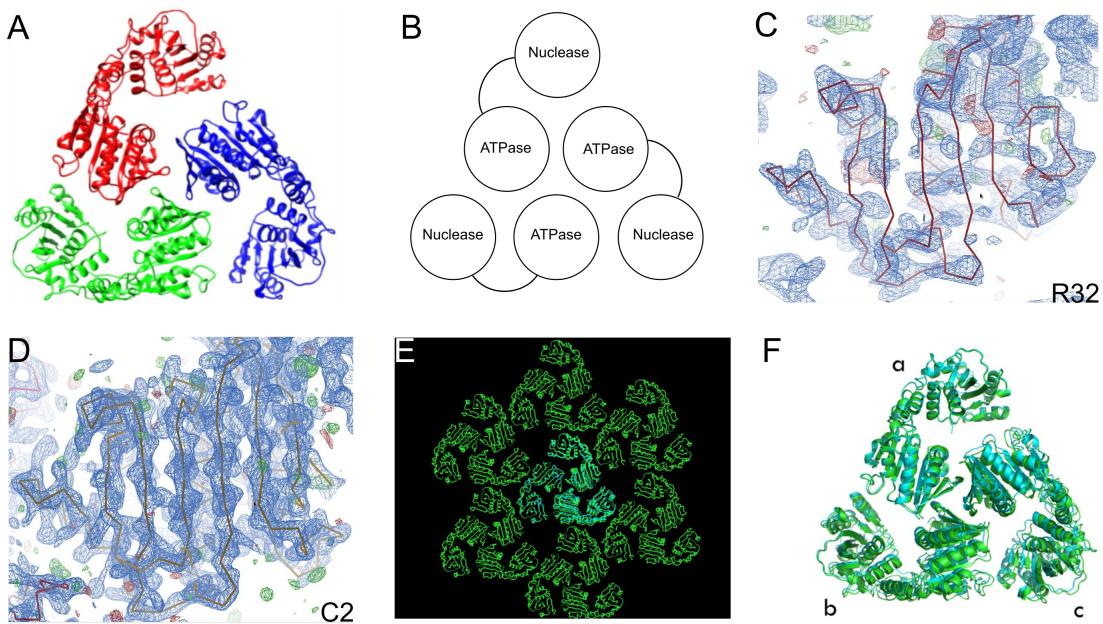
It was found that ammonium sulfate, ammonium phosphate,  $\text{Li}_2\text{SO}_4$ , NaMalonate or Glycerol have no significant effect on the diffraction of the crystal. When Glucose, Sucrose or ammonium acetate were used, a reduction of diffraction quality was observed. In the presence of ethylene glycol or LiCl, the crystals soaked dissolved after a few seconds. Notably, the diffraction power of the crystal was improved by 1 Å when 5 M sodium chloride was used as the cryo-protectant (Figure 6-4). It was later observed that for a successful significant improvement in resolution, triggered by space group transition, it was crucial to make fresh 5.5 M sodium chloride solutions and keep the concentration as high as possible before soaking. Several other salts were then identified which could be used at concentrations that are less close to saturation while achieving similar space group transition and improvement in resolution, including 3.5 M ammonium citrate, 3 M potassium acetate and 4 M sodium formate. Interestingly, the crystal obtained by sodium chloride soaking belonged to a different space group compared to the previous disordered crystals. Matthew's coefficient calculation reveals that there are three molecules in the asymmetric unit. After X-ray data collection at the synchrotron, the structure was solved by molecular replacement, using the structure of the nuclease domain as determined from crystal form 1 (Figure 6-2). Subsequence model building follows the protocol described in Chapter 2.



*Figure 6-4 X-ray diffraction patterns*

The diffraction patterns are shown for a crystal before (A) and after (B) NaCl soaking. The crystals used in this test were harvested from the same drop and were of similar size.

Following space group transition, three molecules were observed in the unit cell. The packing of the molecules is shown in Figure 6-5. Compared to the previous crystal form 1 (R32) which contains a disorder ATPase domain, no disorder in the electron density of the ATPase domain was observed for the new crystal form (space group C2, crystal form 2). Although only one molecule was found in the unit cell for crystal form 1, a trimeric packing like that observed for crystal 2 were also observed (Figure 6-5). This indicates that upon soaking the space group transition from R32 to C2 is likely due to small conformational changes of the molecule which resulted in breakage of the three-fold symmetry. To prove this, molecule (a) of the trimer of the crystal form 2 was superimposed on one of the three molecules in crystal form 1. As expected, small conformational changes between the ATPase and nuclease domains were observed, which must have been caused by soaking with sodium chloride (Figure 6-5). In the structure, around 10 residues at the very N-terminal and C-terminal regions were not visible in the electron density maps. Subsequently, in the structure determined by soaking with 3.5 M ammonium citrate, a bound sulfate ion was found in the ATPase active site. This structure has an ordered C-terminal region which is in close proximity to the bound sulfate (see below).



*Figure 6-5 Structure and comparisons of packing in crystal forms 1 and 2 of the D6E large terminase*

(A) Structure of the trimeric D6E large terminase protein. (B) Representative diagram showing the molecular contacts within an asymmetric unit. (C) and (D): Comparison of the electron density of the ATPase domain between crystal form 1 and 2. (E) and (F): Superposition of one of the three chains (chain A) of the trimer from crystal form 2 (cyan) to the monomer in crystal form 1 (green).

*Table 6-4 X-ray data collection and refinement statistics (D6E large terminase crystal form 2, table reproduced from (186))*

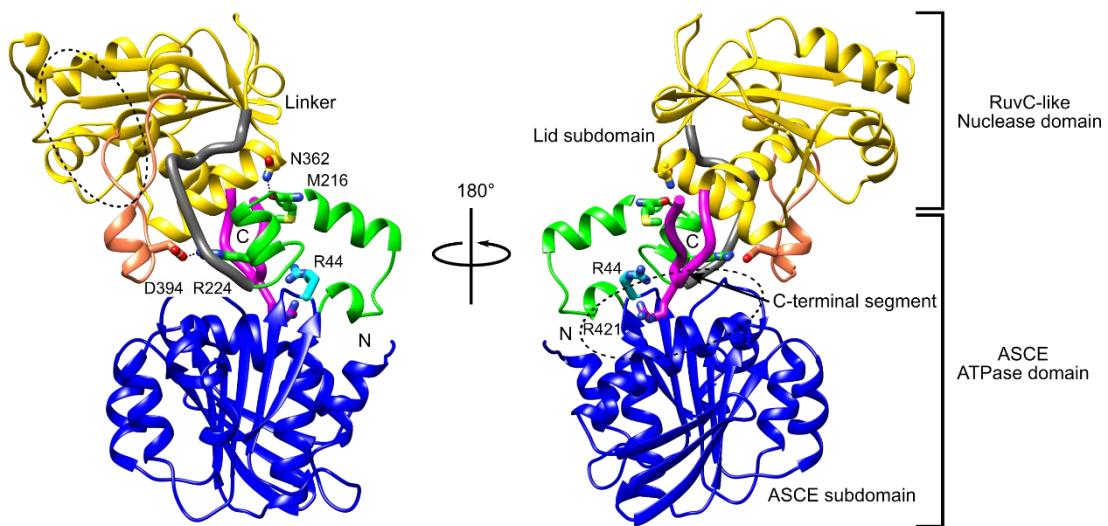
Crystal form	2	2*
	HR	$\text{SO}_4^{2-}$
Wavelength (Å)	0.9763	0.9763
Space group	C2	C2
Unit-cell a, b, c (Å)	180.2, 101.3, 108.5	181.3, 102.3, 110.3
Unit-cell $\alpha$ , $\beta$ , $\gamma$ (°)	90.0, 124.9, 90.0	90.0, 124.9, 90.0
Resolution(Å)	45.90-2.10 (2.14-2.10)	46.65-2.40 (2.46-2.40)
Rmerge (%)	7.6(123.3)	6.2(109.8)
$I/\sigma < I >$	9.5(1.3)	13.6(1.2)
Completeness (%)	99.2(99.4)	99.0(94.1)
Multiplicity	4.1(4.2)	4.2(3.8)
$\text{CC}_{1/2}$	0.997(0.448)	0.999(0.476)
Refinement		
Rwork/Rfree	0.187/0.236	0.195/0.233
Mean B factor(Å <sup>2</sup> )	51	73
R.M.S.D.		
Bond lengths (Å)	0.009	0.010
Bond angles (°)	1.4	1.3
Ramachandran (%)		
Favored	96.43	96.85
Allowed	3.57	3.15
Outliers	0	0

Values in the parentheses are for the outermost resolution shell.

HR: Highest resolution.

### 6.3 Structure of the D6E large terminase protein

As mentioned in the above sections, there are three molecules in the asymmetric unit. Slightly different domain orientations were observed for each of them, with an overall intersubunit C $\alpha$  rmsd of 0.8 -1.6 Å. The N-terminal ATPase and C-terminal nuclease domains are joined through an ordered linker, amino acids 230-241 (thick ribbon). Like other large terminase ATPases (47,56,57), the D6E ATPase domain has a C-terminal lid subdomain above the N-terminal ASCE subdomain while the nuclease domain resembles the RuvC fold, as seen previously (175,194) (Figure 6-6). Interestingly, the equivalent structural element to the nuclease  $\beta$ -hairpin, previously predicted to interact with DNA (67,68,81,195), forms an extended hairpin-like loop with an  $\alpha$ -helix at its tip. The relative position of the ATPase and nuclease domains was stabilised by a salt bridge between R224 on the ATPase lid subdomain and D394 on the extended hairpin-like loop and an interdomain hydrogen bond (M216-N362). In addition, a C-terminal segment (aa. 418-427) from the nuclease domain is found in proximity to the ATPase active site (Figure 6-6).



*Figure 6-6 Overall structure of the D6E large terminase*

*The structure is shown as a ribbon diagram. The ATPase domain is coloured in blue (ASCE subdomain) and green (lid subdomain). The nuclease domain is coloured in gold. The ATPase and nuclease active sites are indicated by dashed ovals. The linker region (grey) and the C-terminal segment (magenta) are shown using thick ribbons. The  $\beta$ -hairpin equivalent (salmon) and the residues (sticks) of the salt bridge formed between this element and the lid subdomain is indicated using dashed lines. The sulfate ion bound in the ATPase active site is omitted for clarity (figure reproduced from (186)).*

### 6.4 Effects of ATP analogues on protein thermostability

As DNA packaging motor proteins, the large terminases use energy generated by ATP hydrolysis to power DNA translocation (83). To characterise the binding of ADP and non-hydrolysable ATP analogues to the D6E large terminase protein, the effect of ADP and non-hydrolysable ATP analogues including ATP- $\gamma$ -S and AMP-PNP on the thermostability of

full-length D6E large terminase were quantified by fluorescence-based thermal shift assays (Figure 6-7). Addition of 5 mM magnesium chloride conferred a 1.2 °C increase in the Tm with subsequent addition of ADP or non-hydrolysable ATP analogues resulting in larger increases in Tm. ATP- $\gamma$ -S, significantly stabilised the protein producing a  $\Delta T_m$  of 7.8, while a modest increase ( $\Delta T_m=2.3$ ) was induced by ADP. Strikingly, addition of AMPPNP produced a much smaller increase in Tm ( $\Delta T_m=1.4$ ) compared to that of ATP- $\gamma$ -S or even ADP, likely attributed to conformational difference of the  $\gamma$ -phosphates in AMP-PNP, which may alter binding affinity (Figure 6-7).

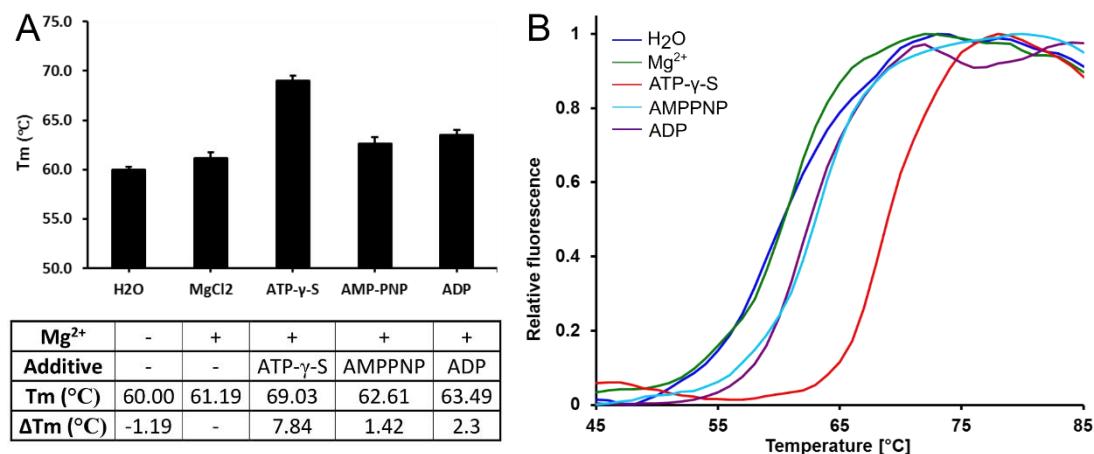


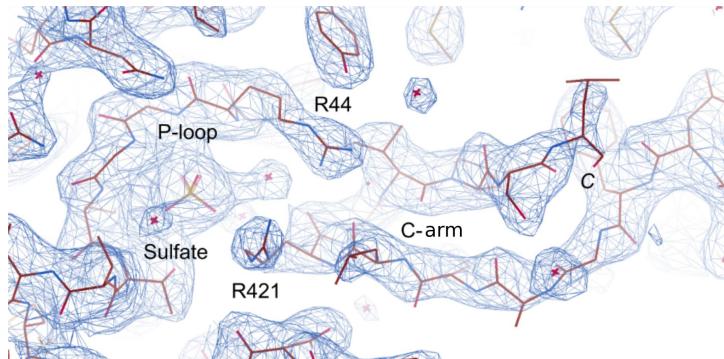
Figure 6-7 Protein stabilisation by binding of ATP analogues

(A) Plot of the observed melting temperatures in the presence and absence of ADP, ATP and nonhydrolyzable ATP analogues, showing the average and standard deviation for three independent experiments. (B) Representative melting curves are shown within a temperature range of 45–85 °C (figure reproduced from (186)).

## 6.5 Crystallisation of full-length protein in complex with ADP and non-hydrolysable ATP analogues

As described earlier, the D6E large terminase can bind non-hydrolysable ATP analogues and hydrolyse ATP (see Chapter 3, malachite green phosphate assays). To further understand the structural basis for ATP binding and hydrolysis, crystallisation experiments were performed in an attempt to obtain structures in complex with ADP and non- hydrolysable ATP analogues. Initial experiments were performed by soaking crystals of wild type protein with ADP and non-hydrolysable ATP analogues. However, co-crystallisation or soaking using ATP- $\gamma$ -S abrogated the diffraction. In the presence of ADP, AMP-PNP or AMP-PCP, the diffracting power was retained but no ligand was observed in the electron density. Subsequent co-crystallisation trials with ADP or non-hydrolysable ATP analogues using the wild type protein resulted in crystals with the same characteristics as obtained in soaking experiments. Investigation of the electron density around the ATPase active site indicates

that a C-terminal arm (418-427) of the nuclease domain may occlude the ATP binding site (Figure 6-8).

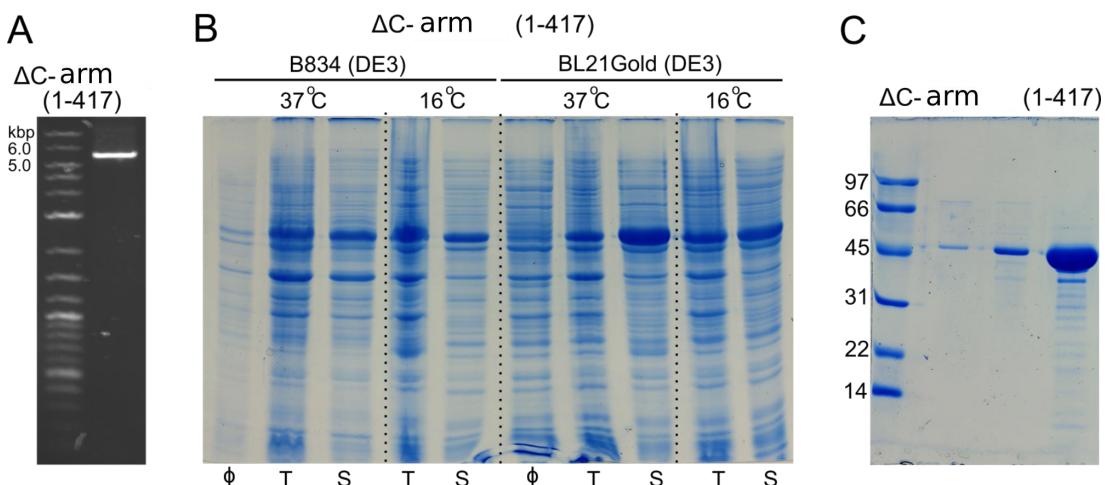


*Figure 6-8 Electron density of the ATPase active and the C-terminal arm*

*The map coloured in blue is contoured at 1  $\sigma$  (2Fo-Fc). The C-terminal arm, including R421A is in proximity to the P-loop of the ATPase active site.*

## 6.6 C-terminus truncation

Initial attempts using full-length protein in obtaining crystal structures of the D6E large terminase in complex with ADP and non-hydrolysable ATP analogues failed. It was observed above that a C-terminal arm (418-427) of the nuclease domain may affect ATP binding at the ATPase active site, hence a new construct with truncated C-terminal arm (1-417) was cloned, expressed, purified and used in further crystallisation trials (Figure 6-9).



*Figure 6-9 PCR amplification, expression test and purification*

(A) Agarose gel electrophoresis showing the PCR amplified recombinant plasmid of the C-terminally truncated D6E large terminase. (B) SDS-PAGE showing the expression and solubility level under tested conditions.  $\Phi$ : before IPTG induction, T: total cell lysate, S: Supernatant after cell lysis. (C) SDS-PAGE after size exclusion chromatography showing the purity and molecular weight of the protein.

## 6.7 Crystallisation and structure determination of complexes with ADP and non-hydrolysable ATP analogues

Crystals were grown in the same conditions as the full-length protein, but were soaked in 4 M sodium formate in the presence of 50 mM ATP- $\gamma$ -S or 100 mM ADP for 16h, with 100 mM MgCl<sub>2</sub> provided to facilitate nucleotide binding. To produce apo crystals, the same soaking method was used but with the omission of nucleotides and MgCl<sub>2</sub>. While the soaked crystals of this new construct retained diffraction power, the crystal form was different to that obtained with the wild type protein (space group P1, crystal form 3).

Data sets were collected at the synchrotron using crystals soaked with and without ADP and non-hydrolysable ATP analogues (Table 6-5). The Matthew's coefficient calculation indicated 6 molecules in the asymmetric unit (Figure 6-10). The structures were determined by molecular replacement, using the crystal form 2 structure as the search model. Of the 6 molecules in the asymmetric unit, density for the ATP- $\gamma$ -S was observed only for 4 molecules while only one molecule contained density corresponding to ADP. In the structure determined by crystal soaking with AMPPNP, density corresponding to AMPPNP was found at the nuclease domain close to a crystal contact region (Figure 6-10). It is likely that due to the low affinity of AMPPNP, observed in thermal shift assays (Figure 6-7), the molecule binds at a non-specific site in the crystal.

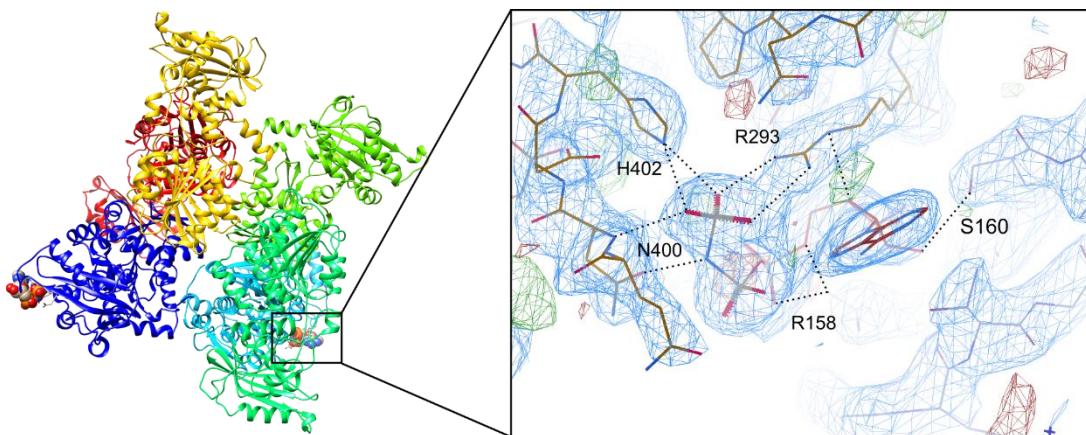


Figure 6-10 Structure of the D6E large terminase (space group P1, crystal 3) in complex with AMPPNP

(left) The structure is shown using ribbon diagram. The 6 molecules in the unit cell are coloured differently. The two AMPPNP found bound to the crystal contact regions are shown in spheres. (right) Salt bridges and hydrogen bonds formed between AMPPNP and residues from the protein are shown together with the 2Fo-Fc map which is contoured at 1  $\sigma$ .

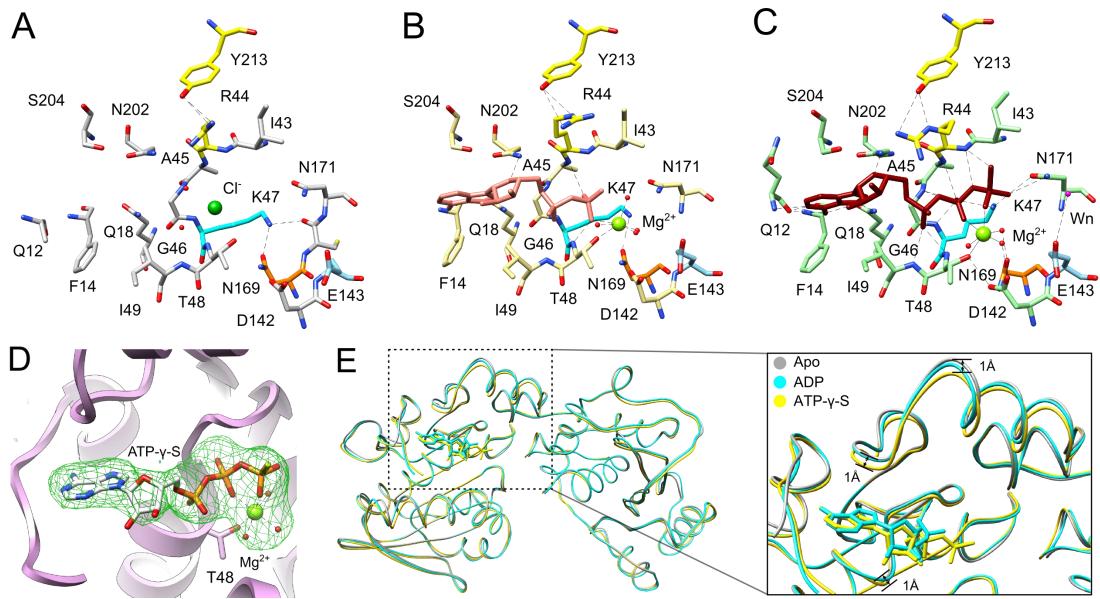
*Table 6-5. X-ray data collection and refinement statistics (D6E large terminase crystal form 3, table reproduced from (186))*

Crystal form	3	3 ATP- $\gamma$ -S	3 ADP
Wavelength (Å)	0.9763	0.9763	0.9763
Space group	P1	P1	P1
Unit-cell a, b, c (Å)	100.8, 102.6, 103.1	101.4, 102.9, 103.0	100.8, 102.8, 103.0
Unit-cell $\alpha$ , $\beta$ , $\gamma$ (°)	90.7, 119.4, 117.4	90.9, 119.3, 117.1	91.1, 119.1, 117.1
Resolution(Å)	45.69-2.60 (2.65- 2.60)	45.90-3.00 (3.08- 3.00)	45.80-3.10 (3.19-3.10)
Rmerge (%)	10.3(115.3)	7.5(81.1)	16.4(83.2)
I/ $\sigma$ <I>	7.7(1.0)	6.3(1.0)	4.0 (1.0)
Completeness (%)	97.7(93.8)	96.5(96.5)	96.4(94.0)
Multiplicity	3.5(3.4)	1.9(2.0)	1.8(1.7)
CC <sub>1/2</sub>	0.995(0.331)	0.996(0.379)	0.972(0.337)
Refinement			
Rwork/Rfree	0.194/0.255	0.189/0.243	0.221/0.281
Mean B factor(Å <sup>2</sup> )	71	79	71
R.M.S.D.			
Bond lengths (Å)	0.011	0.007	0.013
Bond angles (°)	1.5	1.1	1.7
Ramachandran (%)			
Favored	94.11	95.30	95.38
Allowed	5.89	4.70	4.62
Outliers	0	0	0

Values in the parentheses are for the outermost resolution shell.

## 6.8 Structures of complexes with ATP analogues

The structures in complex with ADP and ATP- $\gamma$ -S reveal a novel active site architecture with small overall conformational changes upon binding of ATP analogues. Firstly, the side chain of R44 which hydrogen bonds to Y213 from the lid subdomain adopts different conformations upon the binding of ATP analogues. Secondly, the hydrogen bond formed by the P-loop lysine K47 with side chains of N169 in the apo and ADP bound structures was not observed in the structure with bound ATP- $\gamma$ -S. Instead, the side chain amine of this lysine formed salt bridges with the  $\beta$ - and  $\gamma$ -phosphates of the bound nucleotide. N169 also undergoes a side chain rotation to hydrogen bond to an inner shell water molecule of the catalytic magnesium ion (Figure 6-11). Notably, comparisons of the overall conformational differences of the structures in complex with ATP- $\gamma$ -S or ADP ( $C\alpha$  rmsd 0.3-0.4 Å) were smaller than those observed for the three molecules within an asymmetric unit ( $C\alpha$  rmsd 0.8-1.6 Å). The lid subdomain and the P-loop are the major regions where the nucleotide binding induced conformational changes were observed (up to 1 Å  $C\alpha$  rmsd) (Figure 6-11).



*Figure 6-11 The ATPase active site*

Active site residues are shown as balls and sticks for (A) apo (grey), (B) bound ADP (khaki), (C) bound ATP- $\gamma$ -S (green). Electrostatic interactions of the ATP analogues with active site residues (black dashed lines) are shown along with the P-loop arginine R44 and the lid subdomain residue Y213 (yellow). The P-loop lysine K47 (cyan), N169 (orange) and the catalytic glutamic acid E143 (blue) which binds to a potential nucleophilic water (magenta) in the ATP- $\gamma$ -S complex are highlighted. (D) The omit maps, contoured at  $3\sigma$ , are shown for the structures with bound ATP- $\gamma$ -S. E. Superposition of the apo (grey), ADP (cyan) and ATP- $\gamma$ -S (yellow) complex structures. ADP and ATP- $\gamma$ -S are shown as sticks. Ca rmsd differences are indicated after superposition of the central  $\beta$ -sheet of the ATPase domain. (E) Superposition of the apo (grey), ADP (cyan) and ATP- $\gamma$ -S (yellow) complex structures. ADP and ATP- $\gamma$ -S are shown as sticks. Ca rmsd differences are indicated after superposition of the central  $\beta$ -sheet of the ATPase domain (figure reproduced from (186)).

