Structure and Function of Human Respiratory Syncytial Virus M2-1 Protein

Sian Jean Tanner

Submitted in accordance with the requirements for
the degree of Doctor of Philosophy

The University of Leeds

The Astbury Centre for Structural Molecular Biology

September 2013

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapters 3, 4 and 5 are based on work contributed to a jointly authored manuscript:
Tanner SJ, Ariza A, Richard CA, Kyle HF, Dods RL, Blondot ML, Wu W, Trincão J, Trinh CH, Hiscox JA, Carroll MW, Sillman NJ, Eléouët JF, Edwards TA and Barr JN (accepted December 2013) Crystal structure of the essential M2-1 antiterminator of HRSV and implications of phosphorylation.

Chapter 3: Sian Jean Tanner designed and executed all cloning, expression and purification strategies for all systems used.

Mass spectrometry analysis was performed as a service by the University of Leeds Mass Spectrometry facility (James Ault, Kevin Tipping).

Chapter 4: Sian Jean Tanner performed all crystallisation experiments, collected X-ray diffraction data, helped with data processing and structure solution, performed the initial refinement on molecular replacement data, mutagenesis of phosphorylation mutants, and co-wrote the manuscript and figures therein with John Barr and Thomas Edwards.

Antonio Ariza refined the WT-P422 dataset, helped with molecular replacement and refinement of all other datasets, and contributed to the methods and results sections of the manuscript.

Chi Trinh and José Trincão helped with data collection and the initial SAD phasing at Diamond Light Source.

Chapter 5: Sian Jean Tanner performed all RNA binding assay development and experiments, and mutagenesis of protein for binding assays, featured in this thesis and the subsequent manuscript. Additional RNA binding experiments for the manuscript were performed by Hannah Kyle and Rachel Dods.

Charles-Adrien Richard, Marie-Lise Blondot and Jean-François-Eléouët performed the minigenome analysis and the associated cloning (section 5.2.6 and figure therein).

Mass spectrometry analysis was performed as a service by Helen Beeston (Alison Ashcroft group).

Chapter 5 is based on work contributed to a jointly authored manuscript:
Ariza A, Tanner SJ, Walter CT, Dent KC, Shepherd DA, Wu W, Matthews SV, Hiscox JA, Green TJ, Luo M, Elliot R, Fooks AR, Ashcroft AE, Stonehouse NJ, Ranson NA, Barr JN and Edwards TA. (2013) *Nucleic Acids Research*. Nucleocapsid protein structures from orthobunyaviruses reveal insight into ribonucleoprotein architecture and RNA polymerization.

Sian Jean Tanner performed all RNA binding experiments for the above publication, which contributed to the assay development in Chapter 5 of this thesis, and contributed to the manuscript.

All other work in the above publication was performed by the authors listed above.

ACKNOWLEDGEMENTS

My first acknowledgement of many must go to John, whose unwavering belief that this stupid little protein held all the answers kept the project going through its darker times. I am equally grateful to Ed, who truly rescued my PhD. Thank you both for your guidance and support.

Everyone in Virology at Leeds, and 8.61 especially, has been an invaluable source of knowledge, assistance and amusement. In particular, Cheryl, without whom I would have been utterly lost, and Carsten, without whom I would have drunk a lot less coffee.

My conversion to crystallographer would not have been possible without Tony, Chi, or anyone else who ever picked a crystal, kept me awake at 4am at Diamond, or helped me get over my pathological fear of liquid nitrogen. Enormous thanks must also go to Jean-François Eléouët and his group at the INRA – the mystery collaborators who dropped everything for our paper.

Thank you to my family and friends, for never agreeing with me when I said this PhD was a ridiculous idea, and to Hannah, because at least I never had *that* bad a day in the lab.

And lastly, to Ben. Thank you for sticking with me through my decision to venture 'North' for 4 years – you are the only reason I made it.

ABSTRACT

The M2-1 protein of the important pathogen human respiratory syncytial virus is a transcription antiterminator that is essential for viral gene expression. We present the X-ray crystal structure of full-length M2-1 protein in its native tetrameric form at a resolution of 2.52 Å. The structure reveals M2-1 forms a disk-like assembly with tetramerisation driven by a long helix forming a four-helix bundle at its center, further stabilised by contact between the zinc finger and adjacent protomers. The tetramerisation helix is linked to a core domain responsible for RNA binding activity by a flexible loop on which lie two functionally critical serine residues, 58 and 61, that are phosphorylated during infection. The identity of these residues was confirmed by mass spectrometric analysis of M2-1 protein expressed in baculovirus-assisted insect cell culture. The crystal structure of a phosphomimetic M2-1 variant, S58DS61D revealed altered charge density surrounding this flexible loop, although loop position was unaffected.

Structure guided mutagenesis identified residues that contributed to RNA binding and antitermination activity, revealing a strong correlation between these two activities, and further defining the role of phosphorylation in M2-1 antitermination activity. The data presented here identify surfaces critical for M2-1 function that may be targeted by anti-viral compounds, and allow us to propose a possible model for M2-1 function during respiratory syncytial virus transcription.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AMPV avian metapneumovirus
ARI acute respiratory infection
AU asymmetric unit
bp base pairs
BRSV bovine respiratory syncytial virus
BUNV Bunyamwera virus
cDNA complementary DNA
CR conserved region
CTL cytotoxic T lymphocyte
DD phosphomimetic, double serine to aspartate mutant (also S58DS61D)
EBOV Ebola virus
ESCRT Endosomal Sorting Complexes Required for Transport
ESI electrospray ionisation
eV electron volts
F_{hkl} structure factor
FI-RSV formalin-inactivated RSV
Fl fluorescein
FPA fluorescence polarisation anisotropy
GE gene end
GS gene start
GST glutathione S-transferase
HCV hepatitis C virus
HIV human immunodeficiency virus
HMPV human metapneumovirus
HN haemagglutinin-neuraminidase
HPIV human parainflluenza virus
HRSV human respiratory syncytial virus
HRV human rhinovirus
IAV influenza A virus
ICTV International Committee on Taxonomy of Viruses
IEX ion exchange chromatography
IFN interferon
IL interleukin
IPTG isopropyl β-D-1-thiogalactopyranoside
IRF3 interferon regulatory factor 3
kDa kilo-Dalton
LC liquid chromatography
LCMV lymphocytic choriomeningitis virus
Le leader region
LRI lower respiratory tract illness
mAb monoclonal antibody
MALDI-TOF matrix-assisted laser desorption/ionisation time-of-flight
MeV measles virus
MPV murine pneumonia virus (previously PVM)
MR molecular replacement
MuV mumps virus
NCS non-crystallographic symmetry

NDV Newcastle disease virus
NiV Nipah virus
nt nucleotides
ORF open reading frame
PBS phosphate buffered saline
PDB Protein Data Bank
PEG polyethylene glycol
p.i. post infection
PIV parainfluenza virus
RdRp RNA-dependent RNA polymerase
rmsd route-mean-square deviation
RNAi RNA interference
RNP ribonucleoprotein (or vRNP)
RABV rabies virus
SAD single-wavelength anomalous dispersion
SEC size exclusion chromatography
SeV Sendai virus
TCEP Tris (2-carboxyethyl) phosphine hydrochloride
TLR toll-like receptor
TrC trailer complementary region (3’ antigenome)
URI upper respiratory tract illness
vRNA viral RNA
VSV vesicular stomatitis virus
WT wild-type

1 CHAPTER 1: INTRODUCTION

1.1 GENERAL INTRODUCTION

1.1.1 Discovery of respiratory syncytial virus

Acute respiratory tract disease is the leading cause of infectious disease mortality worldwide, and human respiratory syncytial virus (HRSV) is recognised as the primary viral agent in young children. In recent years, there has been increased focus on HRSV morbidity and mortality in other high-risk groups, including the elderly and the immunocompromised.

In 1956, a previously unidentified virus was isolated from infants suffering from croup (or larynogotracheobronchitis). It caused an unusual ‘sponge-like’ cytopathic effect in monkey kidney tissue culture and was later found to be indistinguishable from a virus named chimpanzee coryza agent, (CCA) recovered from chimpanzees appearing to show symptoms of the ‘common cold’ (rhinitis, coughing) (Chanock et al., 1957, Chanock and Finberg, 1957). The virus was renamed respiratory syncytial virus (RSV) and its high prevalence in infants was realised soon after. Several closely related animal species, including bovine, ovine, and caprine RSV, were also discovered and they are of considerable concern to the livestock industry. As only human RSV is studied in this thesis, the acronym ‘RSV’ shall refer to the human virus unless stated otherwise.

1.1.2 Classification

RSV is a member of the subfamily *Pneumovirinae* in the family *Paramyxoviridae*, order *Mononegavirales* – the non-segmented negative strand RNA viruses (ICTV, 2012). This order contains a number of high profile human pathogens including Ebola and Marburg viruses (family *Filoviridae*), and rabies virus (family *Rhabdoviridae*). The family *Paramyxoviridae* is divided into two subfamilies: the subfamily *Paramyxovirinae*, featuring measles and mumps viruses, and the highly pathogenic Nipah and Hendra viruses; the subfamily *Pneumovirinae*, which is further divided into the genus *Pneumovirus*, containing the respiratory syncytial viruses and murine pneumonia virus (MPV), and the genus *Metapneumovirus*, containing human and avian metapneumoviruses (HMPV, AMPV).

The paramyxoviruses are a family of enveloped viruses with a well-conserved gene order and replication cycle. They are defined by their expression of a fusion protein on the virion surface, their pleomorphic virion morphology, ranging from 150 to 300 nm in diameter, and they contain helically symmetric nucleocapsids comprised of a monopartite, single-stranded, negative-sense RNA genome and a polymerase complex featuring the RNA-dependent RNA-polymerase (RdRp) and its co-factors. The size and shape of these nucleocapsids is considered a distinguishing feature of the family *Paramyxoviridae* subfamilies; the pneumoviruses have narrower nucleocapsids than the paramyxoviruses (Collins, 2007). The pneumoviruses also lack the enhanced coding capacity of the P gene, through RNA editing and alternative translation initiation codons, that is common to the paramyxoviruses – with the exception of murine pneumonia virus (MPV), which has a second P open reading frame (ORF). The nature of the surface proteins is characteristic of each genus within the family *Paramyxoviridae*, with the subfamily *Pneumovirinae* (e.g. RSV and HMPV) lacking both the haemagglutinin and neuraminidase activities that are attributed in various combinations to other genera, expressing instead a glycoprotein, G. The classification of RSV in a separate genus (*Pneumovirus*) from the metapneumoviruses is due to the presence of additional genes and slight differences in genome organisation.

RSV has one serotype containing two antigenically distinct subgroups, A and B, which share 88 % amino acid identity (Biacchesi et al., 2003). The majority of variation occurs within the viral envelope glycoproteins, particularly the attachment glycoprotein (G), which can vary by up to 45 % at the amino acid level between the two subgroups (Meyer et al., 2008). Much of our understanding of RSV molecular biology is based on the archetypal paramyxovirus Sendai virus (SeV) (subfamily *Paramyxovirinae*) and the rhabdovirus vesicular stomatitis virus (VSV), as these were easily adapted to tissue culture and grown to high titres. Despite the high conservation of genome organisation, the RSV genome contains a number of genes not found in these model viruses: non-structural proteins 1 and 2 (NS1, NS2), small hydrophobic protein (SH), and the M2 gene which expresses both M2-1 and M2-2 proteins. It also appears that the viral replication cycle is more complex for RSV, featuring overlapping genes (M2/L) and multiple open reading frames (M2-1/M2-2).

Paramyxoviruses share some similarities with the family *Orthomyxoviridae*, especially influenza A virus (IAV), who also express surface proteins characterised by haemagglutinin and neuraminidase activity. The requirement for a ribonucleocapsid as the template for viral transcription and replication is also a commonality between these two families, however there are many fundamental differences that have resulted in their separate classification: monopartite vs. segmented, cellular (paramyxo) vs. nuclear (orthomyxo) transcription and replication.

Order	**Family**	**Genus**
Mononegavirales	*Paramyxoviridae*	Subfamily *Pneumovirinae*
		Pneumovirus e.g. RSV, MPV
		Metapneumovirus e.g. HMPV, AMPV
		Subfamily *Paramyxovirinae*
		Rubulavirus e.g. mumps virus
		Avulavirus e.g. NDV
		Respirovirus e.g. SeV, PIV
		Henipavirus e.g. Hendra virus
		Morbillivirus e.g. measles virus
		Aquaparamyxovirus
		Ferlavirus
	Filoviridae	*Marburgvirus*
		Ebolavirus e.g. EBOV
	Rhabdoviridae	*Vesiculovirus* e.g. VSV
		Lyssavirus e.g. rabies virus
		Ephemerovirus
		Novirhabdovirus
		Cytorhabdovirus
		Nucleorhabdovirus
		Novirhabdovirus
		Perhabdovirus
		Sigmavirus
		Tibrovirus
	Bornaviridae	*Bornavirus* e.g. Borna disease virus

Figure 1.1 The classification of the order *Mononegavirales*.

Classification determined by the International Committee on Taxonomy of Viruses (ICTV), current update (2012).

1.1.3 Epidemiology

The most recent figures from the World Health Organization (2009) estimate that RSV infects 64 million people each year, killing 160,000, although with 95 % of these cases occurring in developing countries these figures may underestimate the true burden of disease. In Europe, RSV infection accounts for 42-45 % of hospital admissions for lower respiratory tract infection in children <2 years of age, and in the USA in 2005, 2.1 million children <5 years of age required medical attention because of RSV infection (Hall et al., 2009). Consequently, whilst RSV-associated mortality is relatively low, the economic burden of this virus is substantial.

The majority of children are infected within the first 6 months of life, and virtually all have been infected with RSV by the age of 2. Multiple infections throughout childhood, without antigenic variation by the virus, are commonplace despite there being only one serotype, due to inadequate and short-lived protective immunity by neutralising antibodies, compounded by the high levels of exposure in schools and nurseries (Collins and Graham, 2008). In the UK and other temperate countries, RSV epidemics occur annually and show predictable seasonality between October and April (Goddard et al., 2007). Both subgroups of the virus can co-circulate during an epidemic, although one usually predominates.

1.1.4 Pathogenesis and immunology

Transmission of RSV, like many respiratory viruses, is via large-droplet aerosol and direct contact. RSV is extremely contagious, with contamination of the nasal or conjunctival mucosa being an effective means of infection. Viral tropism is for the airway epithelium and infection begins with viral replication in the superficial cells of the nasopharynx. The initial host response to infection is mediated by the innate immune system, triggered by recognition of the pathogen by macrophages and toll-like receptor 4 (TLR-4), which recognises the RSV F protein (Kurt-Jones et al., 2000). This results in upregulation of an array of pro-inflammatory cytokines and chemokines that promote neutrophil and natural-killer cell recruitment, responsible for cytotoxicity and interferon-γ (INF-γ) induction, respectively (reviewed in McNamara and Smyth, 2002). The histology of infected tissues shows necrosis of the epithelium and accumulation of immune cells, such as monocytes and T-cells, leading to obstruction of the small bronchioles by sloughed epithelial cells and mucus (Collins and Graham, 2008).

The adaptive immune response is responsible for both cell-mediated clearance of RSV and the humoral response required for protective immunity. $CD8^+$ cytotoxic T-lymphocytes (CTL) and $CD4^+$ T-helper cells are both recruited to the sites of RSV infection, and they play roles in the pathogenesis of RSV disease as well as viral clearance. The balance between the two types of $CD4^+$ T-helper cell responses is a recurring theme in RSV pathogenesis: 'anti-viral' Th1 responses promote $CD8^+$ CTL-mediated cytotoxicity of infected host cells; Th2 responses promote the humoral adaptive immune response through B-cell activation and production of neutralising antibodies, but are also associated with allergic responses such as asthma. It has been suggested that a bias towards a Th2-type response may be responsible for the bronchiolitis and recurrent wheezing of severe RSV infection, as Th2-bias was responsible for vaccine-enhanced disease after the 1960's formalin-inactivated RSV (FI-RSV) vaccine trial: RSV-naïve children suffered enhanced disease severity upon subsequent infection (Kapikian et al., 1969). Recently it has been suggested that poor stimulation of B-cell-associated TLRs by the FI-RSV led to a lack of affinity maturation – the differentiation and proliferation of antigen-specific B-cells required to produce a high-affinity antibody response – and resulted in the significant but non-protective antibody response. These antibodies failed to neutralise the virus, allowing unrestricted amplification of the Th2 response (Delgado et al., 2009). A link between recurrent RSV infection in early life and the development of allergic asthma has recently been shown to be as a result of Th2-like inflammation in the lungs; this results in permanently biasing T-regulatory cells (a T-cell population involved in control and suppression of inflammatory responses) towards low antigen tolerance, and permits the allergic response to otherwise innocuous antigens that is characteristic of asthma (Krishnamoorthy et al., 2012).

Data to support a Th2-bias in RSV infection are conflicting, due in part to the fact that the mouse models used are poorly representative of the human immune system. A more appropriate theory may be that balancing the different $CD4^+$ T-helper cell responses is required to manage the compromise between the protective and pathogenic effects of the cellular adaptive immune response. A serum neutralising-antibody response to RSV, particularly against the F and G protein antigens, is induced during infection and a strong response is associated with reduced risk of bronchiolitis in children (McNamara and Smyth, 2002). Despite this apparent protective effect, RSV is able to infect infants in the first 2-6 months of life when maternal antibodies transferred across the placenta should still be present, due to the immunological immaturity of young infants preventing them from mounting a robust and specific response.

The relative contributions of the host response and the virus itself to RSV pathogenesis are still debated. There is considerable evidence to suggest that the host response to infection plays a significant role, in particular the FI-RSV vaccine disease-enhancement. Additionally, RSV is not highly cytopathic and there are genetic polymorphisms that are linked to increased RSV disease severity, such as those leading to increased expression of cytokines that recruit and activate immune cells (e.g. IL-4 and IL-8), further suggesting that the host response is a factor in the pathogenesis of infection (Collins, 2007). However, in support of a direct role of the virus in RSV pathogenesis, disease appears to be more severe in the immunocompromised and some cytopathology is observed, such as disruption of ciliary beating and the sloughing of epithelial cells.

Most paediatric RSV infections result in mild symptoms such as rhinorrhoea, fever and a cough. However, 25-40 % of children develop lower respiratory tract diseases such as bronchiolitis, pneumonia or bronchitis, due to a number of risk factors: premature birth, male gender, lack of maternal antibodies, low birth weight, and underlying cardiopulmonary disorders

(Olszewska and Openshaw, 2009). This results in symptoms including wheezing and laboured breathing, because of increased airway resistance. Lower respiratory tract infection (LRI) is as a direct result of RSV infection of the bronchiolar and alveolar epithelium, rather than as a result of the host response to the preceding upper respiratory tract infection (URI).

In contrast to the frequently severe pathogenesis of infection in children, most occurrences of RSV in adults go undiagnosed as the symptoms are non-specific and similar to other respiratory pathogens (Hall, 2001). Particularly severe RSV infection during the first 6 months of life may lead to childhood wheezing and asthma, as described above. In recent years there has been increased focus on the significant morbidity and mortality in the elderly, high-risk adults (those with chronic heart or lung disease) and the immunocompromised, particularly bone marrow transplant patients (Falsey et al., 2005).

1.1.5 Treatment

1.1.5.1 Therapeutics

Despite awareness of the high prevalence of this virus, and continued research efforts over the last 50 years, there is no cost-effective efficacious anti-viral therapy for RSV infection. Current therapeutic options include prophylactic use of palivizumab (MedImmume), a humanised mouse monoclonal neutralising antibody (mAb) against RSV fusion protein (F), in high-risk neonates; this is expensive and only acts to reduce disease severity by preventing viral spread to the lower respiratory tract, resulting in fewer hospitalisations (55 % in a 2004 study) (Fenton et al., 2004). Development of motavizumab (AstraZeneca), a second-generation mAb based on palivizumab with significantly higher neutralisation activity *in vitro* (Wu et al., 2007), was discontinued in 2010 after phase III trials.

The use of ribavirin, a nucleoside analogue with broad anti-viral activity, is technically licenced for use against RSV, however it is no longer recommended due to a lack of proven efficacy and possible toxicity. Alios BioPharma began a phase I clinical trial for a first-in-class structurally novel anti-RSV nucleoside analogue in 2013, following on from their phase I success with a similar strategy for hepatitis C virus.

A number of fusion inhibitors have been found by high-throughput screening of small molecule libraries, and co-crystal structures revealed that these target formation of the six-helix bundle of the fusion protein. However, development of all but one was discontinued by 2009 and new candidates are only in early clinical evaluation (Olszewska and Openshaw, 2009).

An antisense RNA against the RSV nucleocapsid protein, ALN-RSV01 (Alnylam Pharmaceuticals), has been evaluated in phase II trials for the prevention of a severe complication of RSV LRI, bronchiolitis obliterans syndrome – a permanent constriction and fibrosis of the bronchioles. This is an important proof of concept, as it is the first RNAi-based drug to be tested in humans against a microbial pathogen. It works by binding N protein mRNA in the cytoplasm, preventing its expression and consequently stopping the virus from replicating (DeVincenzo et al., 2010).

1.1.5.2 Vaccine candidates

A number of inactivated and live attenuated vaccines for bovine RSV have been available for 20 years (Meyer et al., 2008), but the situation for human RSV is very different. Prior to the last 15 years, the development of vaccines against human RSV has been tentative and slow. Recently reviewed by Collins and Molero (2011), a number of live-attenuated viruses, initially produced by cold passage and later by reverse genetics, have been identified and evaluated in clinical trials and do not show the vaccine-enhanced disease associated with inactivated RSV vaccines. Currently promising candidates are: RSV cps2 live-attentuated (phase I, MedImmune), which is a codon-stabilised version of MEDI-559 – a strain containing mutations in the polymerase, the M2 gene-start signal, and the deletion of the SH gene, which is in phase II trials; MEDI-ΔM2-2 live-attenuated (phase I, MedImmune), a rationally-designed strain with most of the M2-2 ORF removed to increase transcription and, thus, antigen expression.

Other types of vaccine are also being evaluated for RSV. A subunit vaccine of baculovirus-expressed RSV post-fusion F protein, designed for maternal vaccination and consequent placental-transfer of neutralising antibodies, is currently entering phase II trials (Novavax) (Glenn et al., 2013). This strategy aims to circumvent one of the most significant issues with RSV vaccinology – the very short window of opportunity for vaccination. A number of viral-vector vaccines are in early stages of clinical testing, and a chimeric bovine parainfluenza virus 3 (bPIV3) has been successfully evaluated at phase I: the bPIV3 F and haemagglutinin proteins are substituted for human PIV3 proteins, and the vector additionally modified to express the RSV F protein.

1.2 THE VIRUS

1.2.1 The virion

The RSV virion is renowned for its pleomorphism, instability, and relatively poor titres in cell culture – features that have hampered attempts to characterise the 3D structure of the whole virus until recently, when the ultrastructure of the RSV virion was confirmed by electron cryotomography (Liljeroos et al., 2013), shown in figure 1.2. Preparations of purified virus consist of a mixture of spherical and filamentous particles, and particles that show both morphologies, ranging from 100 nm to 1 μm in diameter and up to 2 μm in length. Both spherical and filamentous morphologies have been associated with infectivity, although Liljeroos et al (2013) demonstrated that the loss of infectivity resulting from freeze-thaw of the virus corresponds to loss of the filamentous form. A host-derived lipid bilayer is studded with an irregular arrangement of the viral transmembrane glycoproteins, F and G, which are required for fusion and attachment, respectively, and are the primary antigenic components of the virus. The virion membrane also contains the small hydrophobic (SH) protein, which forms a pentameric ion channel (Gan et al., 2012), but its low abundance means it has not been fully characterised by cryo-EM. The virion has an internal membrane-associated layer of matrix, M, protein that is considered essential for particle assembly (figure 1.2 B). The left-handed helical ribonucleoprotein (RNP) complex forms the core of the virion, and is composed of the genomic RNA in association with a number of proteins: the nucleocapsid, N, protein is tightly bound to the RNA creating the helical structure; the RNA-dependent RNA-polymerase (RdRp), or 'large' L protein, is the key component of the transcription/replication complex; the phosphoprotein, P, protein is an essential co-factor of the viral RdRp; the M2-1 protein, an additional RdRp co-factor that is required for transcription of viral mRNAs.

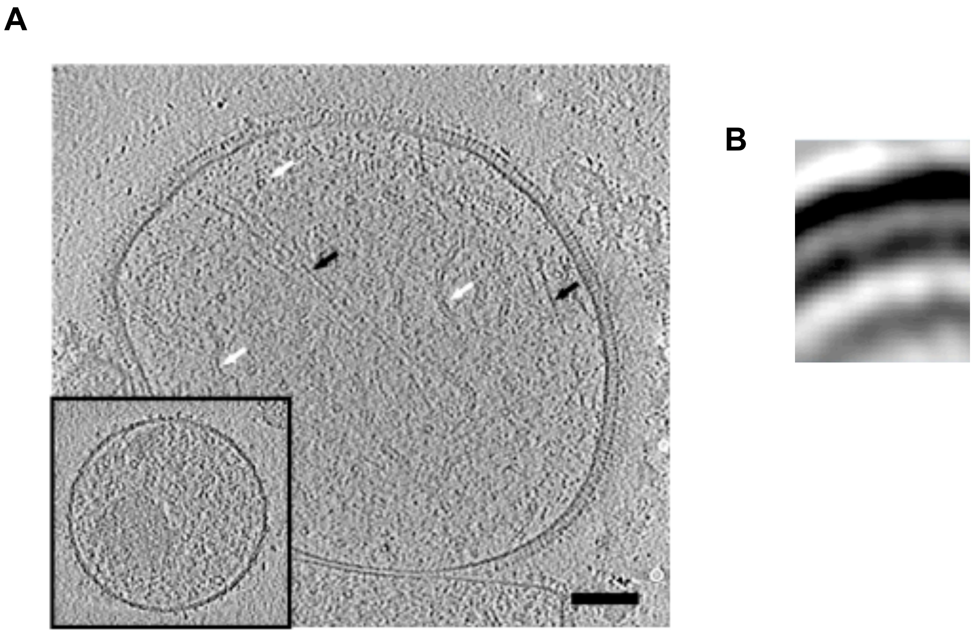


Figure 1.2 Electron micrographs of RSV virions.

A) Purified RSV visualised by cryo-EM shows a particle with an almost uniform layer of glycoprotein spikes, with an inset particle containing only a few spikes. Filamentous actin inside the virion is marked with black arrows and RNPs with white arrows. Scale bar, 100 nm. **B)** Averaged profile of the virion membrane and its associated matrix layers, M and possibly M2-1. *Figures from Liljeroos et al, 2013.*

1.2.2 Genome organisation

1.2.2.1 The Mononegavirales

The RNA genomes of the order *Mononegavirales* are single-stranded, monopartite, negative-sense, vary in length from 8.9 to 19 kb, and are displayed 3' to 5' by convention. The general order of viral genes is conserved across the whole order: 3' – nucleocapsid, P protein, matrix, surface glycoproteins (F/G/HN), and the polymerase – 5'. This is shown in figure 1.3, which compares the genomes of several genera of the family *Paramyxoviridae* and the rhabdovirus VSV. Both genera of the subfamily *Pneumovirinae* contain additional genes not present in the other paramyxoviruses, namely the SH, M2 and, in the case of RSV, NS1 and NS2 genes. Table 1 outlines the basic functions of RSV encoded proteins and they are described in more detail below.