## 6.9 Summary

In this chapter, the structure of the D6E large terminase was determined using three different crystal forms. In the structure determined by multi-wavelength anomalous dispersion (MAD) using crystal form 1, only the structure of the nuclease domain is visible whereas the ATPase domain is disordered. To solve this problem, a cryo-protectant optimisation method by soaking the crystals containing disordered ATPase in high concentration salt solutions was developed which led to interpretable electron density maps for the ATPase domain and allowed improvement of the resolution. Several types of salt additives were found to order the disordered ATPase domain, which caused transition from crystal form 1 to crystal form 2. As an ATPase motor protein which powers DNA translocation by ATP hydrolysis, the protein can bind ADP and non-hydrolysable ATP analogues. Among the three non-hydrolysable ATP analogues tested, ATP- $\gamma$ -S gave significant changes in melting temperature, indicating stronger binding to the protein than that for AMPPNP and AMPPCP. Although the protein binds ATP- $\gamma$ -S, initial attempts to obtain complex with this ATP

analogue were unsuccessful, until the removal of a C-terminal arm that was predicted to be in proximity with the ATPase active site. Structures with bound ADP and non-hydrolysable ATP analogues were finally determined using a protein construct that lacked the C-terminal segment (crystal form 3). Comparison of conformational changes upon binding of ATP analogues revealed insignificant differences, smaller than differences observed between the three molecules within the asymmetric unit. Several interactions with ATP analogues were identified, which are likely to be important for ATP hydrolysis and conformational changes that powers DNA translocation.

## 7 DNA binding properties of G20c and D6E large terminases

This chapter describes characterisation of the DNA binding properties of the G20c and D6E large terminases by size exclusion chromatography (SEC), electrophoretic gel mobility shift assays (EMSA), microscale-thermophoresis (MST), and analytical ultra-centrifugation (AUC). In addition, crystallisation and preliminary X-ray data analysis were also performed for the D6E large terminase complex with DNA.

### 7.1 DNA binding of the G20c large terminase

The DNA binding properties of the G20c large terminase protein and the C-terminal nuclease domain were investigated using EMSAs as described in Chapter 2. Initially, the H427A mutant was designed and used in the *in vitro* nuclease assays because the H427 side chain was found coordinating metal in the G20c nuclease domain structure complexed with Zn<sup>2+</sup> (see chapter 5). It was found later that this mutant shows excellent solubility compared to that of the wild type protein, which aggregates when the sodium chloride concentration is reduced below 200 mM. H427A is soluble in 10 mM potassium glutamate at a protein concentration of more than 50 mg/ml. Additionally, *in vitro* nuclease assays performed in Chapter 5 showed that H427A is nuclease activity deficient, making it suitable for DNA binding characterisation while preventing the DNA from being degraded during the experiment.

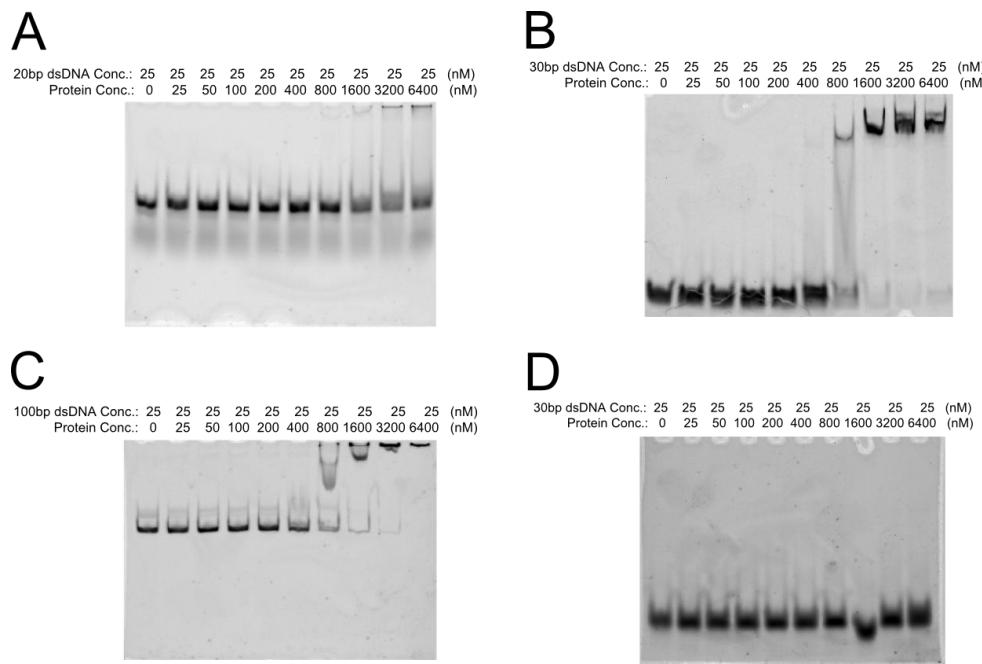
#### 7.1.1 Electrophoretic gel mobility shift assays

5'-6-carboxyfluorescein labelled dsDNAs with lengths of 20, 30 and 100 bp were used in EMSAs to characterise the DNA binding properties of the G20c large terminase protein and its C-terminal nuclease domain. The fluorescent probe was prepared by annealing the chemically synthesized 5'-6-carboxyfluorescein labelled 20 bp, 30 bp and 100 bp ssDNA (Table 7-1) with its non-fluorescent labelled complementary strand in a 1:1 molar ratio in a buffer containing 10 mM Tris pH 7.8, 50 mM NaCl and 1mM EDTA. Each fluorescent probe (25 nM) was mixed with an increasing concentration of protein in a 20  $\mu$ l reaction containing 10 mM HEPES pH 7.5 and 10 mM potassium glutamate. This assay follows the standard protocol described for electrophoretic gel mobility shift assays in Chapter 2. When 20 bp DNA was used in DNA binding assays, weak DNA binding was observed at a protein concentration of above 1600 nM. The DNA binding affinity appeared to be significantly increased when 30 bp DNA was used, at a protein concentration of 800 nM. Similar levels of binding affinity were observed when 100 bp DNA was used (Figure 7-1). The results indicate that for efficient binding, the dsDNA length needs to be more than 20 bp. The

apparent DNA binding affinity of the protein to a longer DNA with a length of 100 bp is similar to that of a 30 bp long dsDNA. The smearable band on the gel is an indication that the binding is likely to be multiple non-specific interactions. Interestingly, the nuclease domain alone does not show any binding, suggesting that the primary DNA binding region of the G20c large terminase protein comes from the ATPase domain.

*Table 7-1 The length and sequence of the dsDNA constructs used in electrophoretic gel mobility shift assays*

dsDNA length (bp)	Sequence
20	5'-GCAGGAGGGACGTAGGGTAGG
30	5'-GAGACCTGGAACGCGATCATTGGCAGCACG
100	5'- GGGACCTATTGGGTGCTTTTGTATAATTAGGTTATAT AAGG TTTATCATTAGATATGAGGTTCAAATATAGTTAAGGAGGT TTTC



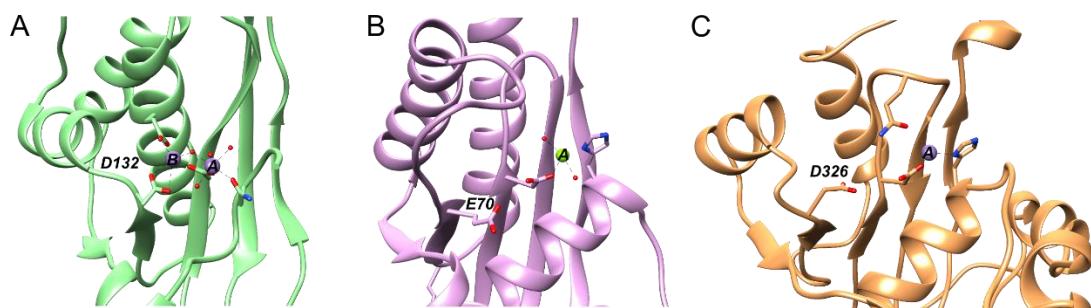
*Figure 7-1 DNA binding assays of the G20c large terminase protein and its C-terminal nuclease domain*

The DNA binding properties were probed by EMSA using fluorescent labelled DNA at (A) 20bp, (B)30bp, (C)100bp. for G20c large terminase and (D) 30bp for its C-terminal nuclease domain.

## 7.2 DNA binding of the D6E large terminase

### 7.2.1 Preparation and nuclease activity of the D6E large terminase nuclease deficient mutant

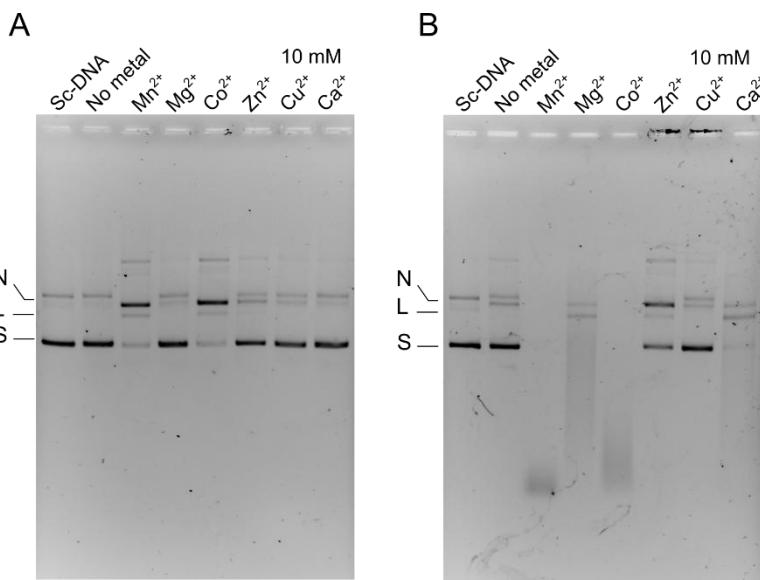
To characterise the DNA binding properties of the D6E large terminase, the nuclease active site mutant D326N was cloned, expressed, and purified to prevent the DNA from being cleaved by the large terminase nuclease during the binding assays. The D326N mutation was introduced at a residue predicted to be important for the catalytic metal ion B coordination, based on structural superposition with the Bh-RNase H (153) and Th-RuvC (184). Equivalent mutations of this residue in Bh-RNase H and Th-RuvC were used to prevent degradation of an RNA/DNA hybrid and holliday junction DNA, respectively (153,175) (Figure 7-2).



*Figure 7-2 Structure comparisons of the nuclease active site*

*Ribbon diagrams showing structure of the metal binding sites of A. Bh-RNase with bound manganese, B. Th-RuvC with bound magnesium. C. D6E large terminase nuclease domain with bound manganese.*

The nuclease activity of the D326N mutant was tested by mixing excess protein with the supercoiled DNA in the presence of different divalent metals. The results show that in the presence of various metal ions, the nuclease activity of the D326N mutant is significantly reduced compared to the WT, although some residual activity was observed when manganese and cobalt were used as metal co-factors.



*Figure 7-3 Comparison of in vitro nuclease activities of the wild type D6E large terminase protein and D326N mutant*

The nuclease activity in the presence of various metal ions at 10 mM concentration were shown for D6E large terminase protein (1 µM) A. D326N mutant, B. wild type.

### 7.2.2 Electrophoretic gel mobility shift assays

To investigate the effects of DNA length on the DNA binding affinity for the D6E D326N large terminase, EMSAs were performed using 5'-6-carboxyfluorescein labelled poly(T) ssDNA of various lengths (5 to 50 bases), ssDNA with a length of 20bp (Table 7-1) and its dsDNA form annealed with non-fluorescent labelled complementary strand. The D6E D326N large terminase was titrated into the DNA solution at increasing concentrations up to 6400 nM. This assay follows the general protocol as used for the characterisation of the DNA binding properties of the G20c large terminase protein. Only weak DNA binding was observed when poly(T) 5 was used. When the DNA length of the poly(T) was increased, an enhancement of DNA binding affinity was observed, as seen for poly(T) 15 and 20. Further increase in length of poly(T) from 20 to 50 gives a moderate increase of the DNA binding affinity.

To investigate the effects of DNA type on DNA binding affinity, 20 bp ssDNA and dsDNA with random sequence were used (Table 7-1). A modest increase in DNA binding affinity was observed when ss or ds DNA with random sequence were used compared to poly(T) at 20 bp under the tested conditions.

In summary, the results of the above DNA binding assays show that the DNA binding affinity of D6E large terminase is effected by the length of the DNA used. The binding affinity is weak for short DNA lengths while stronger binding was observed as the length of

DNA was increased. In addition, it was observed that the protein binds to a random ssDNA and dsDNA sequence more tightly than to the poly(T) ssDNA.

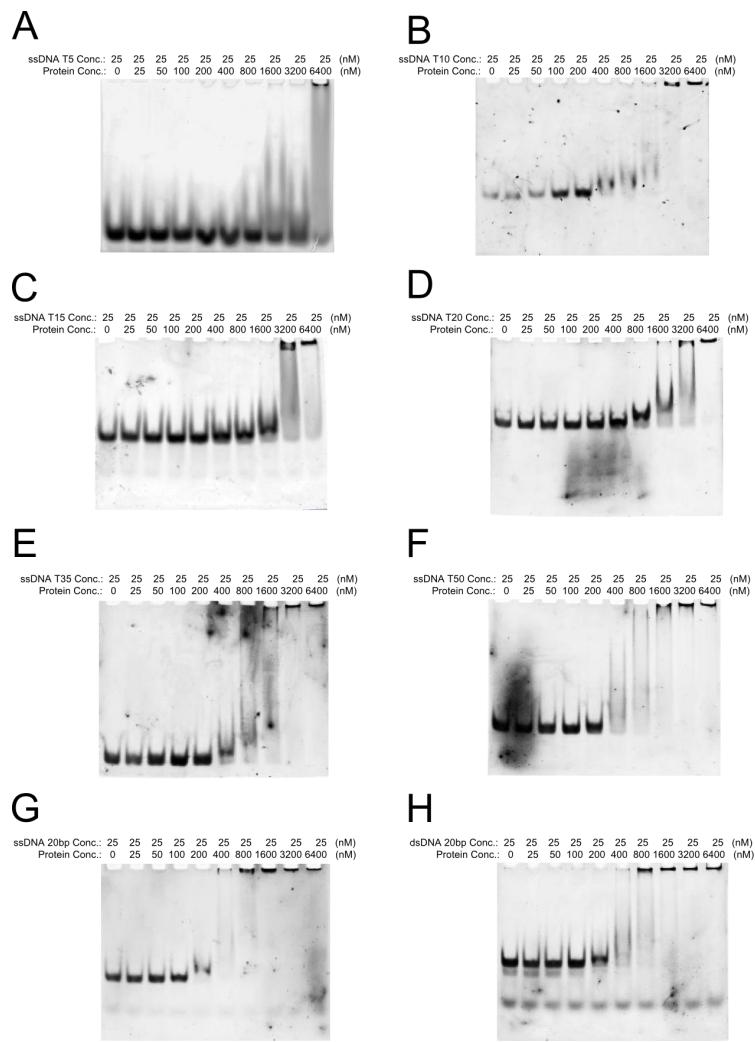


Figure 7-4 DNA binding assays for D6E D326N large terminase

(A-F): Poly(T) ssDNA, (G): random sequence ssDNA and (H): random sequence dsDNA.

### 7.2.3 Microscale thermophoresis

To further quantify the DNA binding affinity of the D6E large terminase, as well as obtain information on binding cooperativity, MST binding assays were performed. In this assay, 5'-6-carboxyfluorescein labelled poly(T) ssDNA at lengths of 10, 15, and 50 were used at a concentration of 25 nM. Samples were prepared in the same way as for EMASAs, except that 16 samples of various protein concentrations were prepared instead of 9 for each binding reaction. All samples were loaded into Monolith NT.115 standard treated capillaries and analysed on a NT.115 series MST instrument (Nano temper).

The temperature jump data under various protein concentrations were normalised and fitted by the Hill equation, from which the half maximal effective concentration (EC50, in this case it refers to the protein concentration where 50% of the protein molecules bind to DNA)

and cooperativity of binding were determined. Consistent with the observations from EMSAs, the binding affinity of the D6E large terminase D326N to DNA shows a dependence on the length of the DNA added. The apparent EC50 for Poly(T) 10, 15, 20 and 50 to the D6E large terminase D326N is 2100, 1000, 850 and 740 nM, respectively (Figure 7-5). Interestingly, the EC50 undergoes a two-fold decrease when the length of the DNA decreased from 10 to 20 nucleotides. Further elongation of the DNA does not produce a massive decrease in the EC50. This indicates that the minimum DNA length for apparent DNA binding is likely between 10 and 15 bases for poly (T) ssDNA. It was also found that the binding is highly cooperative. The Hill constants were determined at 3, 7, 12 and 26 for Poly(T) 10, 15, 20 and 50, respectively (Figure 7-5). Consistent with the observed massive decrease in EC50, there is also a significant increase of the binding cooperativity when the DNA length was increased from 10 to 20. These results support previous the suggestion that the minimum DNA length for apparent binding is between 10 to 20 for the D6E large terminase D326N. It is likely that when DNA with a length of 10 or less was used, the DNA was too short to cover the DNA binding region of the protein. The reduced DNA length could reduce the EC50 and the binding cooperativity.

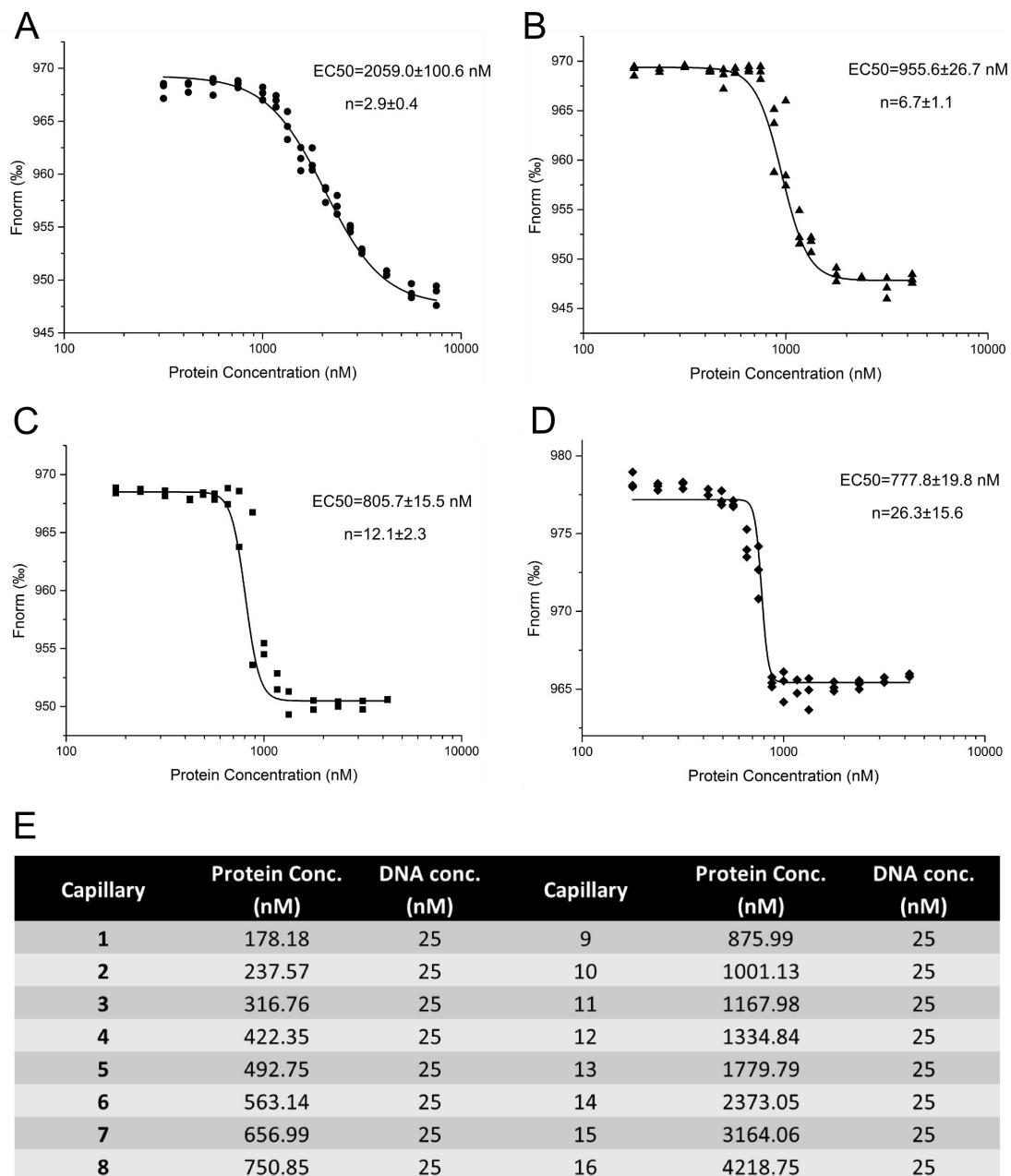


Figure 7-5 MST analysis of the DNA binding affinity and cooperativity of poly(T) ssDNA

(A) 10, (B) 15, (C) 20 and (D) 50 to the D6E large terminase D326N mutant were analysed by micro-scale thermophoresis. The T-jump data (temperature jump) was fitted by Hill equation. The EC<sub>50</sub> and Hill constant (*n*) are shown in each panel. The protein and DNA concentrations in this experiment are shown in (E).

To investigate the stability of the protein-DNA complex formed in this experiment, a competitive assay was developed by titrating non-fluorescent (native) poly(T) ssDNA to protein-bound 5'-6-carboxyfluorescein labelled poly(T) 20 ssDNA. Data were collected in the same way as mentioned above for the determination of EC<sub>50</sub> and binding competitiveness. Even at the maximum concentration of native Poly(T) 20 used, the data points still did not reach plateau (Figure 7-6). It was estimated that EC<sub>50</sub> of competition is around 15 times higher than that observed for complex formation. This is an indication that the 5'-6-

carboxyfluorescein labelled poly(T) ssDNA forms stable protein-DNA complex with the D6E large terminase D326N mutant.

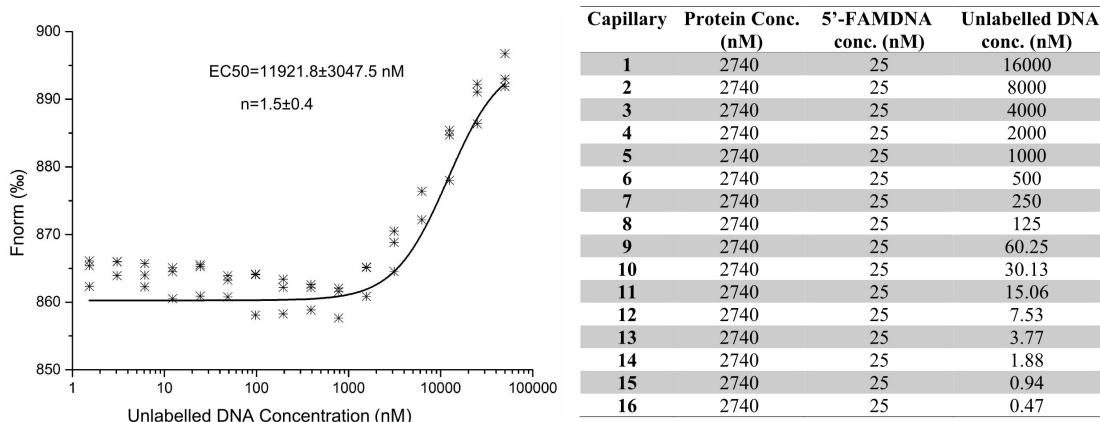


Figure 7-6 MST analysis of the stability of protein-poly(T) ssDNA complex

The pre-formed D6E large terminase D326N-Poly(T) 20 ssDNA labelled with 5'-FAM is titrated using unlabelled Poly(T) 20 ssDNA with no fluorescent label. Samples were analysed using micro-scale thermophoresis. The T-jump data (temperature jump) were fitted by Hill equation. The EC<sub>50</sub> and Hill constant (*n*) are shown in each panel.

#### 7.2.4 Analytical ultra-centrifugation

To probe the stoichiometry of the protein-ssDNA complex, sedimentation velocity experiments were performed using analytical ultra-centrifugation. Based on the observation in gel shift assays and microscale thermophoresis on the minimum length of DNA for apparent protein-DNA interaction for this protein, poly(T) ssDNA at a length of 20 base with either 5'-FAM or 3'-FITC labels were used in this experiment. The experimental set-up has been described in Chapter 2.

Sedimentation profiles of the D6E large terminase protein D326N in the presence of ssDNA show that two species of protein-DNA complex with different sedimentation coefficients were observed when different protein concentrations were used. At 10 µM protein concentration, the sedimentation coefficient of the protein-DNA complex peak corresponded to a molecular mass of a 1:1 protein-DNA complex. A further increase in the protein concentration to 100 µM resulted in a higher sedimentation coefficient than that at 10 µM protein concentration. The molecular mass of the second protein-DNA complex peak is higher than for the 1:1 protein-DNA complex, but lower than for the 2:1 protein-DNA complex, indicating that the second peak might be a rapidly re-associating intermediate with the protein: DNA ratio between 1:1 and 2:1. The data were processed using the method mentioned in Chapter 2 with an example shown in the Appendix 4.

The sedimentation coefficients determined for the protein-ssDNA complex were very similar when using 5'-FAM or 3'-FITC labelled fluorescent DNA. This indicated that the

fluorescent tag at either end of the ssDNA had no significant effect on the stoichiometry of the complex. Despite high similarity in sedimentation coefficients, the observed peak shapes for the 5'-FAM and 3'-FITC labelled DNA were different. In the presence of 3'-FITC labelled fluorescent DNA, the peak is wider than when 5'-FAM labelled fluorescent DNA was used. This is an indication that complex with the 3'-FITC labelled fluorescent DNA is more heterogeneous undergoing more rapid re-equilibration over the time course of the experiment than the complex with the 5'-FAM labelled fluorescent DNA. The frictional ratio was observed to be 2.0 and 1.0-1.1 for DNA and protein-DNA complex, respectively. This indicated that the overall shape of the protein-DNA complex is globular. Subsequent experiments focused on increasing the protein concentration to saturated protein-DNA binding as well as obtaining further stoichiometric information.

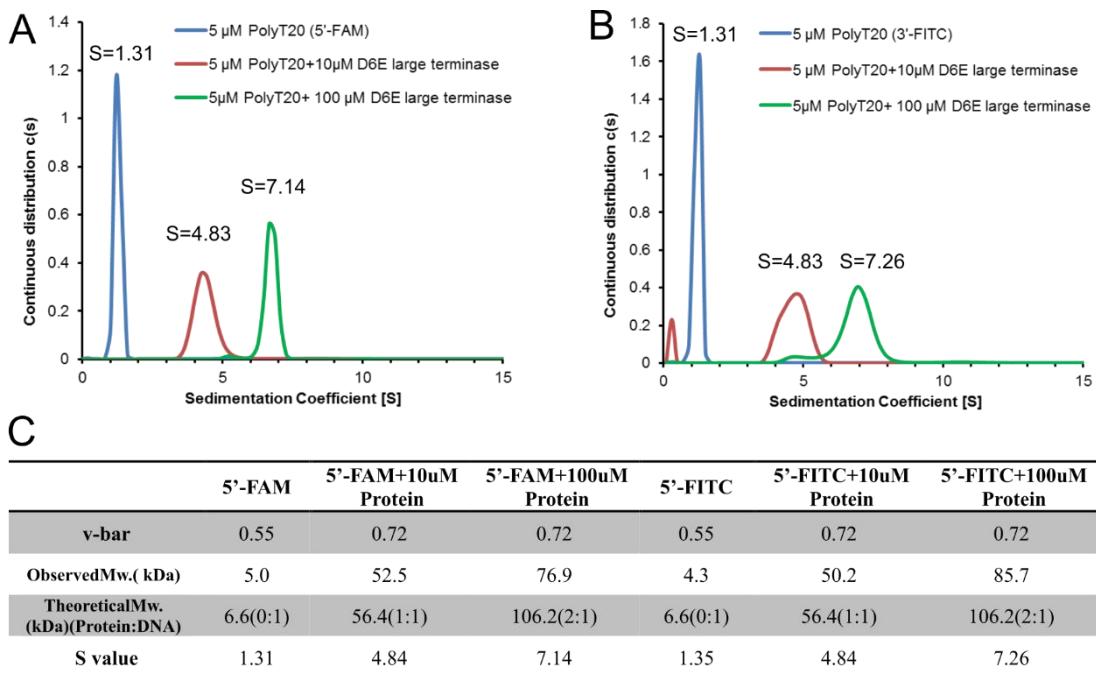


Figure 7-7 Analysis of the stoichiometry of the D6E D326N large terminase complex with ssDNA

The overlay of peaks in the AUC experiment in the presence of (A) 5'-FAM labelled poly(T)20 and (B) 3'-FITC labelled poly(T) 20. (C) Comparison of v-bar, S-value as well as the observed and theoretical molecular weights of the DNA and protein-DNA complex.

As 3'FITC labelling was more likely to produce heterogeneous protein-DNA complexes characterised by broader peaks (Figure 7-7) that were difficult to fit well, further experiments were performed with the 5'-FAM labelled DNA. It was observed in the gel shift assays that DNA with random sequence had higher affinity towards the protein compared to the poly(T) ssDNA. Furthermore, as dsDNA represents the natural substrate more closely, dsDNA with a random sequence was used for further experiments. The 10 bp dsDNA sequence was GCACTGAGCG (Figure 7-8). 20 and 30 bp dsDNA were based on the same

repetitive sequence as 10 bp dsDNA. In these experiments, the protein concentration used to form DNA complex was increased from 100  $\mu$ M to 500  $\mu$ M, to ensure saturation of DNA with bound protein molecules.

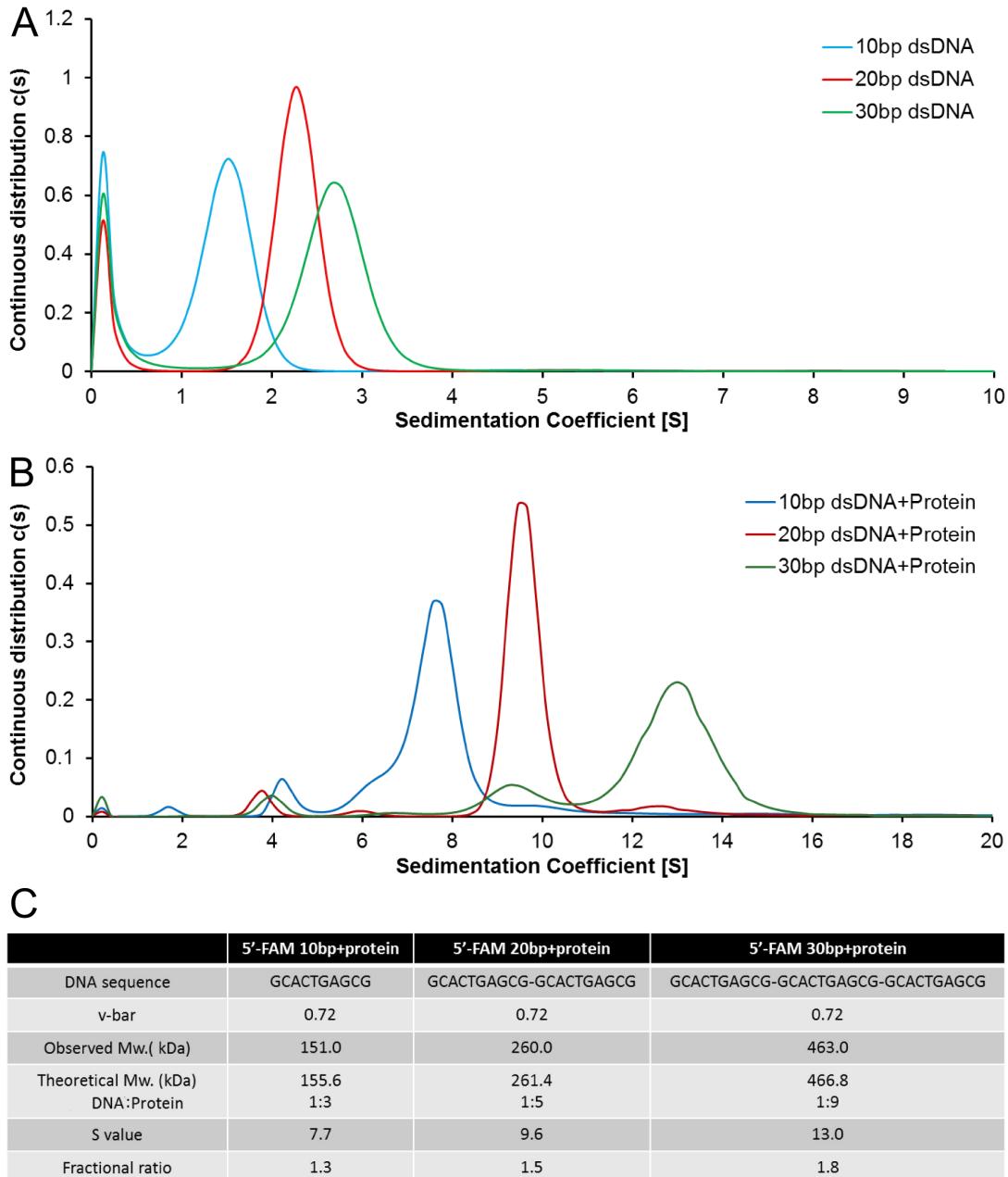


Figure 7-8 Analysis of stoichiometry of the D6E D326N large terminase complex with dsDNA

The overlay plot of the species peaks in the AUC experiment in the presence of (A) 5'-FAM labelled DNA at lengths of 10, 20 and 30 bp, (B) 5'-FAM labelled DNA at lengths of 10, 20 and 30 bp with 500  $\mu$ M protein. (C) Comparison of v-bar, S-value, as well as the observed and theoretical molecular weights of the protein DNA complex, with DNA sequences listed at the top of the Table.

Complexes with protein-DNA stoichiometries of around 3:1, 5:1 and 9:1 were observed when 10 bp, 20 bp and 30 bp dsDNA were added to the protein, respectively. When 20bp

DNA was used, the distribution of the molecular weight of the complex was narrow resulting in a sharp peak. However, the peaks observed for the 10bp and 30 bp DNA were broader indicating that complexes formed with these DNA contained species of different molecular weights. Interestingly, the frictional ratio of the fitting was determined at 1.3, 1.5 and 1.8 respectively, for complexes formed with 10, 20 and 30 bp DNA. This observation indicates that the protein-DNA complex was changing its shape from globular to become more elongated when the length of DNA was increased.

### 7.2.5 Size exclusion chromatography

As mentioned in the above sections, the D6E D326N large terminase interacts with poly(T) 20 ssDNA and forms protein-DNA complexes with stoichiometries of 1:1 or above. In AUC experiments, the protein-DNA complexes were analysed during condensation under the centrifugal force. Different to AUC, SEC analyses samples by washing and eluting the sample through the column. In this process, the protein-DNA complexes formed will be diluted. This section describes further characterisation and purification of this complex using SEC. The data obtained would be useful for evaluating the stability of the protein-DNA complex, its suitability for purification and/or use in crystallisation experiments.

The poly(T) ssDNA containing 20 or 25 nucleotides with no fluorescence label was mixed with D6E D326N large terminase at an equal molar ratio. The mixture was loaded on a S200 10/30 size exclusion column, equilibrated with 20 mM Tris pH 8.0, 200 mM NaCl and 1 mM MgCl<sub>2</sub>, as used in the DNA binding gel shift assays. Samples were eluted at 0.5 ml/min. For comparison of retention volume, the DNA and protein samples used to form the complex were also eluted separately and overlaid with the protein-DNA complex elution data. It was observed that the protein-DNA complex eluted earlier than the DNA or protein alone, suggesting a larger complex formation. It was also observed that no protein-DNA complex formed when poly(T) 10 ssDNA was mixed with the protein at the same molar ratio. It is likely that the binding affinity between poly(T) 10 and the protein is not high enough to form a protein-DNA complex that is stable during elution from the column. This is consistent with observations obtained by the microscale thermophoresis and gel shift assays where the minimum DNA length for apparent protein binding was over 10 nucleotides.

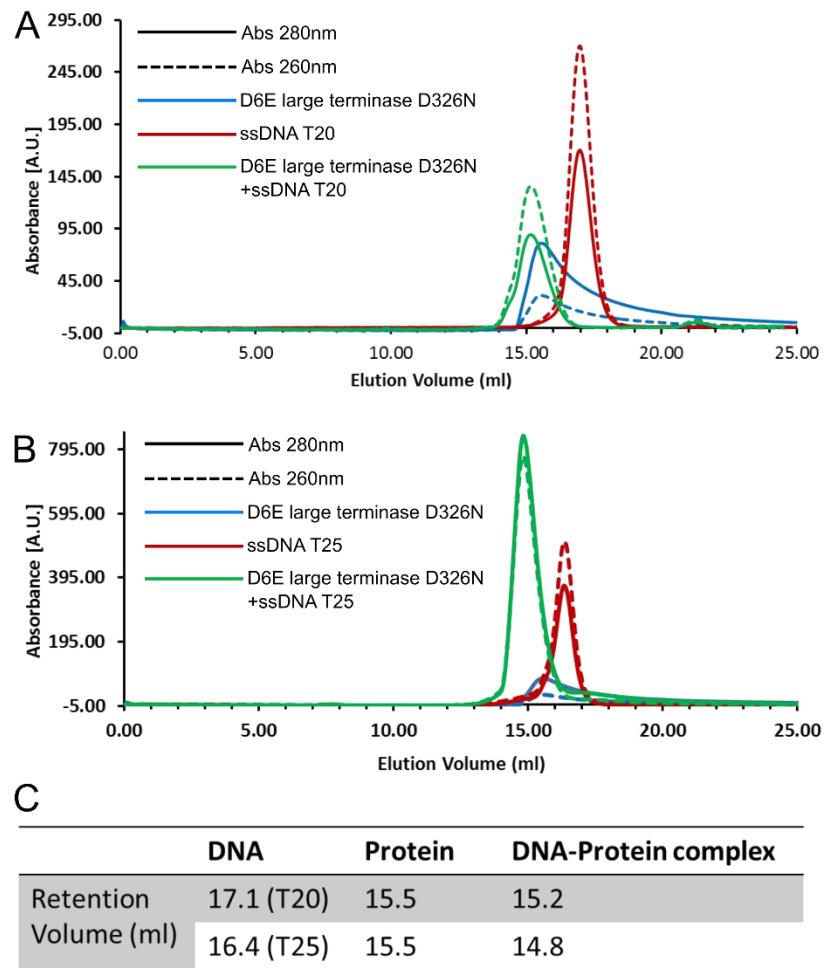


Figure 7-9 Analysis of the stability of protein-DNA complexes

The elution profiles for the D326N protein-DNA complexes with poly(T) ssDNA containing 20 and 25 nucleotides are shown in (A) and (B). The retention volume of the peaks (C).

### 7.3 Crystallisation and preliminary X-ray analysis of the D6E large terminase complex with DNA

In this section, crystallisation and complex preparation of the D6E large terminase D326N with various DNA constructs will be described, followed by analysis of preliminary X-ray diffraction data of the D6E large terminase-DNA complex.

#### 7.3.1 Crystallisation of protein-DNA complex using poly (T) ssDNA

As characterised above, the protein-DNA complex remains stable during SEC. Subsequent crystallisation experiments were set up in attempt to obtain crystals of these protein-DNA complexes. Poly(T) 10, 20 and 25 bp ssDNA were mixed with the D326N D6E large terminase at a molar ratio of 1:1 to form the protein-DNA complex. The final concentration of the protein was 3-5 mg/ml. No crystals were obtained for the mixture with the Poly(T) 10 ssDNA, consistent with the low binding affinity of this DNA to the protein. All crystals obtained using poly(T) 20 and 25 were non-diffracting (data not shown).

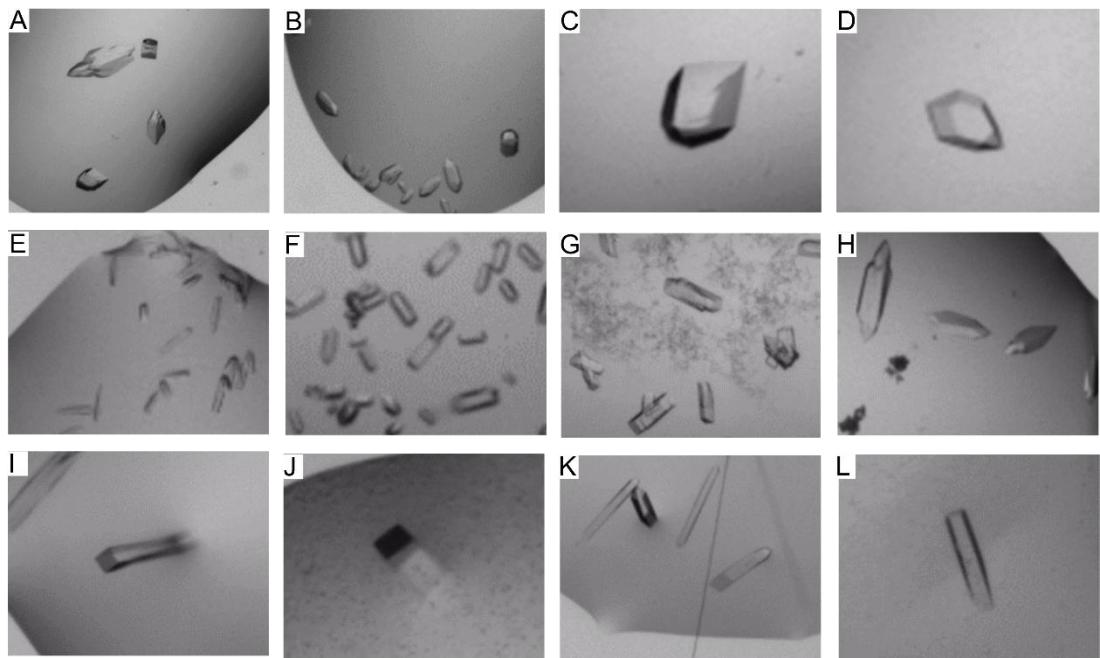
### 7.3.2 Crystallisation of protein-DNA complex using dsDNA

Subsequently, dsDNA at a length of 21 bp (DNA 1 in Table 7-2) was designed and annealed with its complementary strand. The protein-DNA complex was prepared using the same protocol as mentioned above and crystals were found as shown in panel A and B (Figure 7-10). The best crystals diffracted to 8 to 9 Å when tested using in-house X-ray facilities. To facilitate crystal packing and formation of crystal contacts, a palindromic dsDNA at a length of 20 bp (DNA 2 in Table 7-2) was used for further crystallisation. Based on this sequence, a palindromic dsDNA with one or two base overhangs (DNA 3 and 4 in Table 7-2) were also designed and used for complex preparation and crystallisation. Following the same complex preparation protocol, diffracting crystals were obtained when DNA 2 was used (panels C and D on Figure 7-10). These crystals diffracted to around 5 Å when using the in-house X-ray generator. From these crystals, X-ray data to ~3.3-3.5 Å resolution were collected at the I03 beamline at Diamond (Harwell). As crystals obtained using DNA 3 and 4 diffracted only to 8 to 9 Å, X-ray data were not collected from these crystals.

Since dsDNA constructs with lengths of 18 to 21 bp produced no suitable crystals for structure determination, a dsDNA construct of 8 bp was tested. Unlike previous protein-DNA complex crystallisation protocols, a protein-DNA ratio of 1:2 was used in this case. It was expected that the increase in DNA concentration would lead to all protein molecules being occupied with bound DNA. Crystals were obtained, as shown on panels E-L, Figure 7-10. The best crystal diffracted to 3.2 Å at the synchrotron. The space group, cell parameters and estimated number of molecules per asymmetric unit, for each data set collected, are shown in Table 7-3.

*Table 7-2 Sequences of the DNA constructs used in protein-DNA complex crystallisation*

DNA	Sequence	GC content (%)	Tm (°C)
1	5'-CGCCTAGTCGCTCGCATGTCC (21)	66.7	65.7
2	5'-GCTGGTATGCGCATACCAGC (20)	60.0	61.4
3	5'-TATGGTATGCGCATACCAT (19)	42.1	52.4
4	5'-TATGGTATGCGCATACCA (18)	44.4	51.4
5	5'-GCGCGCGC (8)	100	32.0



A	8 w/v polyethylene glycol 20000, 8 v/v polyethylene glycol monomethyl ether 500, 800 mM sodium formate	G	0.2M Calcium Acetate, 0.1 M MES 6.0 , 20% PEG8000
B	10 w/v polyethylene glycol 1000, 10 v/v polyethylene glycol 8000, 200 mM magnesium chloride	H	0.2M Calcium Chloride, 0.1 M HEPES 8.0, 20% PEG6000
C	0.2M MgCl <sub>2</sub> , 0.1M HEPES 7.5, 20% PEG3350	I	1 M Lithium Sulfate, 0.1 M MES 6.0, 20% PEG6000
D	0.2M Na/K tartrate, 0.1M BTP 6.5, 20% PEG3350	J	60% tacsimate
E	2.4M Ammonium Sulfate, 0.1 M Tris pH 8.0	K	0.2M Ammonium Citrate, 20% PEG3350
F	0.2M Calcium Chloride, 0.1 M MES 6.0 , 20% PEG8000	L	1 M Lithium Sulfate, 0.5M Ammonium Sulfate, 0.1 M tri-sodium citrate

*Figure 7-10 Crystals obtained from samples of protein-DNA mixture  
Well solution composition for each drop is listed in the table below the figure.*

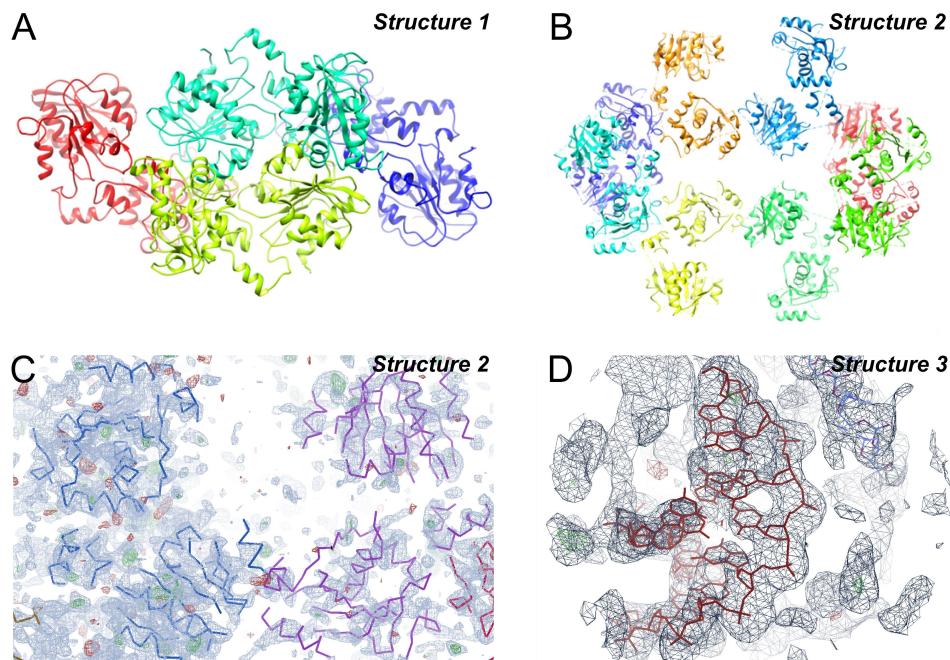
### 7.3.3 Analysis of X-ray diffraction data

Following data processing by XDS, the number of molecules in an asymmetric unit was estimated by Matthew's coefficient (196) calculations. For data sets 1 and 2, the number of molecules in the asymmetric unit were estimated as 4 and 8, respectively. However, in the case of data set 3, the cell volume was too small for the occupation of even one copy of the D6E large terminase (Table 7-3). It is likely that the unit cell contains only a fragment of the D6E large terminase protein monomer or some copies of the 8 bp dsDNA. Structures for data sets 1 and 2 (Figure 7-11, A and B) were determined by molecular replacement using a single chain of D6E large terminase structure (see chapter 6) as the search model. The electron density for several copies of the molecule in the asymmetric unit was poor,

indicating disorder in crystal packing and structural variability (Figure 7-11, C). Likewise, poor electron density was observed for the ATPase domain of the R32 crystal form of the wild type D6E large terminase, as described in Chapter 6. Attempts to determine the structure for data set 3 by molecular replacement, using models of the nuclease or ATPase domain of the D6E large terminase, failed. Solutions were found when using an ideal 8 bp dsDNA as the search model, with two DNA molecules being observed in the unit cell (Figure 7-11, D). Inspection of the density of structure 3 indicated that there is unexplained density in the unit cell. This density is likely to belong to a disordered DNA molecule. Refinement of all three structures determined using data set 1, 2 and 3 yielded  $R_{free}$  values of around 0.4 to 0.5. This is not surprising given the disorder in molecular packing in structures 1, 2 and 3. For structure 3, automatic model building was then used in an attempt to generate a more complete DNA by filling missing parts using Arp/Warp (197) DNA/RNA builder. However, no significant improvement was achieved.

*Table 7-3 Resolution, space group and unit cell parameters of the three data sets collected using DNA 2 (data set 1 and 2) and DNA 5 (data set 3).*

Data set	1	2	3
Resolution	3.5 Å	3.3 Å	3.2 Å
Space group	C 2	C 2	I 422
Unit cell parameters			
a, b, c (Å)	147.0, 126.2, 146.3	95.9, 349.0, 142.6	124.8, 124.8, 76.9
NCS <sub>protein</sub>	4	8	<1



*Figure 7-11 Structure and electron density maps.*

*The structures determined using data-set 1 and 2 are shown in (A) and (B), respectively. The structures and their corresponding 2Fo-Fc electron density map contoured at 1 sigma are shown for structure 2 (C) and structure 3 (D).*

#### 7.4 Summary

In this chapter, the DNA binding properties of the H427A G20c large terminase and D326N D6E large terminase were characterised by biophysical and biochemical approaches, including EMSAs, MST and SV-AUC and SEC. The data suggest that the minimum length for apparent dsDNA binding to the G20c and D6E proteins was around 20 to 30 and 10 to 20 bp, respectively. The D326N D6E large terminase formed complexes with both ssDNA and dsDNA. At a fixed DNA length, the stoichiometry of the complex depended on the protein concentration. Titration of protein at various concentrations resulted in sigmoid curves with Hill constants higher than 1, indicating cooperative binding. The cooperativity of this binding was dependent on the length of DNA used. Compared to short ~10 bp DNAs, longer DNA resulted in an increase in cooperativity as well as modest increase in apparent binding affinity. At a fixed protein concentration that saturates dsDNAs containing repetitive sequence, the increase of DNA length resulted in complexes with larger molecular weights and different hydrodynamic properties, indicating change from globular to rod-like assemblies. Combined with the increase in molecular weight, these observations suggest a minimum binding length of 3 to 4 bases of DNA per one protein molecule. In addition, the D326N D6E large terminase was observed to bind to poly(T) ssDNA, and also random-sequence ss or dsDNA, indicating non-sequence specific binding.

The protein-DNA complexes were stable in competition assays with native DNA under MST, and could be eluted as stable complexes during gel filtration. However, subsequent crystallisation of such protein-DNA complexes yielded crystals that contained either protein or DNA but not both. This observation indicates that protein-DNA complexes were unstable and/or not suitable for crystallisation. Future work could focus on designing a protein-DNA complex with increased stability, for example by using chemically modified DNA and cross-linking strategies.

## 8 ATP hydrolysis and oligomerisation of the D6E large terminase

This chapter describes studies investigating how the large terminase ATPase assembles and hydrolyses ATP. The kinetic parameters of the isolated ATPase domain and the full-length large terminase protein were compared using coupled ATPase assays. Since an oligomeric state was observed experimentally, a molecular model of the ATPase assembly was constructed *in silico* by docking. Based on this model, mutagenesis studies of the residues that are close to the ATPase active site of the full-length protein were performed, from which a *trans*-arginine-finger residue was identified. In addition, the role of the C-terminal arm of the large terminase in ATP binding and hydrolysis was also investigated. These data suggested a model for large terminase assembly and function.

### 8.1 Cloning and expression of the ATPase domain

Observations for large terminase motors of other bacteriophages showed that the protein assembles on the portal vertex as a pentamer (41,56,198). To investigate the oligomeric state of the D6E large terminase ATPase, the isolated ATPase domain was cloned, expressed and purified. Tests compared the amount of protein expressed using *E. coli* cell strains including B834, BL21 and Rosetta pLysS at 37 °C and 16 °C (Figure 8-1). The results show that under tested conditions, B834 cells at 37 °C gave the optimal expression level of this protein. Large scale protein expression was then based on these conditions. The protein was purified using standard protocol described in Chapter 2, except that 1M NaCl was used instead of 250 mM for the full-length protein.