The viral genome is flanked by 3' leader and 5' trailer regions, which show inverse complementarity and contain the promoters to drive mRNA/antigenome and genome synthesis, respectively. Despite this complementarity, none of the paramyxoviridae are predicted to circularise or form a panhandle structure, as with influenza A virus. Each viral gene is delineated by gene start and gene end *cis*-acting signals, which direct the sequential 'stop-start' mechanism of transcription (Kuo et al., 1996). Gene start signals are highly conserved within each virus family; gene end sequences are less well conserved, but contain essential motifs including a poly-uridine tract for the generation of the 3' mRNA poly-adenosine tail. Between these *cis*-acting signals is a non-transcribed intergenic region that can be absolutely conserved in length (3 nucleotides for respiroviruses and morbilliviruses) or variable in length (1-190 for pneumoviruses and rubulaviruses). Many of the mononegaviruses employ mechanisms to increase the coding-capacity of their genomes, including overlapping reading frames (ORFs), such as the M2-1 and M2-2 ORFs of the pneumovirus M2 gene or the P/C ORFs of many of the subfamily *Paramyxovirinae* P genes. The latter also use a mechanism of 'RNA editing', where non-templated insertion of guanosine nucleotides shifts the downstream translational reading frame to produce V/I/W/D proteins from the P gene (Collins and Crowe Jr, 2007). Borna disease virus, the only current member of the family *Bornaviridae*, has many unique characteristics that set it apart from the other mononegaviruses, including its nuclear site of transcription and replication. This allows the virus to use host RNA splicing machinery to generate the complete mRNA transcript of the viral polymerase (Cubitt et al., 1994).

Gene	**Protein function**
NS1	Non-structural proteins: inhibit interferon α/β induction and signalling
NS2	
N	Nucleocapsid: encapsidates genomic and antigenomic RNA, component of the RNP template
P	Phosphoprotein: essential polymerase co-factor
M	Matrix: viral assembly
SH	Small hydrophobic: viroporin of unknown function
G	Glycoprotein: attachment, secreted form binds neutralising Abs i.e. immune evasion
F	Fusion: viral entry, syncytia formation
M2	'Matrix-like' protein 1, M2-1: transcription antitermination factor 'Matrix-like' protein 2, M2-2: RNA regulatory protein, viral transcription/replication balance?
L	Large protein: RNA-dependent RNA polymerase

Table 1 Virally encoded proteins of RSV and their functions.

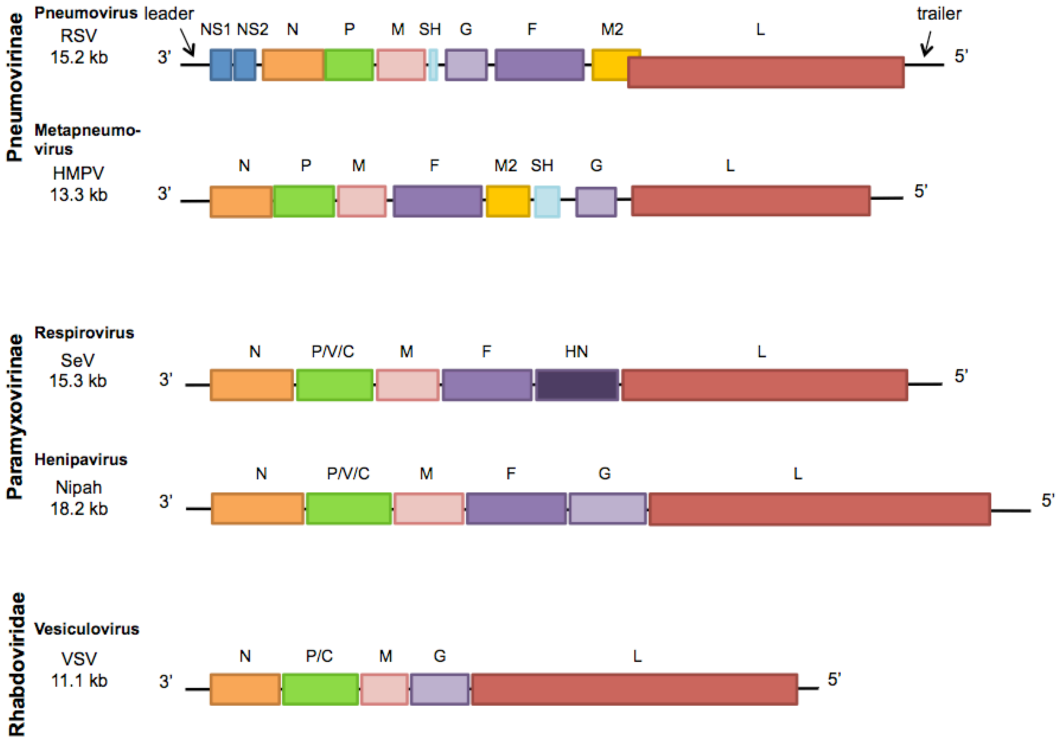


Figure 1.3 Genome comparisons of the order *Mononegavirales.*

Approximately to scale representations of genomes from the subfamilies *Pneumovirinae* and *Paramyxovirinae* and an example from the family *Rhabdoviridae*. Homologous genes are the same colour, with the various glycoproteins shown in shades of purple. The overlap between M2/L of RSV is shown by the offset L gene.

1.2.2.2 RSV

The genome of RSV is 15.2 kb in size and contains 10 genes that are transcribed into 10 5'-capped and 3'-polyadenylated mRNAs, and translated into 11 viral proteins due to the overlapping ORFs of the M2 mRNA. The 3' leader region (Le) is 44 nucleotides long and is well conserved across the subfamily *Pneumovirinae*; in particular the first 11 nucleotides are necessary and sufficient for polymerase recruitment and initiation of RNA synthesis

(Fearns et al., 2002, Cowton and Fearns, 2005), although the 5' end of the leader sequence is required for transcriptional efficiency (McGivern et al., 2005). The first 34 nucleotides are required for antigenome synthesis and encapsidation – a distinguishing characteristic of RSV and the pneumovirinae, as the subfamily *Paramyxoviridae* has a bipartite replication promoter requiring sequences in the first gene. The 3' antigenome contains the 155 nt trailer complementary (TrC) region, which contains the promoter sequences for genome synthesis and the first 26 nt are highly conserved with the Le sequence (Fearns et al., 2000).

The conserved gene start (GS) sequence of RSV is 3' CCCCGUUUA 5', and differs only for the L gene, whose gene start sequence is 3' CCCUGUUUU 5'. The L gene is also unusual in that it overlaps with the previous M2 gene: The L gene start signal is upstream of the M2 gene end signal, creating a 68 nucleotide overlap (Collins et al., 1987). This arrangement is found in all respiratory syncytial viruses, but no other members of the subfamily *Pneumovirinae*. RSV gene end signals vary in length and sequence not only between viral strains, but also between gene junctions within those strains. For RSV A2 strain, the sequence consists of a conserved 3' UCAAU 5' motif, followed by an unconserved region of 3-4 nt and a polyU tract of variable length. The efficiency of termination at a gene junction is controlled by the gene end sequence (Hardy et al., 1999), and results in read-through mRNAs at a level that may be proportional to divergence from the consensus sequence (Sutherland et al., 2001).

1.3 VIRUS LIFE CYCLE

1.3.1 Entry

The life cycle of RSV begins at the apical cell surface of the nasopharyngeal epithelium (figure 1.4). *In vitro,* initial attachment is thought to be by electrostatically driven binding of the G protein to negatively charged cell surface carbohydrates as, unlike the specific binding of sialic acid by the haemagglutinin glycoproteins of many of the other paramyxoviruses, the RSV G protein has been suggested to bind a range of targets, many of which are not expressed at the apical surface of the airway epithelia, e.g. heparin (Mastrangelo and Hegele, 2013). In further contrast to the subfamily *Paramyxovirinae*, the RSV G protein is not thought to bind to the F protein and regulate its fusogenic activity, as it is dispensable for viral infection *in vitro*; ΔG mutants are, however, attenuated *in vivo* (Dutch, 2010).

The only definitive entry receptor for RSV is nucleolin (Tayyari et al., 2011), which is found in the lipid rafts that RSV requires for entry, and F protein has been shown to bind specifically to nucleolin at the cell surface. Fusion proteins of the paramyxoviridae are highly conserved and require cleavage by cellular proteases (furin in the case of RSV) to induce extensive and irreversible conformational change, allowing fusion of the viral and host cell membranes (see below). The F protein of RSV is unique in that it requires two fusion events, rather than one as for the other paramyxoviruses (Gonzalez-Reyes et al., 2001). The recent discovery of the mechanism of RSV entry may explain this requirement: Unlike other members of the family *Paramyxoviridae* who fuse directly at the plasma membrane, RSV is endocytosed intact by macropinocytosis and the second F protein cleavage event occurs post-internalisation, within the endosome, initiating intracellular membrane fusion (Krzyzaniak et al., 2013). Interestingly, macropinocytosis is also the method of entry for the Ebola filovirus. Fusion of viral and host membranes by the F protein does not require acidification of endosomes, and results in the release of the viral ribonucleoprotein complex (vRNP) into the cytoplasm.

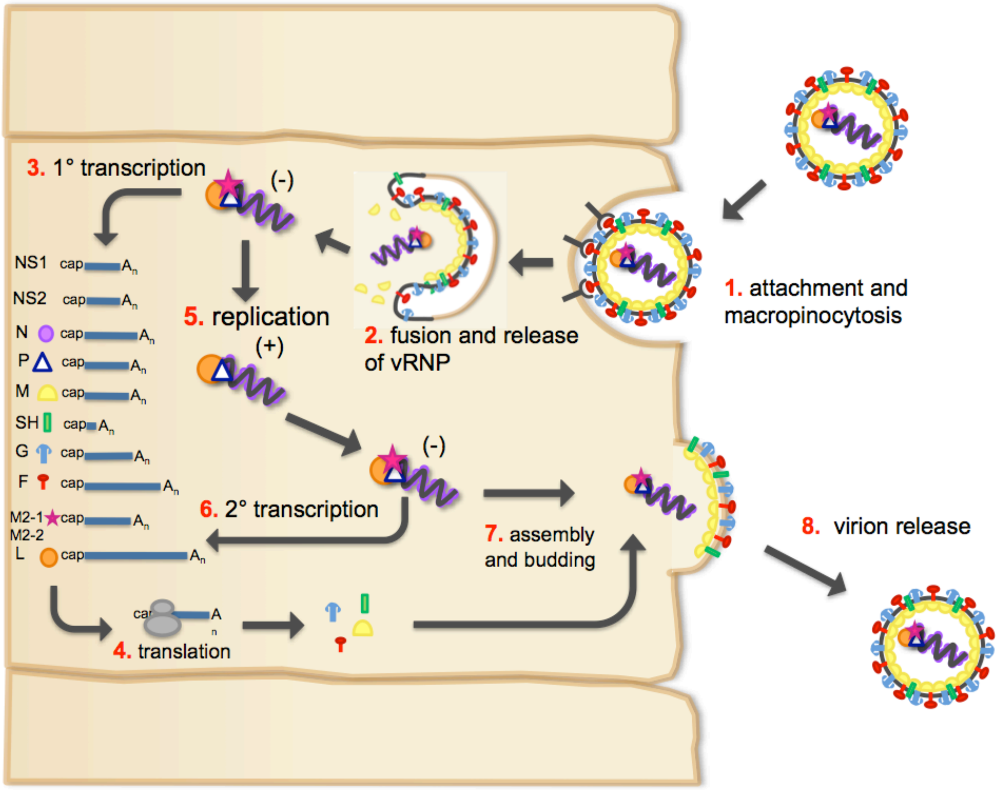


Figure 1.4 The life cycle of RSV.

1) Attachment of RSV virions by G protein to apical cell surface carbohydrates and binding of the F protein to the entry receptor nucleolin, triggers endocytosis by macropinocytosis. **2)** Further cleavage of the F protein and conformational change initiates fusion of viral and host membranes, releasing the vRNP. **3)** Primary transcription of the genome into 10 mRNAs by the viral transcription complex (L, N, P, M2-1). **4)** Translation by host ribosomes into viral proteins, allowing build up of N protein required for replication. **5)** Replication (L, N, P) and concurrent encapsidation of antigenome and genome. **6)** Secondary transcription of newly synthesised genomes to amplify viral protein expression. **7)** Trafficking of viral proteins and genomic vRNP complexes to the apical cell membrane for virion assembly and budding. **8)** Release of RSV virions from the cell surface into the nasopharyngeal lumen.

1.3.1.1 G protein

RSV G protein is a highly glycosylated transmembrane protein with a polypeptide molecular weight of 32 kDa, although the considerable carbohydrate content raises its apparent molecular weight to 3 times this. It bares no sequence or structural similarity to the HN proteins of the family *Paramyxovirinae*, and is the least conserved protein amongst RSV strains. The hydrophobic C-terminal ectodomain of G protein comprises two thirds of the protein, and is made up of two highly variable mucin-like domains. The majority of G protein epitopes recognised by neutralising antibodies are in this region, and its variability and extensive glycosylation may partly explain the virus' ability to evade such a response (Collins and Melero, 2011).

G protein is proposed to bind cell surface carbohydrates to mediate viral attachment, however this protein is not essential for the *in vitro* rescue of RSV, or other pneumovirinae such as the metapneumovirus avian metapneumovirus (AMPV) (Naylor et al., 2004). The importance of the G protein *in vivo* may lie in its alternative secreted form, generated by a second translation initiation codon and resulting in an N-terminal truncation of the transmembrane anchor. Soluble G protein accounts for ~80 % of total G and acts as a decoy for the host adaptive immune response (Bukreyev et al., 2012).

1.3.1.2 F protein

RSV F protein is synthesised as a 67 kDa inactive precursor, which is cleaved in the trans-Golgi network to form two disulphide-linked polypeptides. This creates the N-terminal fusion peptide that directly inserts into the target membrane, and allows homo-trimerisation to form the metastable pre-fusion conformation. The crystal structure of both pre- and post-fusion forms of RSV F protein have allowed the process of membrane fusion to be explained, and reinterpreted in light of the role of endocytosis in RSV entry (Mastrangelo and Hegele, 2013): Pre-fusion F protein binds to nucleolin, initiating macropinocytosis and refolding the F protein into an extended intermediate; within the endosome, the fusion peptide is inserted into the host cell membrane and a second furin cleavage triggers the irreversible conformational change of F protein to the post-fusion six-helix-bundle form – a lower energy state that drives the energetically-unfavourable process of membrane fusion. This 'spring-loaded' mechanism of fusion is thought to be similar to influenza A virus haemagglutinin-mediated fusion (Collins and Crowe Jr, 2007). All other family *Paramyxoviridae* F proteins undergo only one cleavage event, and the triggering of conformational change is instead driven by interaction with the attachment glycoprotein.

Syncytia formation, whereby infected cells express the F protein at their surface and fuse with their neighbours forming giant multinucleated cells and allowing viral spread, occurs by pH-independent fusion at the plasma membrane. Although syncytia are a hallmark of infection *in vitro*, they are rarely seen in histological patient samples, except in very severe disease.

1.3.2 Overview of transcription and replication

Once endosomal fusion has occurred, the viral RNP is released into the cytoplasm and associates with lipid rafts and cytoskeletal proteins, such as actin and profilin, to form cytoplasmic inclusion bodies where viral RNA synthesis occurs (Cowton et al., 2006). A single promoter in the 3' leader region of the genome drives both transcription and replication - two mutually exclusive modes of viral RNA synthesis.

Initially the virus is limited to a phase of primary transcription, when mRNAs are transcribed from the encapsidated genome by the viral RdRp and its cofactors, P and M2-1, in a sequential 'stop-start' mechanism; this requires transcription termination of the upstream gene at the gene end sequence, scanning across the intergenic region, and transcription reinitiation at the downstream gene start sequence – all without disassociation from the template. RSV mRNAs and proteins can be detected in the cytoplasm 4-6 hours post infection (Collins and Crowe Jr, 2007). Host-translated viral proteins accumulate in the cytoplasm until a concentration of free nucleocapsid protein is reached that permits replication of the genome; this process does not require M2-1 and is dependent on *de novo* synthesis of N for the concurrent encapsidation of both the antigenomic intermediate, and replicated genome. The polymerase must ignore the *cis*-acting signals that directed termination and reinitiation during transcription, and synthesise a complete viral antigenome, from which the subsequent synthesis of a viral genome is initiated at the TrC promoter. Secondary transcription of these newly synthesised genomes allows amplification of viral protein production.

1.3.3 Transcription

The mechanism of transcription for RSV is shared by the three main families of the order *Mononegavirales*: *Paramyxoviridae*, *Filoviridae* and *Rhabdoviridae*. Initial insights came from UV target size experiments on VSV and RABV (Abraham and Banerjee, 1976); UV-induced lesions in upstream genes prevented transcription of those downstream, consistent with the mechanism of sequential transcription described above.

The 5' end of nascent mRNA transcripts is capped (i.e. guanylylation and methylation) by the viral RdRp itself (Liuzzi et al., 2005), and is followed by elongation of the mRNA transcript – a process that requires the antitermination ability of M2-1, the subject of this thesis and described in detail below. The 'stop-start' mechanism of transcription links the termination and polyadenylation of the upstream gene, the latter of which is thought to be by reiterative transcription of a polyU tract in the gene end sequence as shown for VSV (Barr and Wertz, 2001), to the transcription initiation of the downstream gene. After transcription of an upstream gene is terminated at the gene end signal and the nascent mRNA released, the polymerase complex scans in either direction until a gene start signal is encountered (Fearns and Collins, 1999a). This bi-directionality is particularly important at the M2/L gene junction, as these two genes overlap by 68 nt and the polymerase must scan backwards to locate the L gene GS signal upstream of the M2 GE signal. 90 % of polymerase initiations at the L gene start sequence will terminate 68 nt later at the M2 gene end sequence, producing a short polyadenylated mRNA of no known function (Collins and Crowe Jr, 2007). Successful L mRNA synthesis requires read-through of the M2 GE signal. VSV, subfamily *Paramyxovirinae*, does not have an overlapping L gene but a very high level of attenuation at the G/L junction has been observed, likely as a result of the polymerase being insufficiently processive to traverse the 6,379 nucleotide L gene in the absence of an M2-1-like protein (Barr et al., 2002). This suggests that downregulation of the RdRp may be essential to maintaining an appropriate level of viral transcription and replication, and may be common to all paramyxoviruses.

Descriptions of each of the proteins required for viral transcription are below, and M2-1 is described in further detail later in the chapter.

1.3.3.1 N protein

The nucleocapsid protein of all mononegaviruses is essential, as the complete encapsidation of the genome is required to form the RNase-resistant template for viral RNA synthesis. RSV N is 45 kDa and its crystal structure bound to bacterial RNA revealed the oligomerisation of N to form the helical RNP: N protein monomers, consisting of N- and C-terminal regions linked by a hinge, oligomerised to form a decameric ring representing one turn of the helix (Tawar et al., 2009), shown in figure 1.5. RNA was buried in a basic groove on the external surface of the ring, in contrast to similar structure of VSV N:RNA. Each monomer bound 7 nucleotides of RNA, in further contrast to many of the paramyxoviruses, such as SeV, whose genomes are obligatorily multiples of six in length to satisfy the binding of N to viral RNA. This structure also showed that the N protein C-terminus extended above the ring, in line with the binding of this region to the viral phosphoprotein, P (Stokes et al., 2003), which is thought to keep a pool of soluble N available for concurrent encapsidation.

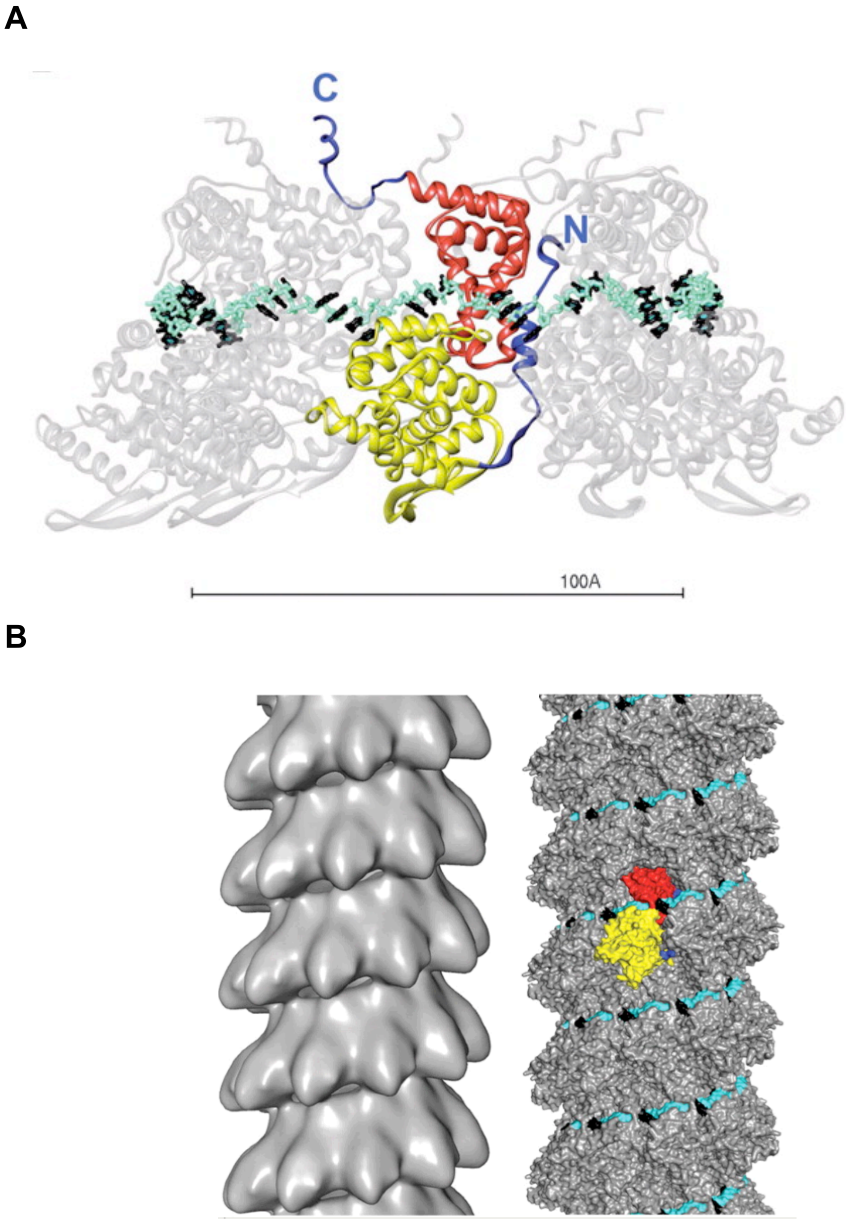


Figure 1.5 Crystal structure of RSV N protein with bacterial RNA.

A) The decameric N ring structure mimicking one helix turn is shown in grey, with one N monomer coloured yellow (N-terminal domain) and red (C-terminal domain), with its termini in blue. RNA is shown wrapped around the outside of the N ring, with the phosphate backbone in cyan and bases in black. **B)** A 26 Å resolution 3D reconstruction of the helical RNP complex, calculated from cryo-EM negative-stain images (left), and modeling of the N crystal structure into the reconstruction using the crystal contacts and the pitch of the EM helix (right). *Figures from Tawar et al. 2009.*

1.3.3.2 L protein

The large protein, L, is the RdRp component of the polymerase complex and is the 5'-most gene for all viruses of the order *Mononegavirales*. At 250 kDa, L comprises almost half of the coding potential of the RSV genome and is similar in size to the L proteins of other paramyxoviruses. Sequence alignment of paramyxovirus L proteins suggests they consist of 6 conserved regions (CR): in SeV, CRI appears to be involved in N recruitment and P binding, and CRII is involved in binding the N:RNA template; CRIII contains the enzymatically active site for all mononegaviruses; CRIV functions as the cap methylase in SeV and VSV; CRV contains the capping enzyme motif in VSV; CRVI has not been assigned (Li et al., 2008b). There are still no crystal structures available for a complete polymerase of any negative strand RNA virus, in stark contrast to the positive strand RNA viruses - although individual structures of the component proteins of the tripartite influenza A polymerase have been built into EM reconstructions of the polymerase complex. Low resolution EM images averaged over many observations of the VSV polymerase, suggest it consists of a large ring domain with several smaller appendage domains (Morin et al., 2013).

1.3.3.3 P protein

The phosphoprotein P is the essential co-factor of the L protein, required for both transcription and replication, possibly by ensuring placement of the polymerase on the RNP template. The RSV P gene encodes a single 33 kDa protein, which is considerably smaller than the subfamily *Paramyxovirinae* P proteins and lacks their alternative ORFs and RNA editing mechanisms. Fellow pneumovirus murine pneumonia virus (MPV) does have an second ORF in its P gene, and the P-2 protein inhibits viral mRNA synthesis in a minigenome system (Dibben et al., 2008). There is little structural information available for P; it is homotetrameric and forms interactions with L, N, and M2-1. RSV P is constitutively phosphorylated at five sites but the role of this modification is not known; unphosphorylated bacterially-expressed P protein cannot form tetramers without *in vitro* phosphorylation, and serine to alanine mutants of key phosphorylation sites reduced N-binding by coimmunoprecipitation. However, ablation of phosphorylation does not seem to affect replication of a minigenome in mammalian cells, and phosphoablatant-P virus can be rescued by reverse genetics (Lu et al., 2002). Phosphorylation of P at threonine 108 has been shown to specifically prevent the P:M2-1 interaction (Asenjo et al., 2006).

1.3.3.4 M2 gene

The M2 gene is unique to the subfamily *Pneumovirinae* and does not appear to have a homologue in the subfamily *Paramyxovirinae,* although VP30 protein of the family *Filoviridae* Ebola virus has some similarities (see below). The M2 gene is transcribed into a single mRNA, which is translated into two proteins from overlapping open reading frames: M2-1 and M2-2. The translation of the 3' M2-2 ORF, whose 3 possible start codons are upstream of the M2-1 stop codon, is by a coupled translation mechanism that had not been previously described in any eukaryotic system prior to its discovery in RSV (Ahmadian et al., 2000). M2-2 translation requires the termination of M2-1 translation and the reverse translocation of the host ribosome to reinitiate translation at the M2-2 AUG. Truncation analysis of the M2-1 ORF showed that virtually the whole M2-1 sequence was required to achieve wild-type levels of reporter expression from the downstream open reading frame, suggesting that an RNA secondary structure requiring most of the M2-1 mRNA may be involved in the coupling process (Gould and Easton, 2005). This was supported by Mfold predictions and the observation that moving the M2-1 stop codon further downstream of the M2-2 AUG codons prevents M2-2 translation, possibly as a ribosomal interaction with an M2-1 mRNA secondary structure element is no longer sterically favourable (Ahmadian et al., 2000). A similar process has been demonstrated in the pneumovirus MPV and the metapneumovirus AMPV (Gould and Easton, 2007), and the variation in M2-2 expression levels appears to be determined by the secondary structure imparted by the upstream M2-1 sequence. Coupled translation has also been reported in two caliciviruses, a positive strand RNA virus family.

The protein products of the two overlapping M2 ORFs are M2-1, a 22.15 kDa transcriptional antiterminator that increases polymerase processivity and is essential for transcription but dispensable for replication (described it detail below), and M2-2, a 11 kDa protein that was proposed to regulate the balance between viral transcription and replication (Bermingham and Collins, 1999). Expression of M2-2 *in vitro* was found to increase viral replication and decrease transcription, although data on the precise role of M2-2 are conflicting and ΔM2-2 viruses are still viable, if attenuated, in cell culture. A regulatory role for M2-2 is in line with the control over its expression that results from the inefficient nature of coupled translation (Gould and Easton, 2007), and a similar role has been identified for M2-2 of MPV (Dibben et al., 2008).

1.3.4 Replication and its differentiation from transcription

Replication of the RSV genome is via a full genomic-length, encapsidated, positive-sense antigenomic intermediate, from which an encapsidated negative-sense genome is synthesised; both antigenome and genome synthesis require *de novo* synthesis of the N protein for concurrent encapsidation of the nascent RNA. The 5' genomic trailer region contains the reverse complement of the 3' leader promoter, resulting in a trailer complementary (TrC) promoter at the 3' end of the antigenomic intermediate. The TrC promoter directs a higher rate of replication than the leader region promoter, resulting in an increased ratio of genome to antigenome synthesis (Fearns et al., 2000). This difference in promoter strength may be important for ensuring the majority of progeny virions contain negative sense genomes, as is the case with rabies virus.

Most importantly, replication differs from transcription in that the viral polymerase complex must bypass the *cis*-acting signals for termination and synthesise a copy of the entire genome – a process that, counter-intuitively, does not require the polymerase processivity factor, M2-1. How the same promoter and RdRp achieve both of these mutually exclusive processes is still debated. Historically, there are two proposed models for how transcription and replication are differentiated:

Model 1 suggests that the polymerase complex initiates RNA synthesis at the first 3' nucleotide and transcribes the leader region. During transcription this sequence will be terminated and released, allowing the polymerase to reinitiate at the first gene start signal. During replication, the leader RNA is concurrently encapsidated and is not terminated, thus synthesis continues along the whole length of the genome. The presence or absence of sufficient soluble N protein is thought to be the 'switch' between the two modes.

Model 2 suggests that there are two pools of polymerase: a 'transcriptase' composed of L, P and M2-1 which scans along to the first GS signal and initiates mRNA synthesis independently of leader transcription; and a 'replicase' composed of L and P, which initiates at the first nucleotide and transcribes the whole genome, including the leader sequence (Curran and Kolakofsky, 2008, Whelan, 2008, Banerjee, 2008).

There is evidence to support both models. Other members of the family *Paramyxoviridae* have demonstrated distinct pools of polymerase complexes (model 2) that vary in their composition and post-translational modifications, such as VSV (Qanungo et al., 2004). The crystal structure of RSV N:RNA shows that the 3' polymerase recruitment sequence and the first gene start signal are spatially close together (figure 1.5), adding support to the idea of leader-independent initiation (model 1). However, recent developments in the methods of analysing RNA synthesis initiation *in vitro* provide new insights into a model 1-type mechanism:

Since the successful expression of recombinant RSV L protein, an *in vitro* RNA synthesis assay has been developed based on the assumption that the N protein flexes to temporarily expose the viral genome to the RdRp (as was revealed by the N:RNA crystal structure for RSV, Tawar et al., 2009), and that this action can be recapitulated using a naked RNA oligonucleotide (Noton et al., 2012). Data from this system are recent and constantly changing, and the latest publications indicate that initiation of RNA synthesis at both leader and trailer complementary promoters is significantly more complex than previously thought. The RSV polymerase is able to not only initiate at the 1+ position of the Le promoter, but also at 3+ position in response to a gene-start-like sequence within the promoter (Tremaglio et al., 2013). This is the first example of non-1+ transcription initiation in a non-segmented negative strand RNA virus, and has led to the suggestion that these different initiation sites may differentiate transcription from replication as follows (possible model outlined in figure 1.6): during replication, the RdRp initiates at 1+ and becomes concurrently encapsidated, synthesising a full genomic length RNA; during transcription, the RdRp initiates at 3+ and begins transcription of the leader region, which is prematurely terminated at an uncharacterised signal in the 5' of the leader (possibly explaining the requirement of this region for transcriptional efficiency), allowing the RdRp to scan along to the first gene start sequence and reinitiate transcription (figure 1.6). This is supported by the identification of abortive Le transcripts by primer extension of RNA purified from RSV infected cells. There is also evidence to suggest that RNA synthesis from the TrC promoter in the 3' antigenome can also initiate at 3+ position during replication, and that the viral polymerase can add additional nucleotides to the 3' antigenome in a back-priming mechanism similar to Borna disease virus (Noton et al., 2012). The implication of either of these two mechanisms is not currently understood.

Taken together, the newly discovered complexity of RSV polymerase interactions with the viral promoters provides an attractive theory for how the mutually exclusive processes of transcription and replication are differentiated on a molecular level, and the functional data presented in this thesis may support this mechanism and suggest an integral role for M2-1 in the process.

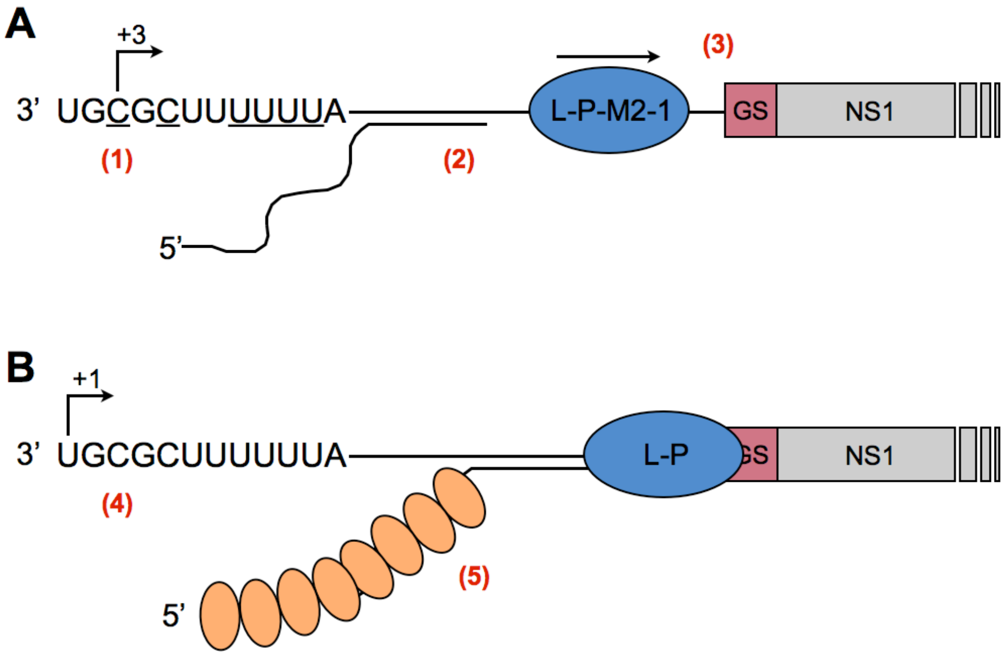


Figure 1.6 Model for transcription and RNA replication.

A) Transcription: the transcriptionally competent polymerase complex (blue; L, P and M2-1 proteins) enters the genome at the 3' end of the template and binds to nucleotides 3 to 12. RNA synthesis is initiated opposite position +3 of the Le (1) at a gene start-like sequence (underlined). After the synthesis of approximately 25 nt, the RdRp reaches a checkpoint and aborts synthesis of the Le RNA (2). The RdRp is now poised to scan the template to locate the first gene start signal at position +45 (red) and reinitiate RNA synthesis (3). **B)** RNA replication occurs if the RdRp initiates at +1 (4) and the RNA becomes concurrently encapsidated with N protein (orange) before the ~25 nt checkpoint (5).

1.3.5 Assembly

The release of progeny virus can be detected 10 to 12 hours post infection, implying that assembly begins not long after viral proteins are detected in the cytoplasm (4-6 hours p.i.). Release peaks after 24 hours, and continues in tissue culture until cell death by 30 to 48 hours p.i. (Collins, 2007).

Assembly of progeny virions occurs at the apical surface of the polarised epithelial cells, using the host apical recycling endosome to transport the matrix protein and RNPs to the viral glycoproteins at the plasma membrane. RSV proteins can be seen to localise at cell surface lipid microdomains linked to the cortical actin network, and the finding that cellular actin is packaged in RSV virions (Radhakrishnan et al., 2010) supports a role for actin in assembly, although this is not well defined (Shaikh et al., 2012). A popular mechanism for enveloped virus assembly and release is through the host cell ESCRT pathway that regulates vesicle budding into endosomes: retroviruses (e.g. HIV), the segmented negative strand RNA arenaviruses (e.g. LCMV), and most of the order *Mononegavirales*, including filoviruses, some rhabdoviruses, and some paramyxoviruses (e.g. PIV5 and MuV), are all ESCRT-dependent. However other paramyxoviruses, including NiV and RSV itself, as well as the rhabdovirus VSV and the orthomyxovirus influenza A, are ESCRT-independent and rely on as yet undefined mechanisms (reviewed in Harrison et al., 2010). The RSV matrix protein is assumed to be central to the coordination of virion assembly, binding the glycoproteins, viral RNPs and the lipid membrane (Ghildyal et al., 2002, Ghildyal et al., 2005b).

1.3.5.1 M protein

The RSV matrix protein, M, is 25 kDa hydrophobic protein that is essential to virion morphogenesis. The crystal structure of RSV M (figure 1.7) shows two compact N- and C-terminal domains with a linker region, and an extensive continuous patch of positive electrostatic surface potential covering one whole face of the monomer, which likely mediates the membrane association (Money et al., 2009).

M has been shown to silence viral transcription in preparation for RNP packaging – a role that has been demonstrated for other M proteins of the family *Paramyxoviridae* (Ghildyal et al., 2002); transcription of purified RSV RNPs in the presence of M protein was increased upon addition of M-specific antibodies that render M non-functional. The interaction between M and the RNPs is also suggested to recruit the RNPs to the plasma membrane for inclusion in nascent virions (Ghildyal et al., 2002), and has been proposed to be mediated by an M:M2-1 interaction. This is based on immunofluorescence co-localisation of M and viral inclusion bodies in transfected cells being dependent on M2-1 expression (Li et al., 2008a), and is supported by the suggestion of an additional matrix-like layer below the M layer in virions visualised by cryo-EM (figure 1.2) (Liljeroos et al., 2013). However this interaction could not be shown by co-immunoprecipitation in transfected cells and the assignment of M2-1 to the second matrix layer is speculative, due to a lack of immunogold staining.

Residues 120-170 of M are required for non-specific RNA binding (Rodríguez et al., 2004), and this has been linked to its nuclear localisation. The M protein has been detected in the nucleus during early infection by confocal microscopy, like other paramyxovirus M proteins (VSV, NDV, SeV), and has been linked to a decrease in host cell transcription (Ghildyal et al., 2003). The transport of RSV M to the nucleus is thought to be mediated by direct binding to importin β1 (a host protein responsible for transport through nuclear pores) by the M protein RNA binding region, which is sufficient to relocalise GFP in transfected cells (Ghildyal et al., 2005a).

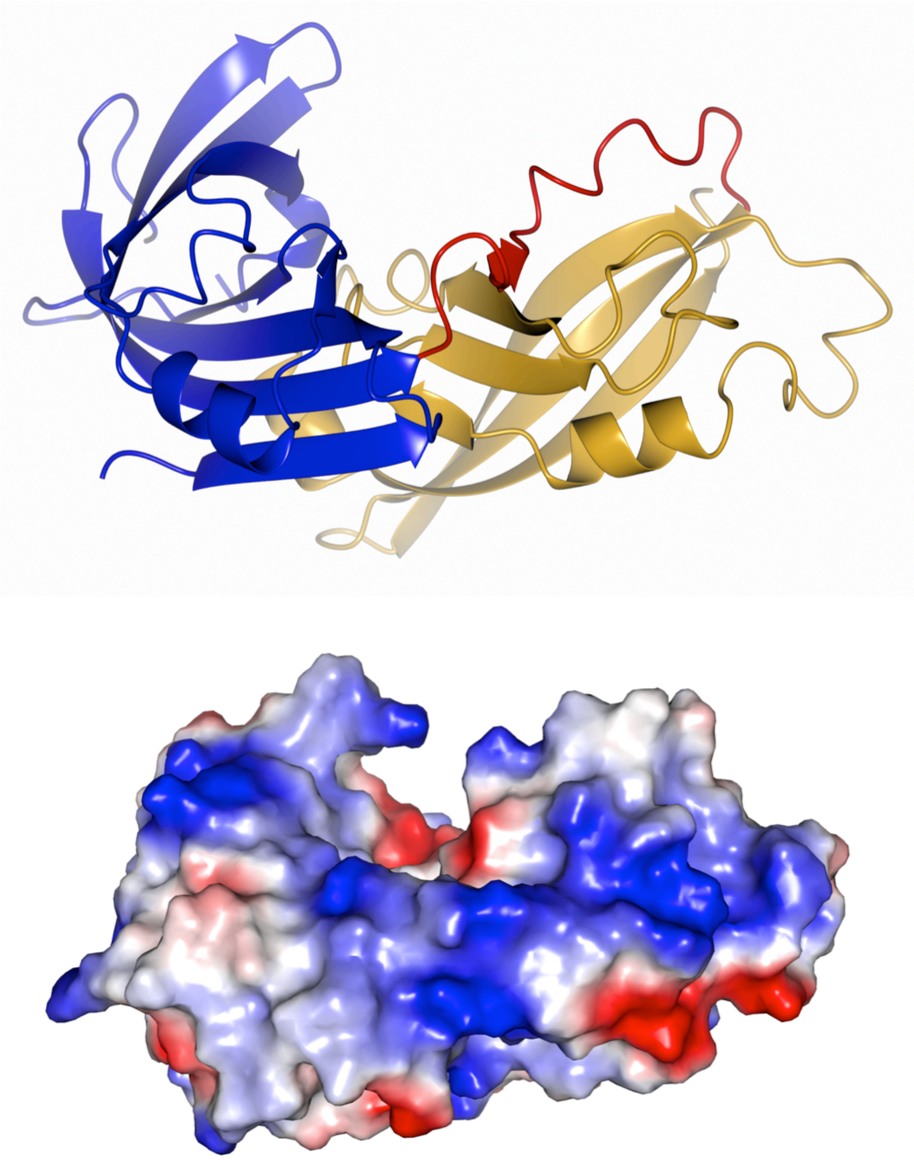


Figure 1.7 Crystal structure of RSV M protein.

Top, ribbon structure displaying the two β-sheet domains, N-terminal in blue and C-terminal in yellow. The linker domain is in red. **Bottom**, approximate electrostatic surface potential of M with charge coloured blue (positive) to red (negative), generated in PyMOL. PDB 2VQP.

1.3.6 Other viral proteins

1.3.6.1 SH protein

The small hydrophobic (SH) protein of RSV is the third transmembrane glycoprotein, encoded by all pneumovirinae and some paramyxovirinae from the *Rubulavirus* genus – parainfluenza virus type 5 and mumps virus (Collins and Crowe Jr, 2007). In RSV, the SH polypeptide is 7.5 kDa but has multiple glycosylated and N-linked-sugar forms. An SH solution NMR structure has recently shown that the protein forms pentameric ion channels with non-specific cation channel activity, demonstrated with whole-cell patch clamping (Gan et al., 2012), and so SH is now classed with other viroporins – such as HCV p7 and IAV M2 – characterised by their transmembrane amphipathic α-helices that form channel pores. Regulation of this channel activity in RSV is likely to be by pH-dependent protonation of a conserved luminal histidine, as with influenza A M2 (figure 1.8). The role of RSV SH in viral infection is not clear, as ΔSH viruses are viable *in vitro* and are the focus of several live-attenuated vaccine candidates. In infected cells, SH accumulates at the Golgi, endoplasmic reticulum and the plasma membrane, and is capable of permeabilising such membranes *in vitro*. Paramyxovirus SH proteins appear to inhibit apoptosis in mammalian cell lines, in direct contrast to the caspase 3-mediated apoptosis induced by other viroporins, e.g. IAV M2 and poliovirus 2B (Madan et al., 2008).

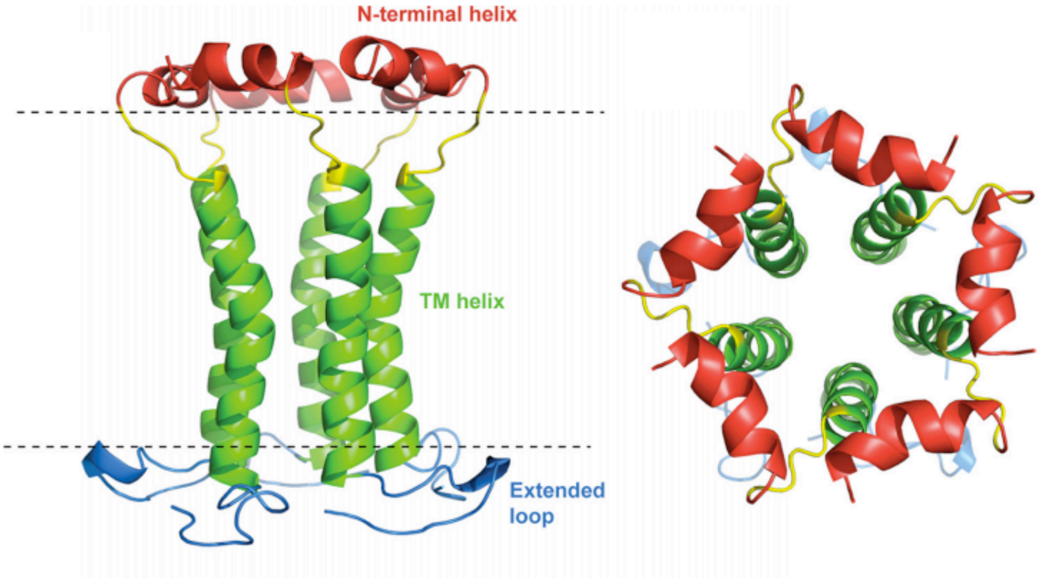


Figure 1.8 NMR structure of SH protein.

Solution NMR structure of the SH monomer, modelled in the most likely pentameric complex, shows the formation of an ion channel structure with a pore of ~3.5 Å at its narrowest. *Figure from Gan et al. 2012*.

1.3.6.2 NS1 and NS2 proteins

Non-structural proteins 1 and 2 are the 3'-most genes of RSV and are unique to the *Pneumovirus* genus. They are not essential for virus growth, but their deletion results in attenuation as their role is in immune evasion; NS1 appears to have a variety of effects on the host immune system, many of which are augmented by NS2. In transfected cells, NS1 and NS2 form both homo- and hetero-mers of unknown stoichiometry, and these oligomers appear to change the localisation of the proteins: NS1 alone is found diffusely in the cytoplasm and in nuclear puncta, but when co-expressed with NS2 it relocates to the mitochondria (Swedan et al., 2011).

NS1 and NS2 are known to inhibit interferon-α/β: NS2 prevents the nuclear translocation of IRF3, a transcription factor for interferon expression; NS1 supresses IFN-induced signalling by targeting components of the signal transduction pathway for proteasomal degradation, and has been suggested to act as a ubiquitin E3 ligase (Collins and Melero, 2011). A similar interferon antagonism is assigned to the C protein of the subfamily *Paramyxoviridae*, such as for SeV. A role for NS1 in suppressing the CD8+ cytotoxic T-cell response, whilst simultaneously promoting the proliferation of IL-4-producing Th2-type CD4+ T-cells (creating a positive feedback amplification of the Th2 response), confirms that NS1 promotes the immune-mediated pathogenesis that is characteristic of RSV infection (Munir et al., 2011).

Interestingly, expression of NS1 has an inhibitory effect on transcription and replication of the minigenome system (Atreya et al., 1998). How and why NS1 might mediate this role is unknown, although an NS1:P interaction has been suggested by yeast-two-hybrid experiments.

1.4 M2-1 PROTEIN

1.4.1 Discovery

The protein product of the 5’ ORF of the RSV M2 gene was first predicted to encode a small, basic polypeptide of 22.15 kDa. Early virion dissociation experiments found it co-purified in the envelope fraction with the matrix protein and so was named ‘envelope-associated 22K protein’, later the ‘matrix-like protein 1’ or M2-1 (Collins and Wertz, 1985), and proposed to be a structural membrane protein. Initially, characterisation of M2-1 proved difficult due to its heterogeneity when analysed by reducing and non-reducing SDS-PAGE; at least five species between 14 and 24 kDa can be visualised in lysates of RSV-infected cells using specific monoclonal M2-1 antibodies, and their electrophoretic mobilities vary depending on the conditions of electrophoresis and the cell type (Routledge et al., 1987). This was suggested to be caused by incomplete reduction of intra-molecular disulphide bridges between the 4 cysteines in M2-1.

Sequence alignments show M2-1 has 92 % amino acid identity between RSV subgroups A and B (Collins et al., 1990), making it one of RSV's more highly conserved proteins, but only 43 % and 36 % identity with M2-1 from the pneumovirus MPV and the metapneumovirus HMPV, respectively.

The first indications that M2-1 was actually involved in viral RNA synthesis, as opposed to being a structural membrane protein, came from early minigenome experiments. The established reverse genetics system for RSV, termed the minigenome system, recapitulates viral transcription and replication of a bi- or tri-cistronic subgenomic replicon: 3' leader and 5' trailer regions flank two/three transcriptional units (viral proteins or reporters such as luciferase) separated by authentic viral gene junctions, containing the *cis*-acting gene end and gene start signals. This minigenome is cloned in antigenome sense into a cDNA under the control of T7 promoter, meaning that primary T7 RNA polymerase transcription results in genome sense RSV minigenome synthesis. Transfection of this construct into mammalian cells expressing T7 RNA polymerase, along with support plasmids coding for the viral proteins required for replication and transcription, results in amplification of the minigenome and expression of its encoded proteins (Yu et al., 1995). Supplying the full-length RSV genome allows rescue of infectious virus by this method.

Support plasmids expressing L, N and P proteins were found to be necessary and sufficient to support replication, but analysis of the RNA products of transcription revealed a smear of prematurely terminated species. Single, discrete products of transcription could only be achieved by superinfection with RSV, implying that another viral protein was necessary for processive transcription (Grosfeld et al., 1995). Then it was found that rescue of viable RSV virions from cDNA co-transfection required M2-1 *in trans* (Collins et al., 1995). Omission of M2-1 from the minigenome system did not, however, reduce the number of transcription initiation events, as determined by northern blot against the 5' mRNA (Collins et al., 1996). Consequently, M2-1 protein was termed a transcriptional elongation or antitermination factor required for complete and efficient viral mRNA transcription, but not for replication or transcription initiation.

1.4.2 Transcription antitermination

M2-1 promotes processive transcription through two distinct manifestations of its antitermination ability: prevention of premature intragenic termination, and the read-through of gene end signals.

The polymerisation of nucleic acid is probabilistic by nature, and both eukaryotic and prokaryotic polymerases have evolved mechanisms to increase their processivity, thus reducing the probability of premature polymerase 'fall-off': DNA-dependent RNA polymerase II – responsible for synthesis of all mRNAs in eukaryotes – employs a complex array of co-factors to enable it to transition to its stable elongation phase, and an equally complex set of proteins to regulate its termination activities (Kuehner et al., 2011); *E. coli* RNA polymerase combats premature termination of its ribosomal RNAs by binding a number of antitermination factors (Greenblatt et al., 1993). Intragenic antitermination by M2-1 is RSV's mechanism for improving transcriptional processivity of the viral polymerase L, and is particularly important for longer genes, such as the L gene itself.

The second aspect of M2-1 antitermination is the read-through of gene end sequences. By preventing termination at a gene end, the polymerase complex continues transcribing the intergenic region and the downstream gene, synthesising 'read-through mRNAs' (Hardy and Wertz, 1998). These products of gene end antitermination have been identified in RSV infected cells and it is estimated that read-through transcription occurs 10 % of the time during natural infection. This is a significantly higher abundance compared to other paramyxoviruses, and allowed them to be used to map the RSV gene order (Collins and Wertz, 1983), but whether these mRNA products have a function during infection is unknown.

The process of reinitiating transcription after scanning across an intergenic region is inefficient: For VSV, only ~70 % of viral polymerases reinitiate, leading to a gradient of mRNA abundance (Barr et al., 2002). Therefore, it is likely that read-through mRNAs created by M2-1 antitermination at gene end sequences serve to increase access to the 5' genome by the polymerase, alleviating the transcriptional gradient. Similar read-through mRNAs are found in the vesiculovirus VSV, despite this virus not encoding M2-1 or a homologue (Masters and Samuel, 1984), suggesting that read-through of gene end sequences is intrinsic to the paramyxovirus polymerase and that RSV M2-1 acts to promote and regulate its occurrence. For the subfamily *Paramyxovirinae*, transcriptional regulation is primarily by taking advantage of the polar attenuation inherent in sequential transcription; gene order has evolved to place genes required in highest abundance at the 5' end of the genome.

How M2-1 prevents intra- or inter-genic termination is not known, but it is assumed that M2-1 somehow allows the polymerase complex (L-P) to bypass destabilising *cis*-acting signals within the genome (Fearns and Collins, 1999b). One possible theory suggests M2-1 binds to L and directly modifies its processivity, however given that M2-1 is dispensable for replication, during which the polymerase must be processive enough to synthesise the entire genome, and that no M2-1:L interaction has been found, this seems unlikely. Further to this, supplying increasing levels of M2-1 *in trans* in the minigenome system did not affect the levels of genome/antigenome vs. mRNA accumulation, confirming that M2-1 does not regulate the role of the polymerase between replication and transcription (Hardy and Wertz, 1998, Collins et al., 1996, Fearns and Collins, 1999b). A role for temporal regulation of gene expression has also been discounted for M2-1, as the relative ratios of viral mRNAs did not change during the course of RSV infection in cell culture, despite an increase in M2-1 expression throughout infection (Fearns and Collins, 1999b)

Different gene junctions appear to have different sensitivities towards the antitermination effects of M2-1 (Hardy et al., 1999), and this may be determined by the level of deviation from the gene end consensus sequence (Sutherland et al., 2001). Mutation of gene end sequences to be less 'gene end-like', i.e. decreasing the length of the polyU tract to 3 Us, had little effect on the intrinsic level of termination in the absence of M2-1, but caused a significant decrease in termination in the presence of M2-1. This suggests that termination efficiency is regulated not by polymerase recognition of gene ends, but by M2-1 recognition.

Events at the leader-NS1 and NS1-NS2 gene junctions, the first two junctions in the RSV genome, appear to be different from other gene junctions; not only are levels of read-through up to 3 times higher in natural RSV infection at these junctions, but processive transcription appears to be M2-1 independent, as significant synthesis of NS1 and NS2 mRNAs has been observed in the absence of M2-1 (Fearns and Collins, 1999b). This had led to the suggestion that this may represent a low-grade transcription program that allows the virus to persist, particularly in light of the immune evasion functions of NS1 and NS2. However, it is not known if transcription at these junctions is truly M2-1 independent in the context of viral infection and there is no evidence of persistent RSV infection in humans.

1.4.3 Binding partners

1.4.3.1 Proteins

During infection, M2-1 localises to the cytoplasmic inclusion bodies with P, N and L proteins, where viral transcription and replication are thought to occur (Carromeu et al., 2007, Garcia et al., 1993). A specific M2-1:P interaction is necessary for M2-1-dependent transcription (Mason et al., 2003), and has recently been described as a high-affinity interaction between tetrameric P and tetrameric M2-1, binding in 1:1 stoichiometry (Esperante et al., 2012). It is thought that binding to P may determine the incorporation of M2-1 in the polymerase complex. The interaction is ablated by phosphorylation of P protein at threonine 108, consequently abolishing M2-1 transcriptional antitermination (Asenjo et al., 2006). The residues required for this interaction have not been precisely determined for either protein, although a number of M2-1 residues have been proposed by observing NMR chemical shifts when monomeric M2-1_{58-177} and tetrameric P were mixed in solution (Blondot et al., 2012).

Binding of M2-1 to the M protein, determined in a cell-free binding assay, has been suggested to mediate the link between viral RNPs and the matrix during virion assembly (Li et al., 2008a), although this interaction could not be detected by coimmunoprecipitation from transfected cells.

A previously suggested M2-1:N interaction was found to be mediated by RNA, due to its sensitivity to disruption by RNase A (Cartee and Wertz, 2001).

1.4.3.2 RNA

M2-1 has been observed to bind RNA *in vitro* in gel-retardation assays (Cuesta et al., 2000), and by coimmunoprecipitation of M2-1 and RNA in RSV infected cells (Cartee and Wertz, 2001). The specificity of this RNA binding is still debated: M2-1 has been proposed to bind to long RNAs (>700 nts) with no sequence specificity and short RNAs (<80 nts) containing the antigenomic leader sequence (Cuesta et al., 2000), RSV-specific mRNAs (Cartee and Wertz, 2001), and A-rich RNAs (Tran et al., 2009). The role of RNA binding in M2-1 transcription antitermination is not clear, but it may be that binding of M2-1 to viral RNA (genome, antigenome or mRNA) may signal the polymerase to ignore *cis*-acting signals.