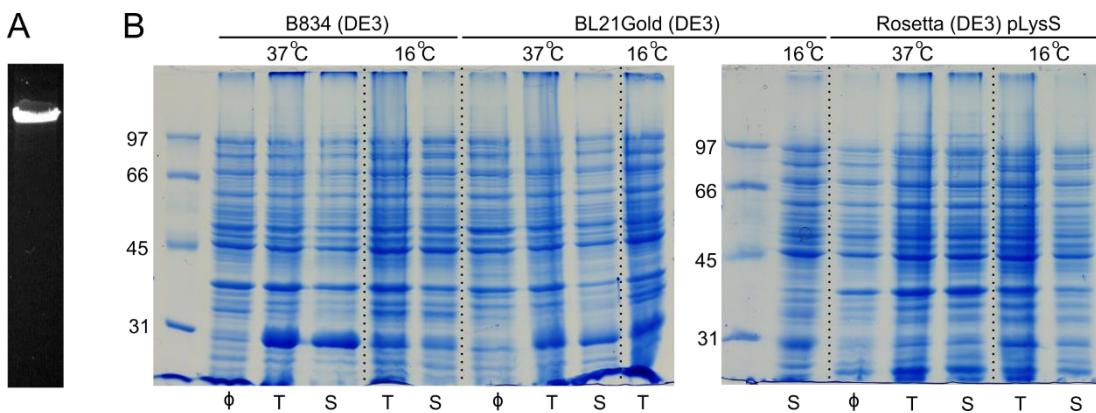
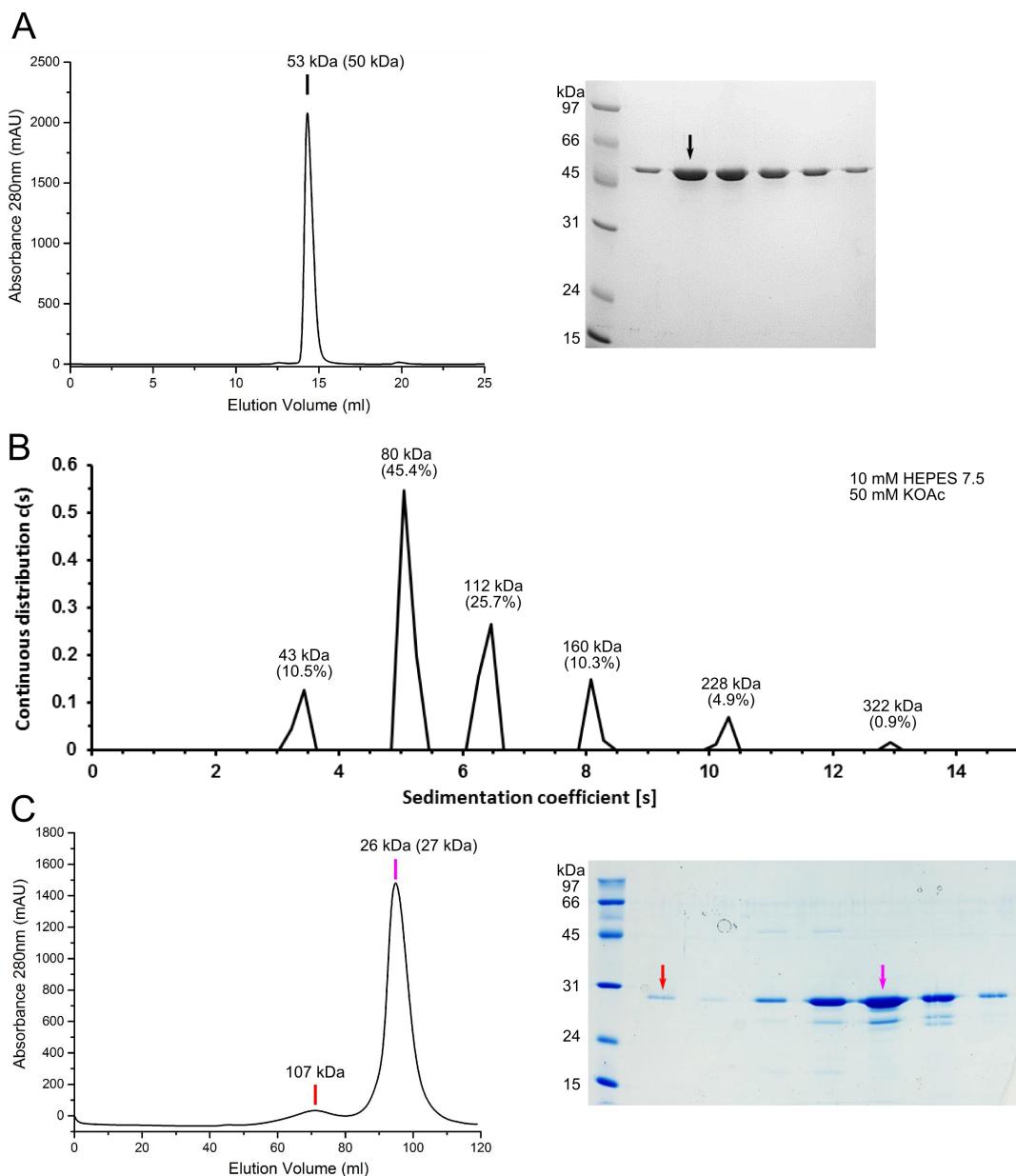


Figure 8-1 PCR amplification and expression tests

(A) Agarose gel electrophoresis showing the PCR amplified recombinant plasmid of the D6E large terminase ATPase domain. (B) SDS-PAGE showing the expression and solubility level under tested conditions. Φ: before IPTG induction, T: total cell lysate, S: Supernatant after cell lysis.

## **8.2 Analysis of the oligomeric state**

In this section, the molecular weight and oligomeric state of the full-length D6E large terminase (1-427) and isolated ATPase domain (1-234) were characterised by size exclusion chromatography and analytical ultra-centrifugation. Like T4 (40), SPP1 (185) and Sf6 (83) large terminases, the full-length protein of the D6E bacteriophage is monomeric in solution, as judged by size exclusion chromatography (SEC). However, subsequent species analysis by sedimentation velocity analytical ultracentrifugation (SV-AUC) revealed the presence of multiple oligomeric species at different molecular weights. The ATPase domain alone was also predominately monomeric but additionally contained a small fraction of an oligomeric species by SEC (Figure 8-2). The molecular weight of this oligomer was estimated as ~110 kDa, around 4-5 times larger than that of the monomer (27 kDa), and is consistent with the pentameric state observed for the ATPase domain of large terminase from another thermophilic bacteriophage (56) and with SV-AUC experiments performed with the full-length protein.



*Figure 8-2 Characterisation of the oligomer state of the full-length D6E large terminase and the N-terminal ATPase domain*

(A) Estimation of the molecular weight of large terminase (l-427) by analytical size exclusion chromatography (left) and SDS-PAGE (right). (B) Oligomeric states of the large terminase in sedimentation velocity analytical ultracentrifugation. The deduced molecular weight and ratio of the monomer in each peak are indicated. (C) Characterisation of the N-terminal ATPase domain (residues 1-234) by size exclusion chromatography. The corresponding fractions of the oligomer and monomer peaks (indicated by coloured arrows) were analysed by SDS-PAGE (right) (figure reproduced from (186)).

### 8.3 ATP hydrolysis by the ATPase domain and full-length terminase

In chapter 3, the ATPase activity of the D6E large terminase was characterised by malachite green phosphate assays. To gain further understanding, the activity of the full-length protein was compared with that of the ATPase domain using a steady-state ATPase activity assay.

Measurement of ATPase activity at various protein concentrations revealed a non-linear increase in the rate of ATP hydrolysis at low protein concentrations, suggesting that assembly between subunits is concentration dependent (Figure 8-3). Interestingly, distinct ATP binding and hydrolysis profiles were observed with the  $k_{cat}$  of the ATPase domain ( $457.7 \pm 9.2 \text{ min}^{-1}$ ) around 60 times higher than that of the full-length protein ( $7.8 \pm 0.1 \text{ min}^{-1}$ ). This indicates that the ATPase domain exhibits a much higher catalytic efficiency than the full-length protein. In addition, the  $K_m$  was measured to be  $68.4 \pm 4.3 \mu\text{M}$ , around 7 times higher than that of the full-length protein ( $9.2 \pm 0.5 \mu\text{M}$ ) suggesting a reduced level of ATP binding. Despite these differences in  $k_{cat}$  and  $K_m$ , a lack of positive cooperativity during ATP hydrolysis was observed for both proteins, with Hill coefficients of 1.1 for both the full-length protein and ATPase domain (Figure 8-3).

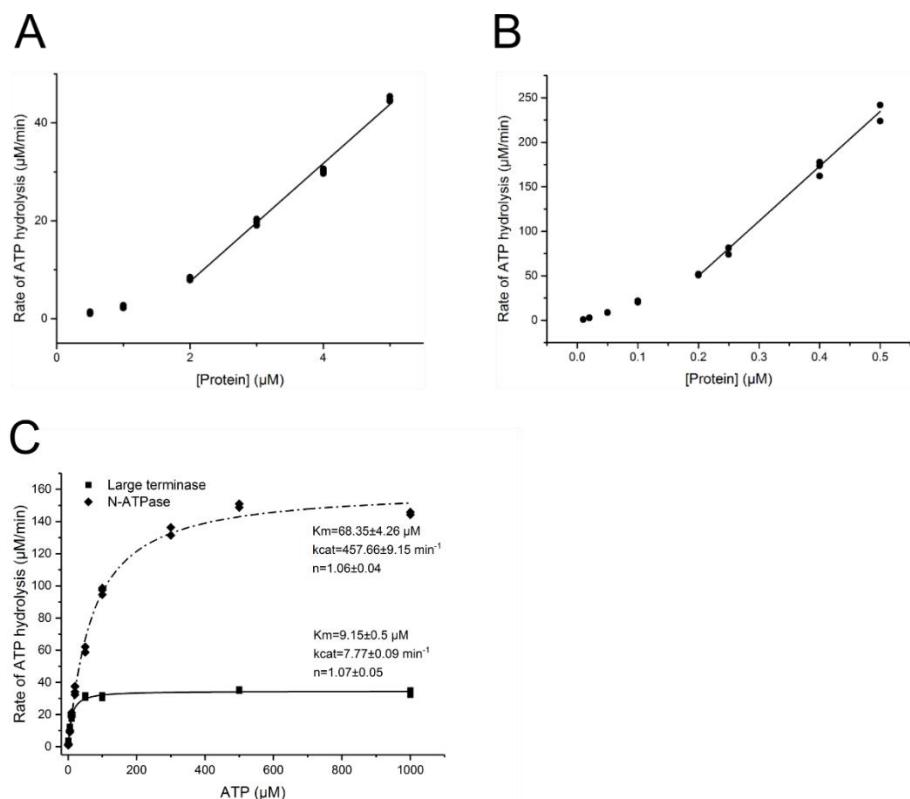


Figure 8-3 Kinetics of ATP hydrolysis

Rate of ATP hydrolysis of the full-length wild type D6E large terminase protein (A) and the ATPase domain (B) determined by coupled ATPase assays at various protein concentrations. (C) Titration of ATP with full-length large terminase (residues 1-427) containing both ATPase and nuclease domains (diamonds) and large terminase ATPase domain, residues 1-234, (rectangles). The data points were globally fitted by with the Hill equation.  $K_m$ ,  $k_{cat}$  and Hill constant ( $n$ ) are indicated (figure reproduced from (186)).

#### 8.4 Structure of the C-terminal arm

In chapter 6, structures of the D6E large terminase in complex with ADP and ATP- $\gamma$ -S were determined after removal of the C-terminal arm (residues N418-Y427) located in the

nuclease domain. This C-terminal arm is predicted to play a role in ATP binding and/or hydrolysis since amino acids N418-Y427 of this arm contribute to the active site (Figure 8-4). An arginine residue, R421, on the tip of the C-terminal arm, is in proximity to the sulfate ion bound in the active site and is predicted to form a hydrogen bond with the  $\alpha$ -phosphate of ATP- $\gamma$ -S. Interestingly, similar extended C-terminal arms appear to be present in P74-26 (180), Sf6 (69), T4 (40), P22 (68), SPP1 (67) and RB49 (76) large terminases (Figure 8-4). In common with the D6E large terminase, this region usually contains several charged residues. For T4, RB49 and P22 large terminases, there is an additional  $\alpha$ -helix connecting to an equivalent C-terminal arm region (Figure 8-4).

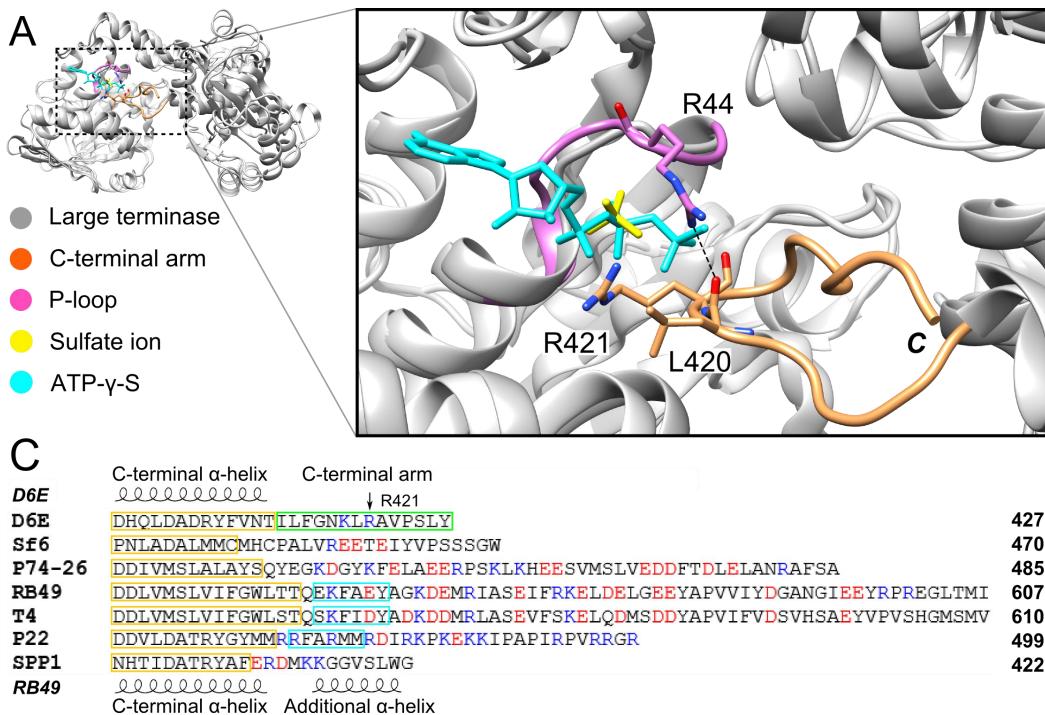


Figure 8-4 Structure of the C-terminal arm and comparisons with equivalent structural elements in large terminases of other bacteriophages

(A) Superposition of the large terminase structures for the sulfate ion complex (white) and the ATP- $\gamma$ -S complex (grey). The C-terminal arm (gold) stabilised by the bound sulfate ion (yellow) and a hydrogen bonding interaction with R44 of the P-loop (magenta) are highlighted. The interactions between the P-loop (R44) and C-terminal arm (L420) are shown in a close-up view. Hydrogen bonds are in black dashed lines. (B) Structure superposition of D6E (grey/green), Sf6 (salmon), RB49 (orchid) and P74-26 (yellow) large terminase nuclease domains. For clarity, only the C-terminal  $\alpha$ -helix is shown for all proteins except D6E terminase which is represented by a whole nuclease domain. (C) A closed-up view at the C-terminal arm with positively charged residues shown in sticks. (D) Sequence comparison of the C-terminal region of D6E with Sf6, P74-26, RB49, T4 and P22 large terminases. The secondary structure elements of D6E and RB49 terminases are shown at the top and bottom of the sequence alignment, respectively. The C-terminal  $\alpha$ -helix is coloured in gold. The position of R421 and C-terminal arm (green) of the D6E large terminase are indicated. An additional C-terminal  $\alpha$ -helix only found in RB49, T4 and P22 large terminases is highlighted in cyan. Positively and negatively charge residues are coloured in blue and red respectively (figure reproduced from (186)).

## 8.5 The role of the C-terminal arm in ATP hydrolysis

To investigate the role of the C-terminal arm in ATP binding and hydrolysis, we determined the  $K_m$  and  $k_{cat}$  values for the C-terminally truncated protein (1-417) and the full-length R421A mutant. Compared to the wild type full-length large terminase, both mutants have reduced ATP binding affinity (increased  $K_m$ ) and catalytic efficiency (decreased  $k_{cat}$ ) (Figure 8-5). Notably, a 10-fold and 20-fold increase in  $K_m$  values were observed for the R421A mutant and the C-terminal arm truncated protein, respectively. Previous studies on bacteriophage T4 indicated that several residues in the C-terminal region of the large terminase interact with the portal clip during DNA translocation (44). As shown in Figure 8-4, this C-terminal region is also present in other bacteriophage large terminases and is equivalent to the observed C-terminal arm of the D6E large terminase. In the ATPase assays, it was observed that this C-terminal arm and its R421 residue play a role in facilitating ATP binding to the full-length D6E large terminase.

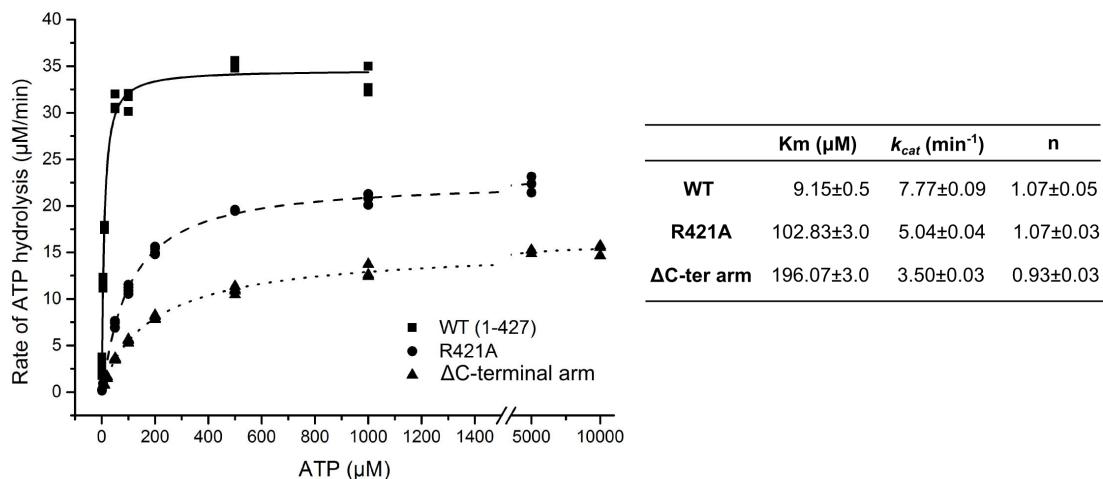


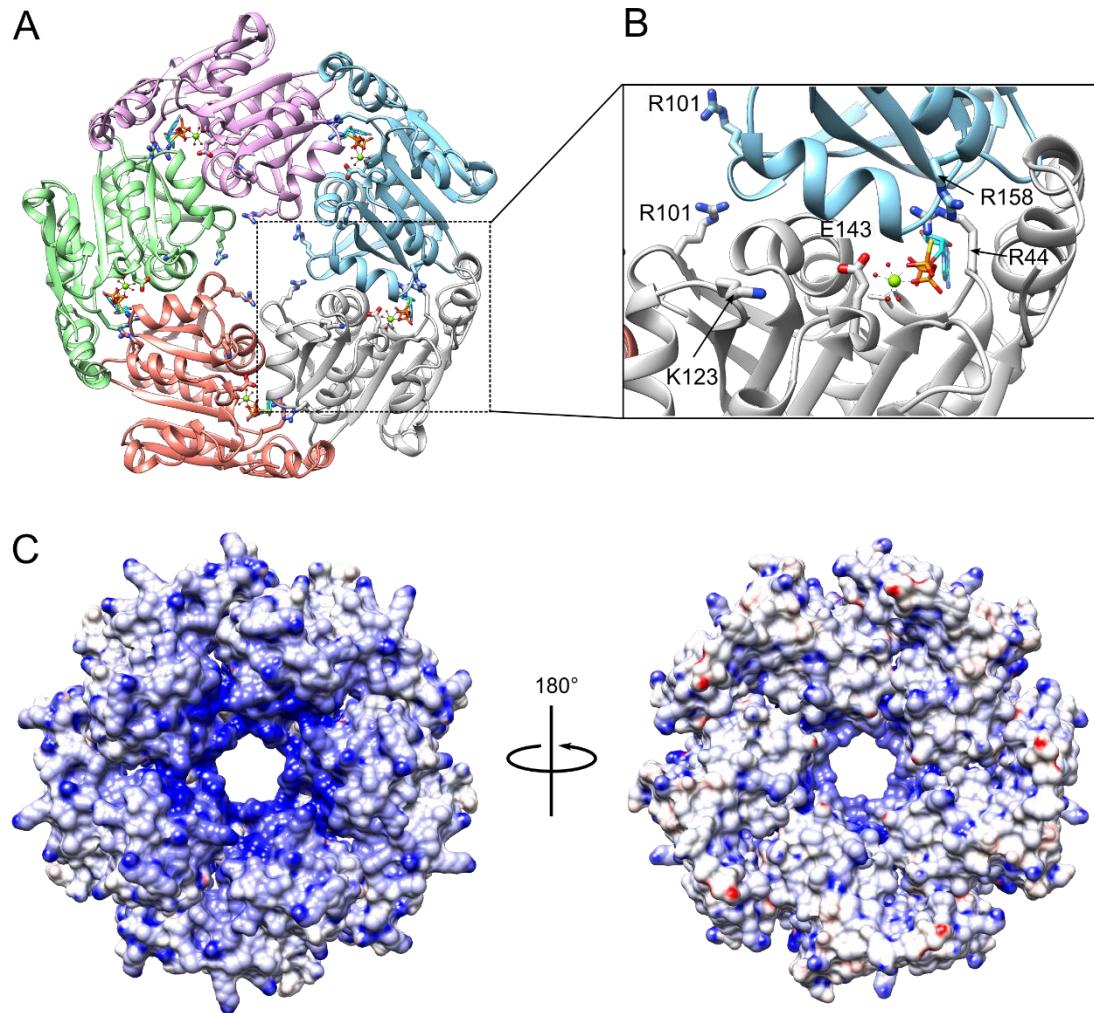
Figure 8-5 Plot of ATP hydrolysis rate versus ATP concentration for the wild type D6E large terminase (squares), R421A mutant (circles) and C-terminal arm truncated protein (triangles)

Data points (three repetitions each) were globally fitted by the Hill equation.  $K_m$ ,  $k_{cat}$  and Hill constant ( $n$ ) are shown in the table (figure reproduced from (186)).

## 8.6 Model of the D6E ATPase pentamer

To gain understanding on how the motor is assembled, we constructed a pentamer model of the D6E ATPase domain, using the molecular docking server (M-ZDOCK (199)) as reported previously for the *Thermus* P74-26 large terminase (56). Two extended loops (residues 24-35 and 159-164) were removed during pentamer construction. In the final model, the ATPase active site is placed at the oligomeric interface, similar to the phi29 structure-based model (200) and the model proposed for the P74-26 large terminase ATPase (56) (Figure 8-6). In the ATPase active site, R44, E143 (same subunit) and R158 (adjacent subunit) are in proximity to the bound nucleotide (Figure 8-6). Two loops, defined as loop L<sub>1</sub> and L<sub>2</sub>

which contain positively charged residues R101 and K123, respectively (Figure 8-6), are found to contribute to the tunnel (18 Å in diameter) that is expected to interact with DNA during translocation. It is possible that the tunnel can accommodate a double stranded DNA following potential conformational adjustments upon assembly. Similar to observations for other multi-subunit ATPase motors (201,202), the internal tunnel surface area is mostly positively charged (Figure 8-6).



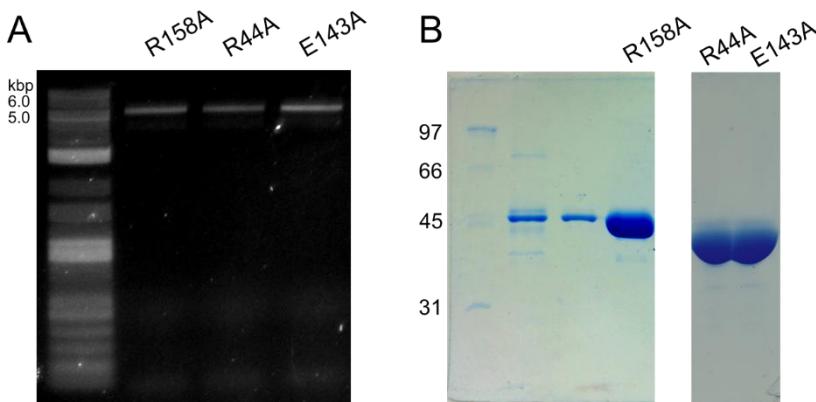
*Figure 8-6 Model of the D6E ATPase domain pentamer*

(A) Pentameric model obtained by molecular docking. (B) Close-up view of an ATPase active site at the interface between two subunits showing ATP- $\gamma$ -S (cyan) and the side chains of R101, K123, R44, E143 and R158 (sticks). (C) Molecular surface of the pentamer coloured according to the electrostatic charge potential, calculated for pH 7.0. Positive charge is shown in blue while negative charge is shown in red (figure reproduced from (186)).

### 8.7 Cloning, expression and purification of R158A, R44A and E143A

Based on the pentamer model of the ATPase, R158, R44 and E143 are predicted to be in proximity to the ATPase active site and may play important roles in ATP binding and hydrolysis. To further investigate the role of these residues, mutagenesis was performed to

produce mutant proteins, including R158A, R44A and E143A. The mutant proteins were expressed and purified following the standard procedures in Chapter 2. The final purified product of these mutant proteins is shown in Figure 8-7.



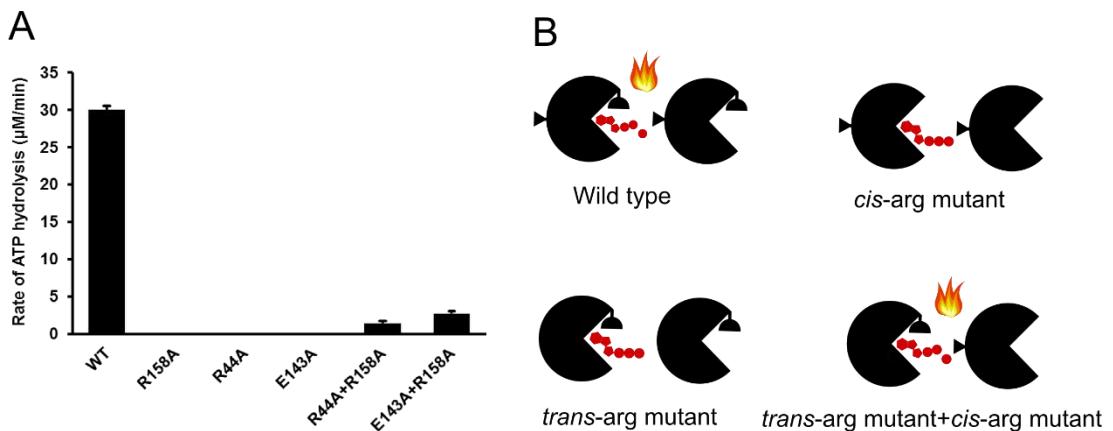
*Figure 8-7 PCR amplification and SDS-PAGE*

(A) Agarose gel electrophoresis showing the PCR amplified gene constructs of D6E large terminase protein. (B) SDS-PAGE showing the purity of mutant D6E large terminase proteins after nickel affinity and size exclusion chromatography.

### 8.8 Identification of the *trans*-acting arginine finger

Previous structure/sequence comparisons within the HerA/Ftsk superfamily of ASCE ATPases revealed a conserved arginine residue indicative of ring-like oligomeric ATPases, likely to promote inter-subunit coordination of ATP hydrolysis *in trans* (50). Recent structural and biochemical studies of the phi29 packaging ATPase and P74-26 large terminase identified the *trans*-acting arginines required for ATPase activity as R146 and R139, respectively, located at the monomer-monomer interface of the proposed pentamer model (56,200). R158, which projects into the active site from the adjacent subunit in our pentamer model (Figure 8-6), is predicted to stabilise the transition state and facilitate ATP hydrolysis *in trans*. Additional active site residues such as the P-loop arginine, or ‘sensor’, and the catalytic glutamate were identified in other large terminases as facilitating ATP catalysis *in cis* (Figure 8-6) (40,56,83). The analogous three residues in D6E large terminase, respectively predicted to be R158 (*trans*-acting), R44 and E143 (*cis*-acting), should be indispensable for catalysis. As expected, mutagenesis of any one of these residues to alanine completely abrogated ATP hydrolysis (Figure 8-8). However, this activity was partially recovered in assays where R158A was complemented with either R44A or E143A (Figure 8-8), as previously found for P74-26 large terminases (56) and other ring-like ATPase motors (200,203,204). This data is consistent with the pentamer model since complementation requires oligomer assembly to create a catalytically competent active site at the interface containing both *trans* and *cis* acting arginines. This catalytically competent active site can be formed between wild type monomers or complemented by a mixture of

trans and *cis* acting arginine mutant monomers (R158A+R44A or R158+E143A) (Figure 8-8).