The region of M2-1 responsible for RNA binding was first assigned to residues 59-153, by observing the ability of truncation mutants to pull-down RNA in cell-free assays (Tran et al., 2009). This assignment also led to the suggestion that RNA and P binding by M2-1 are mutually exclusive as they involve the same regions - a fact which was later confirmed when residues involved in RNA and P binding were more precisely mapped by NMR chemical shift perturbations, using a monomeric truncated M2-1_{58-177} (Blondot et al., 2012). This study suggested that RNA binding was indeed mediated by M2-1 residues between 59 and 153, particularly arginines and lysines located within a region of positive charge. Interestingly, this is the latest in several studies to conclude that the predicted zinc finger motif at the N-terminus of M2-1 (described below) is not required for RNA binding (Cartee and Wertz, 2001, Tran et al., 2009), despite such motifs frequently mediating the nucleic acid binding ability of a range of eukaryotic proteins, e.g. transcription factors, and even other viral proteins, e.g. HIV-1 nucleocapsid.

1.4.4 Phosphorylation

M2-1 is known to require phosphorylation for efficient antitermination function, as phosphoablatant serine to alanine mutations are deleterious to transcription in the minigenome system (Cartee and Wertz, 2001). However, this loss of antitermination function is not due to prevention of RNA binding. Equally, M2-1 phosphorylation does not appear to regulate the interaction with P protein (Mason et al., 2003), and so the role of phosphorylation is not known. M2-1 purified from transfected cells, or from RSV infection, is resolved by SDS-PAGE as two electrophoretically distinct species – the slower of which is doubly phosphorylated (Hardy and Wertz, 2000). The sites of phosphorylation have separately been identified as Thr56 and Ser58 (Cuesta et al., 2000), and Ser58 and Ser61 (Cartee and Wertz, 2001). The mechanism of M2-1 phosphorylation is also unknown, although it is dependent on the predicted zinc finger and can be achieved *in vitro* by casein kinase I (Cartee and Wertz, 2001).

1.4.5 Structure

At the beginning of this project, very little was known about the structure of RSV M2-1. It is a 194 amino acid (22.15 kDa) basic protein that forms stable homo-tetramers in solution, determined by size exclusion chromatography, dynamic light scattering and analytical ultracentrifugation (Tran et al., 2009, Esperante et al., 2011). These analyses also suggest that M2-1 has an extended globular conformation, indicated by its larger than anticipated hydrodynamic volume, and that M2-1 is predominantly alpha-helical by circular dichroism spectroscopy. The M2-1 tetramer was found to be strongly pH-dependent, dissociating into partially folded monomeric intermediates at pH 5, with no significant dimeric or trimeric unfolding intermediates seen (Esperante et al., 2011).

The N-terminus of all M2-1 proteins of the subfamily *Pneumovirinae* contains four highly conserved residues, 3 cysteines and a histidine, that have been predicted to form a zinc finger motif. This motif is essential for transcription antitermination by M2-1 (Esperante et al., 2011, Cartee and Wertz, 2001, Tran et al., 2009) and it is thought that co-ordination of 1 atom of zinc per monomer is required for correct folding and stability of M2-1 (Esperante et al., 2011). The putative Cys_3-His_1 motif has been compared to that of the mammalian protein Nup475 (also called tristetraprolin or ZFP36), which regulates mRNA turnover by binding AU-rich regions of mRNA via its tandem Cys_3-His_1 zinc finger motifs (discussed in more detail in Chapter 4).

The combination of past loss-of-function mutants and truncation analyses has resulted in the assignment of 4 putative M2-1 domains: a predicted Cys_3-His_1 zinc finger (residues 7-25); an oligomerisation domain between amino acids 33 and 62; a globular α-helical domain capable of binding to P and RNA (residues ~58-177); and a predicted unstructured C-terminus containing the last ~20 residues (Tran et al., 2009, Cuesta et al., 2000).

Towards the end of this project, a solution NMR structure of M2-1 residues 58-177 was published, corresponding to the predicted globular α-helical domain (Blondot et al., 2012, Dubosclard et al., 2011). It revealed a single globular α-helical fold with 6 helices and extensive positive surface charge that could mediate RNA or P binding (figure 1.9).

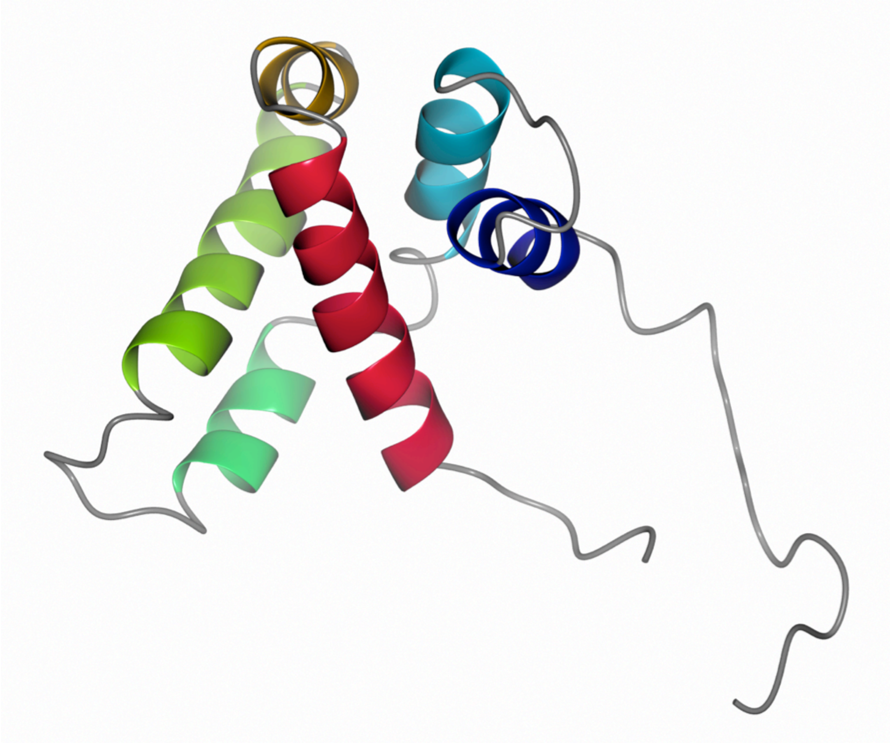


Figure 1.9 NMR structure of RSV M2-1$_{58\text{-}177}$.

A ribbon representation of the average of the 20 lowest energy models. Residues 58-177 of M2-1 formed a monomeric globular alpha-helical domain, coloured blue through to red (N to C terminus).

1.4.6 Related proteins

1.4.6.1 Other M2-1 proteins

The M2 gene is unique to the genus *Pneumovirus* amongst the family *Paramyxoviridae.* However, there is evidence to suggest that the M2-1 protein of pneumovirus murine pneumonia virus (MPV, or pneumonia virus of mice) does not perform the same function in its cognate virus life cycle as RSV M2-1. MPV M2-1 is expressed as the first of two overlapping ORFs in the M2 mRNA, as with RSV, shares ~40 % amino acid identity, and is functionally similar at low concentrations. At high concentrations, however, MPV M2-1 has an inhibitory effect on viral replication – not seen with RSV M2-1 or the M2-1 protein of metapneumovirus AMPV, which has RSV-like function (Dibben et al., 2008, Naylor et al., 2004). There has been no biophysical analysis of either of these proteins, however.

1.4.6.2 Ebola VP30

Ebola virus, a filovirus of the order *Mononegavirales*, encodes a protein called VP30, described as a minor nucleoprotein as it forms part of the viral RNP complex. VP30 appears to share many of the same properties are RSV M2-1: it is predominantly basic, with clusters of arginine and lysine residues suggested to bind RNA; it has a highly conserved Cys_3-His_1 zinc finger motif; it is phosphorylated and this plays a role in regulating VP30 function; and it functions as a transcriptional regulator that is not required for replication (Biedenkopf et al., 2013). There are some differences, however, as VP30 forms hexamers in solution, binds the L protein directly, and is suggested to be involved in transcription at the level of initiation, a role that has been discounted for RSV M2-1.

The two proteins have only 9 % sequence conservation, but an X-ray crystal structure of the VP30 C-terminal domain (Hartlieb et al., 2007) revealed similarities to the helical arrangement of the RSV M2-$1_{58\text{-}177}$ core domain, with a route-mean-square deviation (rmsd) of 3.9 Å for the carbon alphas of 92 aligned residues (Blondot et al., 2012). This has led to the suggestion that the reflection of functional similarity in structural identity, may result in a common mechanism of transcriptional regulation for RSV M2-1 and EBOV VP30.

1.5 MACROMOLECULAR X-RAY CRYSTALLOGRAPHY

1.5.1 Introduction and comparable techniques

High resolution X-ray crystallography is a technique for probing macromolecular structures, and accounts for ~90 % of all structures deposited in the Protein Data Bank (PDB). It uses the diffraction of X-rays by electrons within a crystal to calculate an electron density map, into which an atomic model of the molecule can be built. There are a number of difficulties with X-ray crystallography, some of which are practical (e.g. crystal growth and manipulation) and others are fundamental (e.g. the 'phase problem' described in Section 1.5.4), however the ability to achieve resolutions that are theoretically only limited by X-ray wavelength and crystal quality (the current record for a protein structure is 0.48 Å for crambin, a small plant antimicrobial, Schmidt et al., 2011) means it is currently the preferred choice for high resolution structure solution.

There are several other methods for determining protein structure, each of which have their advantages and disadvantages over X-ray crystallography. The most popular alternative is nuclear magnetic resonance spectroscopy (NMR), which uses the fact that certain NMR-active magnetic nuclei (particularly ^{1}H, ^{15}N and ^{13}C) absorb electromagnetic radiation at the resonant frequency determined by the isotope, exciting these nuclei to a high energy state. When these nuclei relax back to their equilibrium state they release radiation, the frequency of which is characteristic of their chemical environment and allows the mapping of interatomic distances (through bonds and through space). This results in an ensemble of possible structures and typically the quality of the 20 most probable structures is assessed by their root mean square deviation (rmsd). NMR has the advantage that proteins are observed in solution and flexible regions that may not be resolved by X-ray crystallography can be sampled in multiple conformations, providing information about the dynamics of protein folding. NMR also requires less protein at lower purity, and samples can be reused for several experiments (Egli, 2010). However there are several disadvantages to using NMR, the most important of which are the size limitation (current technology means solution NMR is restricted to proteins <30 kDa in the majority of cases) and the resolution; NMR structures are not assigned a resolution in the same way X-ray crystal structures are, but a good NMR structure is considered to equate to a 3-3.5 Å X-ray crystal structure i.e. inadequate resolution for structure-guided drug design. Proteins must also be labelled with ^{15}N or ^{13}C, which can result in altered expression and purification requirements, and they must be stable at room temperature in buffers that are typically below physiological pH (5-6). In the case of RSV M2-1, the large size of the physiologically relevant tetramer (90 kDa), its poor stability at room temperature, and its characterised unfolding at acidic pH (Esperante et al., 2011), meant X-ray crystallography was the obvious choice over solution NMR spectroscopy.

Single-particle reconstruction from cryo-electron microscopy (EM) images can provide low resolution information about the general shape of molecules (e.g. VSV polymerase, Rahmeh et al., 2010) and tomography allows 3D reconstruction by combining a series of 2D images at a wide range of viewing angles. However, EM analysis alone is limited to molecules >150-300 kDa and achieves resolutions lower than 10 Å; the power of EM now lies in its complementarity with X-ray crystallography, as crystal structures of subunits can be modelled into EM densities (e.g. the rotavirus double-layer particle, Zhang et al., 2008), and its ability to simultaneously sample multiple protein conformations within a population of molecules.

New techniques for macromolecular structure determination are constantly being developed. One option that is already beginning to contribute structural data to the PDB is neutron diffraction. In contrast to X-ray crystallography, this makes use of neutrons diffracted from protons within a crystal, instead of electrons. This has the advantage of causing considerably less radiation damage to the crystal and allows hydrogens to be visualised, providing information about protonation states and hydrogen bonding. However, this technique is currently not widely available due to the lack of neutron sources and the often-prohibitive experimental requirements (large crystals, 1-7 day data collection compared to minutes with synchrotron X–rays). X-ray free-electron lasers (X-FEL) have been described as the future of X-ray crystallography. Able to produce ultrashort intensely brilliant X-ray pulses (10x brightness of synchrotron radiation), it is thought they could allow the collection of diffraction data from single crystals before the onset of radiation damage - 'diffraction-before-destruction' – a considerable problem with current X-ray methodologies (Schlichting and Miao, 2012).

1.5.2 Crystallisation

1.5.2.1 Crystal growth

Once a high-yield (mg/mL) source of homogeneous, >95 % pure, soluble protein has been optimised, an X-ray crystallography experiment begins with crystallisation. The requirement for well-ordered, high quality (i.e. well-diffracting) protein crystals is the remaining bottleneck to structure solution by X-ray crystallography. Crystals are necessary because the intensity of diffracted X-rays from a single protein molecule would be very weak; multiple copies of a protein arranged in a periodic, ordered lattice allows amplification of this signal. To grow protein crystals, an undersaturated solution of protein and precipitant must be concentrated to supersaturation, beyond the concentration where the protein is soluble in the defined mixture of other solutes (the solubility line), but not allowed to reach the point of precipitation (disordered aggregation). The concentration of precipitant changes the position of the solubility line at a given protein concentration. Figure 1.10 shows a solubility phase diagram, illustrating the relationship between protein and precipitant concentrations – which varies depending on the specific protein. Spontaneous nucleation (ordered aggregation) of crystals occurs in the labile zone, which reduces the protein concentration and moves the phase back into the metastable zone, where sustained crystal growth can occur. Crystals will continue to grow in the metastable zone until either the protein is depleted, or until impurities are incorporated into the lattice at the surfaces of the growing crystal.

The most common method of growing protein crystals is vapour-diffusion: The evaporation of water from a protein:crystallant drop as it approaches equilibrium with a reservoir of crystallant (in a closed system), concentrates both protein and crystallant to the point where nucleation and crystal growth can occur. Crystallant solutions (or 'mother liquor') typically contain combinations of precipitants (organic polyalcohols e.g. polyethylene glycol (PEG)), salts (e.g. ammonium sulphate or sodium chloride), additives (e.g. detergents, metal ions, reducing agents etc.) and buffers. Protein:crystallant drops are either sat on platforms over a reservoir (sitting drop) or suspended below a coverslip over a reservoir (hanging drop) and sealed.

Sufficiently large protein crystals can take hours or weeks to grow, but on average it takes 3-7 days. Manipulation of these crystals is with small loops, in which the crystal is suspended in its mother liquor. To reduce the radiation damage induced by X-rays, crystals are cryo-cooled in liquid nitrogen. This can be done directly from drops if sufficient cryo-protectant is present in the mother liquor, or by first placing crystals in cryo-protectant solution such as dimethyl sulfoxide (DMSO), low molecular weight PEG, or glycerol. The cryo-protectant allows water in and around the crystal to cool into vitreous ice (an amorphous glass) rather than freezing as ice crystals, which can disrupt the crystal lattice and cause high background diffraction.

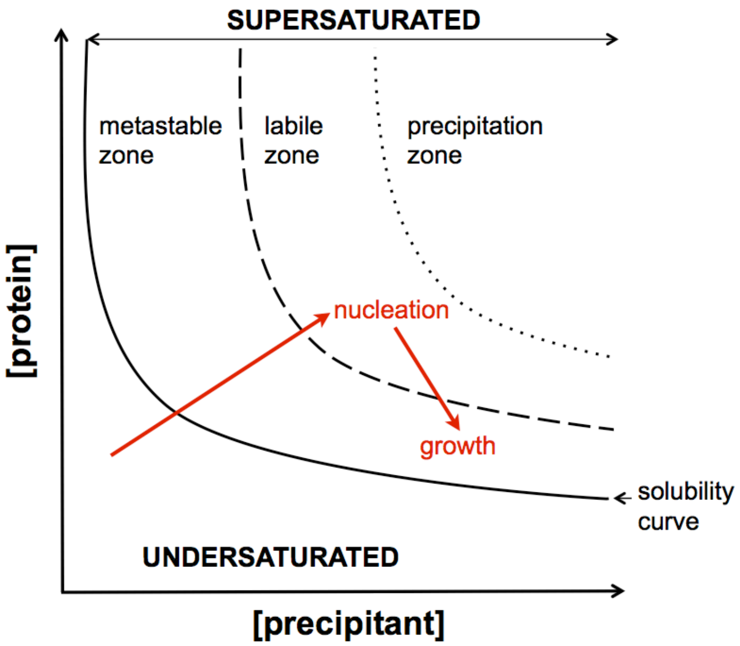


Figure 1.10 A solubility phase diagram.

As concentrations of free protein and precipitant increase, the phase moves from undersaturated to supersaturated. The supersaturated labile zone, where spontaneous nucleation occurs, is followed by the metastable zone, which supports sustained crystal growth. The precipitation zone results in disordered aggregation and no crystal growth.

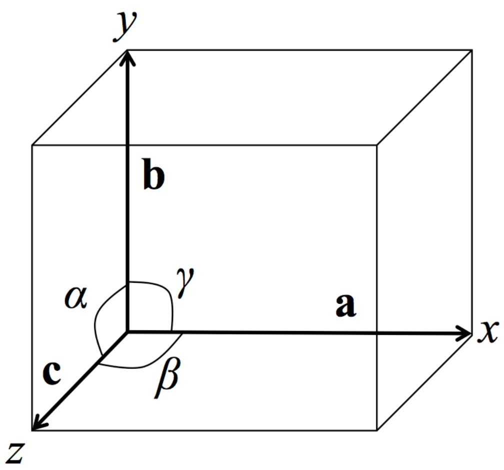


Figure 1.11 The unit cell.

The minimum unit that describes a crystal lattice only in terms of translational symmetry, and is annotated by vectors **a**, **b** and **c** along axes *x, y, z,* and angles α, β and γ.

1.5.2.2 Crystal lattices and space groups

Protein crystals are made of periodically arranged, identical repeating units that create a weakly-interacting 3D lattice of protein and solvent molecules. The minimum unit of volume that can describe the crystal lattice only in terms of translational symmetry, i.e. the repeating unit of the crystal, is called the unit cell and is described in figure 1.11. The unit cell is further divided into asymmetric units (AU), which define the smallest unit of volume that can describe the unit cell in terms of translational and rotational symmetry. In terms of a protein crystal, the asymmetric unit contains the smallest object that can be repeated to make the entire crystal and the unit cell describes how multiple asymmetric units are symmetrically related so that they form a building block that can be simply 'stacked' to describe the entire crystal. Asymmetric units often contain oligomers of proteins and the relationship between these molecules is called non-crystallographic symmetry (NCS), as the symmetry axis does not extend throughout the whole unit cell. The relationship between the asymmetric units and the unit cell is determined by the space group of the crystal, described below.

The chirality of proteins means only translation, rotation, and screw axes can describe their symmetry, not mirror planes or inversions. There are 14 possible crystal lattices (Bravais lattices) that describe the crystal system (the relationship between axis lengths and angles in the unit cell) and their centering (whether they have lattice points at their vertices, faces, or internally). Combined with the 11 chiral Laue groups that describe 1-3 intersecting perpendicular axes of symmetry within the unit cell, and any possible screw axes generated by combined translation and rotation, this results in 65 possible space groups for protein crystals. Consequently, the space group of a protein crystal defines all the symmetry operations needed to describe the crystal in terms of the asymmetric unit. Knowing the correct space group for a crystal is essential to correctly indexing the diffraction data (see Chapter 4).

1.5.3 Crystal diffraction

X-ray diffraction data is collected by placing a protein crystal in the path of a beam of monochromatic (of equal wavelength) X-rays, either from a laboratory source (a rotating anode, usually copper) or from a synchrotron (a high-brilliance, tuneable source). X-rays are used because their wavelength is in the same order of magnitude as the interatomic distances we are trying to resolve (1 Å = 0.1 nm). A photon from the incident X-rays interacts with the electron density around atoms and is scattered by each electron it encounters, resulting in emerging partial X-ray waves with different scattering angles and phases. The way in which these waves combine either constructively or destructively has an additive or subtractive affect, respectively, on the amplitude of the resultant X-ray wave and results in recording a pattern of discrete intensities on the detector in front of the crystal; these are known as reflections and they form the diffraction pattern – the raw data of an X-ray diffraction experiment.

The relationship defining when a particular set of waves will combine constructively to produce a reflection is described by Bragg's Law. Imagine a set of parallel planes that intersect each unit cell of a crystal at a fixed point along their axes: if these planes are a distance (d) apart, with an angle of incidence (θ) for two incoming X-ray waves of wavelength λ, then the two waves will only remain in phase after reflecting from these planes if the additional distance the second wave must travel ($d \sin\theta$) is a multiple (n) of the wavelength:

$$n\lambda = 2d \sin\theta$$

This is also depicted in figure 1.12.

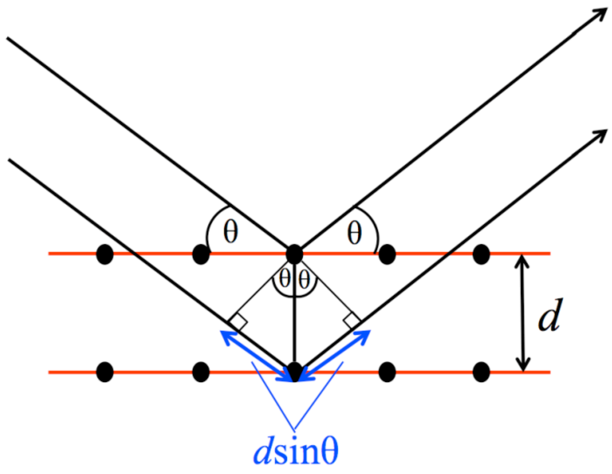


Figure 1.12 Bragg's Law.

This defines the diffraction angle (θ) and lattice spacing (*d*) required for scattered X-rays to remain in phase with each other. The resultant X-ray is generated by constructive interference and results in a diffraction spot, or reflection.

This relationship also explains why the highest resolution (smallest *d*) data is observed at the largest scattering angle (θ) i.e. at the edge of the detector. Each unit cell contributes equivalent constructive interference as the Bragg planes intersect every unit cell in the same way, creating a resultant wave that has contributions from each diffracting electron in every unit cell throughout the whole crystal. These planes are assigned Miller indices (*hkl*) according to where they intersect the unit cell, and matching indices are given to the corresponding reflections during indexing of diffraction data (see Chapter 4). The summation of all contributing waves across the whole crystal for a particular reflection is called a structure factor, F_{hkl}, and the structure factor amplitude ($|F_{hkl}|$) is proportional to the square root of the reflection intensity. A single diffraction pattern provides information about the lattice from one direction only, and so the crystal must be rotated to observe the lattice from all angles; at each angle, a different set of planes will satisfy Bragg's law and a different set of resultant waves will be diffracted.

However, not all scattering electrons lie directly on these planes – many are in between them at fractional values of *d*. These electrons will result in diffracted waves that are out of phase, causing a degree of destructive interference that will result in a decrease in the amplitude of the emerging X-ray and, consequently, the intensity of the reflection. In this way, the intensity of a reflection provides information about the arrangement of atoms in a unit cell relative to the Bragg planes. Equally, the symmetry of reflections in a diffraction pattern provides information about how asymmetric units are related to each other i.e. the space group. Together we have all the information required to recreate the structure of our crystal, except for one vital piece: To translate the information encoded in the diffraction pattern into an electron density map, a Fourier transformation is required. A Fourier transformation is a way of describing a wave as a sum of its component waves, but the relative phases of these waves are required to ensure they are added correctly. The way diffraction data is currently recorded, i.e. photon counts on a detector, results in the loss of this phase information and causes the fundamental 'phase problem' that must be solved in any X-ray diffraction experiment.

1.5.4 Solving the phase problem

There is no formal relationship between phases and the amplitudes recorded in the X-ray diffraction experiment – except through the molecular structure, which is unknown. Recovering the lost phase information can be achieved in one of two ways: experimental phasing, where the phases for a small substructure of atoms are calculated, then used to estimate phases for the whole molecule; or molecular replacement, where the phases from a previously known similar structure are used as a starting estimate for the unknown structure. Both methods were used in this work, however the primary method used was experimental phasing and so this will be described here. The subsequent molecular replacement is described in brief in Chapter 4.

In the absence of a structurally similar model, as was the case for RSV M2-1, phases must be derived experimentally. The method chosen for this thesis was single-wavelength anomalous dispersion (SAD), which can make use of the natural coordination of a heavy atom (e.g. zinc in the predicted zinc finger motif of M2-1), or the incorporation of selenomethionine into the protein. Under normal circumstances, reflections in diffraction patterns come in pairs that have equal intensities but opposite phases, because of the symmetry of the Laue group. The waves responsible for these reflections can be represented as vectors of length F_{PA} (equivalent to the structure factor amplitude) and phase φ, which have an anomalous component (F_{Anom}) and a non-anomalous component (F_{Prot}) on an Argand diagram (figure 1.13 A) – these diagrams allow X-ray waves to be represented as 2-dimensional vectors and to describe how component waves can be combined to create a resultant wave by simple vector addition. When heavy atoms (cadmium and zinc in this instance) are excited with X-rays at their resonant wavelength – at the absorption edge of the element – the phase of the anomalous component for these atoms changes, resulting in a change in the vector length F_{PA} and, consequently, a change in the intensity of the reflection. This breaks the symmetry of these pairs and is known as the anomalous difference.

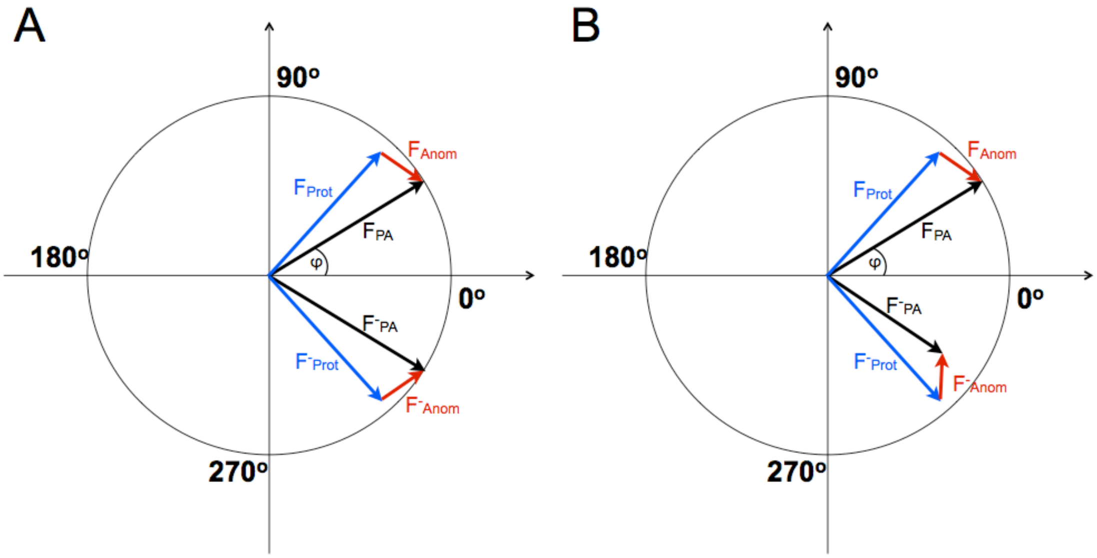


Figure 1.13 Argand diagrams of reflection pairs.

Vector representation of diffracted X-ray waves, with length F_{PA} (equivalent to structure factor amplitude) and phase φ, and comprised of a non-anomalous component (F_{Prot}) and an anomalous component (F_{Anom}). The paired reflection is denoted by a minus sign. **A)** In the absence of anomalous scatterers, vectors F_{PA} and F^{-}_{PA} are of equal length i.e. reflections have equal intensities. **B)** At the absorption edge, the phase of the anomalous component F^{-}_{Anom} is shifted, and vector F^{-}_{PA} has a different length, and thus the reflection has a different intensity.

The anomalous difference between intensities of certain pairs of reflections can be expressed in terms of a Patterson map; the same Fourier summation that is used to calculate electron density is used (below) but using only measured intensities not phases, resulting in a map of interatomic vectors between heavy atoms. This results in large peaks corresponding to the heavy atoms that can be assigned coordinates. Once the positions of the heavy atom substructure are known, i.e. the location of their electron density, their phases are calculated (φ_A) as their location, structure factor amplitude and phase are related through the electron-density equation:

$$\rho(x,y,z) = \frac{1}{V}\sum_{hkl} F_{hkl} e^{-2\pi i(hx+ky+lz)}$$

where $\rho(x,y,z)$ is the electron density at a point x,y,z, V is the volume of the repeating unit of the crystal lattice, F_{hkl} is the structure factor of the scattered wave summed over all reflections (contains amplitude and phase of resultant X-ray wave), and the final term is an operator of the Fourier summation that relates real space (electron density) to reciprocal space (the diffraction pattern). The heavy atom phases, φ_A, are then used to estimate the initial phases for the rest of the protein structure, φ_P, using the following relationship:

$$\varphi_P = \varphi_A + \alpha$$

where α is the shift in phase from heavy atom phase to protein phase, and is related to the anomalous difference and heavy atom structure factor magnitude (Dauter et al., 2002). However, Patterson maps are symmetrical, and the possible substructure solutions do not distinguish between enantiomorphic structures, i.e. you cannot tell which solution provides the correct protein 'handedness'. This ambiguity is solved during the process of density modification, described in Chapter 4.2.6, as only one enantiomorph will result in electron density that is 'protein-like' in its distribution. Once initial phase estimates for the whole protein have been calculated, the process of model building and refinement begins, as described in Chapter 4.2.7.

1.6 PROJECT AIMS

The essential nature of M2-1 protein to the RSV life cycle warrants continued investigation of its role. A lack of biophysical or structural information is a significant barrier to the understanding of how M2-1 functions, and the primary aim of this project was to address this by determining the X-ray crystal structure of RSV M2-1 protein. This would allow us to infer mechanisms of action through structural homology with domains of known function, and to interpret the extensive mutagenesis of M2-1 performed in the literature.

A key function of M2-1 protein is RNA binding, but whether this is a specific interaction, what sequence or secondary structure this specificity consists of, and what the purpose of an M2-1:RNA interaction might be, are all unknown. This project aimed to assess the RNA binding ability of purified RSV M2-1 protein, both in terms of sequence specificity and residues of the protein responsible for the interaction, and relate this function to the crystal structure. Development of an *in vitro* binding assay using fluorescence polarisation anisotropy (FPA) would allow high-throughput analysis.

M2-1 protein is phosphorylated during viral infection, although the implications of this post-translational modification are not understood. Phosphorylation sites have been reported differently by two groups, and clarification of the residues involved was an early target for this project. We aimed to express and purify physiologically-relevant protein for analysis by mass spectrometry. The knowledge of the correct phosphorylation sites of M2-1 would direct X-ray crystallographic analysis of a phosphomimetic mutant, to determine if the result of phosphorylation is conformational change. We also aimed to assess the consequences of phosphorylation on RNA binding using the FPA assay.

To relate our *in vitro* functional analysis to the viral life cycle, we aimed to complement our findings on RNA binding and phosphorylation with an assessment of their roles in M2-1 function in a minigenome system: M2-1-

dependent transcription is reconstituted in mammalian tissue culture and the phenotypic impact of M2-1 mutagenesis observed. This would allow us to further interpret our structural findings in terms of the contribution of certain residues to M2-1 activity.

In summary, the aims of this project were to combine structural studies with functional analysis to create a model for M2-1 function in the RSV life cycle, thus allowing structure-guided drug design to inhibit M2-1 action – a potential therapeutic avenue for the treatment of RSV infection.

2 CHAPTER 2: MATERIALS AND METHODS

2.1 MATERIALS

2.1.1 Vectors

A cDNA of the RSV A2 strain M2-1 open reading frame (ORF) in pGEM3 was a kind gift from Gail Wertz and was described previously (Hardy and Wertz, 1998). See Appendix III for sequence. This ORF was subcloned into a range of vectors:

pTriEx1.1 Neo (Novagen) - contains promoter sequences for mammalian cell expression, mediated by a hybrid promoter composed of the cytomegalovirus (CMV) immediate early enhancer fused to the chicken β-actin promoter, and baculovirus-infected insect cell expression, as it contains both the sequences required to generate recombinant baculovirus and the very late p10 promoter for expression of the ORF of interest. Results in the addition of a C-terminal 8x histidine tag (His-tag). This vector also has T7 RNA promoter sequences for IPTG-induced bacterial expression

pETSUMO 28a – a derivative of pET28a (Novagen) that creates an N-terminal 6x histidine-SUMO protein tag, expressed from the *SMT3* gene (Life Technologies).

pGEX6P2 (GE Healthcare) – glutathione S-transferase (GST) fusion vectors used to N-terminally tag the ORF of interest with GST and PreScission protease (GST-tagged human rhinovirus 3C protease, GE Healthcare) cleavage site.

All RSV nucleotide sequences were confirmed by DNA sequencing and compared to RSV A2 strain (accession number: M74568).

2.1.2 Bacterial strains

Amplification of plasmid DNA constructs or expression of recombinant protein was by transformation of competent *Escherichia coli* (*E.coli*) cells. Strains used were: Subcloning Efficiency DH5α Competent Cells (Life Technologies) for initial transformation of ligation reactions; DH5α, prepared in the laboratory using the standard Inoue method for routine re-transformation; BL21 (DE3) derivatives for isopropyl β-D-1-thiogalactopyranoside (IPTG) inducible protein expression, including Gold, Star, Rosetta, Rosetta II, and OverExpress C41 (Lucigen); B834(DE3) cells, a methionine auxotroph strain for production of selenomethionine protein.

2.2 METHODS

2.2.1 Manipulation of recombinant DNA

2.2.1.1 Polymerase chain reaction (PCR)

PCR was used to amplify the M2-1 ORF whilst incorporating terminal restriction sites for further subcloning. Reactions were performed in 50 μL volumes and contained 10 x Thermopol buffer (New England Biolabs), 0.3 mM of each deoxyribonucleotide triphosphate (dNTPs; Roche), 0.4 μM of each forward and reverse primer, 100 ng of template DNA and 1 unit Vent polymerase (New England Biolabs). Reaction cycles were performed in a thermocycler (Eppendorf) as follows: initial denaturation 95 °C 2 min; 22 cycles of denaturation at 95 °C 30 sec, annealing at 50-65 °C 30 sec (dependent on primer sequence), and elongation at 72 °C for 1 min; final extension at 72 °C for 5 min. PCR products were isolated and purified by agarose gel electrophoresis followed by gel extraction (see below).

2.2.1.2 Site directed mutagenesis

Complementary mutagenic primers were designed manually and consisted of 10-15 nucleotides either side of the encoded mutation, with terminal guanine (G) or cytosine (C) nucleotides where possible, and are listed in Appendix I. Reactions were performed in 50 μL volumes and contained 10 x Pfu Turbo reaction buffer (Agilent), 0.2 mM of each deoxyribonucleotide triphosphate (dNTPs; Roche), 0.2 μM of each forward and reverse primer, 100 ng of template DNA, 5 % dimethyl sulfoxide (DMSO), 5 mM magnesium chloride ($MgCl_2$), and 2.5 unit Pfu Turbo polymerase (Agilent). Reaction cycles were performed in a thermocycler (Eppendorf) as follows: initial denaturation 95 °C 50 sec; 18 cycles of denaturation at 95 °C 50 sec, annealing at 45-47.5 °C 60 sec, and elongation at 72 °C for 60 sec per kb DNA; final extension at 72 °C for 8 min.

Immediately following completion, 1 uL *DpnI* restriction enzyme (New England Biolabs) was added to the reaction and incubated at 37 °C for 1 hour to digest methylated template DNA. Reactions were then transformed into DH5α cells and plated out (see below); a minimum of 3 single colonies were picked and screened for the correct mutation using a Spin Miniprep kit (Qiagen) followed by DNA sequencing.

2.2.1.3 Restriction digest

Restriction enzyme digests were performed on plasmid DNA for cloning and diagnostic purposes. In a 50 μl volume, reactions contained 1 μg DNA, 1x compatible NEBuffer (New England Biolabs), ±1x bovine serum albumin (BSA; enzyme dependent) and 2 units of each enzyme. Reactions were incubated at 37°C for 2-3 hours. Purification of the products of digestion was either by gel extraction after agarose gel electrophoresis (see 2.2.1.4) using QIAquick Gel Extraction Kit (Qiagen), or by removal of short polynucleotides from digestion of linear DNA using a QIAquick Nucleotide Removal Kit (Qiagen). Both were used as per the manufacturer's protocol.

2.2.1.4 Agarose gel electrophoresis

Size and integrity of plasmid DNA, PCR products, restriction digests and ligations was determined using 1% agarose gels: 0.5 g analytical grade agarose, 50 mL 1x TAE buffer [40 mM Tris-acetate, 1 mM EDTA], 1:10,000 SYBR Safe DNA stain (Life Technologies). Samples were mixed with 1x GelPilot DNA Loading Dye (Qiagen) and run on gels at 100V for 40-60 minutes in 1x TAE buffer, alongside molecular marker Hyperladder I (Bioline). DNA was visualised by UV transillumination.

2.2.1.5 Ligations

Ligations were performed in 20 μL volumes containing 1x Ligase Reaction Buffer (Life Technologies), 3:1 or 10:1 molar ratio of insert:vector DNA, and 1 unit T4 DNA Ligase (Life Technologies). Reactions were incubated overnight at 16 °C then transformed into DH5α cells.

2.2.1.6 Bacterial transformation

Transformations were performed as per the manufacturers' guidelines, or as follows for laboratory-made cells: 50 μL *E. coli* cells were thawed on ice and 1 μL of either 25-50 ng/μL plasmid DNA, ligation mix, or site directed mutagenesis reaction, was added. Samples were incubated on ice for 30 minutes, followed by heat shock at 42°C for 45 sec, then rapid cooling on ice for 2 mins. 450 μL pre-warmed Luria-Bertani (LB) medium [in 1 liter: 10 g tryptone, 10 g NaCl, 5 g yeast extract] was added and samples were incubated at 37 °C for 1 hour with shaking. 100 μL of the transformation was spread onto antibiotic-selective LB agar plates and incubated at 37 °C overnight.

2.2.1.7 Starter cultures

Single colonies were picked from transformation plates and used to grow starter cultures in LB medium containing appropriate antibiotic (final concentration 50 μg/mL kanamycin, or 100 μg/mL ampicillin) overnight at 37 °C. Small-scale cultures were 5 mL, large scale cultures were 50 mL. These were then used to inoculate overnight cultures for either plasmid DNA amplification or protein expression. Aliquots of starter cultures were mixed 1:1 with 40 % sterile-filtered glycerol and frozen at -80 °C as glycerol stocks.

2.2.1.8 Plasmid DNA amplification

Starter cultures were used to inoculate overnight cultures of appropriate volume in LB medium containing antibiotics, and DNA isolated using Plasmid Maxiprep or Spin Miniprep kits (Qiagen) based on the alkaline-lysis method followed by ethanol precipitation (maxiprep) or adsorption to a silica membrane under high-salt conditions (miniprep). DNA was quantified by spectrophotometry using a NanoDrop 1000 (Thermo Scientific).

2.2.2 Protein expression in *E.coli*

2.2.2.1 Culture growth, IPTG induction and lysis

5 or 50 mL starter cultures were grown from stabs of glycerol stocks of transformed *E. coli* BL21-derivative cells, and used to inoculate 100 mL (small scale) or 1L (large scale) LB overnight cultures with appropriate antibiotic.

Cultures were grown to OD_{600} 0.6-0.8 at 37 °C and induced with 0.1-1 mM IPTG, ±0.05-1 mM zinc sulphate ($ZnS0_4$), for 6-16 hours at 18-37 °C. Cells were recovered by centrifugation at 4,000 xg for 30 mins and either resuspended immediately, or frozen at -20 °C until lysis. Resuspension was in 5 mL lysis buffer per 100 mL pelleted culture. Lysis buffer contained 50 mM Tris-Cl pH 7-8 or 10 mM sodium phosphate dibasic/1.8 mM potassium phosphate monobasic, 0.1-1 M NaCl, 1-2 % (v/v) Triton-X100, 1x complete protease inhibitor cocktail EDTA-free (Roche), 1 mg/mL chicken egg white lysozyme, ±1 mM dithiothreitol (DTT), ±1 mM calcium chloride ($CaCl_2$), ±10 mM magnesium sulphate ($MgS0_4$), ± 10 % (v/v) glycerol, ± 50mM L-arginine and 50 mM L-glutamic acid. Cells were incubated on ice for 30 mins, then lysis was completed by sonication on ice with 8 cycles of 20 sec on, 30 sec off at 10 micron amplitude. Lysates were clarified by centrifugation at 18,000 xg for 30 mins and the supernatant retained for purification by affinity chromatography.

2.2.2.2 Autoinduction

5 mL starter cultures were grown overnight at 37 °C in minimal medium containing the following:
2mM $MgS0_4$
1 μL 1000x trace elements [stock concentrations: 50 mM $FeCl_3$, 20 mM $CaCl_2$, 10 mM $MnCl_2$, 2 mM $CoCl_2$, 2 mM $CuCl_2$, 2 mM $NiCl_2$, 2 mM Na_2MoO_4, 2 mM Na_2SeO_3, 2 mM H_3Bo_3]
0.5 % (w/v) glucose
0.25 % (w/v) aspartate
100 μL 50x M [stock concentrations: 1.25 M Na_2HPO_4, 1.25 M KH_2PO_4, 2.5 M NH_4Cl, 0.25 M Na_2SO_4]
50 mg/mL kanamycin

50 μL of starter culture was used to inoculate 50 mL autoinduction medium growth cultures containing:
47.85 mL ZY medium [5 g/L yeast extract, 10 g/L tryptone]
2 mM $MgSO_4$
10 μL 1000x trace elements (as above)
1 mL 50x 5052 [stock concentrations: 25 % (w/v) glycerol, 2.5 % (w/v) glucose, 10 % (w/v) lactose]
1 mL 50x M (as above)
50 mg/mL kanamycin

Cultures were grown at 37 °C for 24 hours. Cells were recovered and lysed as above (2.2.2.1).

2.2.3 Purification by glutathione affinity chromatography

Clarified supernatant containing soluble GST-M2-1 protein was applied in batch to glutathione sepharose 4B resin (GE Healthcare) at 4 °C overnight with rotation. Resin was pelleted at 700 xg for 5 mins and washed 3 times for 15 mins at 4 °C with 4 volumes of lysis buffer. Additional washes (2x 30 mins) in 10 mM Tris-Cl pH 7.4, 1 M NaCl were performed to remove M2-1-bound RNA from the expression host (Tran et al. 2009). Resin was then washed and equilibrated in an equal volume of protease cleavage buffer [50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM DTT] and M2-1 was cleaved from GST on the resin overnight at 4 °C by HRV-3C protease, which was expressed and purified in-house from glycerol stocks. Cleaved M2-1 was eluted by pelleting the resin at 700 xg for 5 mins and collecting the supernatant. Resin was then washed twice for 10 mins at 4 °C in 1 volume of protease cleavage buffer, pelleted by centrifugation, and the supernatant collected. Supernatants were pooled and M2-1 concentration determined by adding 1-20 μL (depending on concentration range) protein to 1:5 dilution of Protein Assay reagent (Bio-Rad) and reading the absorbance at 595 nm.

2.2.4 Size exclusion chromatography

Size exclusion chromatography was used initially analytically to determine the oligomeric status of M2-1 expressed in this system, and then preparatively to isolate tetrameric M2-1. A 320 mL HiLoad 26/600 Superdex 75 pg column (GE Healthcare) was used with an Akta prime pump system at 4 °C, and absorbance at 280 nM was used to monitor protein elution. All buffers used were 0.2 μm filtered and degassed. The column was equilibrated with running buffer [50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM DTT], and the protein (concentrated to just over 5 mL) was 0.2 μm filtered and injected using a 5 mL loop. The column was run at approximately 2 mL/min and 3 mL fractions collected once the void volume had eluted (~100 mL). The column was calibrated using a gel filtration standard kit (Bio-Rad) containing thyroglobulin 670,000 Da, γ-globulin 158,000 Da, ovalbumin 44,000 Da, myoglobin 17,000 Da, vitamin B_{12} 1,350 Da.

Appropriate fractions were pooled at concentrated to approximately 12 mg/mL using 10,000 Dalton molecular weight cut-off centrifugal concentrators (Sartorius) and stored at 4 °C.

2.2.5 Ion exchange chromatography

A 1 mL Resource S cation exchange column (GE Healthcare) was used at pH 7.5 (below the isoelectric point (pI) of M2-1: 9.1), with an Akta Purifier pump system at 4 °C, and absorbance at 280 nM was used to monitor protein elution. All buffers used were 0.2 μm filtered and degassed. The column was equilibrated in buffer A [20 mM Tris-Cl pH 7.5, 50 mM NaCl] and protein sample diluted into the same buffer. Protein was applied to the column and run until the 280 nM trace returned to baseline. A 20 CV gradient was applied with 0-100% buffer B [20 mM Tris-Cl pH 7.5, 1 M NaCl] and 1 mL fractions collected.

2.2.6 SDS-polyacrylamide gel electrophoresis (SDS-PAGE)

5 mL SDS-PAGE gels were made with a 12 % resolving gel [4 mL 30% bis-acrylamide, 2.5 mL 1.5 M Tris-HCl pH 8.8, 3.3 mL ddH_20, 100 µL 10 % SDS, 100 µL 10 % ammonium per sulphate (APS), 10 µl TEMED] and a 5 % stacking gel [0.83 mL 30 % bis-acrylamide, 0.63 mL 1 M Tris-HCl pH 6.8, 3.4 mL ddH_20, 50 µL 10 % SDS, 50 µL 10 % APS, 5 µL TEMED]. Samples were mixed 1:1 with 2x denaturing sample buffer [60 mM Tris pH 6.8, 25 % (v/v) glycerol, 2 % (w/v)

SDS, 5 % β-mercaptoethanol, 0.01 % (w/v) bromophenol blue] and heated at 95 °C for 5 mins immediately prior to loading onto the gel. Electrophoresis was performed in 1x SDS running buffer [25 mM Tris, 192 mM glycine, 0.1 % (w/v) SDS] at 200V for 40-60 minutes. Broad range molecular weight markers (New England Biolabs), covering 7-175 kDa, were used. Unless western blotting was required, proteins were visualised using Coomassie stain.

2.2.7 Protein identification, accurate mass and oligomeric state determination by mass spectrometry

The lack of an M2-1 antibody resulted in the use of mass spectrometry to confirm the identity of bacterially expressed M2-1 protein. Protein was separated by SDS-PAGE and the appropriate band excised from the gel into 30 % (v/v) ethanol. The sample was given to the mass spectrometry facility at University of Leeds for in-gel trypsin digestion and peptide sequencing by LC-MS/MS. The analysis was performed as a service by facility manager James Ault.

Denaturing mass spectrometry was used to determine the accurate mass of monomeric M2-1. Protein was diluted to 10-20 μM in 50 % (v/v) acetonitrile, 10 % (v/v) acetic acid. Intact mass analysis by native mass spectrometry was used to determine the oligomeric state of M2-1. Protein was buffer exchanged into 10 mM ammonium acetate pH 7.5. Both analyses were conducted by Helen Beeston (Alison Ashcroft group, University of Leeds) using nano-electrospray-ionization mass spectrometry (nanoESI-MS) with an LCT premier (Waters UK Ltd.) mass spectrometer. Samples were infused using a Nanomate (Advion) using the following instrument parameters: ionization voltage +1.75 kV and sample cone voltage 30 V.

2.2.8 Baculovirus expression system

2.2.8.1 Insect cell culture

The insect cell line Sf9, derived from pupal ovarian tissue of the moth *Spodoptera frugiperda,* is highly susceptible to baculovirus infection. Cells were maintained in TC-100 Insect Medium (GIBCO® Invitrogen) with L-glutamine, supplemented with 10 % fetal bovine serum (FBS) and 100 unit/ mL penicillin and 0.1 mg/mL streptomycin (Sigma) ('complete' TC-100 media), at 27 °C. Cells were grown in adherent culture immediately after thawing, with passaging at 70-80 % confluency and renewing the media every 3-4 days due to toxicity of waste products. Once adherent culture had produced sufficient viable cells, they were adapted to suspension culture: cultures were maintained in 50 mL volumes in 125 mL volume spinner flasks, with stirring at 75 rpm. Once cells reached a density of ~ $2x10^6$ cells/mL, cultures were diluted with fresh, room temperature medium 1:10.

2.2.8.2 Bacmid linearization

BAC10:KO_{1629} bacmid DNA used was kindly donated by Professor Ian Jones' group from the University of Reading, and consisted of the genome of baculovirus strain Autographa californica Multicapsid Nucleopolyhedrovirus (AcMNPV) with a targeted gene knock-out of the essential ORF1629 adjacent to the polyhedrin recombination locus (polh), in an E.coli:BAC10 vector (Zhao et al., 2003). When transfected, this creates a virus that is non-viable unless the ORF1629 sequence is supplied by homologous recombination with the pTriEx1.1 Neo transfer vector, removing the need for several time-consuming rounds of plaque-assay purification. Bacmid DNA was linearised before transfection to increase recombination efficiency. A restriction digest using 5 units Bsu361 (New England Biolabs) with 2 μg bacmid DNA and 1x BSA was performed as previously described, followed by heat inactivation at 80 °C for 20 mins and storage at -20 °C.

2.2.8.3 Generation of a P_0 generation recombinant baculovirus

Suspension culture of Sf9 cells was diluted to 0.5 x 10^6 cells/mL and 2 mL added to a single well of a 6-well tissue culture plate. Cells were allowed to settle and were checked for 50 % confluency. Transfection mixes were prepared by incubating 500 ng linearised bacmid, 500 ng pTriex1.1 Neo M2-1, 8 μL Lipofectin (Life Technologies) and sterile water to 24 μL, at room temperature for 20 minutes. Media was removed from the settled cells and replaced with the transfection mix plus 1 mL serum-free, antibiotic-free TC-100 medium. The plate was placed inside a plastic bag to prevent drying-out and incubated for 16 hours at 27 °C. Media was changed for serum-containing media and cells incubated for 3 days. Cells were scraped and centrifuged at 13,000 rpm for 2 minutes, and the supernatant (containing P_0 baculovirus) was retained in a sterile tube and stored at 4 °C in foil, to prevent degradation.

2.2.8.4 Infections and virus amplification

Small-scale infections to confirm expression of M2-1-His were performed as follows: 0.5 x 10^6 cells per well were seeded into a 6-well plate and infected with 0.5 mL P_0 virus for 1 hour at 27 °C, with gentle rocking to prevent drying out. Virus was replaced with fresh 'complete' TC-100 media and the plate was incubated for 3 days at 27 °C.

To amplify the virus stock, subsequent generations of virus (P_1, P_2 etc.) were produced by reinfecting 15 x 10^6 Sf9 cells in a T175 tissue culture flask with 0.5 mL P_0 generation virus for 1 hour at 27 °C, followed by removal of the virus and incubation in fresh media for 5 days at 27 °C. P_1 generation virus was recovered from the supernatant of pelleted cells.

Large-scale infections were performed in 200 mL Sf9 suspension culture (1 x 10^6 cells/mL) with 3 mL P_1 baculovirus and incubation at 27 °C for 3 days, with stirring.

2.2.8.5 Harvesting whole cell lysates

To assess overall expression during optimisation stages, whole cell lysates were prepared. Infected cells were recovered by centrifugation at 2000 xg 20 mins and a sample of the supernatant was retained for SDS-PAGE analysis. Whole cell lysates were prepared by resuspension of pelleted cells in 100 μL or 10 mL (small- and large-scale infections, respectively) RIPA buffer [50 mM Tris-Cl pH 7.5, 150 mM NaCl, 1 % NP-40, 0.5 % sodium-deoxycholate, 0.1 % SDS, 1x complete protease inhibitor cocktail EDTA-free (Roche)].

2.2.9 Purification of M2-1-His by cobalt affinity chromatography

Protocol is adapted from Mason et al. (Mason et al., 2003). Cells were recovered by centrifugation at 2000 xg 20 mins and resuspended in binding buffer, 5 mL per gram wet-weight of pellet [20 mM Tris-Cl pH 7.9, 500 mM NaCl, 5 mM imidazole, 1x complete protease inhibitor cocktail EDTA-free (Roche)]. Cells were lysed by sonication on ice with 5 cycles of 10 sec on, 10 sec off, and lysates were clarified by centrifugation at 100,000 xg 30 mins.

A HiTrap IMAC FF 1 mL column (GE Healthcare) was washed and charged with 0.1 M colbalt chloride, as per the manufacturer's guidelines using a manual syringe method. The column was equilibrated in 5 column volumes (CV) binding buffer and the supernatant applied, collected, and reapplied to ensure maximum binding. A 10 CV wash with binding buffer, followed by stepwise elution with 5 CV steps of elution buffer [20 mM Tris-Cl pH 7.9, 500 mM NaCl, increasing concentrations of imidazole: 50 mM, 100 mM, 150 mM, 250 mM, 500 mM]. Eluted protein was stored at 4 °C.

2.2.10 Calf intestinal alkaline phosphatase (CIP) treatment

CIP treatment was used to observe dephosphorylation of cobalt-affinity purified M2-1-His protein. A volume of 50 μL contained 8 units CIP (New

England Biolabs), 1x NEBuffer 3 (New England Biolabs), 44 μL eluted protein. Reactions were incubated for 2 hours at 37 °C.

2.2.11 Western blotting

Proteins were transferred from SDS-PAGE gels (see 2.2.2.6) to polyvinylidene fluoride (PVDF) membrane (Immobilon-P Transfer Membrane; Millipore) using a Trans-Blot semi-dry cell (Bio-Rad) in Towbin buffer [25 mM Tris, 192 mM glycine, 20 % (v/v) methanol]. Membranes were blocked overnight at 4 °C in 10 % (w/v) non-fat powdered milk in 1x Tris-buffered saline with Tween-20 (TBS-T) [50mM Tris-HCl pH 7.6, 150mM NaCl, 1 % Tween-20]. Blocking buffer was then replaced with primary antibody (rabbit polyclonal anti-His probe G-18:sc-804, Santa Cruz Biotechnology Inc.) diluted 1:500-1000 in 5 % (w/v) non-fat powdered milk in 1x TBS-T for 1 hour at room temperature. Membranes were washed twice for 5 minutes in 1x TBS-T, then incubated with horseradish peroxidise (HRP) conjugated goat anti-rabbit secondary antibody (A6154 Sigma), under the same conditions and dilution as for the primary incubation, followed by 4 washes in 1x TBS-T.

Proteins were visualised using enhanced chemiluminescence on photographic film. Enhanced chemiluminescence solution (ECL) was made with 1:1 mix of each of the following solutions and applied to the membrane: solution 1, 25 mM Luminol (3-aminophthalydazide), 0.3 mM *p*-coumaric acid, 100 mM Tris-HCl pH 8.5; solution 2, 0.01 % (v/v) hydrogen peroxide, 100 mM Tris-HCl pH 8.5. Membranes were incubated for 1 min, then exposed to film and developed with a Konica SRX-101A processor.

2.2.12 Phosphopeptide mapping by mass spectrometry

Identification of phosphorylated residues in M2-1 was by mass spectrometric analysis performed by Kevin Tipping (Sheena Radford group, University of Leeds) and James Ault (Mass Spectrometry Facility Manager, University of

Leeds). Briefly, M2-1 protein purified from baculovirus-infected insect cells was separated into phosphorylated and unphosphorylated species using SDS-PAGE. Both bands were excised and analysed separately. Proteins were in-gel digested with trypsin, eluted in acetonitrile:water:formic acid (60:35:5 v/v) and lyophilised by vacuum centrifugation. Samples were then enriched for phosphorylated peptides using titanium oxide (TiO_2) beads (GL Sciences, Japan) and desalted into 50 % (v/v) acetonitrile. Mass analysis was by nanoESI-MS and peptide sequencing by tandem mass spectrometry using a quadrupole-IMS-orthogonal time-of-flight MS (Synapt HDMS, Waters UK Ltd.) with ionization voltage of +1.5 kV and cone voltage of 20 V. Data processing was performed using the MassLynx v4.1 suite of software supplied with the mass spectrometer.