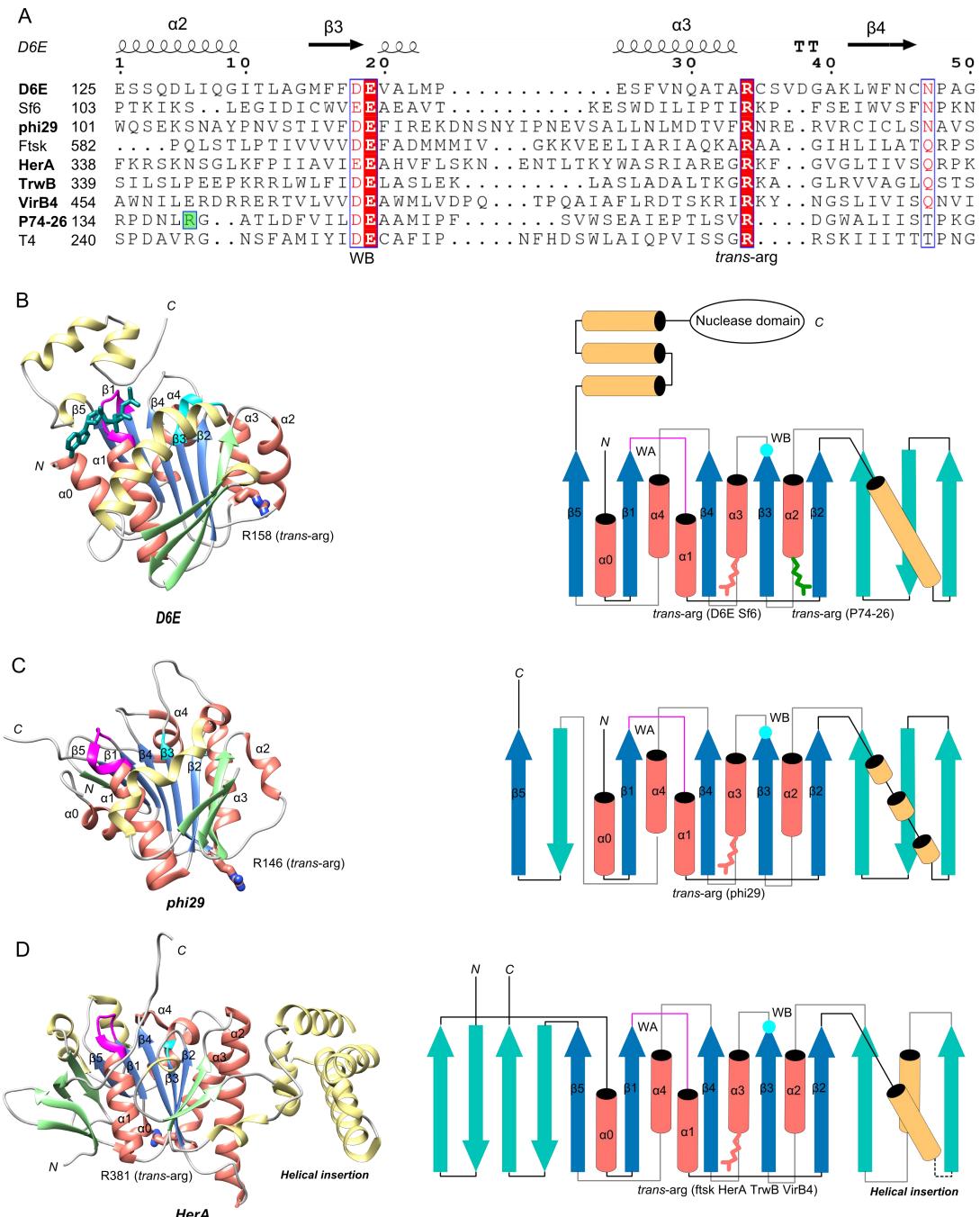
*Figure 8-8 ATP hydrolysis by the wild type D6E large terminase and its mutants*

(A) Steady-state ATPase activity of wild type, R44, E143, R158, R44A+R158, and E143A+R158A (equal molar mixture). (B) Schematic representations of the ATPase complementation assays. The ATPase subunit is represented by an open circle with the active site located at the subunit interface. The *cis*- and *trans*-acting arginines are shown as semi-circles and triangles, respectively. A mixture of *trans* and *cis*-acting arginine mutants leads to the formation of an intact active site able to hydrolyse ATP (figure reproduced from (186)).

## 8.9 The *trans* arginine finger

In multi-subunit ASCE ATPases, a *trans*-acting arginine finger residue is generally required to couple ATP hydrolysis between the subunits of the active ring-like assembly (49,50,205). Comparative genomic analysis has indicated that although the presence of the *trans*-arginine finger is commonly found for multi-subunit ASCE ATPases, the position of the *trans*-arginine finger varies across different superfamilies (50). Within the HerA/Ftsk superfamily, conserved arginine residues identified between  $\alpha$ 3 and  $\beta$ 4 of the canonical ASCE fold have been proposed to act as *trans*-arginine fingers (50). Interestingly, the biochemically confirmed *trans*-acting arginines, R158 for D6E large terminases and R146 for phi29 packaging ATPase (200), are found at the same locations on protein fold as those observed for the HerA/Ftsk superfamily ATPases (Figure 8-9). Equivalent arginines have also been found at this position for T4, Sf6 and P74-26 large terminases (57) (Figure 8-9). Conversely, studies on P74-26 (56) large terminase identified a *trans*-arginine finger, R139, that is located at a different position, between  $\alpha$ 2 and  $\beta$ 3 elements of the canonical ASCE fold, to that observed for D6E large terminase and phi29 packaging ATPase. Interestingly, positively charged residues are found at this position in T4 (one arginine) and Sf6 (two lysines) large terminases (Figure 8-9). Further work is necessary to confirm whether these residues are *trans*-arginine fingers in T4 and Sf6 large terminases.

Despite the observed variations in the position of the *tran*-arginine finger, these data support the premise that large terminase proteins are multi-subunit ring-like ATPases, with a *trans*-arginine finger that completes the ATPase active site. It is possible that the distinct positions of the *trans*-arginine finger in D6E and P74-26 large terminases indicate that there are two branches of ring-like large terminases utilising slightly different oligomer interfaces for ATP hydrolysis. These data also suggest that the large terminase superfamily of ATPases including phi29 ATPase are more closely related to the HerA/Ftsk superfamily compared to other superfamilies of ASCE ATPases than proposed previously (200).



*Figure 8-9 ATPase comparison*

(A) Structure based sequence alignment of the D6E large terminase ATPase with packaging ATPases of other bacteriophages (Sf6, phi29) and with the Ftsk/HerA superfamily ATPase. Proteins for which the trans-acting arginine finger residues have been characterised biochemically are highlighted in bold. The structure and topological diagrams of the D6E large terminase ATPase, phi29 ATPase and HerA are shown in (B) (C) and (D). Secondary structures are coloured in salmon ( $\alpha$ -helix), blue ( $\beta$ -strand) and grey (coil) for the canonical ASCE fold, while yellow ( $\alpha$ -helix) and marine blue ( $\beta$ -strand) are used for the insertions. ATP is coloured in dark cyan (B). Walker A (magenta) and Walker B (cyan) represent the Walker A and B motifs, respectively. The position of the trans-acting arginine is indicated (figure reproduced from (186)).

## 8.10 The *cis* arginine finger

The P-loop arginine found in many bacteriophage large terminases, including T4 (40), Sf6 (83), P74-26 (56),  $\lambda$  (55) and other multi-subunit ATPases (200,206,207), corresponds to R44 in the D6E large terminase (Figure 8-7). It has been proposed that this arginine may act *in cis* to trigger ATP hydrolysis through subdomain rotation and/or chemo-mechanical coupling to coordinate DNA translocation (40,56,83). Recent studies for  $\lambda$  large terminase support the proposed *cis-acting* mechanism implicated previously, but further suggest that it is also likely to play a role in motor assembly (55). Our structural and biochemical observations on the D6E large terminase are consistent with the proposed mechanism that R44 is a *cis-acting* arginine that couples the motion of the lid subdomain to ATP hydrolysis. However, in common with structural data obtained for the Sf6 large terminase (57), no substantial conformational differences were induced upon binding of different ATP analogues to monomeric form of the large terminase. It is possible that larger conformational changes that can generate the DNA translocating power stroke (208,209) can only take place once a complete motor is assembled and require the participation of the *trans* arginine finger from the adjacent subunit within the ATPase pentamer.

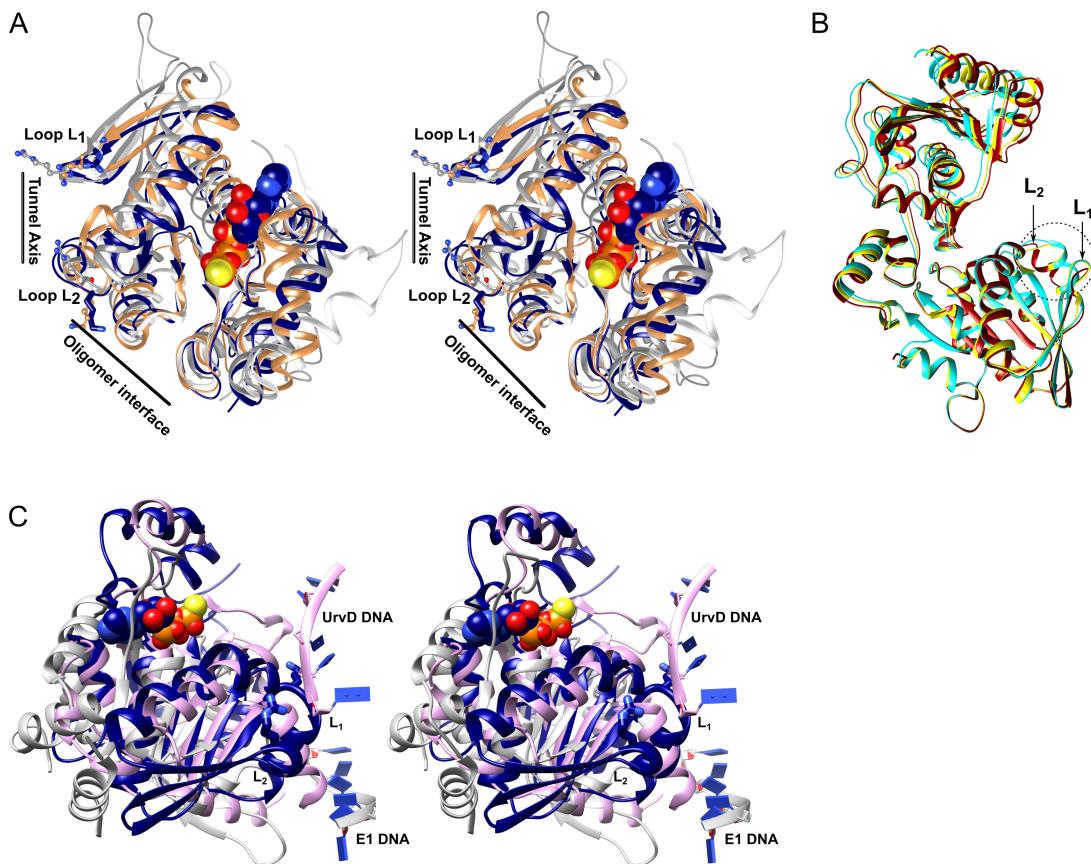
## 8.11 Insights into ATP hydrolysis by large terminase

The results above provide insight into mechanism of ATP hydrolysis of the large terminase ATPase motor. In the assembled ATPase ring, R158 facilitates ATP hydrolysis *in trans* while R44 acts *in cis* to transmit the conformational changes of the P-loop upon ATP hydrolysis through lid subdomain. Consistent with previous studies (56), only modest or low positive cooperativity for ATP hydrolysis (Hill coefficient 1.1 for D6E and 1.7 for P74-26) was observed. Because the Hill coefficient indicates the minimum number of ATP bound active sites needed for activity, the observed low cooperativity suggests that the motor is likely to be partially assembled (1-2 active site formed by 2-3 subunits) without the stabilisation by DNA and/or portal ring. This is supported by the intermediate oligomeric states that were observed by SV-AUC (Figure 8-2). However, it is also likely that, if as proposed for the phi29 DNA packaging motor (Hill constant 1.2 measured by ATPase

assays during DNA packaging) (208), binding of ATP in the fully assembled motor is independent within the five subunits (209), this would also be consistent with a Hill coefficient close to unity.

### **8.12 DNA translocating “tunnel” loops**

In the model of assembled ATPase ring (Figure 8-6), loops L<sub>1</sub> and L<sub>2</sub> are placed in proximity to the central tunnel. It is likely that these two loops form part of the DNA translocation region in the assembled motor. In loop L<sub>1</sub>, a positively charged residue was observed (R101) for the D6E large terminase and may interact directly with DNA. Consistent with this, the equivalent residue R101 for the P74-26 large terminase is also predicted to be in the DNA translocation ring pore based on the pentamer model of this large terminase (56). This is further supported by the observation that mutagenesis of this residue abrogated DNA binding. In addition, a positively charged residue is also found in loop L<sub>2</sub> (K123). This loop is connected by an  $\alpha$ -helix that adopts different conformations in the crystal structure (Figure 8-10). This conformational flexibility is likely to be required for DNA binding or translocation. Furthermore, loops L<sub>1</sub> and L<sub>2</sub> are also predicted to interact with DNA based on structural superposition with super family 1 helicase UrvD (210) (Closest structural homologues by DALI search (174)) and the hexameric papillomavirus helicase E1 (211) (Figure 8-10, Table 8-1). Notably, positively charged residues equivalent to R101 and K123 are also present in P74-26 and Sf6 large terminases (Figure 8-10), supporting the expected role of loops L<sub>1</sub> and L<sub>2</sub> in DNA translocation.



*Figure 8-10 Identification of DNA tunnel loops*

(A) Superposition of D6E (navy blue), T4 (white), Sf6 (coral) and P74-26 (grey) large terminase ATPases showing the bound ATP- $\gamma$ -S (spheres) and the tunnel axis and oligomer interface (black sticks). The position of L<sub>1</sub> and L<sub>2</sub> are indicated. R101 and K123 for the D6E large terminase (thick sticks) and the equivalent residues in T4, Sf6 and P74-26 (ball and sticks) are shown. (B) Superposition of the D6E large terminase ATPase domains from the different molecules within an asymmetric unit. An  $\alpha$ -helix which adopts different conformations is highlighted using a dashed circle. The positions of loops L<sub>1</sub> and L<sub>2</sub> are indicated. (C) Stereo view of the superposition of D6E large terminase ATPase (navy blue) with the E1 helicase (grey) and UrvD helicase (light pink) with bound DNA. Note that loops L<sub>1</sub> and L<sub>2</sub> (thick ribbons) of the D6E large terminase ATPase are in proximity to the DNA (square bases). R101 and K123 are shown in thick sticks. The bound ATP- $\gamma$ -S is shown as a van der Waals model (figure reproduced from (186)).

*Table 8-1 PDBe fold alignment of D6E large terminase ATPase with UvrD and E1 helicases (177).*

	Q-score	Z-score	RMSD(Å)	N <sub>align</sub>	N <sub>sse</sub>	Seq-%	N <sub>res-Q</sub>
UvrD	0.084	5.9	3.1	160	11	0.066	649
E1	0.067	4.8	3.0	89	6	0.101	269

### 8.13 Comparison of inter-domain conformations and interactions

Although the structures of the individual ATPase and nuclease domains show high similarities to those found in other large terminases, the D6E large terminase has a distinct inter-domain arrangement. Superposition of the ATPase domain of the T4, Sf6 and D6E large terminase structures reveal that a significant rotation is required to move the D6E

nuclease domain in order for it to occupy a similar position to that observed in T4 or Sf6 (Figure 8-11). The observed domain orientations are stabilised by intra-domain electrostatic and VDW interactions. Compared to T4 and Sf6, more intra-domain interactions were observed for D6E. Specifically, a larger VDW interaction area was observed for D6E than for T4 and Sf6 (362 vs 212 and 78) (Table 8-2). In addition, three intra-domain electrostatic interactions were found for the D6E terminase, including two salt bridges, while only one for Sf6 and none for T4 (Table 8-3). It is possible that the stabilisation of the distinct domain arrangement of the D6E large terminase by these interactions represents a biologically important conformational state of the large terminase, adopted by thermophilic viruses.

*Table 8-2 Comparison of intra-domain electrostatic interactions (table reproduced from (186))*

Electrostatic interactions		ASCE subdomain-nuclease domain	Lid subdomain-nuclease domain
Large terminase			
D6E	M216-N362 (1)	R224-D394 (2)	
T4	-	-	
Sf6	-	K434-T197 (1)	

*Table 8-3 Comparison of intra-domain VDW interaction areas (table reproduced from (186))*

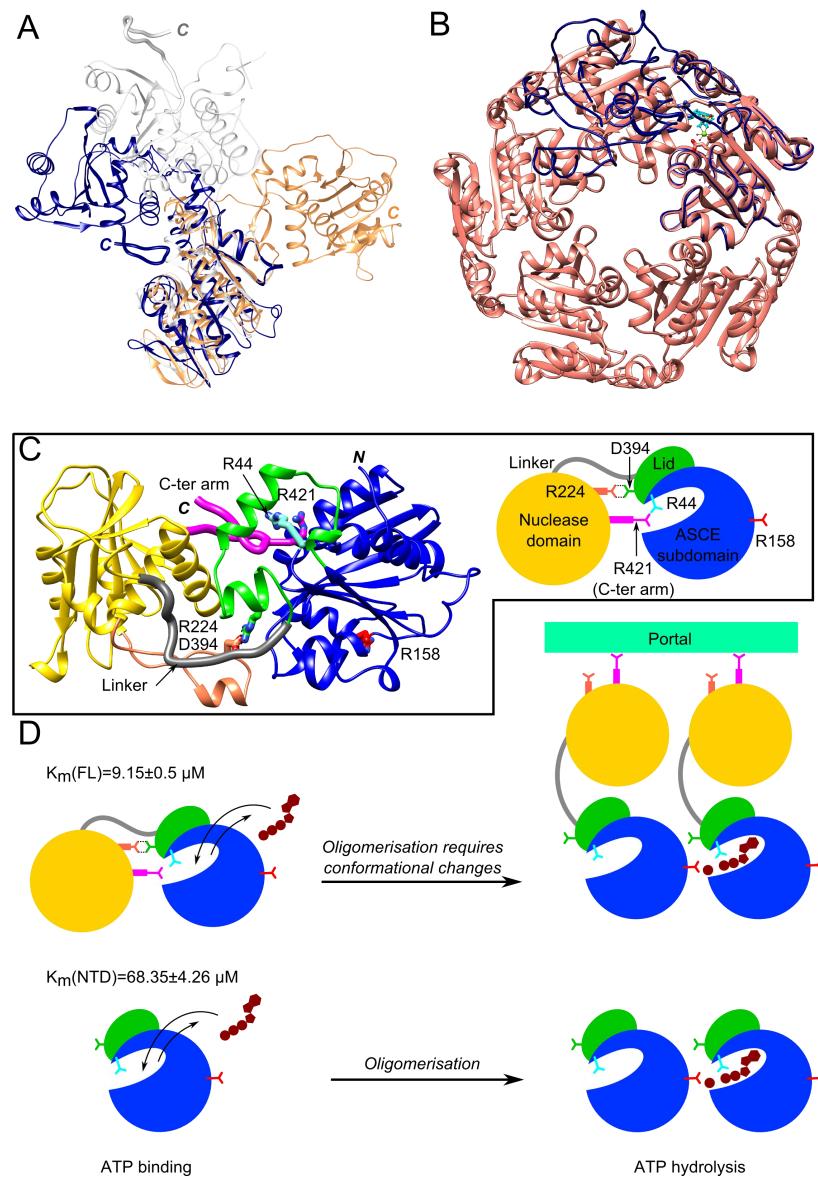
Areas of VDW interaction		ASCE subdomain-nuclease domain	Lid subdomain-nuclease domain	Total between ATPase and Nuclease domains
Large terminase				
D6E	119	243	362	
T4	196	6	212	
Sf6	30	48	78	

## 8.14 Formation of oligomeric assembly of large terminase requires large conformational changes

Based on the above analysis, it is likely that motor assembly requires significant rearrangement in the relative orientation of the ATPase and nuclease domains. This is supported by the observation that the domain orientation seen in the crystal structure is incompatible with the pentameric ATPase model, with the nuclease domain clashing with the ATPase domain of adjacent subunit of the pentamer (Figure 8-11). Taken all the data together, a model for the oligomeric assembly of the D6E large terminase was proposed below.

Before assembly, ATP binds in the active site of the large terminase monomer, facilitated by the C-terminal arm and R421. The bound ATP cannot be hydrolysed without the *trans*-acting arginine finger from an adjacent subunit (Figure 8-11). The monomeric conformation is stabilised by inter-domain contacts (Table 8-2, Table 8-3), described above, hindering the self-assembly of the pentameric ring. However, in the absence of the nuclease domain, the

ATPase domain has 1) increased  $k_{cat}$ , suggesting that ATP hydrolysis becomes more efficient within the pentamer assembly; 2) increased  $K_m$ , indicating decreased ATP binding affinity, which could be explained by the enhanced conformational flexibility of the lid subdomain (and associated P-loop) due to increased solvent exposure upon nuclease domain truncation (Figure 8-3). Given that the C-terminal arm of the nuclease domain has been implicated in the assembly of the active motor complex onto the portal vertex of viral capsids (44), it is likely that this interaction facilitates rotation of the nuclease domain and the exposure of the domain interface for efficient assembly of the pentameric motor (Figure 8-11).



*Figure 8-11 Proposed mechanism for large terminase motor assembly*

(A) Superposition of the ATPase domain of D6E (navy blue), T4 (coral) and Sf6 (white) large terminases. (B) Structure superposition of the ATPase domain of full-length D6E large terminase (navy blue) onto a subunit of the pentameric molecular docking model of the ATPase ring (salmon). (C) Schematic representations of the D6E large terminase. The

*ATPase and nuclease domain of the D6E large terminase protein are represented by coloured spheres. The side chain of R44 (cyan), R158 (fire brick) and R421 (magenta) are shown in sticks. (D) Schematic model for large terminase motor assembly and coupled ATP hydrolysis. The scheme describes the proposed mechanism of ATP binding, hydrolysis and oligomerisation for the full-length large terminase (top) and ATPase domain alone (bottom) (figure reproduced from (186)).*

## 8.15 Summary

In this chapter, the oligomeric state and ATP hydrolysis properties of the full-length D6E large terminase as well as the isolated ATPase domain were investigated. After the removal of the nuclease domain, pentameric assembly was observed for the ATPase domain, with a significant increase in  $k_{cat}$  upon ATP hydrolysis. These data suggest a more efficient assembly and ATP hydrolysis profile of the isolated ATPase domain compared to the full-length large terminase. A C-terminal arm found in close proximity to the ATPase active site in the crystal structure, is playing role in facilitating ATP binding. An arginine residue (R158) predicted to contribute to the active site of an adjacent subunit in the ATPase pentamer, generated by molecular docking, is indispensable for catalysis. Complementary biochemical assays show that this arginine promotes ATP hydrolysis *in trans*, as found for the phi29 packaging ATPase and HerA/FtsK translocase. This suggests that the previous classification of large terminases as a member of the ASCE division along with the HerA/FtsK, PilT, RecA and helicase superfamilies (49,50) may need to be reconsidered. Structure comparison with the T4 and Sf6 large terminases demonstrates large variation in mutual position and orientation of the ATPase and nuclease domains. The observed incompatibility of the nuclease domain orientation with the pentameric model suggests that formation of the oligomeric assembly requires significant adjustments in relative positions of the two domains, supported by the SV-AUC and ATPase kinetics data.

## 9 Conclusions

Most of the dsDNA bacteriophages and viruses utilise a powerful DNA packaging machinery that is capable of translocating their genomic DNA into an empty preformed procapsid. The key components of the machinery are the portal, small terminase and large terminase proteins. The large terminase plays an essential role in the DNA packaging process, by generating force through ATP hydrolysis to power DNA translocation while its nuclease activity cleaves the DNA concatemer before and after packaging.

This work focused on the structural studies as well as biochemical and biophysical characterisations of large terminases from two thermophilic bacteriophages, *Thermus* phage G20c and *Geobacillus* phage D6E. These data provide important mechanistic insights into how large terminase proteins binds to- and cleave- DNA, and how they assemble and hydrolyse ATP. Additionally, large terminases from thermophilic bacteriophages GVE2, GBSV1 and phiOH2 were also cloned, expressed and purified as soluble proteins. An efficient solubilisation and purification method was developed for the GVE2 large terminase protein using CHES buffer and arginine-hydrochloride. Preliminary characterisation suggests that these proteins are active nucleases while the phiOH2 terminase also has an apparent ATPase activity.

Structural and functional characterisation of the G20c large terminase protein identified two metal binding sites and provided a structural basis for the metal dependent nuclease activity of the large terminase. A subsequent structural similarity search revealed that the protein is most similar to RuvC endonucleases of the RNase H-like superfamily. The similarity suggests that the large terminase is likely to use the same two-metal catalysis mechanism as proposed for RuvC endonucleases (74). Based on this, a detailed mechanism of DNA cleavage and model for a transition state, can be proposed (81). Future work could focus on obtaining X-ray structure of the nuclease-DNA complex. Although the protein has nuclease activity, unlike RNase H and RuvC proteins, the isolated nuclease domain does not form a stable complex with DNA under test conditions (Chapter 7). Cross-linking experiments, using backbone or base modified DNA with a thiol group and large terminase with a cysteine mutation at suitable location could be used to produce a stable protein-DNA complex which can be used for further structural work.

Structural studies and characterisation of the D6E large terminase showed that the full-length protein oligomerised in the sedimentation velocity AUC experiment while the formation of pentamer was observed for the isolated ATPase domain. Based on these observations, a pentameric model of the ATPase domain was constructed by molecular docking. This model placed the ATPase active site at an interface between adjacent subunits

with arginine R158 from the adjacent subunit contributing to the catalytic site. This arginine was later confirmed to be a *trans*-acting arginine finger, in common with functionally equivalent arginine residues of other ring-like ASCE ATPases. The arginine finger is located at the same position of the canonical ASCE fold, as that identified for the phi29 packaging ATPase and HerA/Ftsk superfamily ATPases. The C-terminal arm of the nuclease domain, occluding the ATPase active site in the crystal structure, plays a role in facilitating ATP binding. Compared to T4 and Sf6 large terminases, the crystal structure of the large terminase protein of D6E reveals a completely distinct mutual orientation of the ATPase and nuclease domains. This orientation appears to be stabilised by the C-terminal arm and is incompatible with the pentamer model of the isolated ATPase domain. This notion is supported by the increase in  $k_{cat}$  for ATP hydrolysis upon removal of the C-terminal nuclease domain. Taken together, structural and biochemical data suggest a model where transition from the “initiation state” with high affinity towards ATP, into a catalytically competent pentameric state, is accompanied by substantial domain rearrangements. In this model, the C-terminal arm promotes ATP binding while stabilising the resting state. Protein activation upon pentamer formation is predicted to occur with the C-terminal arm occupying a new position, possibly allowing interaction with the portal protein, as has been suggested for large terminase of several bacteriophages. Future work could be directed at: 1) producing the D6E portal protein to verify the interaction between the C-terminal arm and the portal protein using FRET. 2) probing the role of residues predicted to be important for motor assembly and DNA translocation, including positively charged residues on L<sub>1</sub> and L<sub>2</sub> loops, by site directed mutagenesis; 3) producing stable pentameric assemblies by GraFix for cryo-EM studies.

The characterisation of DNA binding properties of the D6E large terminase suggested that the apparent binding affinity becomes higher as the length of the DNA increases. The protein-DNA interaction was highly cooperative and a similar increase in cooperativity was also observed at longer DNA length. According to the frictional ratio obtained from AUC data fitting, the shape of the protein-DNA complex was undergoing a change from globular, in the presence of short DNA fragments, to rod-like when longer DNA were used. Taken these data together, the protein appears to bind to the DNA non-specifically, forming a non-biological assembly. The increased length of the DNA provided more binding sites for the protein which may explain the observed increase in binding affinity and cooperativity. The observed high binding cooperativity can be due to protein-protein interactions, given the concentration dependent oligomer formation, as observed by AUC experiments (Chapter 8). Future work could be directed at preparing a biologically relevant protein-DNA complex in

the presence of other motor components, such as the portal protein and/or capsid; and at cryo-EM studies of such assemblies.