2.2.13 Crystallographic techniques

2.2.13.1 Sparse matrix screening

Commercially available sparse matrix screens were used for initial crystal trials. The following kits were used: Crystal Screen I and II, Matrix, Index, Salt RX (Hampton Research); Wizard I and II (Emerald Biosystems). Sitting drops were set up in 96-well MRC 2-drop plates (Molecular Dimensions) using an Oryx 6 robot (Douglas Instruments) with 60 μL reservoir solution. 1 μL drops at 50:50 or 70:30 protein:reservoir ratios were dispensed, using protein at 10-20 mg/mL. Plates were sealed with Viewseal pressure adhesive clear seals (Greiner Bio-One) and incubated at 4 °C, 18 °C or 25 °C.

2.2.13.2 Crystal optimisation

Screening drops were monitored for crystal growth after 24 hours, and periodically for two weeks thereafter. Hits were optimised in both sitting drop (96-well plate) and hanging drop (24-well plate) orientations using 1-2 μL drops at 50:50 and 70:30 protein:reservoir ratios. Matrices were set up to vary, typically, concentration of precipitant and pH of buffer. Temperature and protein concentration were also varied, and Hampton Research Additive Screen used to further improve optimised hits.

2.2.13.3 Crystal harvesting and data collection

Crystals were picked in appropriately sized loops (Hampton Research) and cryo-cooled in liquid nitrogen, either straight from drops when grown in low molecular weight polyethylene glycol (PEG), or by first transferring the crystal into a 1:1 drop of reservoir and cryoprotectant [5 % glycerol, 5 % PEG-400, 5 % 2-Methyl-2, 4-pentanediol (MPD), 5 % ethylene glycol].

All data were collected at 100 K at Diamond Light Source on beamlines I24, I02 and I04. Crystals were screened for diffraction by taking test images at ϕ = 0° and 90°, using 12,658 eV X-rays (0.9795 Å). Crystals that diffracted to better than 4 Å were fluorescence scanned to identify the presence of any zinc (from the zinc finger motif) or other atoms (either naturally occurring or from crystallisation conditions), and their absorption edges calculated for SAD data collection (approximately 9658 eV for zinc K-edge).

2.2.13.4 Data processing and structure solution

X-ray data were integrated using iMOSFLM (Battye et al., 2011, Leslie, 2006) from the Collaborative Computer Project 4 (CCP4) suite (Winn et al., 2011, Potterton et al., 2003) or XDS (Kabsch, 2010). POINTLESS was used to test for twinning, confirm the space group identity, and provide statistics for the significance of any anomalous signal (Evans, 2006, Evans, 2011) and reduced data from all crystals were then scaled with AIMLESS (Evans, 2006). Experimental phasing of the M2-1 WT-P422 data using the naturally-bound zinc and cadmium atoms, was carried out with AUTOSOL (Terwilliger et al., 2009) using the single-wavelength anomalous dispersion (SAD) method. A density-modified map was produced using RESOLVE (Terwilliger and

Berendzen, 1999) and was then used with BUCCANEER (Cowtan, 2006) to build an initial model. This was followed by consecutive cycles of manual building in COOT (Emsley and Cowtan, 2004) and structure refinement in REFMAC5 (Murshudov et al., 2011). Molecular replacement into other datasets was performed using PHASER (McCoy et al., 2007) The structures were fully refined following cycles of manual building in COOT and refinement cycles in REFMAC5, coupled with automatically generated local non-crystallographic symmetry restraints. Intermolecular interactions were analysed using PISA (Krissinel and Henrick, 2007) and the MolProbity server. Models were examined in COOT and figures made using PyMOL (Schrödinger) and CCP4mg (McNicholas et al., 2011).

2.2.14 RNA binding

2.2.14.1 Phenol-chloroform extraction of nucleic acid

To visualise any nucleic acid bound to purified protein phenol-chloroform extraction, followed by ethanol precipitation and agarose gel electrophoresis, was used. Protein was mixed with 1 volume TE buffer-equilibrated phenol [TE buffer: 10 mM Tris-Cl pH 8, 1 mM EDTA] and vortexed. Samples were centrifuged at 21,000 xg 10 mins, the nucleic acid-containing aqueous phase was removed, and the above procedure repeated. The aqueous layer was then vortexed with an equal volume of chloroform, centrifuged, and the aqueous layer retained as above. Ethanol precipitation of nucleic acid was performed by adding NaCl to a final concentration of 0.1 M to the aqueous phase, followed by 2.5 volumes of 96 % ice-cold ethanol, and incubation at -20 °C overnight. Samples were centrifuged at 21,000 xg for 30 mins at 4 °C, the supernatant decanted, and the pellet washed in 2 volumes of 70 % ice-cold ethanol. Pellets were air-dried and resuspended in 10-20 μL water with 0.1 % (v/v) diethyl pyrocarbonate (DEPC). Addition of RNase A was used to distinguish RNA from DNA.

Samples were added to 5 μL 2x RNA ladder loading dye (New England Biolabs) and electrophoresed in 1.5 % agarose gels made with, and run in, Tris-borate-EDTA buffer [TBE: 89 mM Tris, 89 mM Borate, 2 mM EDTA, in DEPC-treated water] and 3 μL ethidium bromide, with a single-stranded low range RNA marker (New England Biolabs) at 100 V for 40 mins. Nucleic acids were visualised with UV transillumination.

2.2.14.2 Mass spectrometry

Mass spectrometry was used to determine if short synthetic RNAs could be bound to purified, RNA-free tetrameric M2-1. Protein and RNA was buffer exchanged into 100 mM ammonium acetate pH 7, and 40 μM protein was mixed with 10 μM RNA (a number of sequences were used). Analysis was performed by Helen Beeston (Alison Ashcroft group, University of Leeds) using nano-electrospray-ionization mass spectrometry (nanoESI-MS) with an LCT premier time-of-flight mass spectrometer (Waters UK Ltd.). Samples were infused using a Nanomate (Advion) using the following instrument parameters: ionization voltage +1.75 kV and sample cone voltage 30 V.

2.2.14.3 Fluorescence polarisation anisotropy

Binding of M2-1 to short RNAs (8-44 nucleotides) was tested using fluorescence polarisation anisotropy (FPA). 3'-fluorescein (3'-Fl)-labelled RNAs representing a number of RSV cis-acting signals in both positive and negative senses, as well as poly-G, -A, -U or –C 8mers, were synthesised (Thermo Scientific). Reactions were carried out in 96- and 384-well format, in 20 mM Tris-Cl pH 7.5, 150 mM NaCl and 0.01 % (v/v) Triton X-100, with 20 nM 3'-Fl-labelled RNA and increasing concentrations of protein (0.1 nM to 300 μM). Plates were allowed to equilibrate at room temperature for 30 mins and polarisation measured using an EnVision Multilabel Plate Reader (PerkinElmer) with a 480 nm excitation filter, and 530 nm S- and P-channel emission filters. Polarisation (*mP*) is defined as follows:

$$mP = 1000 \times \frac{S - GP}{S + GP}$$

where S and P are background subtracted fluorescence count rates and G (grating) is an instrument and fluorophore dependent factor.

Experiments were performed in triplicate and data were expressed as the fraction of RNA bound, plotted against protein concentration and fitted by standard logistic regression using OriginPro 8.6 (Origin Lab) according to the following equation:

$$y = \frac{A_1 - A_2}{1 + (\frac{x}{x_0})^p} + A_2$$

where y is the fraction of RNA bound, x is protein concentration, A_1 is initial value of y (0 % RNA bound), A_2 is the final value of y (100 % RNA bound), x_0 is the centre of the curve (Kd), and p is the Hill coefficient. Dissociation constants (Kd) were averaged from the Kd calculated from each triplicate dataset.

3 CHAPTER 3: MULTISYSTEM EXPRESSION AND CHARACTERISATION OF M2-1

3.1 CHAPTER INTRODUCTION

This chapter describes the development of high-yield expression systems for full-length RSV M2-1, as an optimised method had not been published prior to the start of this project.

The first sections (3.2 to 3.4) discuss a recombinant baculovirus-assisted insect cell expression system, and how this system was used to confirm the phosphorylation status of the M2-1 protein and accurately assign the residues involved. This system was found to be unsuitable for the large scale, homogeneous M2-1 expression required for crystallographic studies, and so the remainder of the chapter (sections 3.5 to 3.8) outline the optimisation of bacterially expressed M2-1 protein for downstream use in structural and RNA binding studies.

3.2 INTRODUCTION TO PHOSPHORYLATION OF RSV M2-1 IN A BACULOVIRUS-ASSISTED INSECT CELL EXPRESSION SYSTEM

M2-1 protein has long been observed as heterogeneous when resolved by SDS-PAGE (Routledge et al., 1987), with as many as five different molecular weight forms being detected by SDS-PAGE analysis. When over-expressed in mammalian cells or by *in vitro* transcription/translation, as well as in the context of viral infection, M2-1 appears predominantly as two distinct species: phosphorylated and unphosphorylated protein, with lower and higher electrophoretic mobility, respectively (Lambert et al., 1988, Hardy and Wertz, 2000). The function of this post-translational modification is unknown, although it is necessary for efficient M2-1 anti-termination (Cartee and Wertz, 2001). The phosphorylated residues have been separately reported as follows: threonine 56 and serine 58 (Cuesta et al., 2000), by determining $^{[32]}$P-orthophosphate incorporation into peptide fragments of mammalian cell-expressed M2-1, followed by site directed mutagenesis; serine 58 and serine 61 (Cartee and Wertz, 2001), using protease digestion of mammalian cell-expressed protein followed by mass spectrometric analysis by MALDI-TOF MS/MS. To confirm which, if either, assignment is accurate, we optimised a recombinant baculovirus expression system and used the phosphopeptide mapping service at the mass spectrometry facility of the University of Leeds.

Using insect cell culture, a eukaryotic expression system, greatly increases the likelihood of producing appropriately folded protein due to the expression of molecular chaperones and the mammalian-like cytoplasmic environment, in terms of pH and compartmentalisation (Sokolenko et al., 2012, Kost et al., 2005). There is also extensive precedent for the expression of post-translationally modified and immunologically relevant proteins using baculovirus-infected insect cells, including the virus-like particle human papilloma virus (HPV) vaccine (Harper et al., 2004), and a panel of human cellular kinases where a 99 % success rate in expressing soluble, folded proteins was achieved in Sf9 cells compared to 54 % in *E. coli* (Chambers et al., 2004).

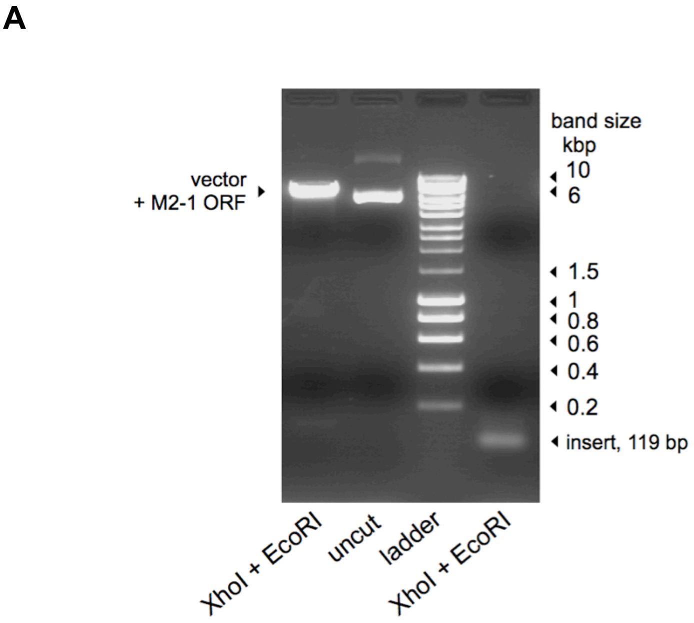
Generation of a recombinant baculovirus used to rely solely on inefficient homologous recombination between the baculovirus genome and a transfer vector, co-transfected into insect cells, with a recombination frequency of 0.1 % (Kost et al., 2005). This frequency was gradually improved with the addition of improvements to the baculovirus construct, such as a unique *Bsu36I* restriction site to linearise the genome, and further *Bsu36I* sites that knocked out the essential baculoviral gene ORF_{1629}, preventing the replication of parental baculovirus and resulting in 100 % recombinant virus (Zhao et al., 2003). This is the method used here, although there are others based around site-specific bacterial transposition (Bac-to-Bac, Life Technologies) where the gene of interest is transposed into the baculoviral genome in *E. coli* cells and a single plasmid is transfected into insect cell culture (Luckow et al., 1993).

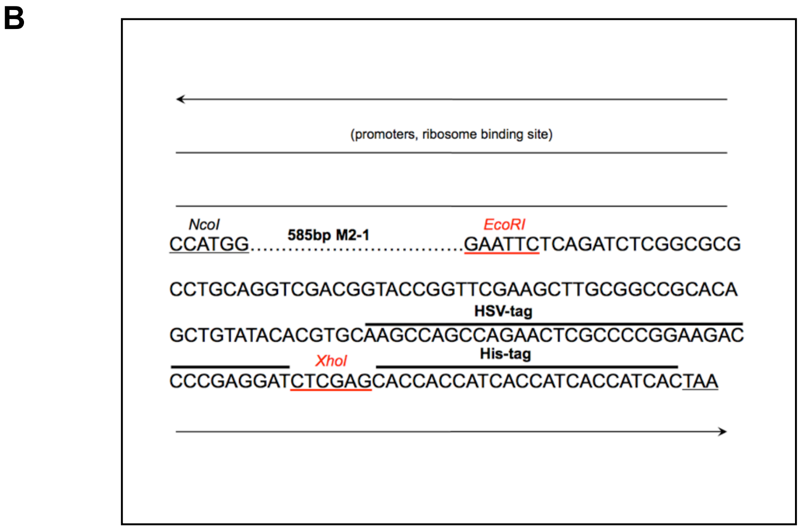
3.3 RESULTS

3.3.1 Optimisation of pTriEx1.1 Neo M2-1 construct

Subcloning of the M2-1 ORF cDNA into NcoI and *EcoRI* sites of pTriEx1.1 Neo, the transfer vector required for recombinant baculovirus formation, was performed previously in our group. However, the resulting plasmid was not optimally constructed; it contained extraneous vector sequences encoding an additional 40 amino acids, including a redundant HSV-tag (11 amino acids from Herpes Simplex Virus glycoprotein D), prior to the C-terminal 8xHis-tag (figure 3.1 B). This region of the vector was removed and replaced with a thrombin cleavage site [L-V-P-R-G-S], resulting in expression of a M2-1 fused to a minimal number of non-native amino acids and with provision for post-expression removal of the His-tag (figure 3.1 C).

Optimisation of the vector construction was achieved using two complementary oligonucleotides designed to form the thrombin cleavage site flanked by *EcoRI* and *XhoI* restriction sites (Appendix I). Equimolar amounts of each oligonucleotide were annealed together at 95 °C for 5 minutes followed by slow cooling to room temperature. Both this insert and pTriEx1.1 Neo M2-1 were digested with *EcoRI* and *XhoI*, and the desired fragments purified by agarose gel electrophoresis and gel extraction (figure 3.1 A). Vector and insert were ligated overnight, transformed into DH5α cells, DNA isolated from single colonies by miniprep purification, and successful recombinant plasmids identified by DNA sequencing. This construct was renamed ‘pST01’ (figure 3.1 C).


A
vector
+ M2-1 ORF
band size
kbp
10
6
1.5
1
0.8
0.6
0.4
0.2
insert, 119 bp
XhoI + EcoRI
uncut
ladder
XhoI + EcoRI


B
(promoters, ribosome binding site)
NcoI
CCATGG
585bp M2-1
EcoRI
GAATTCTCAGATCTCGGCGCG
CCTGCAGGTCGACGGTACCGGTTCGAAGCTTGCGGCCGCACA
HSV-tag
GCTGTATACACGTGCAAGCCAGCCAGAACTCGCCCCGGAAGAC
XhoI
His-tag
CCCGAGGATCTCGAGCACCACCATCACCATCACCATCACTAA

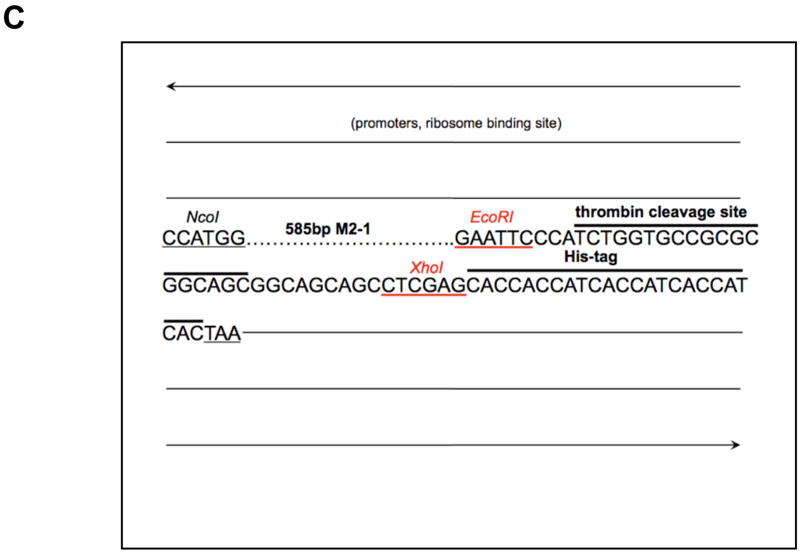

C
(promoters, ribosome binding site)
NcoI
CCATGG
585bp M2-1
EcoRI
thrombin cleavage site
GAATTCCCATCTGGTGCCGCGC
His-tag
XhoI
GGCAGCGGCAGCAGCCTCGAGCACCACCATCACCATCACCAT
CACTAA

Figure 3.1 Optimisation of pTriEx1.1 Neo M2-1 clone.
A) shows results of a double *EcoRI/XhoI* restriction digest on both pTriEx1.1 Neo M2-1 and the thrombin site-containing insert created by annealing complementary oligonucleotides. The vector+M2-1 ORF and insert bands were isolated by gel extraction. **B)** and **C)** The arrangement of the M2-1 ORF (585bp cloned between *NcoI* and *EcoRI* downstream of the expression promoters and ribosome binding site) in pTriEx1.1 Neo before **B)** and after **C)** removal of extraneous sequence (multiple cloning site and HSV-tag sequences) and insertion of thrombin cleavage site. The construct shown in C) was successfully generated and referred to as 'pST01'.

3.3.2 Generating recombinant baculovirus

Initial recombinant baculovirus, P_0, for use in small scale expression trials, was generated by co-transfection by lipofection of Sf9 insect cells with the linearised BAC10:KO_{1629} bacmid, containing the AcMNPV genome, and the transfer vector pST01 containing the M2-1 ORF. Virus was harvested 3 days post infection (p.i.) and used to infect small-scale adherent cultures to confirm expression of M2-1-His. At 3 days p.i. virus infected cells were harvested and, to confirm that M2-1 protein was not secreted from insect cells in this system, a sample of the supernatant was also retained for SDS-PAGE analysis. Whole cell lysates were prepared and analysed by SDS-PAGE alongside the supernatant, followed by western blotting using an anti-His-tag primary antibody. Blots showed the presence of a ~26 kDa band in the whole cell lysate but not the supernatant, suggesting the M2-1 fusion protein was expressed in these cultures, but was not secreted into the supernatant (figure 3.2).

3.3.3 Purification of M2-1 from large-scale infections

Due to the lack of an adequate M2-1-specific antibody and the high non-specific binding of anti-His-tag antibodies, we used cobalt affinity chromatography to purify the 8x-His-tagged M2-1 protein and confirm its identity. P_1 generation baculovirus was created by amplification of P_0 virus via a round of infection and replication. This virus was used to infect 200 mL suspension cultures of Sf9 cells for 3 days, and cells were recovered and lysed in binding buffer containing 5 mM imidazole to prevent non-specific binding to the HiTrap IMAC FF column. Protein was bound to the column, washed, and eluted by stepwise imidazole gradient. Samples were taken at each stage for analysis by SDS-PAGE followed by Coomassie staining (figure 3.3 A) and anti-His-tag western blotting (figure 3.3 B). Both analyses show a doublet of protein, presumably M2-1, was successfully purified at ~26-27 kDa, eluting primarily between 150 mM and 250 mM imidazole, although the presence of two species is very difficult to resolve by western blot with this antibody.

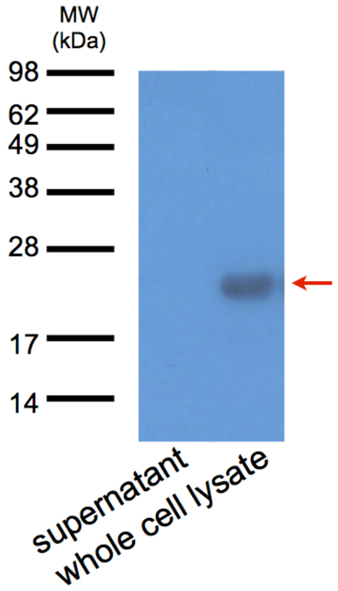


Figure 3.2 P_0 baculovirus expression trial.

Anti-His-tag western blot of supernatant and whole cell lysate shows that M2-1 (red arrow) is present only in whole cell lysates i.e. it is not secreted.

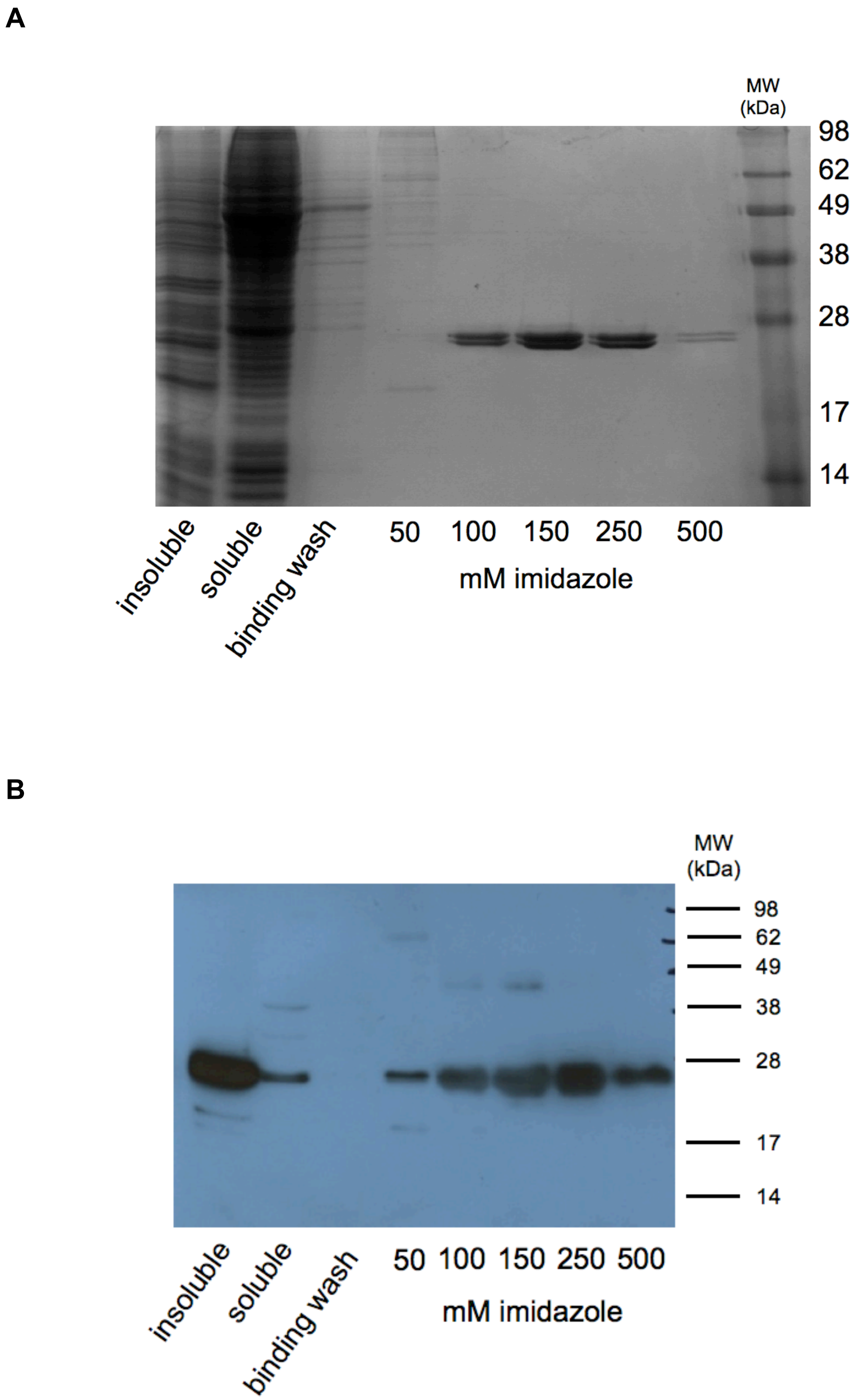

A
MW (kDa)
98
62
49
38
28
17
14
insoluble
soluble
binding wash
50
100
150
250
500
mM imidazole
B
MW (kDa)
98
62
49
38
28
17
14
insoluble
soluble
binding wash
50
100
150
250
500
mM imidazole

Figure 3.3 Cobalt affinity purification of M2-1 from baculovirus infected cells.

(Above) Samples from multiple steps throughout purification were analysed by SDS-PAGE followed by A) Coomassie staining, or B) anti-His-tag western blotting. M2-1 appears as a doublet of approximately 26-27 kDa; although the two species are difficult to distinguish in this blot, they were visible at shorter exposures.

3.3.4 Insect cell expressed M2-1 is phosphorylated

To confirm that the presence of a doublet on SDS-PAGE was due to differential phosphorylation, as has been shown in the literature for mammalian cell expressed protein (Cartee and Wertz, 2001), purified M2-1 was calf intestinal alkaline phosphatase (CIP) treated to remove any phosphorylation present. Samples were taken before and after CIP treatment and analysed by SDS-PAGE and Coomassie staining. The gel showed CIP treatment resulted in a shift in protein abundance to the species with higher electrophoretic mobility, assumed to be unphosphorylated protein, suggesting that the species with lower electrophoretic mobility was indeed phosphorylated in this system (figure 3.4).

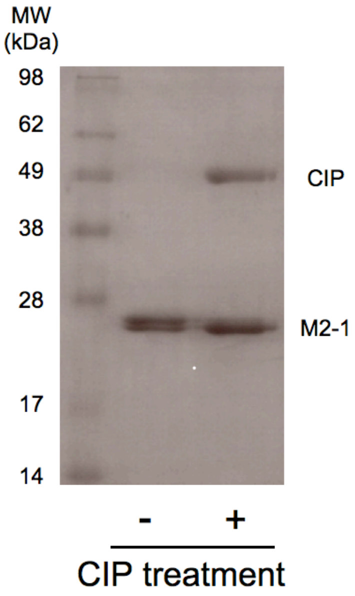


Figure 3.4 Phosphorylation of insect cell expressed M2-1.
CIP treatment of purified M2-1 shows the expected shift in abundance to the higher electrophoretic mobility species of the M2-1 doublet (26-27 kDa).