Preliminary characterisation of the phiOH2 large terminase showed that this protein has nuclease and ATPase activities *in vitro*. Distinct to G20c, D6E and other reported large terminases which mainly use Mn<sup>2+</sup> and Mg<sup>2+</sup> for DNA cleavage, this protein is able to utilise Zn<sup>2+</sup>, Cu<sup>2+</sup> and Ni<sup>2+</sup> for nuclease activity. This is unusual, because Zn<sup>2+</sup>, Cu<sup>2+</sup> and Ni<sup>2+</sup> have been reported as non-active co-factors for proteins belonging to the RNase H-like superfamily (162). Future work could be directed at optimising the protein construct for crystallisation and determining the crystal structure of the phiOH2 large terminase protein in the presence of divalent metals and DNA.

# Appendices

## Appendix 1

*Appendix 1 Oligonucleotide sequences used for PCR amplification and colony PCR.*

Gene construct	Vector	Primers 5'----3'
D6E WT 1-485	YSBL3CLIC+	CCAGGGACCAGCAATGAAAAAGCTGAAGCCGG GAGGAGAAGGCGCGTTAATATAAACTCGAACGGC GC
G20c WT 257-443	YSBL3CLIC+	CCAGGGACCAGCAATGAACTCGGTCTTCTCCGG GAGGAGAAGGCGCGTTAGCCCTCGTACTGGCTG
GBSV1 WT 1-573	YSBL3CLIC+	CCAGGGACCAGCAATGCGTAAATATAAAAATTATG ATGCAGTGATG GAGGAGAAGGCGCGTTAACCGCCTTCAGCCACTTG
GVE2 WT 1-567	YSBL3CLIC+	CCAGGGACCAGCAATGATGAGAGTTACTTGAGC GAGGAGAAGGCGCGTTACCATCCATCATTAGA TAATC
phiOH2 WT 1-420	YSBL3CLIC+	CCAGGGACCAGCAATGAAGCAGATTAGGCTTCTG GAGGAGAAGGCGCGTTACTATTCAATATCGATAACG GACGG
1-1 WT 1-666	YSBL3CLIC+	CCAGGGACCAGCAATGGGCAAGGAAAAGGAAAC GAGGAGAAGGCGCGTTACTAAAATCCATGGATAT CCTCATG
GVE2 WT 1-567	SUMO	GAACAGATTGGTGGTATGATGAGAGTTACTTGAG C TACCTAAGCTGTCTACCATCCATCATTAGA TAA TC
1-1 WT 1-666	SUMO	GAACAGATTGGTGGTATGGGCAAGGAAAAGGAAAC TACCTAAGCTGTCTAAAATCCATGGATATCCTC ATG
phiOH2 WT 1-420	SUMO	GAACAGATTGGTGGTATGAAGCAGATTAGGCTTCT G TACCTAAGCTGTCTATTCAATATCGATAACGGAC GG

## Appendix 2

*Appendix 2 Oligonucleotide sequences used for site-directed mutagenesis.*

Gene construct	Vector	Primer 5'----3'
D6E D326N 1-427	YSBL3CLIC+	GGCGTGATCGTAATCCGAGTGCCGC GCGGCACTCGGATTACGATCACGCC
D6E E143A 1-427	YSBL3CLIC+	CGGCATCAGTGCCACTGCGTCAAAGAACATG GGCATGTTCTTGACGCAGTGGCACTGATGC
D6E R44A 1-427	YSBL3CLIC+	GCAGCATCGCTGCCGGCAAGACC GGTCTGCCGGCAGCGATGCTGC
D6E R421A 1-427	YSBL3CLIC+	CTGTTCGGCAATAAACTGGCCGCCGTTCCG AACTCGGAACGGCGGCCAGTTATTGCCG
D6E R158A 1-427	YSBL3CLIC+	CGTCAACGCTACAGGCTGCCGGCTG ATCAGGCCACCGCAGCCTGTAGCGTTG
D6E WT 1-417	YSBL3CLIC+	CGGAACGGCGCGCAGTTCTAGCCGAACAGAATTGT ATT AATACAATTCTGTTGGCTAGAAACTGCGCGCCGTT CCG
D6E WT 1-234	YSBL3CLIC+	GCCTGTGGGTTCTGGCATAGGGTATTATTATGATAT

		G CATATCATAAATAATACCCATGCCAGAACCCACAG GC
G20c D347N 1-485	YSBL3CLIC+	CTTACGTGGTGGCGAACACCTGGGCG CGCCCCAGGTGTTGCCACCACGTAAG
G20c D429N 1-485	YSBL3CLIC+	GACGCGGGCACGACAACATCGTCATGTCC GGACATGACGATGTTGTCGTGCCCGCGTC
G20c D300N 1-485	YSBL3CLIC+	CTTCGGTAAGAACCGAAACTACTCCGTGTTAGC GCTAACACCGAGTAGTTCTGGTTCTTACCGAAG
G20c D300A 1-485	YSBL3CLIC+	CGCTAAACACGGAGTAGGCCTGGTTCTTACCGAAG CTTCGGTAAGAACCGAGGCCTACTCCGTGTTAGCG
G20c H427N 1-485	YSBL3CLIC+	CTTACGGACGCGGGAACGACGACATCGTC GACGATGTCGTCGTTCCCGCGTCCGTAAG
G20c D428A 1-485	YSBL3CLIC+	GGACGCAGGCACGCCACATCGTCATG CATGACGATGTCGGCGTGCCCGCGTCC
G20c D428N 1-485	YSBL3CLIC+	CGGACGCAGGCACGCCACATCGTCATGTC GACATGACGATGTCGTTGTGCCCGCGTCCG
G20c H427A 1-485	YSBL3CLIC+	CTTACGGACGCGGGGCCACGCCACATCGTCATG TGACGATGTCGTCGCCCCCGCGTCCGTAAGCG
G20c WT 1-443	YSBL3CLIC+	CTTGTACCCGTCCTAGCCCTCGTACTGGC GCCAGTACGAGGGCTAGGACGGGTACAAG

### Appendix 3

*Appendix 3 Structure based protein sequence alignment of T4, Sf6, G20c, D6E, GBSV1, phiOH2, GVE2 and 1-1 large terminases*

Conservation:  
1-1 9  
T4 1 MGKE-----KETLSM---TLEEDLSKDAPRVENFLKYLGGS----VSKR--ELYSIVARI 46  
phiOH2 1 MEQPINVNLDFHPLNEAGKILIKHPSIAERK-----DEDGIHWIKSQ-----WDGKWYPEKFSDYLRL 58  
GVE2 1 MKQ-----IQLSE-----VFTP----- 12  
Sf6 1 MIE-----SILLERIVHYAQDVVDGRTIACQ-----KHKWACERF 34  
GBSV1 1 MTS-----INP----- 6  
G20c 1 MRKY-----KNYDAVMEYAKSIVDGRKVACR-----ELIQAAKRF 35  
D6E 1 MKRL-----RPSDKF----- 10  
1-1 1 MKKL-----M.p...... 4  
Consensus aa: .....  
Consensus ss: .....  
hhhhhhh h h hhhhh

Conservation:  
1-1 47 -----LRYE-STPPEIDEFISDP-----AYLGGV----- 69  
T4 59 HKIVKIPNN-SDKPELFQTYKDKRSRYMGMLKRANIKTQWTREMVEEWKKCRDDIVYFAETYCAI 127  
phiOH2 -----  
GVE2 35 -----LNDIERITED-----DFPYYFDGEELYR-----FYRWARMF-KH 67  
Sf6 -----  
GBSV1 36 -----FKDLEN-----PKYDFNPREAEF-----VIQIIEKTFVH 64  
G20c -----  
D6E -----  
1-1 -----  
Consensus aa: .....  
Consensus ss: .....

Conservation:  
1-1 70 -----LDGGKAV--YEYW-KAALRAIYPNPF---QSPYLEVVATGAIGIGKTTLAKIGSSYDI 121  
T4 128 THID-----YGVIVKVQ---LRDYQRDMLKIM-----SSKRMTVCNLSQLGKTTVVAI----- 172  
phiOH2 13 -----TFQKVWALVK-----QQRYLRYVLKGGRASAKSTHIAM----- 45  
GVE2 68 NKGI-----LAGQPIE---LTDFQLFIVGNIFGWKRKENGLRFRKAYIQLARKNAKSQLIAL----- 122  
Sf6 7 -----IPEPFI-----EAHRYKVAKGGRGSGKSWAIAR----- 34  
GBSV1 65 DQGERLTDGTPLRGEPFL---LEFWQKFIIYNLLGFYLKGTKRRFKEAFIFIPRKNGKTRLIAA----- 125  
G20c 11 -----FELLGYK---PHHVQLAIHRS-----TAKRRVACLGRQSGKSEAASV----- 49  
D6E 5 -----KPAPFYFQPFSKQQLKVLTWWWRKASP---VSDKDGIICDGSIRAGKTIVMSF----- 53  
1-1 -----hb..hh.....p.....hb...t...tKo.hhh.....  
Consensus aa: .....  
Consensus ss: .....  
hhhhhhhhhh eeeeeee hhhhhh

**Walker A**

Conservation:		5      5	
1-1_	122	AKMTMLSNPQKIYIVKEE--KIQYAIINAT-MSLAKSVIYDELISWWHASPYFSNL-----M	176
T4	173	-----FLAHFVCFNK--DKAVGILAHK-GSMSA-EVLDRTKQAIELLPDFLQPGI-----V	219
phiOH2	46	-----MVLLVMRYP--VTALVVRRVG-NTLAD-SVLEQLKEAMEILG-VTEY-----F	89
GVE2	123	-----IASYECFLSPE-QSEVYIAAGWG-REQSS-IVYNEVLAQVACDILLKGK-----Y	168
Sf6	35	-----LLVVAARRQP--VRILCARELQ-NSISD-SVIRLLEDTIEREG-YSAE-----F	78
GBSV1	126	-----LAWALALLERQSGSKIYITSAA-IQQSL-QSFEPILFNLRQMGEEQNFRILNN-----N	177
G20c	50	-----EAVFELFARP--GSQGWIAAPT-YDQAE-IIFGRVVEKVERLSEVFPTTEVQLQRRLRL	106
D6E	54	-----SYVMWAMDTF-NEQNFGMAGKTIGALRR-NVITPLKRMILKSRG-YRVK-----D	99
<u>Consensus aa:</u>		.....bh <sub>hh</sub> .h <sub>h</sub> .p....p.h.h.....pb <sub>t</sub> ...1hp..lb..hc..s.h.....	
<u>Consensus ss:</u>		hhhhhhh        eeeeeee        hhhh        hhhhhhhhhhh        hhhhh	

Conservation:		5      6      6      67	
1-1_-	177	KKKNKS--KDTLF----PKRVHLTGVSR--PQHFIGK---HIFGAIMDEMNQDRVANQAYDNYT	229
T4	220	EWNKG--SIELD----NGSSTGAYASS--PDAVRGN---SFAMIYIECAFI PN---FHDSWL	268
phiOH2	90	QVINITPMRITYLP----RGNRILFRGADD--PQKIKISKAFPLAIMWI EELAEFKA---EEEVS	146
GVE2	169	TDSYG--RIRHKK---SGSIIQPLSKEA--RKTGDGK---NPSLAVIDEYHTHET---SEIYD	218
Sf6	79	EIQRS--MIRHLG---TNAEMFYGIKNN--PTKIKSLE---GIDICWVEEAEAVT---KESWD	129
GBSV1	178	QENSI--SGEFSD--GSIYIRALAAN--PDKQDSL---NCNVGIADEMHAYKT---PKQYN	226
G20c	107	VHYHD--RPVNAPGAKRVTASERFRGKSADR--PDNLNRGA---TLDLVILLEAMIPF---SVWSE	161
D6E	100	HRADNYLTITFKG---KTNYFYLFGGKDESSQDLIQQGI---TLAGMFFDEVALMPE---SFVN	153
<u>Consensus aa:</u>		p.p.s..p.ph.....s..h..t.p.....p.pt.....shs.hhh-Eh.hh.....p.hp	
<u>Consensus ss:</u>		ee        eeee        eeeeeee        eeeee        hh        hhhh	

**Walker B**

Conservation:		55	
1-1_-	230	NVLRMKSRFGNVSENFFPGHLWLSSKRRD-TDPLQEHIERSNDPTTIIFDAALWEVKRE---RLN	293
T4	269	AIQPVISSGRR----SKIITTTPNGL--NHFYDIWTA-----VEG---KSG	307
phiOH2	147	VIEKSVLRLGELPDGLR--YTFFYSYNNPKR-RQSWSVNQKYET-----QFL---PEN	191
GVE2	219	VLVSGMVARQN-----PLIVIITTAGFNLASPCYTYEQYQYSKVIDP-----GSPIE---NEE	267
Sf6	130	ILIPTIRKPF----SEIWVSNFPKNI-LDDTYQRFVVNP-----PDD	166
GBSV1	227	IIKEAMKAYTN--KLMIGITTAGDDMTSFCYQLQYCK----KILDE--TVT---DEA	273
G20c	162	AIEPTLSVRD--GWALIISTPKG--LNWFYEFFLGMW---RGGLKEGIPNSGINQTHP	212
D6E	154	QATARCSVDG--AKLWFNCNPAGP-YHWFKVVEYLKD-----EKN	191
<u>Consensus aa:</u>		.....h.....hh..ss....hp.hbp.h.s.....ps	
<u>Consensus ss:</u>		hhhhhh        eeeeeee        hhhhhhhh	

**Sensor I**

Conservation:		5	
1-1_-	294	LSGKYFYVVAGDGSRDPFILGDRYDHKLHFLDDSRIVRVPVEFREDFE----KDFNALARLAGVSV	357
T4	308	-FEPYTAIWNS-----VKERL---YNDEDIFD----DGWQWSIQTING---	344
phiOH2	192	-TFVHHSTYLD-----NP-----FLS-----RDFIEEAHETKRTN-	220
GVE2	268	-YFVMICELDK---DDD----IKDERNWVKANP-----IVATYEEGMNFLRSELQTAIDVPE	316
Sf6	167	-ICLLTVNYTD-----NP-----HFP-----EVRLLEMEECKRRN-	195
GBSV1	274	-YFVFICKADE--DENG--EVDYTNPIEHEKANP-----NYGVT-IRPDDIMNDALQAQNPDQ	325
G20c	213	DFESFHAAASWD-----VM-----PER-----REWYMERRLYI	239
D6E	192	-LLHLHFTMDD-----NL-----SLS-----SLS-----KQVKERYQRMYK--	218
<u>Consensus aa:</u>		.....h..hh..h.p.....s.....h.....c.h..phph...s.	
<u>Consensus ss:</u>		eeeeeee        hhhhhhhh	

Conservation:		5	
1-1_-	358	QGVHKFIPSVEAVESVTGKN-----ITEREVLIEDFHDK--DSLMSYLLPNIERIRNPMLPHSPRYI	418
T4	345	-----LAQFRQEHTAAFEGTSGT-LISGMKLAVMDIEVTPTDDHGHQ--FKKPE---PDRKYIA	398
phiOH2	221	-----EMKYRHEYLGEPIGSGV--PFDNIVLVRTITDE-----EIK---QFDNIRQ	261
GVE2	317	K-----MRSFLTKNMNIWVDQKDNGYLPLDKWRACAIK-----EQIDL---DVRDCYV	361
Sf6	196	-----PTLYRHILGEPVASASDM-AIIKREWLEAATDH-----KKGWK---AKGAVVS	241
GBSV1	326	Q-----RKDFLAKSINIYTSPMKA-YFNIDEFKKSDRKY----DWTI-EQLAK---LGI DWFG	374
G20c	240	P-----DLEFRQEYGAEFVSHSN-S-VFSGLDMLTILPYE--RRGTRLV--VEDYR---PDHIYCI	291
D6E	219	-----GVFYQRYILGLWVLAEGI-IYDMFDQDEHV-----PTVPR---PYEKYYV	260
<u>Consensus aa:</u>		.....@.pb.hsb.hs..s..h.....h.....p.c.....h.	
<u>Consensus ss:</u>		hhhhhhh        hhhhhh	

eeee

Conservation:	9    5    7		
1-1_-	419	HVDLA--LTDYAAIASTYVMEYIEVYREDPSTGTVA <sub>S</sub> YREPILFVEWVIYI-KPK-----	471
T4	399	-----TLCSEGRGQDYHALIIDVTD-----DVWQVGVLH-----	430
phiOH2	262	GIDWG--YGVDPFAFVWHYDK-----TRRTIYAIDEIYG-----	294
GVE2	362	GIDLS--MRIDLTSVSGIVPM-----DGRFVVWSHSFI-PEDTLAEKRRTDKVPYD	410
Sf6	242	AHDPS-DTGPDAKGYASRHGS-----VVKRIAEG-----	269
GBSV1	375	GADLS--KLHDLTAAALYGNYN-----GVIAITHAWF-PIVAATRKAEDNIPLF	422
G20c	292	GADFG--KNQDYSVFSVSLDDL-----GAIACLERMN-----	321
D6E	261	SCDYGG--TQNPTTFLGLWGLYN-----GWMYKVKEYHYDRKEN-----	296
<u>Consensus aa:</u>		shdht.....D..shs.h..p.....sh.b.....	
<u>Consensus ss:</u>		eee        eeeeeeeee        eeeeeeee	

Conservation:		6      6	
1-1_-	472	-----PNQEIPYKIRDIFIIELKKS---KFLKIVSAD-GYQSALLQDRLAGIK--TDV	521
T4	431	-----SNTISHLILP DIVMRYLVE---YNECPVYIELNSTGVSVAKSLYMDLEYGVIC-----	481
phiOH2	295	-----VKLNSREVAEKIKAKN-----YHVEPIIADS-----PKSVDEMKEHGIPR-----	337
GVE2	411	LWVKQGWITVTPGAIVDQYQFIQAYIKRЛАEDK-MWN <sub>I</sub> KEICYDPYN-----ATHFAHEMEAEGYVM-----	470
Sf6	270	-----LLMDINEGADWATSLAIE---DGADHYLWDG <sub>D</sub> G-----GAGLRRQTTAEAFSGKKITA	319
GBSV1	423	GWRDGWLTMNTPTVNFSVVWKFETMRAK--GFKIKQVGFDRKF-----GREFFMAMKQKRFN-----	481
G20c	322	-----GATWSQDVARLKALSED--YGHAYVVAUTWGV---GDAIAEELDAQGIN--YTP	368

D6E 297 -----KQKT**DQEYYEDLMKFIEDIEKHKFKGVIVDPSA**-----**ASFIALLRQKGKIV**---- 343  
Consensus aa: .....hs.....hb.h.c...@....l.h-.....t..h..hc...h.....  
Consensus ss: hhhhhhhhhhhhhh eeeee hhhhhhhhhhhh e

**Conservation:**

1-1 522 **VSV**D-----RDY---NPYL**NLKRAIYEKR**---- 542  
T4 482 **DSYT**DLGMK-----QTK**R**TKAVGCSTLKD**LIEKDK**---- 511  
phiOH2 338 IKGAK-----KGPG---SVEYGEKWL**DLEA**---- 360  
GVE2 471 VEIR-----QGF**R**TLSEPTKRFRELVLSGK---- 495  
Sf6 320 TM**PK**-GSESP**FDEDA**PYQAGAWADEVVQGD**NVRT**IGDVFRNK**R**AQFYYALADRLYRAVV**H**EYADPD 388  
GBSV1 482 VDQP-----QYYKKSEGFRRIEK**QVKD**GK---- 506  
G20c 369 LPVK-----SSV**K**EQLISN**LALLMEKQ**---- 392  
D6E 344 IKAK-----ND**VLD**GIRNVAT**ALNKKM**---- 365  
Consensus aa: h.....p.h..h.phl.c.p.....  
Consensus ss: eee hhhhhhhhhhhhhh

**Conservation:**

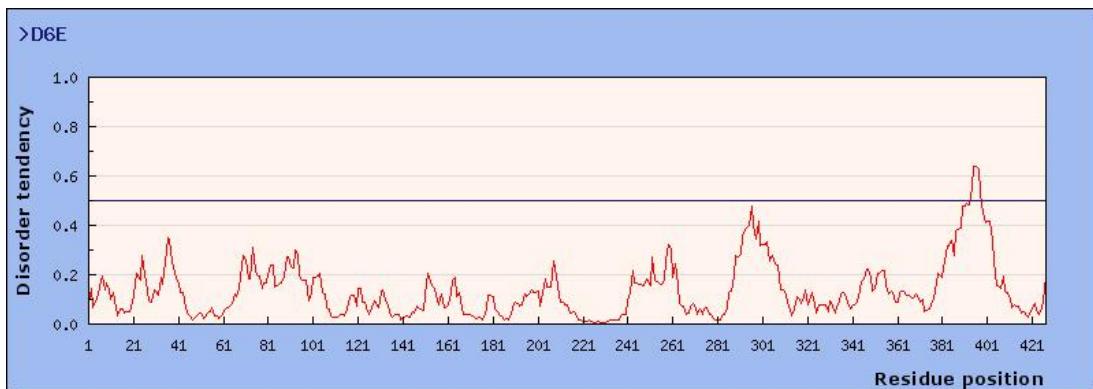
1-1 543 --IVL**PR**-----VSRLKEELLNLIDTGKKIDHPVTGSKDGADAVAG**S**VWLATV**DQSSSR**----KNP 597  
T4 512 --LIIH-----HRATIQEFRTFSEKGVSWA-----AEEGY**H**----DDL 543  
phiOH2 361 --IVIDP-----KRT**PNI**AREFESIDYQVDADG-----NPKPKLEDKN-----NHT 399  
GVE2 496 --ILHDD-----NPV**LNWAI**GNAVVKQDHNE-----NIMLD**DK**STGR----IDP 534  
Sf6 389 DMLSD**D**KEAIGEK**M**LEKLF**AEL**TQI**Q**R**K**FNNNG-----KLELMT**K**VE**M**K**Q**KLGIPSPNL 442  
GBSV1 507 --FYYLH-----SQA**FEYCV**QN**V**HAE**K**TDD-----M**IQF**E**K**I**E**D**K**H**R**----IDI 545  
G20c 393 --VAVPN-----DKT**IL**DEL**R**N**F**RY**R**TASG-----N**QVM**RAY**GRGH**----DDI 430  
D6E 366 --ILYND-----CC**K**ET**FRE**YSSYV**W**DEKAAB-----RGED**K**PVKQN----DHQ 403  
Consensus aa: .lhh.p.....h..h.sh...p.....b.p...p.....p.  
Consensus ss: eee hhhhhhhhhh eee eeeee h

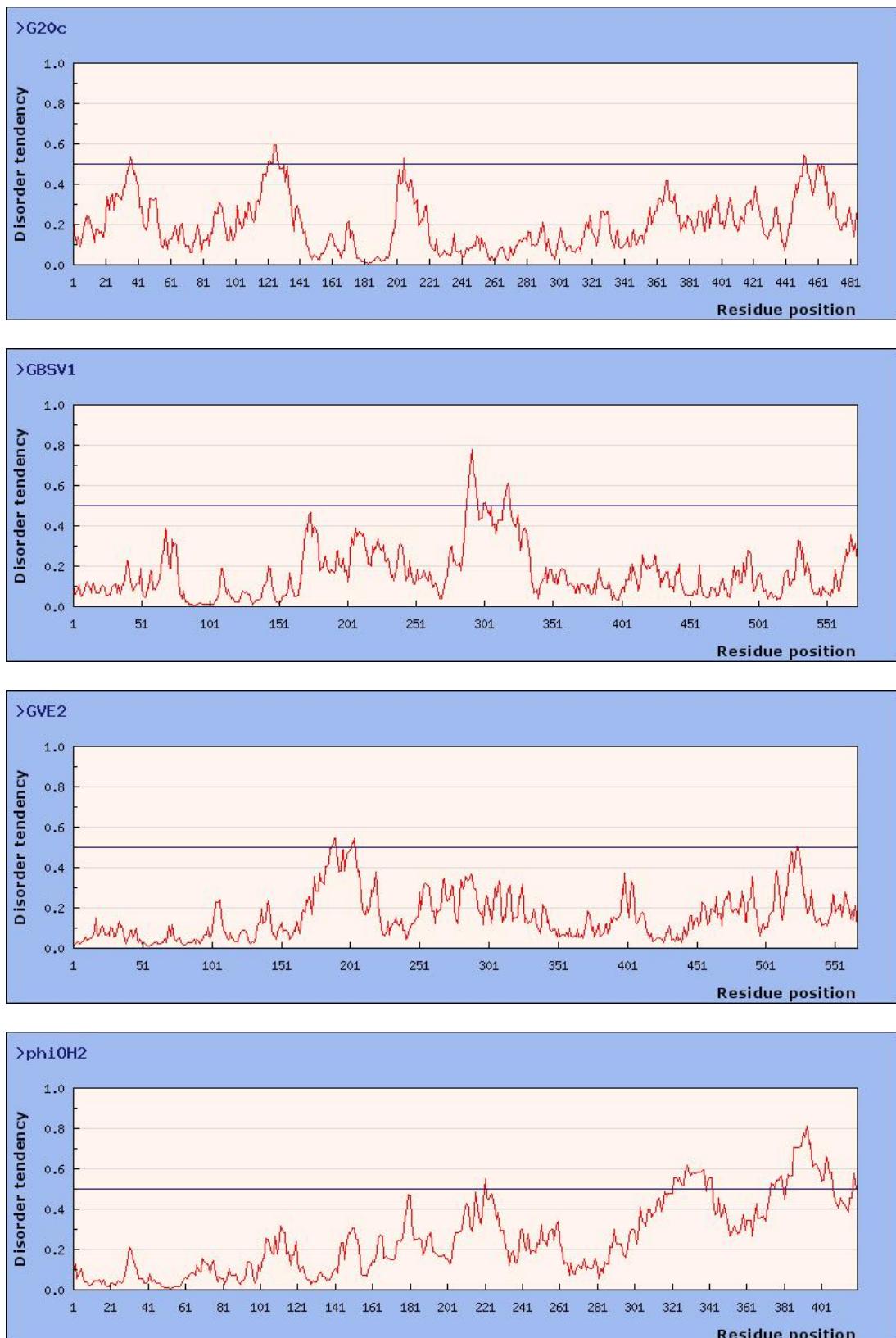
**Conservation:**

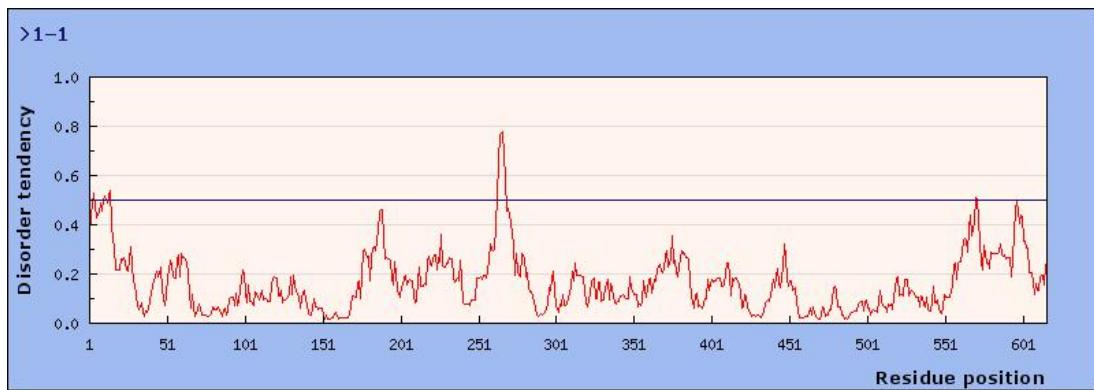
1-1 598 LSALASL**TEEDIM**--RIS-----MGF----- 616  
T4 544 VMSLV**I**F**GWL**STQS**KFIDY**-----ADKDD----- 567  
phiOH2 400 IDATRYAFEDDMK--RPS-----V**SILK**----- 420  
GVE2 535 IAALINAM**VRAIL**--KSD-----EV**DINE**V**TEDY**L**KM**MGW---- 567  
Sf6 443 ADALMM**C**M**HCPA**L--V**R**E-----ETE**I**YV--PSS**SGW**----- 470  
GBSV1 546 FDATVFAAIR**M**LE--N**M**E-----KAATA--TKWL**KGG**----- 573  
G20c 431 VMSL**A**LAYS**Q**Y**E**G--K**DGY**K**F**EL**A**E**E**R**P**SKL**K**HE**E**S**V**MSL**V**E--DDFTD**LE**AN**R**AFSA 485  
D6E 404 LDADRYF**VNTI**I**F**--GNK-----LRAVP-----SLY----- 427  
Consensus aa: h.th..hb.....s.h.....  
Consensus ss: hhhhhhhhhhhh

## Appendix 4

Appendix 4 Protein disorder prediction for T4, Sf6, G20c, D6E, GBSV1, phiOH2, GVE2 and 1-1 large terminases







## Appendix 5

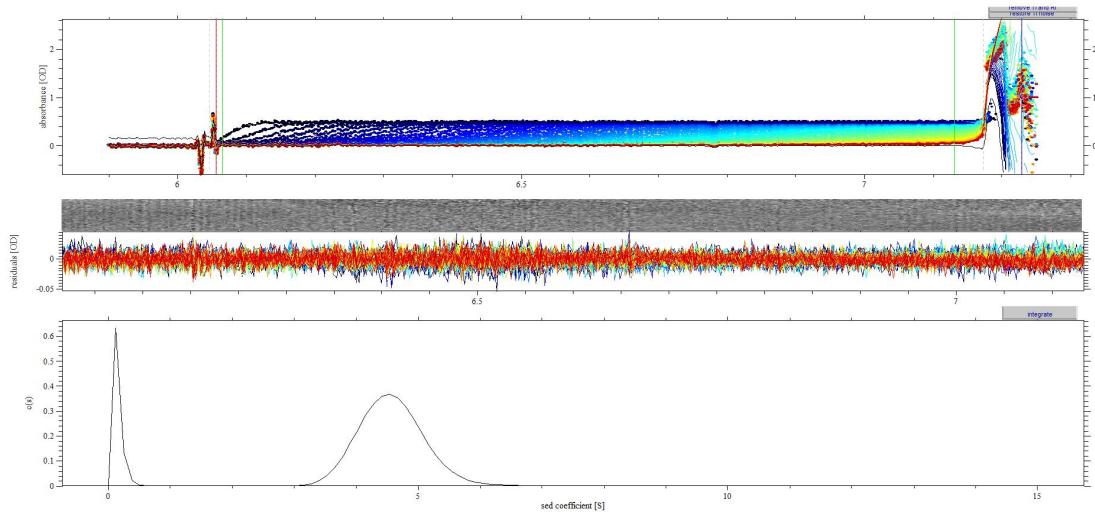
### *Appendix 5 Protein sequence for the 1-1 large terminase*

10	20	30	40	50	60
MGKEKETLSM	TLEEDLSKDA	PRVENFLKYLY	GGSVSKRELY	SIVARIILRYE	STPPEIIDEFI
70	80	90	100	110	120
SDPAYLGGVL	DGGKAVYEW	KAALRAIYPN	PFQSPYLEVV	ATGAIGIGKT	TLAKIGSSYD
130	140	150	160	170	180
IAKMTMLSNP	QKIYGIVKEE	KIQYAIINAT	MSLAKSVIYD	ELISWWHASP	YFSNLMKKNK
190	200	210	220	230	240
SKDTLFPKRV	HTVGSRPQH	FIGKHIFGAI	MDEMNFQDRV	ANQAYDNYTN	VLRRMKSREFG
250	260	270	280	290	300
NVSENFFPGH	LWLLSSKRRD	TDPLQEHEIER	SRNDPTTIIF	DAALWEVKRE	RLNLSGKYFY
310	320	330	340	350	360
VYAGDGSRDP	FILGDRYDHK	LHFHLDSSRI	VRVPVEFRED	FEKDIFNALR	DLAGVSVQGV
370	380	390	400	410	420
HKFIPSVEAV	ESVMTGKNIT	EREVLEIDFH	DKDSLMSYLL	PNIERIRNPM	LPHSPRYIHV
430	440	450	460	470	480
DLALTRDYAA	IASTYVMEYI	EVYREDPSTG	TVASYREPIL	FVEWVIYIKP	KPNQEIPYK
490	500	510	520	530	540
IRDFIIIELKK	SKFPLKIVSA	DGYQSALLLQ	DLRLAGIKTD	VVSVDRDYNP	YLNLKRAIYE
550	560	570	580	590	600
KRIVLPRVSR	LKEELLNLID	TGKKIDHPVT	GSKDGADAVA	GSVWLATVDQ	SSSRKNPLSA
610					
LASLTEEDIM	RISMGF				

## Appendix 6

### *Appendix 6 An example of AUC data fitting using Sednertp*

*The sample contains 10 µM D6E large terminase D326N and 5 µM 5'-FITC labelled Poly(T) 10 ssDNA.*



## Abbreviations

AEBSF 4-(2-Aminoethyl) benzenesulfonyl fluoride hydrochloride

ADP Adenosine diphosphate

APS Ammonium Persulphate

AMPPCP Adenosine 5'-[ $\beta$ ,  $\gamma$ ]-methylene]triphosphate

AMPPNP Adenosine 5'-[  $\beta$ ,  $\gamma$  -imido]triphosphate

ATP Adenosine-5'-triphosphate

ATP- $\gamma$ -S Adenosine 5'-[ $\gamma$ -thio]triphosphate

ASCE Additional Stranded Catalytic glutamate acid (E)

AUC Analytical ultra-centrifugation

Bicine 2-(Bis(2-hydroxyethyl)amino)acetic acid

Bis-tris 2-Bis(2-hydroxyethyl)amino-2-(hydroxymethyl)-1,3-propanediol

Bis-tris propane 2-Bis(2-hydroxyethyl)amino-2-(hydroxymethyl)-1,3-propanediol

Bh- *Bacillus halodurans*

bp Base pairs

CAPS N-cyclohexyl-3-aminopropanesulfonic acid

CCP4 Collaborative Computational Project, number 4

CHES N-Cyclohexyl-2-aminoethanesulfonic acid

*cos* cohesive end site

DIPSO 3-(N,N-Bis[2-hydroxyethyl]amino)-2-hydroxypropanesulfonic acid

DMSO Dimethyl sulfoxide

DNA Deoxyribonucleic acid (ssDNA for single-stranded, dsDNA for double-stranded)

DNase Deoxyribonuclease

RNase Ribonuclease

dNTP Deoxyribonucleotides dATP, dTTP, dCTP and Dgtp

DLS Diamond Light Source

DTT Dithiothreitol

EDTA Ethylenediaminetetraacetic Acid

EM Electron Microscopy  
FAM Fluorescein amidite  
FITC Fluorescein isothiocyanate  
HCMV Human cytomegalovirus  
HEPES 2-[4-(2-hydroxyethyl) piperazin-1-yl] ethanesulfonic acid  
HIV Human immunodeficiency virus  
HRV Human Rhinovirus  
HSV Herpes simplex Virus  
IPTG Isopropyl- $\beta$ -D-thiogalacto-pyranoside  
kb kilobases  
LB Luria–Bertani / Lysogeny broth  
MAD Multiple-wavelength anomalous dispersion  
MALDI-ISD Matrix Assisted Laser Desorption Ionisation- In Source Decay  
MES 2-(N-morpholino) ethanesulfonic acid  
MPD 2-methylpentane-2,4-diol  
MST Micro-scale thermophoresis  
NADH nicotinamide adenine dinucleotide (NAD) + hydrogen (H)  
*pac* packaging site  
PCR Polymerase Chain Reaction  
PDB Protein Data Bank  
PEG Polyethylene Glycol  
PGA Polyglutamic Acid  
RMSD Root Mean Square Deviation  
SAD Single-wavelength anomalous dispersion  
SDS PAGE Sodium dodecyl sulfate polyacrylamide gel electrophoresis  
Th- *Thermus thermophilic*  
Tricine N-(2-Hydroxy-1,1-bis(hydroxymethyl)ethyl)glycine  
Tris 2-Amino-2-(hydroxymethyl)-1,3-propanediol

VDW van der waal

WT wild type

## References

1. Twort, F.W. (1915) An investigation on the nature of ultra-microscopic viruses. *The Lancet*, **186**, 1241-1243.
2. d'Herelle, F. (1917) An invisible microbe that is antagonistic to the dysentery bacillus. *Comptes rendus Acad. Sciences*, **165**, 373-375.
3. Luria, S.E., Delbrück, M. and Anderson, T.F. (1943) Electron Microscope Studies of Bacterial Viruses. *Journal of Bacteriology*, **46**, 57-77.
4. Matsuzaki, S., Rashel, M., Uchiyama, J., Sakurai, S., Ujihara, T., Kuroda, M., Ikeuchi, M., Tani, T., Fujieda, M., Wakiguchi, H. *et al.* (2005) Bacteriophage therapy: a revitalized therapy against bacterial infectious diseases. *J Infect Chemother*, **11**, 211-219.
5. Echols, H. (1972) Developmental Pathways for the Temperate Phage: Lysis VS Lysogeny. *Annual Review of Genetics*, **6**, 157-190.
6. Bruynoghe, R., J. Maisin. (1921) Essais de the 'raeutique au moyen du bacteriophage. *C. R. Soc. Biol.*, **85**, 1120-1121.
7. Sulakvelidze, A., Alavidze, Z. and Morris, J.G. (2001) Bacteriophage therapy. *Antimicrobial agents and chemotherapy*, **45**, 649-659.
8. Slopek, S., Weber-Dabrowska, B., Dabrowski, M. and Kucharewicz-Krukowska, A. (1987) Results of bacteriophage treatment of suppurative bacterial infections in the years 1981–1986. *Arch Immunol Ther Exp*, **35**, 569-583.
9. Yu, M., Slater, M. and Ackermann, H.-W. (2006) Isolation and characterization of *Thermus* bacteriophages. *Archives of virology*, **151**, 663-679.
10. Rice, G., Stedman, K., Snyder, J., Wiedenheft, B., Willits, D., Brumfield, S., McDermott, T. and Young, M.J. (2001) Viruses from extreme thermal environments. *Proceedings of the National Academy of Sciences*, **98**, 13341-13345.
11. Liu, B. and Zhang, X. (2008) Deep-sea thermophilic *Geobacillus* bacteriophage GVE2 transcriptional profile and proteomic characterization of virions. *Appl Microbiol Biotechnol*, **80**, 697-707.
12. Wang, Y. and Zhang, X. (2010) Genome analysis of deep-sea thermophilic phage D6E. *Applied and environmental microbiology*, **76**, 7861-7866.
13. Liu, B., Wu, S. and Xie, L. (2010) Complete genome sequence and proteomic analysis of a thermophilic bacteriophage BV1. *Acta Oceanol. Sin.*, **29**, 84-89.
14. Liu, B., Zhou, F., Wu, S., Xu, Y. and Zhang, X. (2009) Genomic and proteomic characterization of a thermophilic *Geobacillus* bacteriophage GBSV1. *Res Microbiol*, **160**, 166-171.
15. Hjorleifsdottir, S., Aevarsson, A., Hreggvidsson, G., Fridjonsson, O. and Kristjansson, J. (2014) Isolation, growth and genome of the *Rhodothermus* RM378 thermophilic bacteriophage. *Extremophiles*, **18**, 261-270.
16. Matsushita, I. and Yanase, H. (2009) The genomic structure of *Thermus* bacteriophage φIN93. *Journal of biochemistry*, **146**, 775-785.
17. Doi, K., Mori, K., Martono, H., Nagayoshi, Y., Fujino, Y., Tashiro, K., Kuhara, S. and Ohshima, T. (2013) Draft genome sequence of *Geobacillus kaustophilus* GBlys, a lysogenic strain with bacteriophage φOH2. *Genome announcements*, **1**, e00634-00613.

18. Naryshkina, T., Liu, J., Florens, L., Swanson, S.K., Pavlov, A.R., Pavlova, N.V., Inman, R., Minakhin, L., Kozyavkin, S.A., Washburn, M. *et al.* (2006) Thermus thermophilus Bacteriophage φYS40 Genome and Proteomic Characterization of Virions. *Journal of Molecular Biology*, **364**, 667-677.
19. Minakhin, L., Goel, M., Berdygulova, Z., Ramanculov, E., Florens, L., Glazko, G., Karamychev, V.N., Slesarev, A.I., Kozyavkin, S.A., Khromov, I. *et al.* (2008) Genome Comparison and Proteomic Characterization of Thermus thermophilus Bacteriophages P23-45 and P74-26: Siphoviruses with Triplex-forming Sequences and the Longest Known Tails. *Journal of Molecular Biology*, **378**, 468-480.
20. Tamakoshi, M., Murakami, A., Sugisawa, M., Tsuneizumi, K., Takeda, S., Saheki, T., Izumi, T., Akiba, T., Mitsuoka, K., Toh, H. *et al.* (2011) Genomic and proteomic characterization of the large Myoviridae bacteriophage φTMA of the extreme thermophile Thermus thermophilus. *Bacteriophage*, **1**, 152-164.
21. Blöndal, T., Thorisdottir, A., Unnsteinsdottir, U., Hjorleifsdottir, S., Ævarsson, A., Ernstsson, S., Fridjonsson, O.H., Skirnisdottir, S., Wheat, J.O., Hermannsdottir, A.G. *et al.* (2005) Isolation and characterization of a thermostable RNA ligase 1 from a Thermus scotoductus bacteriophage TS2126 with good single-stranded DNA ligation properties. *Nucleic Acids Research*, **33**, 135-142.
22. Lin, L., Han, J., Ji, X., Hong, W., Huang, L. and Wei, Y. (2011) Isolation and characterization of a new bacteriophage MMP17 from Meiothermus. *Extremophiles*, **15**, 253-258.
23. Lin, L., Hong, W., Ji, X., Han, J., Huang, L. and Wei, Y. (2010) Isolation and characterization of an extremely long tail Thermus bacteriophage from Tengchong hot springs in China. *Journal of basic microbiology*, **50**, 452-456.
24. Chelikani, V., Ranjan, T. and Kondabagil, K. (2014) Revisiting the genome packaging in viruses with lessons from the “Giants”. *Virology*, **466**, 15-26.
25. Feiss, M. and Rao, V. (2012) In Rossmann, M. G. and Rao, V. B. (eds.), *Viral Molecular Machines*. Springer US, Vol. 726, pp. 489-509.
26. Casjens, S.R. (2011) The DNA-packaging nanomotor of tailed bacteriophages. *Nature Reviews Microbiology*, **9**, 647-657.
27. Roos, W., Ivanovska, I., Evilevitch, A. and Wuite, G. (2007) Viral capsids: mechanical characteristics, genome packaging and delivery mechanisms. *Cellular and molecular life sciences*, **64**, 1484.
28. Isidro, A., Henriques, A.O. and Tavares, P. (2004) The portal protein plays essential roles at different steps of the SPP1 DNA packaging process. *Virology*, **322**, 253-263.
29. Catalano, C.E., Cue, D. and Feiss, M. (1995) Virus DNA packaging: the strategy used by phage λ. *Molecular microbiology*, **16**, 1075-1086.
30. Kufer, B., Backhaus, H. and Schmieger, H. (1982) The packaging initiation site of phage P22. *Molecular and General Genetics MGG*, **187**, 510-515.
31. Rao, V.B. and Black, L.W. (2005), *Viral genome packaging machines: genetics, structure, and mechanism*. Springer, pp. 40-58.
32. Bravo, A., Alonso, J.C. and Trautner, T.A. (1990) Functional analysis of the *Bacillus subtilis* bacteriophage SPP1 pac site. *Nucleic acids research*, **18**, 2881-2886.
33. Büttner, C.R., Chechik, M., Ortiz-Lombardía, M., Smits, C., Ebong, I.-O., Chechik, V., Jeschke, G., Dykeman, E., Benini, S. and Robinson, C.V. (2012) Structural basis for DNA recognition and loading into a viral packaging motor. *Proceedings of the National Academy of Sciences*, **109**, 811-816.

34. Zhao, H., Finch, C.J., Sequeira, R.D., Johnson, B.A., Johnson, J.E., Casjens, S.R. and Tang, L. (2010) Crystal structure of the DNA-recognition component of the bacterial virus Sf6 genome-packaging machine. *Proceedings of the National Academy of Sciences*, **107**, 1971-1976.
35. Roy, A., Bhardwaj, A., Datta, P., Lander, G.C. and Cingolani, G. (2012) Small terminase couples viral DNA binding to genome-packaging ATPase activity. *Structure*, **20**, 1403-1413.
36. Greive, S.J., Fung, H.K., Chechik, M., Jenkins, H.T., Weitzel, S.E., Aguiar, P.M., Brentnall, A.S., Glousieau, M., Gladyshev, G.V. and Potts, J.R. (2016) DNA recognition for virus assembly through multiple sequence-independent interactions with a helix-turn-helix motif. *Nucleic acids research*, **44**, 776-789.
37. Gao, S., Zhang, L. and Rao, V.B. (2016) Exclusion of small terminase mediated DNA threading models for genome packaging in bacteriophage T4. *Nucleic acids research*, **44**, 4425-4439.
38. McNulty, R., Lokareddy, R.K., Roy, A., Yang, Y., Lander, G.C., Heck, A.J., Johnson, J.E. and Cingolani, G. (2015) Architecture of the complex formed by large and small terminase subunits from bacteriophage P22. *Journal of molecular biology*, **427**, 3285-3299.
39. Sun, L., Zhang, X., Gao, S., Rao, P.A., Padilla-Sánchez, V., Chen, Z., Sun, S., Xiang, Y., Subramaniam, S. and Rao, V.B. (2015) Cryo-EM structure of the bacteriophage T4 portal protein assembly at near-atomic resolution. *Nature communications*, **6**.
40. Sun, S., Kondabagil, K., Draper, B., Alam, T.I., Bowman, V.D., Zhang, Z., Hegde, S., Fokine, A., Rossmann, M.G. and Rao, V.B. (2008) The structure of the phage T4 DNA packaging motor suggests a mechanism dependent on electrostatic forces. *Cell*, **135**, 1251-1262.
41. Daudén, M.I., Martín-Benito, J., Sánchez-Ferrero, J.C., Pulido-Cid, M., Valpuesta, J.M. and Carrascosa, J.L. (2013) Large terminase conformational change induced by connector binding in bacteriophage T7. *Journal of Biological Chemistry*, **288**, 16998-17007.
42. Oliveira, L., Cuervo, A. and Tavares, P. (2010) Direct interaction of the bacteriophage SPP1 packaging ATPase with the portal protein. *Journal of Biological Chemistry*, **285**, 7366-7373.
43. Hegde, S., Padilla-Sánchez, V., Draper, B. and Rao, V.B. (2012) Portal-large terminase interactions of the bacteriophage T4 DNA packaging machine implicate a molecular lever mechanism for coupling ATPase to DNA translocation. *Journal of virology*, **86**, 4046-4057.
44. Dixit, A.B., Ray, K., Thomas, J.A. and Black, L.W. (2013) The C-terminal domain of the bacteriophage T4 terminase docks on the prohead portal clip region during DNA packaging. *Virology*, **446**, 293-302.
45. Ortega, M.E., Gaussier, H. and Catalano, C.E. (2007) The DNA Maturation Domain of gpA, the DNA Packaging Motor Protein of Bacteriophage Lambda, Contains an ATPase Site Associated with Endonuclease Activity. *Journal of Molecular Biology*, **373**, 851-865.
46. Kanamaru, S., Kondabagil, K., Rossmann, M.G. and Rao, V.B. (2004) The functional domains of bacteriophage t4 terminase. *Journal of Biological Chemistry*, **279**, 40795-40801.

47. Sun, S., Kondabagil, K., Gentz, P.M., Rossmann, M.G. and Rao, V.B. (2007) The structure of the ATPase that powers DNA packaging into bacteriophage T4 procapsids. *Molecular cell*, **25**, 943-949.
48. Rao, V.B. and Mitchell, M.S. (2001) The N-terminal ATPase site in the large terminase protein gp17 is critically required for DNA packaging in bacteriophage T4. *Journal of molecular biology*, **314**, 401-411.
49. Burroughs, A., Iyer, L. and Aravind, L. (2007), *Gene and protein evolution*. Karger Publishers, Vol. 3, pp. 48-65.
50. Iyer, L.M., Makarova, K.S., Koonin, E.V. and Aravind, L. (2004) Comparative genomics of the FtsK–HerA superfamily of pumping ATPases: implications for the origins of chromosome segregation, cell division and viral capsid packaging. *Nucleic acids research*, **32**, 5260-5279.
51. Rao, V.B. and Feiss, M. (2008) The Bacteriophage DNA Packaging Motor. *Annual Review of Genetics*, **42**, 647-681.
52. Walker, J.E., Saraste, M., Runswick, M.J. and Gay, N.J. (1982) Distantly related sequences in the alpha-and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *The EMBO journal*, **1**, 945.
53. Hanson, P.I. and Whiteheart, S.W. (2005) AAA+ proteins: have engine, will work. *Nature Reviews Molecular Cell Biology*, **6**, 519-529.
54. Mitchell, M.S. and Rao, V.B. (2004) Novel and deviant Walker A ATP-binding motifs in bacteriophage large terminase–DNA packaging proteins. *Virology*, **321**, 217-221.
55. Ortiz, D., Ordyan, M., Sippy, J., Oh, C.-S., Keller, N., Feiss, M., Catalano, C.E. and Smith, D.E. (2016) Walker-A Motif Acts to Coordinate ATP Hydrolysis with Motor Output in Viral DNA Packaging. *Journal of molecular biology*, **428**, 2709-2729.
56. Hilbert, B.J., Hayes, J.A., Stone, N.P., Duffy, C.M., Sankaran, B. and Kelch, B.A. (2015) Structure and mechanism of the ATPase that powers viral genome packaging. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, E3792-3799.
57. Zhao, H., Christensen, T.E., Kamau, Y.N. and Tang, L. (2013) Structures of the phage Sf6 large terminase provide new insights into DNA translocation and cleavage. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 8075-8080.
58. Mitchell, M.S. and Rao, V.B. (2006) Functional analysis of the bacteriophage T4 DNA-packaging ATPase motor. *Journal of Biological Chemistry*, **281**, 518-527.
59. Mitchell, M.S., Matsuzaki, S., Imai, S. and Rao, V.B. (2002) Sequence analysis of bacteriophage T4 DNA packaging/terminase genes 16 and 17 reveals a common ATPase center in the large subunit of viral terminases. *Nucleic acids research*, **30**, 4009-4021.
60. Tsay, J.M., Sippy, J., Feiss, M. and Smith, D.E. (2009) The Q motif of a viral packaging motor governs its force generation and communicates ATP recognition to DNA interaction. *Proceedings of the National Academy of Sciences*, **106**, 14355-14360.
61. Dhar, A. and Feiss, M. (2005) Bacteriophage λ terminase: alterations of the high-affinity ATPase affect viral DNA packaging. *Journal of molecular biology*, **347**, 71-80.

62. Cordin, O., Tanner, N.K., Doère, M., Linder, P. and Banroques, J. (2004) The newly discovered Q motif of DEAD - box RNA helicases regulates RNA - binding and helicase activity. *The EMBO journal*, **23**, 2478-2487.
63. Gorbalyena, A.E. and Koonin, E.V. (1993) Helicases: amino acid sequence comparisons and structure-function relationships. *Current Opinion in Structural Biology*, **3**, 419-429.
64. Draper, B. and Rao, V.B. (2007) An ATP Hydrolysis Sensor in the DNA Packaging Motor from Bacteriophage T4 Suggests an Inchworm-Type Translocation Mechanism. *Journal of Molecular Biology*, **369**, 79-94.
65. Duffy, C. and Feiss, M. (2002) The large subunit of bacteriophage  $\lambda$ 's terminase plays a role in DNA translocation and packaging termination. *Journal of Molecular Biology*, **316**, 547-561.
66. Alam, T.I. and Rao, V.B. (2008) The ATPase domain of the large terminase protein, gp17, from bacteriophage T4 binds DNA: implications to the DNA packaging mechanism. *Journal of molecular biology*, **376**, 1272-1281.
67. Smits, C., Chechik, M., Kovalevskiy, O.V., Shevtsov, M.B., Foster, A.W., Alonso, J.C. and Antson, A.A. (2009) Structural basis for the nuclease activity of a bacteriophage large terminase. *EMBO reports*, **10**, 592-598.
68. Roy, A. and Cingolani, G. (2012) Structure of p22 headful packaging nuclease. *The Journal of biological chemistry*, **287**, 28196-28205.
69. Zhao, H., Lin, Z., Lynn, A.Y., Varnado, B., Beutler, J.A., Murelli, R.P., Le Grice, S.F. and Tang, L. (2015) Two distinct modes of metal ion binding in the nuclease active site of a viral DNA-packaging terminase: insight into the two-metal-ion catalytic mechanism. *Nucleic acids research*, **44**, 776-789.
70. Nowotny, M. and Yang, W. (2006) Stepwise analyses of metal ions in RNase H catalysis from substrate destabilization to product release. *The EMBO Journal*, **25**, 1924-1933.
71. Ghosh-Kumar, M., Alam, T.I., Draper, B., Stack, J.D. and Rao, V.B. (2011) Regulation by interdomain communication of a headful packaging nuclease from bacteriophage T4. *Nucleic acids research*, **39**, 2742-2755.
72. Zhao, H., Lin, Z., Lynn, A.Y., Varnado, B., Beutler, J.A., Murelli, R.P., Le Grice, S.F. and Tang, L. (2015) Two distinct modes of metal ion binding in the nuclease active site of a viral DNA-packaging terminase: insight into the two-metal-ion catalytic mechanism. *Nucleic acids research*, **43**, 11003-11016.
73. Majorek, K.A., Dunin-Horkawicz, S., Steczkiewicz, K., Muszewska, A., Nowotny, M., Ginalski, K. and Bujnicki, J.M. (2014) The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. *Nucleic acids research*, **42**, 4160-4179.
74. Yang, W. (2011) Nucleases: diversity of structure, function and mechanism. *Quarterly reviews of biophysics*, **44**, 1-93.
75. Steiniger-White, M., Rayment, I. and Reznikoff, W.S. (2004) Structure/function insights into Tn5 transposition. *Current opinion in structural biology*, **14**, 50-57.
76. Alam, T.I., Draper, B., Kondabagil, K., Rentas, F.J., Ghosh - Kumar, M., Sun, S., Rossmann, M.G. and Rao, V.B. (2008) The headful packaging nuclease of bacteriophage T4. *Molecular microbiology*, **69**, 1180-1190.