3.3.5 Confirmation of protein ID and identification of phosphorylation sites by mass spectrometry

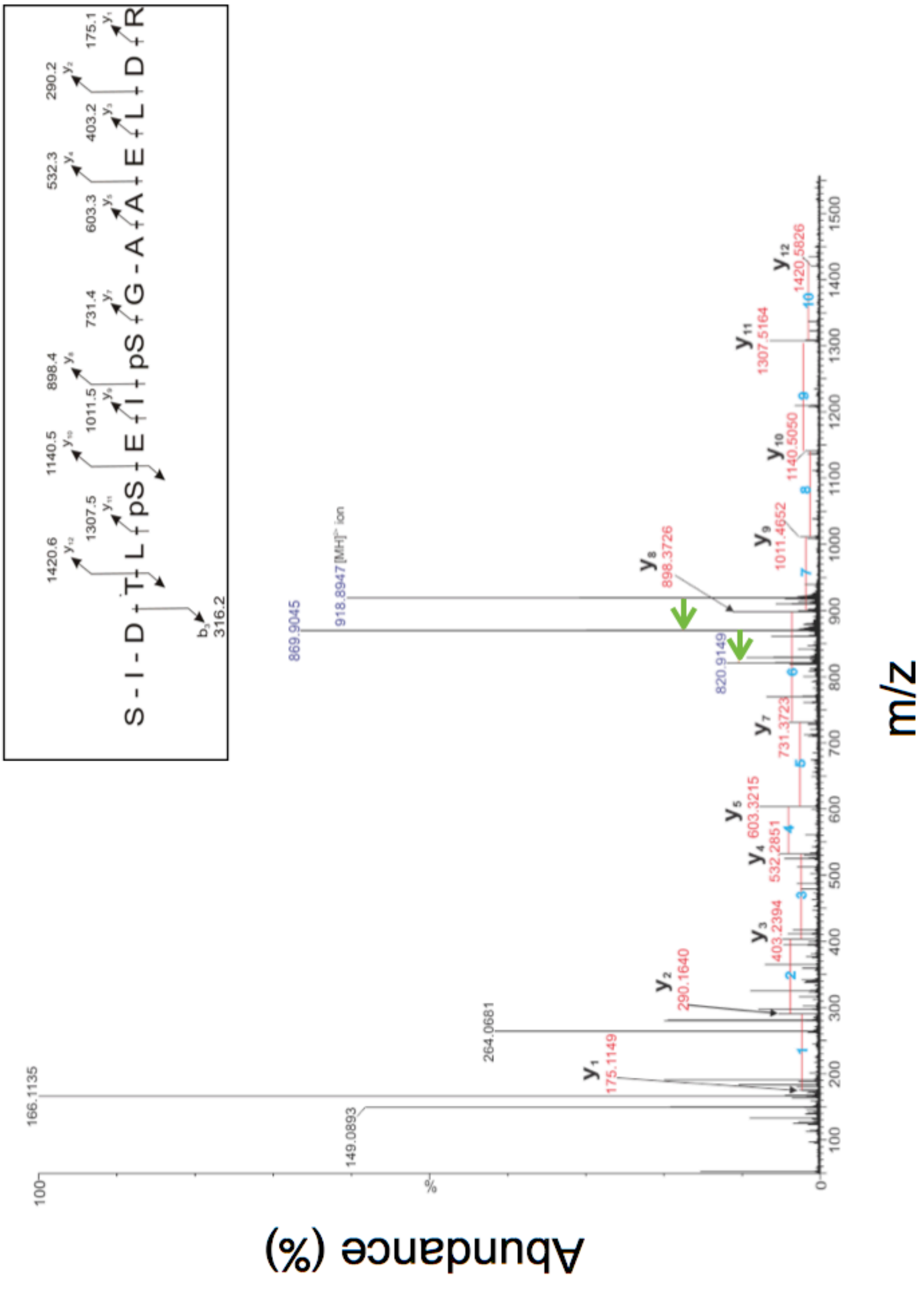
Both bands from the purified M2-1 doublet were excised from 100, 150, and 250 mM imidazole elutions and pooled as two samples: phosphorylated and unphosphorylated protein. Samples were analysed as a service by the University of Leeds mass spectrometry facility, by the facility manager James Ault and a PhD student Kevin Tipping (Prof Sheena Radford group) who was developing an in-house phosphopeptide mapping protocol.

Firstly, protein from both bands was identified as M2-1 by in-gel trypsin digestion followed by analysis of the peptides by LC-MS/MS. The primary database ‘hit’ for both species was RSV M2-1, with sequence coverage of over 50 % for both samples.

Secondly, peptides were prepared for phosphomapping by first enriching for phosphorylated peptides using titanium oxide chromatography. Peptides were then analysed by tandem mass spectrometry using a quadrupole-IMS-orthogonal time-of-flight MS. The mass/charge spectrum shows the presence of a doubly phosphorylated peptide, as indicated by the two neutral loss peaks from the parent ion (figure 3.5). The sequence of the phosphopeptide was calculated from the mass difference between peaks of the y-ion series, generated by retention of protonation on the C-terminal end of a fragmented CO-NH bond. Analysis indicated phosphorylation events at serine 58 and serine 61. This is in line with the previously published result from Cartee et al. (Cartee and Wertz, 2001) and not from Cuesta et al. (Cuesta et al., 2000), who reported threonine 56 and serine 58.

Figure 3.5 Phosphopeptide mapping mass spectrum.

(Below) MSMS acquisition of $[918.8947]^{2+}$ parent ion from direct infusion of tryptically digested RSV M2-1 protein enriched for phosphopeptides using TiO_2 chromatography. The two neutral loss peaks (~49Da each) from the parent ion (green arrows) indicate the presence of a doubly-phosphorylated peptide. The following sequence was determined from the y-ion series by following the addition of mass from y_1 Arg, as indicated by the light blue numbers: 1) 114.9532, **D.** 2) 113.0754, **L.** 3) 129.0457, **E.** 4) 71.0364, **A.** 5) 128.0508, **A & G.** 6) 167.0003, **S^p.** 7) 113.0926, **I.** 8) 129.0398, **E.** 9) 167.0114, **S^p.** 10) 113.0662, **L.**


S - I - D - T - L - pS - E - I - pS - G - A - A - E - L - D - R
b3 316.2
y12 1420.6
y11 1307.5
y10 1140.5
y9 1011.5
y8 898.4
y7 731.4
y5 603.3
y4 532.3
y3 403.2
y2 290.2
y1 175.1
Abundance (%)
m/z
166.1135
149.0893
264.0681
869.9045
918.8947 [MH]2+ ion
820.9149
y1 175.1149
y2 290.1640
y3 403.2394
y4 532.2851
y5 603.3215
y7 731.3723
y8 898.3726
y9 1011.4652
y10 1140.5050
y11 1307.5164
y12 1420.5826
1
2
3
4
5
6
7
8
9
10
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
100
%

3.4 SECTION SUMMARY

This section has confirmed that the M2-1 species that resolves with a lower electrophoretic mobility is a doubly phosphorylated protein with phospho-serines at positions 58 and 61. These sites match those published by Cartee et al. (Cartee and Wertz, 2001).

The baculovirus system of expression was chosen for this section of the project for its ease of scale-up relative to mammalian cell culture, and for its physiologically similar post-translational modifications to mammalian tissue culture systems. Phosphopeptide mapping by tandem mass spectrometry was used for its ability to accurately pinpoint post-translational modifications without the need for time consuming site directed mutagenesis and loss of phenotype analysis. This is the same technique used by Cartee et al., and its superior accuracy when compared to the techniques employed by Cuesta et al., namely $^{[32]}$P-orthophosphate incorporation and site directed mutagenesis, is why serines 58 and 61 are more frequently quoted in the literature as the phosphorylation sites of M2-1. Confirmation of these sites from an orthogonal expression system settles any remaining doubt over their assignment.

Knowledge of the residues involved in phosphorylation is important for our understanding of the possible roles of this modification during the virus life cycle. Cellular kinases involved may be predicted by consensus sequence; serine 58 lies within a casein kinase I (CKI) sequence, and phosphorylation at this site forms a consensus sequence for CKI phosphorylation at serine 61. Cartee et al. showed that CKI could phosphorylate bacterially-expressed M2-1 *in vitro* (Cartee and Wertz, 2001). More importantly for this research, interpretation of the location of the exact residues involved in phosphorylation in terms of a crystal structure may indicate a role in conformational change or regulation of access to nucleic acid or protein binding sites. It was based on this information that expression of both wild type and phosphomimetic (S58DS61D) M2-1 was optimised for structural analysis and RNA binding studies in the remainder of this project.

Although we have presented a system for large-scale expression of functionally relevant M2-1 protein, several factors made this method impractical for further structural studies. Homogeneous protein was required for crystallisation, and separation of the different phospho-species of M2-1 on a large scale was not easily achievable. Equally, bulk dephosphorylation of insect cell expressed protein to create a homogeneous sample was not an option due to incomplete removal of phosphorylation and the cost-ineffectiveness of this process. Finally, whilst yield was high relative to mammalian cell expression, the time scales involved in culturing large volumes of insect cell suspension culture prohibit the high-throughput expression possible with a bacterial system. As a result, optimisation of bacterial expression of RSV M2-1 was performed for crystallographic studies.

3.5 INTRODUCTION TO BACTERIAL EXPRESSION OF M2-1

Over-expression in prokaryotic *E. coli* cells is the most common method of producing high-yield, homogeneous protein for structural studies. The two vectors used here use different promoter systems: pET28a-SUMO places the target protein under the control of a T7 RNA polymerase promoter and requires DE3 *E.coli,* which have a chromosomally-integrated phage (λDE3) encoding the T7 RNA polymerase gene under the control of the *lac*UV5 promoter; pGEX vectors use the *tac* promoter, a strong hybrid promoter containing *lac*UV5 elements, meaning the protein of interest itself is under lac repression. IPTG induction either allows T7 RNA polymerase expression, which in turn transcribes the target gene (pET), or directly relieves repression of the target gene allowing transcription by the *E.coli* RNA polymerase (pGEX).

Advantages of bacterial protein expression include homogeneity, high yield, low cost, and ease of scale-up. However, these advantages are at the expense of correct post-translational modifications, such as phosphorylation, and this must be taken into account when analysing the structure or function of proteins expressed this way. This section outlines the process of optimising bacterial expression of RSV M2-1 protein.

3.6 RESULTS

3.6.1 Cloning of M2-1 ORF into His-SUMO-tag expression vector

To improve bacterial expression of RSV M2-1 the ORF was subcloned into pET-28aSUMO, a vector that encodes an N-terminal 6xHis-SUMO. This tag allows for nickel affinity chromatography for purification, whilst the SUMO protein has been shown to aid in solubility, folding, and prevention of degradation of recombinant protein.

Forward and reverse primers were designed to amplify the M2-1 ORF from pST01, flanking it N-terminally with *BamHI* site and C-terminally with a TAA stop codon followed by *XhoI* site (Appendix I). PCR was performed with these primers and both the PCR product (the insert) and pET-28aSUMO vector were digested with *BamHI* and *XhoI* restriction enzymes. Insert and empty vector fragments were purified by agarose gel electrophoresis and gel extraction. Fragments were ligated together and transformed into DH5α cells, before isolating 9 single colonies for DNA purification by miniprep. Presence of the M2-1 ORF was confirmed by diagnostic restriction enzyme digest with *BamHI* and *XhoI* (figure 3.6). All 9 clones produced a 609 bp fragment corresponding to the PCR product. The clone with the highest yield was amplified by maxiprep and sent for DNA sequencing to confirm correct insertion of the M2-1 ORF. This construct is called ‘pETSUMO-M2-1’.

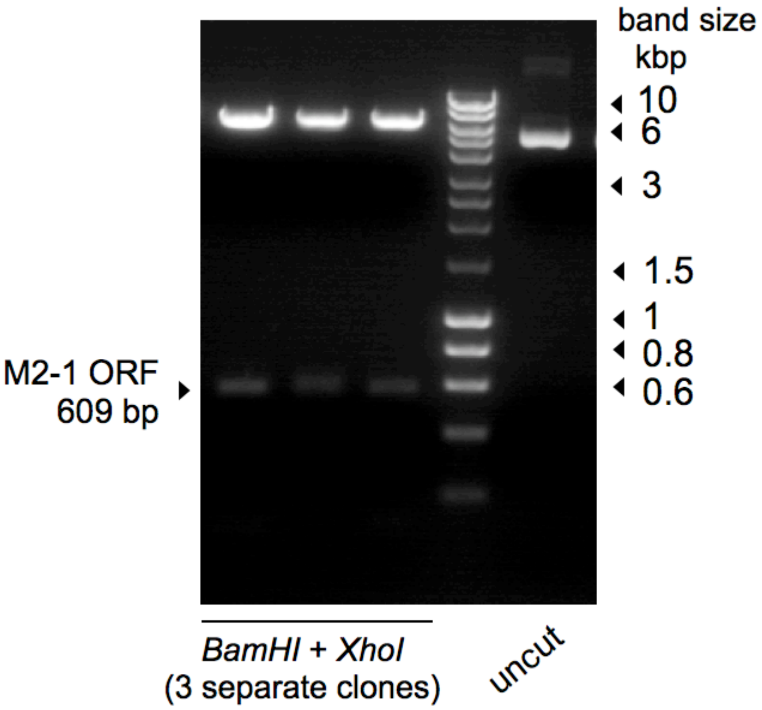


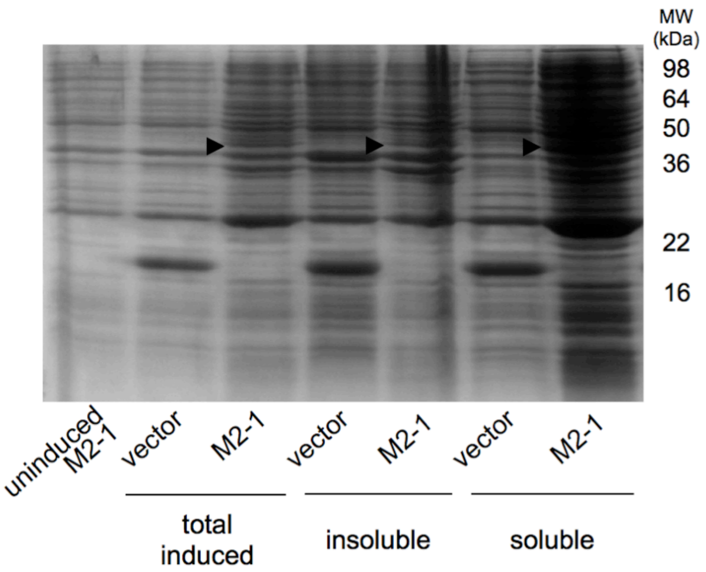
Figure 3.6 Identification of M2-1 ORF subcloned into pET-28aSUMO. Double restriction enzyme digests were performed for all 9 clones with *BamHI* and *XhoI.* 3 clones are shown here as examples, and uncut vector is shown for reference. All clones produced the expected 609 bp fragment for the M2-1 ORF.

3.6.2 Preliminary expression and purification of His-SUMO-M2-1

pETSUMO-M2-1 was transformed into Rosetta and Rosetta 2 cells, BL21(DE3) derivatives designed to supply additional tRNAs for codons rarely used in *E. coli* and remove the need for codon optimisation of the recombinant protein, which is often necessary when attempting to express viral proteins in bacterial systems. Empty pET28aSUMO vector was also transformed as a negative control. Overnight cultures of LB with kanamycin and chloramphenicol were inoculated with starter cultures for empty vector and pETSUMO-M2-1, grown to OD_{600} 0.6-0.8, then induced with 100 μM IPTG overnight at either 18 °C, 30 °C, or 37 °C. Cells were pelleted by centrifugation, resuspended in 20 mM Tris pH 7.8, 500 mM NaCl, and lysed by sonication. Lysates were clarified by centrifugation and samples of insoluble and soluble fractions were analysed by SDS-PAGE. The abundance of bacterial proteins, and either low or no expression of M2-1, resulted in being unable to identify a band uniquely expressed by pETSUMO-M2-1.

100 μM IPTG was a relatively mild induction condition, and so increased IPTG concentrations were tried. Additionally, $ZnSO_4$ was added at induction, as supplementing cultures with zinc has been proposed to help M2-1 expression by aiding folding of the predicted N-terminal zinc finger motif (Tran et al., 2009). Inductions were performed with 300 μM IPTG and 1 mM $ZnSO_4$ overnight at 18 °C, 30 °C, or 37 °C in Rosetta and Rosetta 2 cells. Samples of insoluble and soluble fractions were analysed by SDS-PAGE, and samples from the 30 °C induction are shown in figure 3.7. For both bacterial strains a band with an apparent molecular weight of 43 kDa could be seen in M2-1 lanes (indicated by triangles in figure 3.7), in all fractions, but not empty vector lanes. Although His-SUMO-M2-1 had an expected molecular weight of 35 kDa, both published literature and our insect cell expression suggest that M2-1 may have a higher apparent molecular weight than predicted when analysed by SDS-PAGE, as does SUMO itself. This species was expressed at all induction temperatures, but with higher solubility at temperatures below 30 °C – little difference was observed between 18 °C and 30 °C.

A Rosetta, 300 μM IPTG



B Rosetta 2, 300 μM IPTG

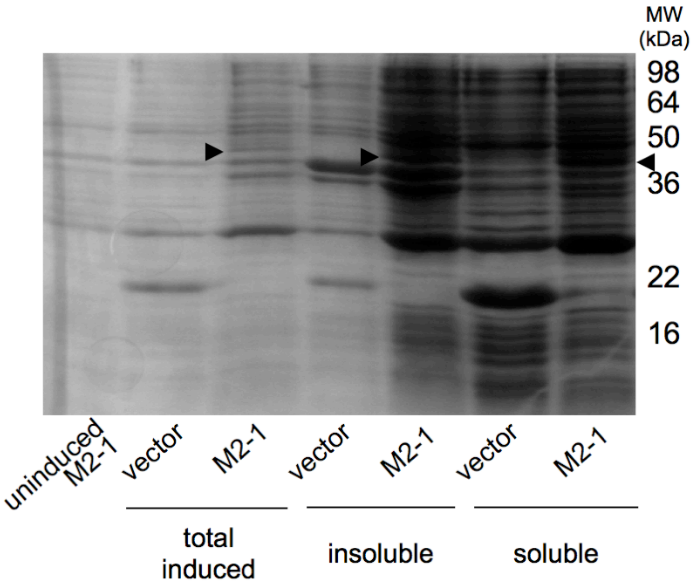


Figure 3.7 Preliminary expression of His-SUMO-M2-1.

Cultures were induced overnight at 30 °C with 300 μM IPTG and 1mM $ZnSO_4$ in **A)** Rosetta and **B)** Rosetta 2 cells. Triangles indicate the 43 kDa species suspected to be His-SUMO-M2-1. Lane labels: vector, empty pET28aSUMO; M2-1, pETSUMO-M2-1

The 43 kDa species could be enriched by Ni-affinity chromatography. Purification was as in Methods 2.2.9, except that the HiTrap IMAC FF 1 mL column (GE Healthcare) was charged with 0.1 M nickel chloride. Inductions were performed with 300 μM IPTG and 1 mM $ZnSO_4$ overnight at 30 °C, for pETSUMO-M2-1 and empty vector transformed Rosetta 2 cells. Pellets were resuspended in 20 mM Tris pH 7.8, 500 mM NaCl, and the soluble fraction was applied to the column. 6xHis-tagged protein was eluted from the column by stepwise imidazole gradient. Figure 3.8 shows the elutions from both inductions: the empty vector purification showed a protein of ~20 kDa eluting between 50 and 150 mM imidazole (figure 3.8 left), likely to be the His-SUMO-tag fused to the multiple cloning site; the M2-1 purification showed a protein of ~43 kDa eluting between 150 and 250 mM imidazole that was not present in the empty vector purification. This molecular weight corresponded to the 43 kDa band suspected to be His-SUMO-M2-1 in figure 3.7 and confirmed, at least, that this species was His-tagged.

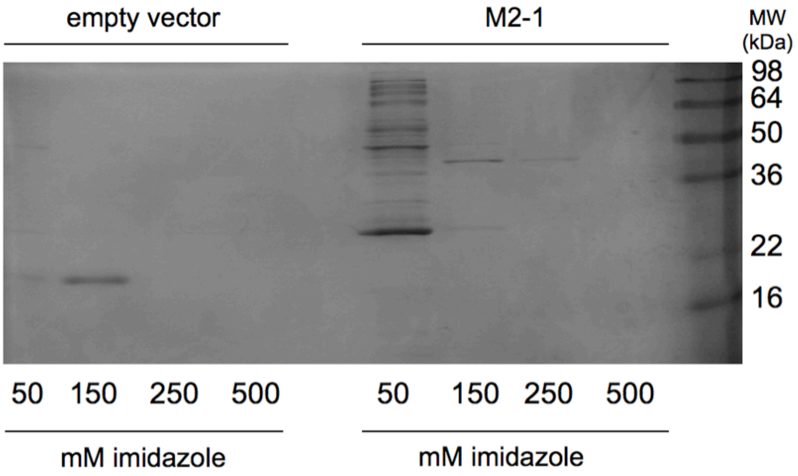


Figure 3.8 Enrichment of 6xHis-tagged proteins by Ni-chromatography. The gel shows the stepwise imidazole elution of protein expressed from empty pET28aSUMO vector and pETSUMO-M2-1: His-SUMO-M2-1 elutes between 150 and 250 mM imidazole at ~43 kDa (right), His-SUMO fused to the multiple cloning site elutes between 50 and 150 mM imidazole at ~20 kDa (left).

To confirm that M2-1 expressed from the pET28SUMO vector was folded and formed the tetramer predicted in the literature, the His-SUMO-tag was first removed from the N-terminus of M2-1. The elution from the Ni-chromatography was incubated with SUMO-protease; a tertiary structure-specific cysteine protease that cleaves SUMO without leaving any extraneous residues and does not demonstrate erroneous cleavage (Butt et al., 2005). Figure 3.9 shows that addition of SUMO protease resulted in cleavage of the 43 kDa species into a 28 kDa species, suspected to be M2-1, and a 20 kDa species, suspected to be the His-SUMO tag as this matched the protein expressed during vector only inductions. This reaction was then further purified by size exclusion chromatography using a Superdex S75 column. Figure 3.10 shows a representative trace monitoring the elution from the size exclusion column by absorbance at 280 nm (figure 3.10 A) and SDS-PAGE analysis of the fractions collected during the purification (figure 3.10 B). M2-1 protein eluted in both peaks 1 and 2, however the SDS-PAGE gels showed that fractions corresponding to peak 1 also contained additional bands; possibly co-eluting contaminants or higher order oligomers that have not been disrupted by the denaturing gel conditions. Comparison with the absorbance trace showed considerably higher absorbance for peak 1, despite the similar distribution of protein according to SDS-PAGE. This may indicate the presence of nucleic acid, which also absorbs near 280 nm wavelengths, and this may be mediating some of the higher order species seen. The His-SUMO tag eluted at 185 mL. Using a calibration curve for the size exclusion column, generated by running a set of standards with known molecular weights, the elution volumes for peaks 1-3 were equated to approximate molecular weights. Peak 1 was ~150 kDa and peak 2 was ~107 kDa, and whilst both of these values were higher than the predicted 90 kDa for an M2-1 tetramer, they agreed with the higher-than-expected values seen with techniques such as dynamic light scattering and analytical ultracentrifugation which have previously suggested an extended conformation is responsible for the high observed molecular weights (Tran et al., 2009). Peak 3 was ~ 20 kDa, which corresponded to the expected molecular weigh for His-SUMO and the size of the band observed by SDS-PAGE.

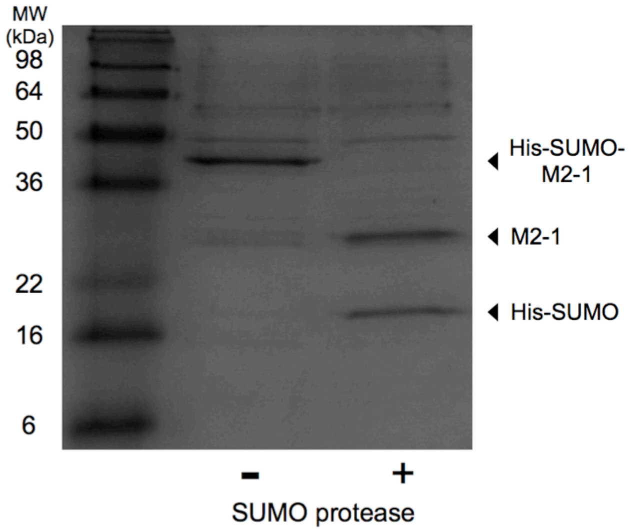


Figure 3.9 Removal of His-SUMO-tag by SUMO protease cleavage.

Ni-chromatography purified His-SUMO-M2-1 was incubated without (-) and with (+) SUMO protease. Cleavage of His-SUMO-M2-1 (43 kDa) results in M2-1 (~28 kDa) and His-SUMO (~20 kDa).

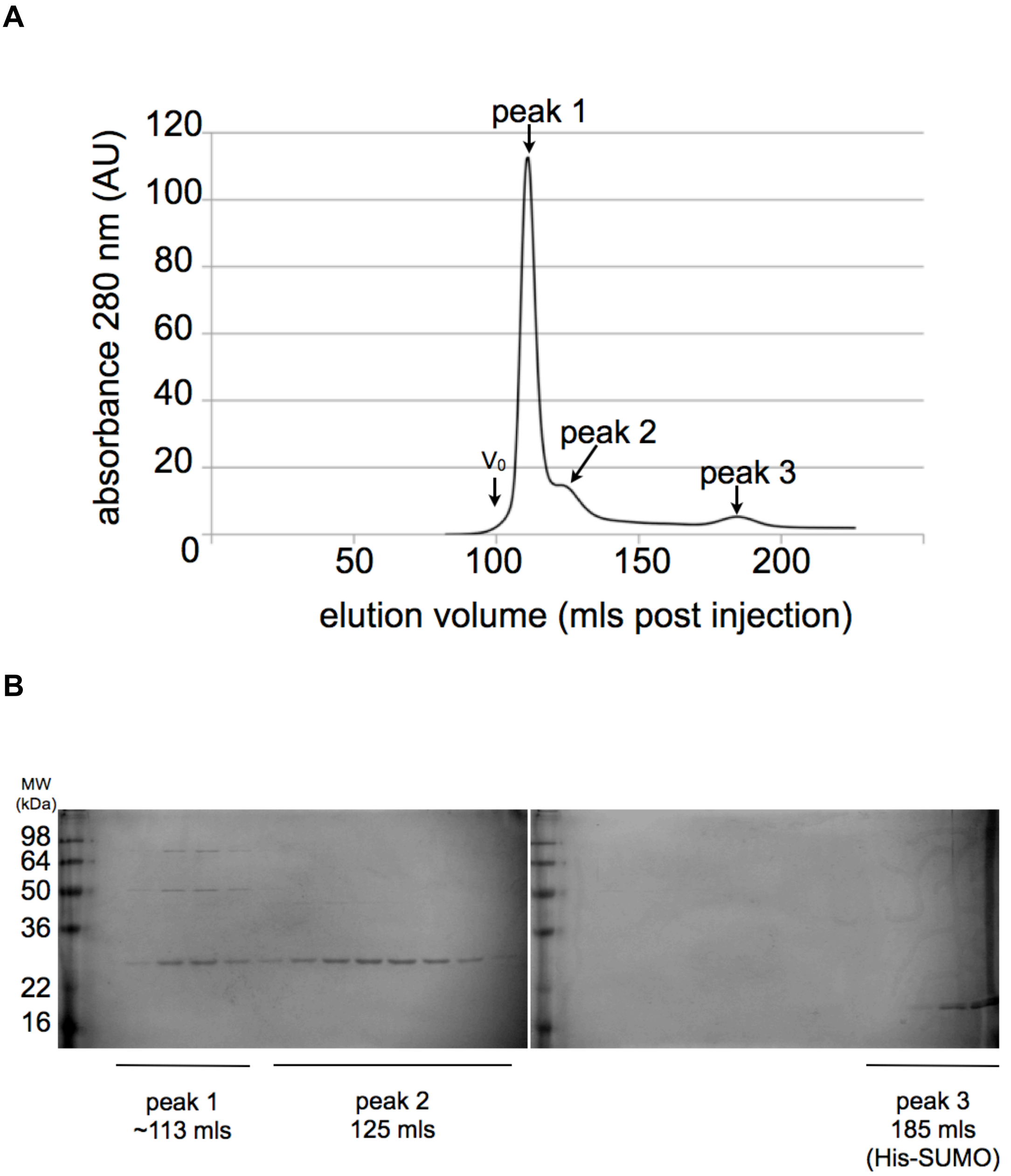


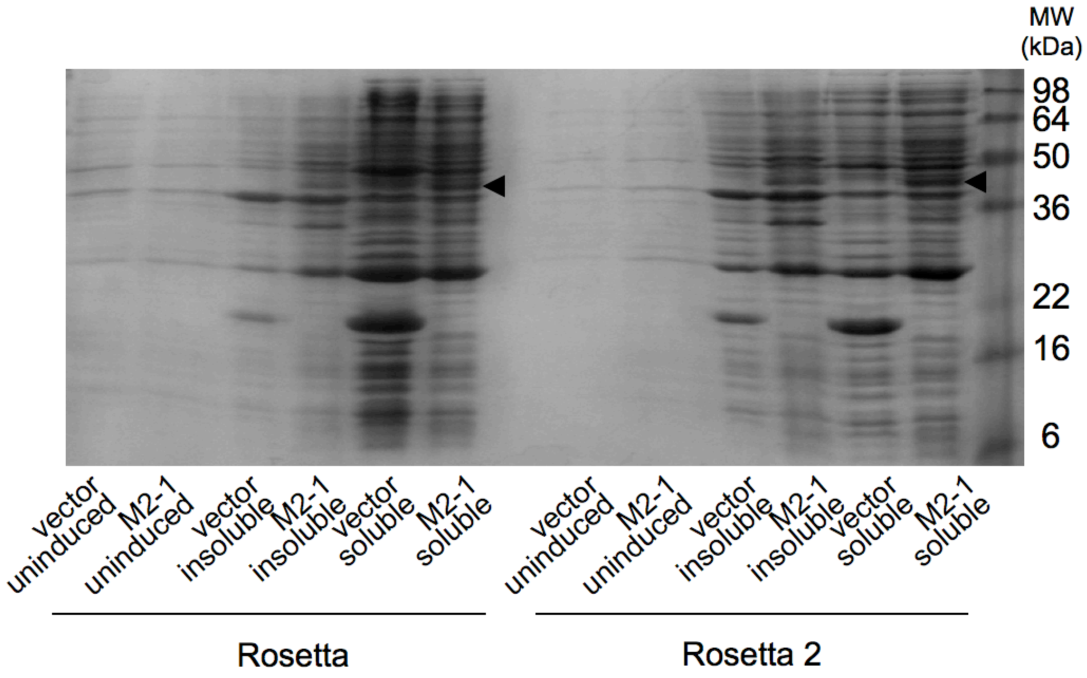
Figure 3.10 Size exclusion chromatography of cleaved His-SUMO-M2-1. Size exclusion chromatography (SEC) using a Superdex S75 column. **A)** A_{280} trace of eluting protein from size exclusion column. V_0 indicates void volume. **B)** SDS-PAGE analysis of SEC fractions labelled with the corresponding peak from A) and the elution volume. A 28 kDa protein (M2-1) eluted in peaks 1 and 2 at 113 and 125 mL, respectively, and a 20 kDa protein (His-SUMO-tag) eluted in peak 3 at 185 mL.