77. Deichelbohrer, I., Messer, W. and Trautner, T.A. (1982) Genome of *Bacillus subtilis* bacteriophage SPP1: structure and nucleotide sequence of pac, the origin of DNA packaging. *Journal of virology*, **42**, 83-90.
78. Ghosh-Kumar, M., Alam, T.I., Draper, B., Stack, J.D. and Rao, V.B. (2011) Regulation by interdomain communication of a headful packaging nuclease from bacteriophage T4. *Nucleic Acids Res*, **39**, 2742-2755.
79. Nadal, M., Mas, P.J., Blanco, A.G., Arnan, C., Sola, M., Hart, D.J. and Coll, M. (2010) Structure and inhibition of herpesvirus DNA packaging terminase nuclease domain. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 16078-16083.
80. Steitz, T.A. and Steitz, J.A. (1993) A general two-metal-ion mechanism for catalytic RNA. *Proceedings of the National Academy of Sciences*, **90**, 6498-6502.
81. Xu, R.-G., Jenkins, H.T., Chechik, M., Blagova, E.V., Lopatina, A., Klimuk, E., Minakhin, L., Severinov, K., Greive, S.J. and Antson, A.A. (2017) Viral genome packaging terminase cleaves DNA using the canonical RuvC-like two-metal catalysis mechanism. *Nucleic acids research*, **45**, 3580-3590.
82. Casjens, S.R. and Gilcrease, E.B. (2009) Determining DNA packaging strategy by analysis of the termini of the chromosomes in tailed-bacteriophage virions. *Bacteriophages: Methods and Protocols, Volume 2 Molecular and Applied Aspects*, 91-111.
83. Sigamani, S.S., Zhao, H., Kamau, Y.N., Baines, J.D. and Tang, L. (2013) The structure of the herpes simplex virus DNA-packaging terminase pUL15 nuclease domain suggests an evolutionary lineage among eukaryotic and prokaryotic viruses. *Journal of virology*, **87**, 7140-7148.
84. Ray, K., Sabanayagam, C.R., Lakowicz, J.R. and Black, L.W. (2010) DNA crunching by a viral packaging motor: Compression of a procapsid-portal stalled Y-DNA substrate. *Virology*, **398**, 224-232.
85. Waters, J.T., Kim, H.D., Gumbart, J.C., Lu, X.-J. and Harvey, S.C. (2016) DNA Scrunching in the Packaging of Viral Genomes. *The Journal of Physical Chemistry B*, **120**, 6200-6207.
86. Dieffenbach, C., Lowe, T. and Dveksler, G. (1993) General concepts for PCR primer design. *PCR Methods Appl*, **3**, S30-S37.
87. Fogg, Mark J. and Wilkinson, Anthony J. (2008) Higher-throughput approaches to crystallization and crystal structure determination. *Biochemical Society Transactions*, **36**, 771-775.
88. Saitoh, H., Pu, R.T. and Dasso, M. (1997) SUMO-1: wrestling with a new ubiquitin-related modifier. *Trends in Biochemical Sciences*, **22**, 374.
89. Studier, F.W. and Moffatt, B.A. (1986) Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *Journal of molecular biology*, **189**, 113-130.
90. Rosenberg, A.H., Lade, B.N., Dao-shan, C., Lin, S.-W., Dunn, J.J. and Studier, F.W. (1987) Vectors for selective expression of cloned DNAs by T7 RNA polymerase. *Gene*, **56**, 125-135.
91. Studier, F.W., Rosenberg, A.H., Dunn, J.J. and Dubendorff, J.W. (1990) [6] Use of T7 RNA polymerase to direct expression of cloned genes. *Methods in enzymology*, **185**, 60-89.
92. Aslanidis, C. and de Jong, P.J. (1990) Ligation-independent cloning of PCR products (LIC-PCR). *Nucleic acids research*, **18**, 6069-6074.

93. Johnston, C., Martin, B., Fichant, G., Polard, P. and Claverys, J.-P. (2014) Bacterial transformation: distribution, shared mechanisms and divergent control. *Nature Reviews Microbiology*, **12**, 181-196.
94. Tegel, H., Tourle, S., Ottosson, J. and Persson, A. (2010) Increased levels of recombinant human proteins with the Escherichia coli strain Rosetta(DE3). *Protein Expression & Purification*, **69**, 159-167.
95. Sun, Y., Zhang, P. and Ru, B. (1999) Study on the expression of a kind of metallothionein and SZ51 recombinant protein in the E. coli BL21(DE3)pLysS strain. *Chinese Journal of Biochemistry & Molecular Biology*, **15**.
96. Wood, W.B. (1966) Host specificity of DNA produced by Escherichia coli: bacterial mutations affecting the restriction and modification of DNA. *Journal of Molecular Biology*, **16**, 118.
97. Doherty, A.J., Ashford, S.R., Brannigan, J.A. and Wigley, D.B. (1995) A superior host strain for the over-expression of cloned genes using the T7 promoter based vectors. *Nucleic Acids Research*, **23**, 2074-2075.
98. Davanloo, P., Rosenberg, A.H., Dunn, J.J. and Studier, F.W. (1984) Cloning and expression of the gene for bacteriophage T7 RNA polymerase. *Proceedings of the National Academy of Sciences of the United States of America*, **81**, 2035-2039.
99. Santos, S.P., Bandeiras, T.M., Pinto, A.F., Teixeira, M., Carrondo, M.A. and Romão, C.V. (2012) Thermofluor-based optimization strategy for the stabilization and crystallization of *Campylobacter jejuni* desulforubrerythrin. *Protein expression and purification*, **81**, 193-200.
100. Ericsson, U.B., Hallberg, B.M., DeTitta, G.T., Dekker, N. and Nordlund, P. (2006) Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Analytical biochemistry*, **357**, 289-298.
101. Niesen, F.H., Berglund, H. and Vedadi, M. (2007) The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nature protocols*, **2**, 2212-2221.
102. Inglese, J., Johnson, R.L., Simeonov, A., Xia, M., Zheng, W., Austin, C.P. and Auld, D.S. (2007) High-throughput screening assays for the identification of chemical probes. *Nature chemical biology*, **3**, 466-479.
103. Baykov, A., Evtushenko, O. and Avaeva, S. (1988) A malachite green procedure for orthophosphate determination and its use in alkaline phosphatase-based enzyme immunoassay. *Analytical biochemistry*, **171**, 266-270.
104. Schlegel, R.H.J. (2003) pUC18 vector. *Encyclopedic Dictionary of Plant Breeding & Related Subjects*.
105. Cole, J.L., Lary, J.W., Moody, T.P. and Laue, T.M. (2008) Analytical ultracentrifugation: sedimentation velocity and sedimentation equilibrium. *Methods in cell biology*, **84**, 143-179.
106. Ralston, G. (1993) *Introduction to analytical ultracentrifugation*. Beckman California:.
107. Brown, P.H. and Schuck, P. (2008) A new adaptive grid-size algorithm for the simulation of sedimentation velocity profiles in analytical ultracentrifugation. *Computer Physics Communications*, **178**, 105-120.
108. Philo, J., Hayes, D. and Laue, T. (2006) Sednertp. *Alliance Protein Laboratories, Thousand Oaks, CA*.

109. Jose, D., Weitzel, S.E., Baase, W.A., Michael, M.M. and von Hippel, P.H. (2015) Mapping the interactions of the single-stranded DNA binding protein of bacteriophage T4 (gp32) with DNA lattices at single nucleotide resolution: polynucleotide binding and cooperativity. *Nucleic acids research*, **43**, 9291-9305.
110. Kneale, G.G. (1994) *DNA-Protein interactions: principles and protocols*. Springer Science & Business Media.
111. Jerabek-Willemsen, M., André, T., Wanner, R., Roth, H.M., Duhr, S., Baaske, P. and Breitsprecher, D. (2014) MicroScale Thermophoresis: Interaction analysis and beyond. *Journal of Molecular Structure*, **1077**, 101-113.
112. Seidel, S.A., Dijkman, P.M., Lea, W.A., van den Bogaart, G., Jerabek-Willemsen, M., Lazic, A., Joseph, J.S., Srinivasan, P., Baaske, P. and Simeonov, A. (2013) Microscale thermophoresis quantifies biomolecular interactions under previously challenging conditions. *Methods*, **59**, 301-315.
113. Dessau, M.A. and Modis, Y. (2011) Protein crystallization for X-ray crystallography. *Journal of visualized experiments: JoVE*.
114. Asherie, N. (2004) Protein crystallization and phase diagrams. *Methods*, **34**, 266-272.
115. Bergfors, T. (2003) Seeds to crystals. *Journal of structural biology*, **142**, 66-76.
116. Harada, A., Sato, Y., Kamimura, N., Venugopalan, N., Masai, E. and Senda, T. (2016) Overcoming a hemihedral twinning problem in tetrahydrofolate-dependent O-demethylase crystals by the microseeding method. *Acta Crystallographica Section F: Structural Biology Communications*, **72**, 897-902.
117. Oswald, C., Smits, S.H., Bremer, E. and Schmitt, L. (2008) Microseeding—A Powerful Tool for Crystallizing Proteins Complexed with Hydrolyzable Substrates. *International journal of molecular sciences*, **9**, 1131-1141.
118. Heras, B. and Martin, J.L. (2005) Post-crystallization treatments for improving diffraction quality of protein crystals. *Acta crystallographica. Section D, Biological crystallography*, **61**, 1173-1180.
119. Russo Krauss, I., Sica, F., Mattia, C.A. and Merlino, A. (2012) Increasing the X-ray diffraction power of protein crystals by dehydration: the case of bovine serum albumin and a survey of literature data. *International journal of molecular sciences*, **13**, 3782-3800.
120. Arndt, U.W. and Wonacott, A. (1977) *Rotation method in crystallography*. North-Holland Pub. Co.
121. Powell, H.R. (2017) X-ray data processing. *Bioscience Reports*, BSR20170227.
122. Kabsch, W. (2010) Xds. *Acta Crystallographica Section D: Biological Crystallography*, **66**, 125-132.
123. Taylor, G.L. (2010) Introduction to phasing. *Acta Crystallographica Section D: Biological Crystallography*, **66**, 325-338.
124. Collaborative, C.P. (1994) The CCP4 suite: programs for protein crystallography. *Acta crystallographica. Section D, Biological crystallography*, **50**, 760.
125. Schneider, T.R. and Sheldrick, G.M. (2002) Substructure solution with SHELXD. *Acta Crystallographica Section D: Biological Crystallography*, **58**, 1772-1779.
126. Sheldrick, G.M. (2010) Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. *Acta Crystallographica Section D: Biological Crystallography*, **66**, 479-485.

127. Terwilliger, T.C. and Berendzen, J. (1999) Automated MAD and MIR structure solution. *Acta Crystallographica Section D: Biological Crystallography*, **55**, 849-861.
128. McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C. and Read, R.J. (2007) Phaser crystallographic software. *Journal of applied crystallography*, **40**, 658-674.
129. Cowtan, K.D. and Zhang, K.Y.J. (1999) Density modification for macromolecular phase improvement. *Progress in Biophysics and Molecular Biology*, **72**, 245-270.
130. Terwilliger, T.C. (2002) Automated structure solution, density modification and model building. *Acta Crystallographica Section D: Biological Crystallography*, **58**, 1937-1940.
131. Vonrhein, C., Blanc, E., Roversi, P. and Bricogne, G. (2007) Automated structure solution with autoSHARP. *Macromolecular Crystallography Protocols: Volume 2: Structure Determination*, 215-230.
132. Terwilliger, T.C. (2000) Maximum-likelihood density modification. *Acta Crystallographica Section D: Biological Crystallography*, **56**, 965-972.
133. Terwilliger, T.C. (2003) Automated main-chain model building by template matching and iterative fragment extension. *Acta Crystallographica Section D: Biological Crystallography*, **59**, 38-44.
134. Langer, G.G., Cohen, S.X., Lamzin, V.S. and Perrakis, A. (2008) Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nature protocols*, **3**, 1171.
135. Cowtan, K. (2006) The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallographica Section D: Biological Crystallography*, **62**, 1002-1011.
136. Emsley, P. and Cowtan, K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D: Biological Crystallography*, **60**, 2126-2132.
137. Murshudov, G.N., Skubák, P., Lebedev, A.A., Pannu, N.S., Steiner, R.A., Nicholls, R.A., Winn, M.D., Long, F. and Vagin, A.A. (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D: Biological Crystallography*, **67**, 355-367.
138. Young, J.Y., Westbrook, J.D., Feng, Z., Sala, R., Peisach, E., Oldfield, T.J., Sen, S., Gutmanas, A., Armstrong, D.R. and Berrisford, J.M. (2017) OneDep: unified wwPDB system for deposition, biocuration, and validation of macromolecular structures in the PDB Archive. *Structure*, **25**, 536-545.
139. Berman, H., Henrick, K. and Nakamura, H. (2003) Announcing the worldwide protein data bank. *Nature Structural & Molecular Biology*, **10**, 980-980.
140. Pei, J. and Grishin, N.V. (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics*, **23**, 802-808.
141. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M. and Söding, J. (2011) Fast, scalable generation of high - quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, **7**, 539.
142. Dosztányi, Z., Csizmok, V., Tompa, P. and Simon, I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433-3434.

143. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, **74**, 5463-5467.
144. Ericsson, U.B., Hallberg, B.M., Detitta, G.T., Dekker, N. and Nordlund, P. (2006) Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Analytical biochemistry*, **357**, 289-298.
145. Cherfils, J., Ménétrey, J., Le Bras, G., Le Bras, G., Janoueix - Lerosey, I., de Gunzburg, J., Garel, J.R. and Auzat, I. (1997) Crystal structures of the small G protein Rap2A in complex with its substrate GTP, with GDP and with GTP  $\gamma$  S. *The EMBO Journal*, **16**, 5582-5591.
146. Dittrich, M., Hayashi, S. and Schulten, K. (2003) On the mechanism of ATP hydrolysis in F 1-ATPase. *Biophysical journal*, **85**, 2253-2266.
147. Rosta, E., Nowotny, M., Yang, W. and Hummer, G. (2011) Catalytic mechanism of RNA backbone cleavage by ribonuclease H from quantum mechanics/molecular mechanics simulations. *Journal of the American Chemical Society*, **133**, 8934-8941.
148. Autry, J.M., Rubin, J.E., Svensson, B., Li, J. and Thomas, D.D. (2012) Nucleotide activation of the Ca-ATPase. *Journal of Biological Chemistry*, **287**, 39070-39082.
149. Winter, J.A., Patoli, B. and Bunting, K.A. (2012) DNA binding in high salt: analysing the salt dependence of replication protein A3 from the halophile haloferax volcanii. *Archaea*, **2012**.
150. Park, C. and Raines, R.T. (2001) Quantitative analysis of the effect of salt concentration on enzymatic catalysis. *Journal of the American Chemical Society*, **123**, 11472-11479.
151. Zheng, H., Chordia, M.D., Cooper, D.R., Chruscz, M., Müller, P., Sheldrick, G.M. and Minor, W. (2014) Validating metal binding sites in macromolecule structures using the CheckMyMetal web server. *Nature protocols*, **9**, 156.
152. Zheng, H., Chordia, M.D., Cooper, D.R., Chruscz, M., Müller, P., Sheldrick, G.M. and Minor, W. (2014) Validation of metal-binding sites in macromolecular structures with the CheckMyMetal web server. *Nature protocols*, **9**, 156-170.
153. Nowotny, M., Gaidamakov, S.A., Crouch, R.J. and Yang, W. (2005) Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis. *Cell*, **121**, 1005-1016.
154. Grinsztejn, B., Nguyen, B.-Y., Katlama, C., Gatell, J.M., Lazzarin, A., Vittecoq, D., Gonzalez, C.J., Chen, J., Harvey, C.M. and Isaacs, R.D. (2007) Safety and efficacy of the HIV-1 integrase inhibitor raltegravir (MK-0518) in treatment-experienced patients with multidrug-resistant virus: a phase II randomised controlled trial. *The Lancet*, **369**, 1261-1269.
155. Dragan, A., Casas-Finet, J., Bishop, E., Strouse, R., Schennerman, M. and Geddes, C. (2010) Characterization of PicoGreen interaction with dsDNA and the origin of its fluorescence enhancement upon binding. *Biophysical journal*, **99**, 3010-3019.
156. Rentas, F.J. and Rao, V.B. (2003) Defining the bacteriophage T4 DNA packaging machine: evidence for a C-terminal DNA cleavage domain in the large terminase/packaging protein gp17. *Journal of molecular biology*, **334**, 37-52.
157. Cornilleau, C., Atmane, N., Jacquet, E., Smits, C., Alonso, J.C., Tavares, P. and Oliveira, L. (2013) The nuclease domain of the SPP1 packaging motor coordinates DNA cleavage and encapsidation. *Nucleic Acids Res*, **41**, 340-354.

158. Nadal, M., Mas, P.J., Blanco, A.G., Arnan, C., Solà, M., Hart, D.J. and Coll, M. (2010) Structure and inhibition of herpesvirus DNA packaging terminase nuclease domain. *Proceedings of the National Academy of Sciences*, **107**, 16078-16083.
159. Harding, M.M. (2001) Geometry of metal-ligand interactions in proteins. *Acta Crystallographica Section D: Biological Crystallography*, **57**, 401-411.
160. Sarakatsannis, J.N. and Duan, Y. (2005) Statistical characterization of salt bridges in proteins. *Proteins: Structure, Function, and Bioinformatics*, **60**, 732-739.
161. Kumar, S. and Nussinov, R. (2001) How do thermophilic proteins deal with heat? *Cellular and Molecular Life Sciences CMLS*, **58**, 1216-1233.
162. Alla, N.R. and Nicholson, A.W. (2012) Evidence for a dual functional role of a conserved histidine in RNA.DNA heteroduplex cleavage by human RNase H1. *FEBS J*, **279**, 4492-4500.
163. Rosta, E., Woodcock, H.L., Brooks, B.R. and Hummer, G. (2009) Artificial reaction coordinate “tunneling” in free - energy calculations: The catalytic reaction of RNase H. *Journal of computational chemistry*, **30**, 1634-1641.
164. Ho, M.-H., De Vivo, M., Dal Peraro, M. and Klein, M.L. (2010) Understanding the effect of magnesium ion concentration on the catalytic activity of ribonuclease H through computation: does a third metal binding site modulate endonuclease catalysis? *Journal of the American Chemical Society*, **132**, 13702-13712.
165. Starnes, M.C. and Cheng, Y. (1989) Human immunodeficiency virus reverse transcriptase-associated RNase H activity. *Journal of Biological Chemistry*, **264**, 7073-7077.
166. Berkower, I., Leis, J. and Hurwitz, J. (1973) Isolation and characterization of an endonuclease from Escherichia coli specific for ribonucleic acid in ribonucleic acid-deoxyribonucleic acid hybrid structures. *Journal of Biological Chemistry*, **248**, 5914-5921.
167. Tan, C.K., Rigal, C., Mian, A.M., So, A.G. and Downey, K.M. (1991) Inhibition of the RNase H activity of HIV reverse transcriptase by azidothymidylate. *Biochemistry*, **30**, 4831-4835.
168. Yang, W., Lee, J.Y. and Nowotny, M. (2006) Making and Breaking Nucleic Acids: Two-Mg<sup>2+</sup>-Ion Catalysis and Substrate Specificity. *Molecular Cell*, **22**, 5-13.
169. Cowan, J. (1998) Metal activation of enzymes in nucleic acid biochemistry. *Chemical reviews*, **98**, 1067-1088.
170. Bolt, E.L., Lloyd, R.G. and Sharples, G.J. (2001) Genetic analysis of an archaeal Holliday junction resolvase in Escherichia coli. *Journal of molecular biology*, **310**, 577-589.
171. Yang, Y., Ishino, S., Yamagami, T., Kumamaru, T., Satoh, H. and Ishino, Y. (2012) The OsGEN-L protein from Oryza sativa possesses Holliday junction resolvase activity as well as 5' -flap endonuclease activity. *Journal of biochemistry*, **151**, 317-327.
172. Domanico, P.L. and Tse-Dinh, Y.C. (1991) Mechanistic studies on E. coli DNA topoisomerase I: divalent ion effects. *Journal of inorganic biochemistry*, **42**, 87-96.
173. Jongruja, N., You, D.J., Kanaya, E., Koga, Y., Takano, K. and Kanaya, S. (2010) The N - terminal hybrid binding domain of RNase HI from Thermotoga maritima is important for substrate binding and Mg<sup>2+</sup> - dependent activity. *FEBS journal*, **277**, 4474-4489.

174. Holm, L. and Sander, C. (1995) Dali: a network tool for protein structure comparison. *Trends in biochemical sciences*, **20**, 478-480.
175. Górecka, K.M., Komorowska, W. and Nowotny, M. (2013) Crystal structure of RuvC resolvase in complex with Holliday junction substrate. *Nucleic acids research*, **41**, 9945-9955.
176. Yang, W. and Steitz, T.A. (1995) Recombining the structures of HIV integrase, RuvC and RNase H. *Structure*, **3**, 131-134.
177. Krissinel, E. and Henrick, K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D: Biological Crystallography*, **60**, 2256-2268.
178. Zhao, H., Christensen, T.E., Kamau, Y.N. and Tang, L. (2013) Structures of the phage Sf6 large terminase provide new insights into DNA translocation and cleavage. *Proceedings of the National Academy of Sciences*, **110**, 8075-8080.
179. Selvarajan Sigamani, S., Zhao, H., Kamau, Y.N., Baines, J.D. and Tang, L. (2013) The structure of the herpes simplex virus DNA-packaging terminase pUL15 nuclease domain suggests an evolutionary lineage among eukaryotic and prokaryotic viruses. *J Virol*, **87**, 7140-7148.
180. Hilbert, B.J., Hayes, J.A., Stone, N.P., Xu, R.-G. and Kelch, B.A. (2017) The large terminase DNA packaging motor grips DNA with its ATPase domain for cleavage by the flexible nuclease domain. *Nucleic acids research*, **45**, 3591-3605.
181. Torres, R.A. and Bruice, T.C. (2000) The mechanism of phosphodiester hydrolysis: near in-line attack conformations in the hammerhead ribozyme. *Journal of the American Chemical Society*, **122**, 781-791.
182. Yang, W. (2008) An equivalent metal ion in one-and two-metal-ion catalysis. *Nature structural & molecular biology*, **15**, 1228-1231.
183. Li, H., Hwang, Y., Perry, K., Bushman, F. and Van Duyne, G.D. (2016) Structure and Metal Binding Properties of a Poxvirus Resolvase. *Journal of Biological Chemistry*, **291**, 11094-11104.
184. Chen, L., Shi, K., Yin, Z. and Aihara, H. (2012) Structural asymmetry in the *Thermus thermophilus* RuvC dimer suggests a basis for sequential strand cleavages during Holliday junction resolution. *Nucleic acids research*, **41**, 648-656.
185. Gual, A., Camacho, A.G. and Alonso, J.C. (2000) Functional analysis of the terminase large subunit, G2P, of *Bacillus subtilis* bacteriophage SPP1. *The Journal of biological chemistry*, **275**, 35311-35319.
186. Xu, R.-G., Jenkins, H.T., Antson, A.A. and Greive, S.J. (2017) Structure of the large terminase from a hyperthermophilic virus reveals a unique mechanism for oligomerization and ATP hydrolysis. *Nucleic Acids Research*, doi: 10.1093/nar/gkx1947.
187. Hu, T.-C., Korczyńska, J., Smith, D.K. and Brzozowski, A.M. (2008) High-molecular-weight polymers for protein crystallization: Poly- $\gamma$ -glutamic acid-based precipitants. *Acta Crystallographica Section D: Biological Crystallography*, **64**, 957-963.
188. Weber, P.C. (1990) A protein crystallization strategy using automated grid searches on successively finer grids. *Methods*, **1**, 31-37.
189. Gorrec, F. (2009) The MORPHEUS protein crystallization screen. *Journal of applied crystallography*, **42**, 1035-1042.

190. Heras, B. and Martin, J.L. (2005) Post-crystallization treatments for improving diffraction quality of protein crystals. *Acta Crystallographica Section D: Biological Crystallography*, **61**, 1173-1180.
191. Newman, J. (2006) A review of techniques for maximizing diffraction from a protein crystal in stilla. *Acta Crystallographica Section D: Biological Crystallography*, **62**, 27-31.
192. Russo Krauss, I., Sica, F., Mattia, C.A. and Merlino, A. (2012) Increasing the X-ray diffraction power of protein crystals by dehydration: the case of bovine serum albumin and a survey of literature data. *Int J Mol Sci*, **13**, 3782-3800.
193. Boutron, P. (1987), *The biophysics of organ cryopreservation*. Springer, pp. 201-236.
194. Ariyoshi, M., Vassylyev, D.G., Iwasaki, H., Nakamura, H., Shinagawa, H. and Morikawa, K. (1994) Atomic structure of the RuvC resolvase: a Holliday junction-specific endonuclease from *E. coli*. *Cell*, **78**, 1063-1072.
195. Hilbert, B.J., Hayes, J.A., Stone, N.P., Xu, R.-G. and Kelch, B.A. (2017) The large terminase DNA packaging motor grips DNA with its ATPase domain for cleavage by the flexible nuclease domain. *Nucleic Acids Res*, **45**, 3591-3605.
196. Kantardjieff, K.A. and Rupp, B. (2003) Matthews coefficient probabilities: improved estimates for unit cell contents of proteins, DNA, and protein–nucleic acid complex crystals. *Protein Science*, **12**, 1865-1871.
197. Langer, G., Cohen, S.X., Lamzin, V.S. and Perrakis, A. (2008) Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat. Protocols*, **3**, 1171-1179.
198. Sun, S., Kondabagil, K., Draper, B., Alam, T.I., Bowman, V.D., Zhang, Z., Hegde, S., Fokine, A., Rossmann, M.G. and Rao, V.B. (2008) The Structure of the Phage T4 DNA Packaging Motor Suggests a Mechanism Dependent on Electrostatic Forces *Cell*, **135**, 1251-1262.
199. Pierce, B., Tong, W. and Weng, Z. (2004) M-ZDOCK: a grid-based approach for C n symmetric multimer docking. *Bioinformatics*, **21**, 1472-1478.
200. Mao, H., Saha, M., Reyes-Aldrete, E., Sherman, M.B., Woodson, M., Atz, R., Grimes, S., Jardine, P.J. and Morais, M.C. (2016) Structural and molecular basis for coordination in a viral DNA packaging motor. *Cell reports*, **14**, 2017-2029.
201. Erzberger, J.P. and Berger, J.M. (2006) Evolutionary relationships and structural mechanisms of AAA+ proteins. *Annu. Rev. Biophys. Biomol. Struct.*, **35**, 93-114.
202. Thomsen, N.D. and Berger, J.M. (2009) Running in reverse: the structural basis for translocation polarity in hexameric helicases. *Cell*, **139**, 523-534.
203. Moreau, M.J., McGeoch, A.T., Lowe, A.R., Itzhaki, L.S. and Bell, S.D. (2007) ATPase Site Architecture and Helicase Mechanism of an Archaeal MCM. *Molecular Cell*, **28**, 304-314.
204. Zhao, Z., De-Donatis, G.M., Schwartz, C., Fang, H., Li, J. and Guo, P. (2016) An Arginine Finger Regulates the Sequential Action of Asymmetrical Hexameric ATPase in the Double-Stranded DNA Translocation Motor. *Molecular and Cellular Biology*, **36**, 2514-2523.
205. Ogura, T., Whiteheart, S.W. and Wilkinson, A.J. (2004) Conserved arginine residues implicated in ATP hydrolysis, nucleotide-sensing, and inter-subunit interactions in AAA and AAA+ ATPases. *Journal of structural biology*, **146**, 106-112.

206. Paschall, C.O., Thompson, J.A., Marzahn, M.R., Chiraniya, A., Hayner, J.N., O'Donnell, M., Robbins, A.H., McKenna, R. and Bloom, L.B. (2011) The Escherichia coli clamp loader can actively pry open the  $\beta$ -sliding clamp. *Journal of Biological Chemistry*, **286**, 42704-42714.
207. Bason, J.V., Montgomery, M.G., Leslie, A.G. and Walker, J.E. (2015) How release of phosphate from mammalian F1-ATPase generates a rotary substep. *Proceedings of the National Academy of Sciences*, **112**, 6009-6014.
208. Chemla, Y.R., Aathavan, K., Michaelis, J., Grimes, S., Jardine, P.J., Anderson, D.L. and Bustamante, C. (2005) Mechanism of force generation of a viral DNA packaging motor. *Cell*, **122**, 683-692.
209. Chistol, G., Liu, S., Hetherington, C.L., Moffitt, J.R., Grimes, S., Jardine, P.J. and Bustamante, C. (2012) High degree of coordination and division of labor among subunits in a homomeric ring ATPase. *Cell*, **151**, 1017-1028.
210. Lee, J.Y. and Yang, W. (2006) UvrD helicase unwinds DNA one base pair at a time by a two-part power stroke. *Cell*, **127**, 1349-1360.
211. Enemark, E.J. and Joshua-Tor, L. (2006) Mechanism of DNA translocation in a replicative hexameric helicase. *Nature*, **442**, 270-275.