3.6.3 Optimising expression of His-SUMO-M2-1

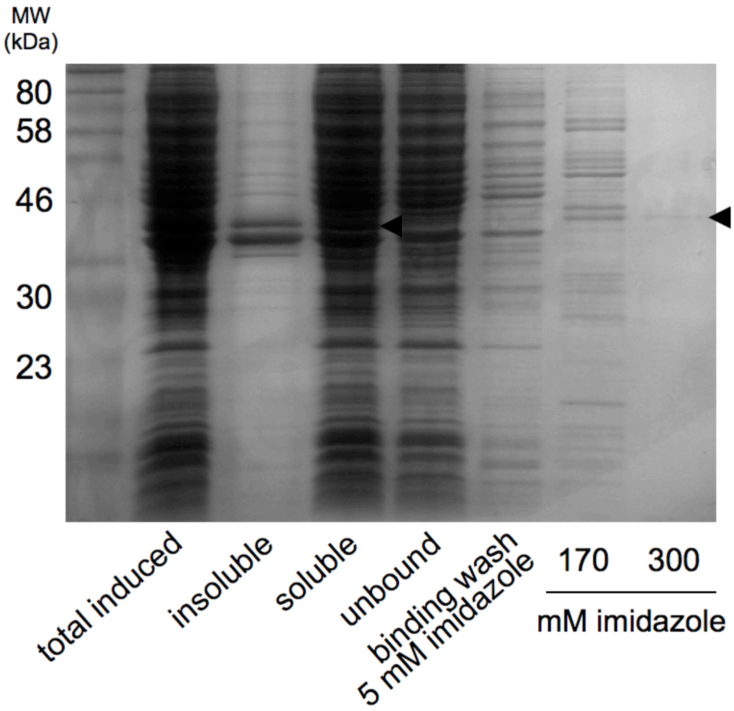
Induction with 500 μM and 1 mM IPTG was also tested, with only 1 mM IPTG resulting in slightly higher levels of soluble M2-1 expression compared to 300 μM IPTG induction as determined by gel band size (comparing figures 3.7 and 3.11 A). Slightly more His-SUMO-M2-1 could be seen in the soluble fraction of Rosetta 2 cells induced at 1 mM IPTG, than with Rosetta cells (figure 3.11 A). Multiple induction temperatures, late-log phase induction (OD_{600} 1-1.2), and various molarities of IPTG (100 μM to 1 mM) and $ZnSO_4$ (50 μM to 1 mM) were all tested to increase total expression levels of His-SUMO-M2-1. High-level expression, i.e. the predominant species in the total induced sample, could not be achieved in Rosetta or Rosetta 2 cells.

Next, pETSUMO-M2-1 was then retransformed into two other *E. coli* expression strains: OverExpress C41 (DE3) ('C41'), engineered to withstand over expression-induced toxicity; BL21 (DE3) Gold ('BL21 Gold'), a standard protease-deficient expression strain. Work using C41 cells to express M2-1 was published during this project (Esperante et al., 2011) and so expression was tested as in this publication: 18 hour 37 °C induction with 300 μM IPTG. Low expression levels required Ni-chromatography purification before a band presumably corresponding to His-SUMO-M2-1 could be identified by SDS-PAGE analysis (figure 3.11 B), and so these cells were not optimised further. Expression in BL21 Gold cells was tested by inducing at 30 °C overnight or for 6 hours, with 1 mM IPTG and 1 mM $ZnSO_4$ (the conditions optimal up to this point). Cells were pelleted and resuspended in 20 mM Tris pH 7.8, 500 mM NaCl, and lysed by sonication. Considerably higher total expression of His-SUMO-M2-1 was observed (figure 3.11 C) although the majority of protein was insoluble. It was noted that this *E. coli* strain was significantly more difficult to sonicate, and so incomplete lysis may have been responsible for low abundance of M2-1 in the soluble fraction.

A Rosetta and Rosetta 2, 1 mM IPTG



B C41, 300 μM IPTG



C BL21 Gold, 1 mM IPTG

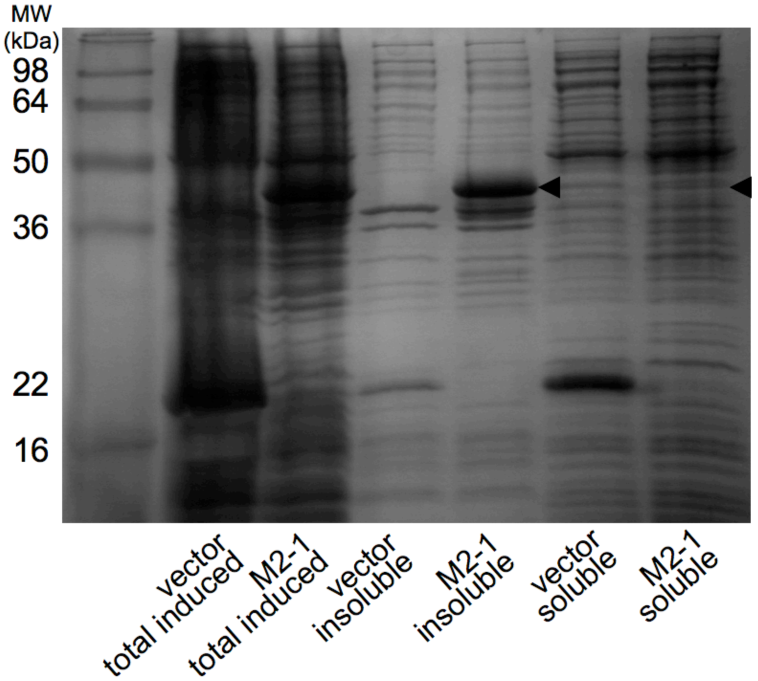


Figure 3.11 Expression optimisation for His-SUMO-M2-1.

A) Slightly improved expression was achieved in Rosetta 2 cells with 1 mM IPTG induction in. **B)** Low expression and poor solubility was seen using C41 cells. **C)** Highest levels of total expression were achieved with BL21 Gold cells, however protein is mostly insoluble. Lane labels: vector, empty pET28aSUMO; M2-1, pETSUMO-M2-1. Triangles indicate His-SUMO-M2-1 at ~43 kDa.

3.6.4 Confirmation of His-SUMO-M2-1 identity

Confirmation of the identity of M2-1 expressed in this system was by mass spectrometric analysis. Protein identification by LC-MS/MS was performed as a service by James Ault (MS Facility, University of Leeds). A SDS-PAGE band of His-SUMO-M2-1 was excised and subjected to in-gel trypsin cleavage. Identification was made by database searching for proteins containing the peptides identified in the mass spectrometry. Figure 3.12 shows a sequence coverage map for both HRSV strain A M2-1 protein and SMT3 SUMO protein.

```
VMA2_HRSVA Coverage Map

1       MSRRNPCKFE   IRGHCLNGKR   CHFSHNYFEW   PPHALLVRQN   FMLNRILKSM
51      DKSIDTLSEI   SGAAELDRTE   EYALGVVGVL   ESYIGSINNI   TKQSACVAMS
101     KLLTELNSDD   IKKLRDNEEL   NSPKIRVYNT   VISYIESNRK   NNKQTIHLLK
151     RLPADVLKKT   IKNTLDIHKS   ITINNPKEST   VSDTNDHAKN   NDTT
```

```
SMT3_YEAST Coverage Map

1       MSDSEVNQEA   KPEVKPEVKP   ETHINLKVSD   GSSEIFFKIK   KTTPLRRLME
51      AFAKRQGKEM   DSLRFLYDGI   RIQADQTPED   LDMEDNDIIE   AHREQIGGAT
101     Y
```

Regions of the protein sequence that match peptides are highlighted in colour, according to the key below

Matched to a peptide	🟦
Matched to a partial peptide	🟥
Matched to a modified peptide	🟩
Matched to a partial modified peptide	🟨

Highlights are transparent, so that regions where peptides overlap are visible on the coverage map. E.g. an overlap between a standard peptide and a modified peptide would appear as shown: 🟦🟩

Figure 3.12 Sequence coverage maps for LC-MS/MS protein identification. Mass spectrometry confirmed the identity of the 43 kDa species as His-SUMO-M2-1. Peptides from both HRSV strain A M2-1 and SMT3 (SUMO) from *S. cerevisiae* were identified.

3.6.5 Increasing solubility of His-SUMO-M2-1 in BL21 Gold cells

Higher total expression of His-SUMO-M2-1 had been observed in BL21 Gold cells and so, despite poor solubility, this strain was used for further optimisation.

3.6.5.1 Additives to induction and buffers

The issue of insufficient lysis was addressed first, by adding 1 mg/mL lysozyme to the lysis buffer and incubating on ice for 30 mins, as well as increasing sonication. A number of solubility aids were tested: 0.1 % (v/v) Triton X-100 and 10 % (v/v) glycerol in the lysis buffer (figure 3.13 A) improved solubility, as did adding 1 mM DTT (figure 3.13 B); 10 mM $MgSO_4$ and 1 mM $CaCl_2$ in the lysis buffer did not improve solubility, however they were found to improve stability of M2-1 through the purification process. Reducing $ZnSO_4$ at induction to 50 μM also increased the proportion of M2-1 in the soluble fraction (figure 3.13 B). Finally, optimisation of IPTG levels for BL21 Gold cell inductions showed that a lower concentration, 340 μM, increased solubility of His-SUMO-M2-1.

Several changes to the expression or purification of His-SUMO-M2-1 were found to have no positive effect: substituting Triton-X100 for 10 mM CHAPS; adding 10 mM $MgSO_4$ to LB medium during culture growth; adding 50 mM free L-arginine and L-glutamine, which have been shown to improve solubility for some proteins (Golovanov et al., 2004).

3.6.5.2 Autoinduction

Autoinduction, a technique commonly used to increase solubility of recombinant proteins, was attempted as in Methods 2.2.2.2. Minimal media starter cultures were used to inoculate autoinduction media cultures 1:1000 and these were grown for 24 hours at 37 °C. Cultures were processed as IPTG-induced cultures and samples from insoluble and soluble fractions were analysed by SDS-PAGE. His-SUMO-M2-1 was found to be more insoluble by this method than standard IPTG-induced expression (figure 3.14).

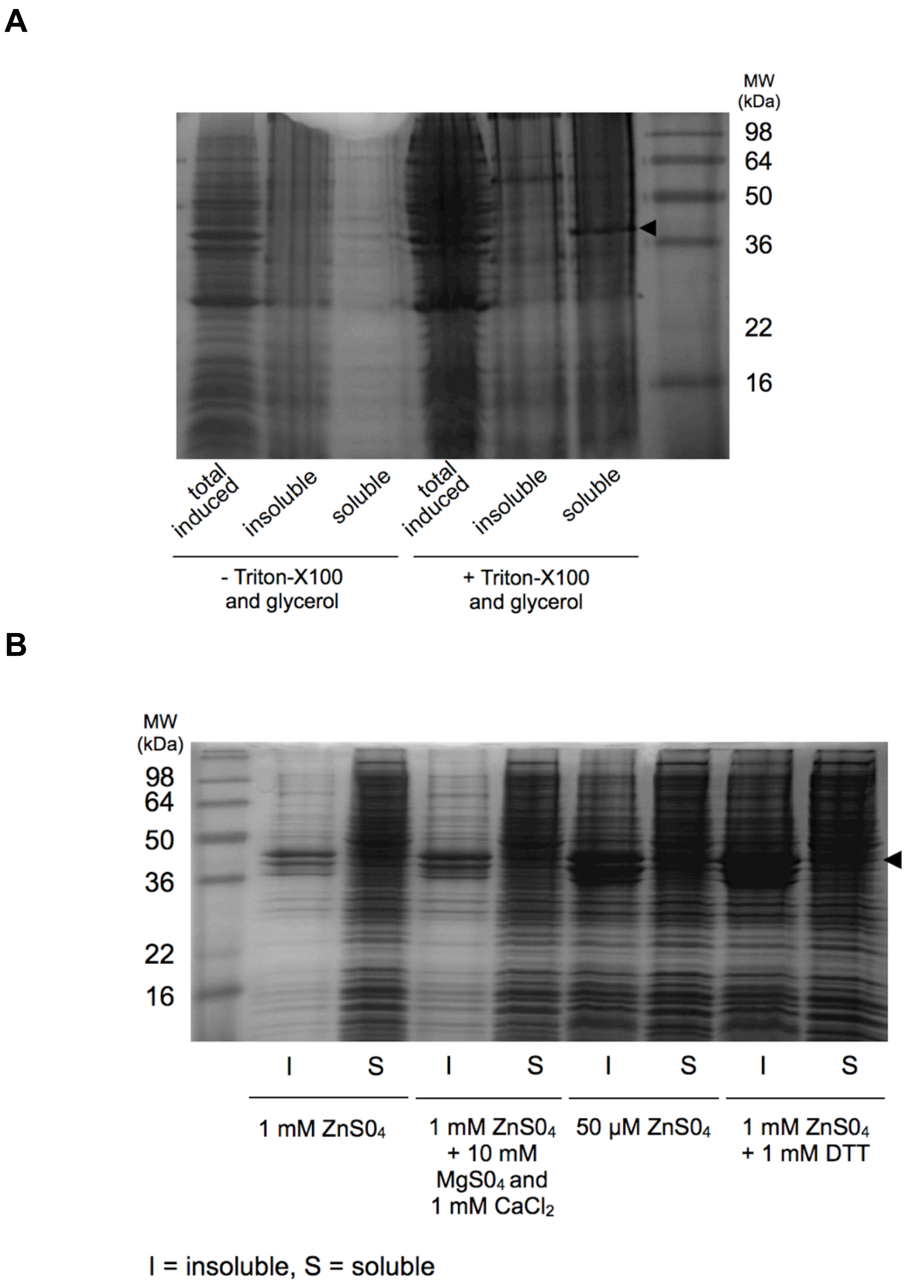


Figure 3.13 Improving His-SUMO-M2-1 solubility.

A) Addition of 0.1 % (v/v) Triton-X100 and 10 % (v/v) glycerol increase the proportion of M2-1 found in the soluble fraction, however glycerol caused smearing of SDS-PAGE gels. **B)** Reducing $ZnSO_4$ levels at induction and including 1 mM DTT both improved M2-1 solubility. 10 mM $MgSO_4$ and 1 mM $CaCl_2$ did not affect solubility, but were found to improve long-term stability of M2-1.

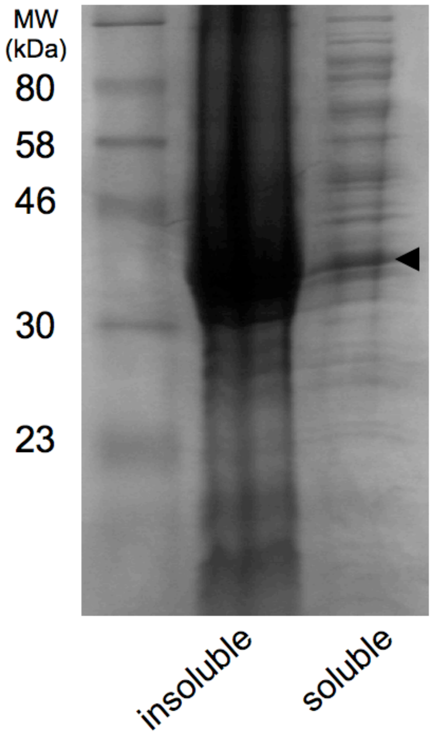


Figure 3.14 Expression of His-SUMO-M2-1 by autoinduction.

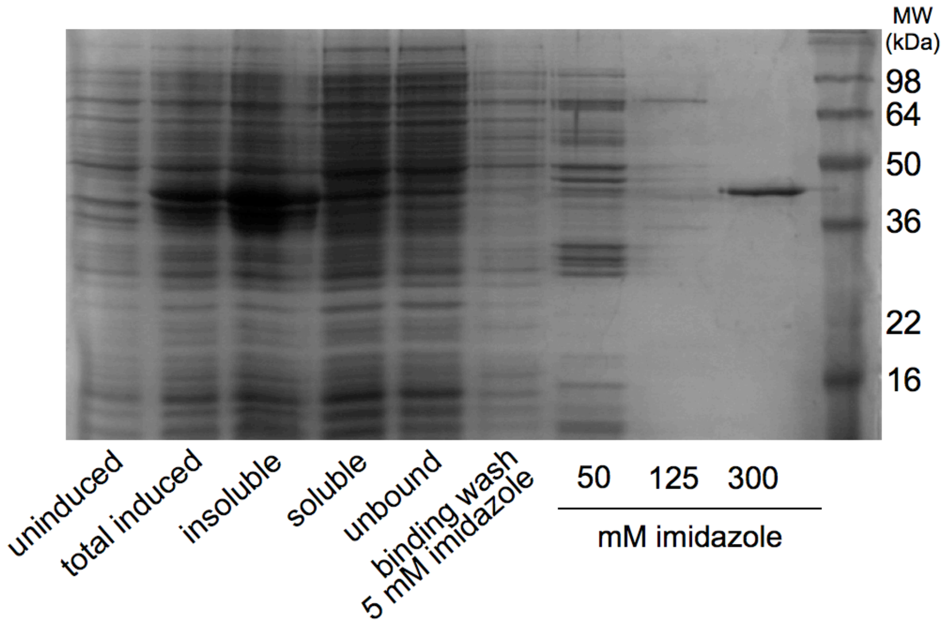
Induction of protein expression follows exhaustion of the limited glucose supply and metabolism of lactose instead, relieving repression of the lac operator element in the promoter. This method reduced solubility of M2-1 in BL21 Gold cells, as shown by the majority of protein found in the insoluble fraction. Triangle indicates His-SUMO-M2-1.

3.6.5.3 C96S mutant

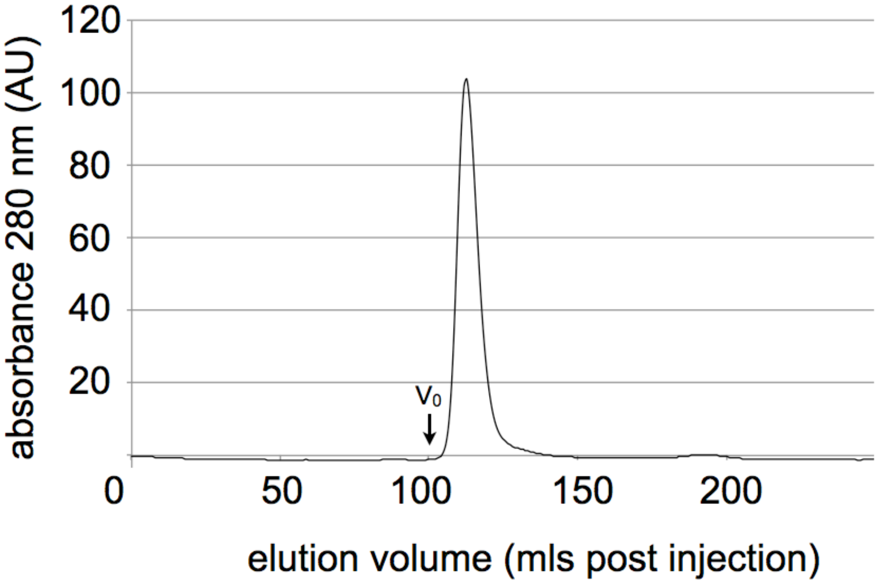
RSV M2-1 contains four conserved cysteine residues, only one of which is not assigned to the Cys_3-His_1 zinc finger motif; C96. This unpaired cysteine may be capable of forming disulphide bonds with either other M2-1 monomers or host cell proteins. Mutation of unpaired cysteine has been shown to improve solubility of recombinant proteins in some instances, and so a C96 to serine mutant was cloned into pETSUMO-M2-1 (mutant referred to as 'C96S'). C96S expressed to similar levels as wild type (WT) M2-1 but did not show a significant increase in solubility. The protein could be successfully purified by Ni-chromatography (figure 3.15 A) but the mutation conferred no additional long-term stability, as C96S precipitated after 3-5 days at 4 °C. Further purification of C96S by size exclusion chromatography showed that all protein eluted at ~113 mL (figure 3.15 B) and SDS-PAGE analysis of the fractions collected show significant contaminants, including an unknown band of ~33 kDa present in all fractions (figure 3.15 C). C96S M2-1 was therefore deemed to be unsuitable for further optimisation.

This result is unsurprising in light of the M2-1 crystal structure presented in Chapter 4, which revealed that C96 is buried within the core domain.

A


MW
(kDa)
98
64
50
36
22
16
uninduced
total induced
insoluble
soluble
unbound
binding wash
5 mM imidazole
50
125
300
mM imidazole

B


absorbance 280 nm (AU)
120
100
80
60
40
20
0
50
100
150
200
V0
elution volume (mls post injection)

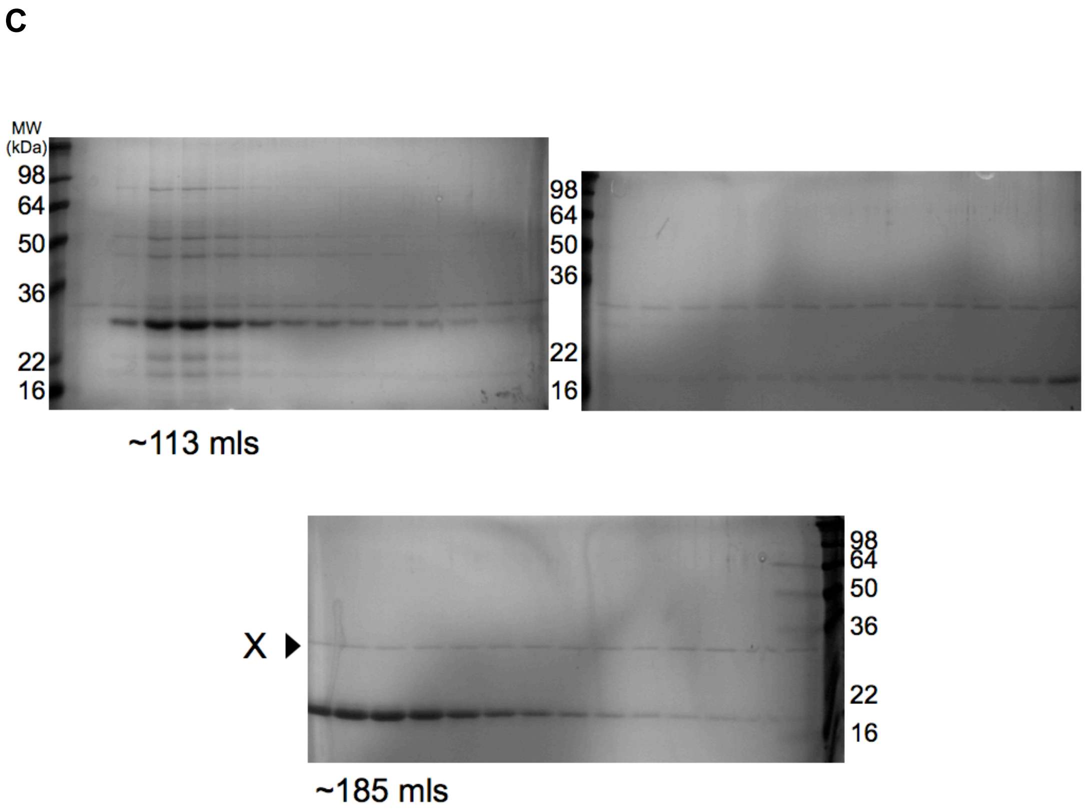


Figure 3.15 C96S M2-1 expression and size exclusion chromatography.

A) BL21 Gold expression of C96S, overnight at 30 °C with 1 mM IPTG, and purification by Ni-chromatography. **B)** and **C)** Size exclusion chromatography using a Superdex S75 column: B) shows the 280 nm absorbance trace, with C96S M2-1 eluting after 113 mL. V_0 indicates void volume; C) shows SDS-PAGE analysis of the corresponding fractions, and highlights the considerable contaminants including an unknown species 'X' of 33kDa eluting in all fractions.

3.6.6 Improving purification of His-SUMO-M2-1

Increasing the proportion of soluble His-SUMO-M2-1 that was purified by Ni-affinity chromatography was first addressed by trying to increase binding to the Ni column. Passing clarified lysates through the column multiple times had no affect on increasing binding of M2-1. Ammonium sulphate precipitation was used to try and reduce the amount of bacterial protein present in the lysate, to prevent any possible competition from histidine-rich proteins. 0.67 volumes of saturated ammonium sulphate solution were added dropwise to the clarified lysate at 4 °C with stirring for 2 hours. Precipitated protein was pelleted by centrifugation and SDS-PAGE analysis revealed that His-SUMO-M2-1 had been successfully precipitated. The pellet was resuspended in 50 mM Tris pH 7, 200 mM NaCl, 1 mM DTT, and dialysed against the same buffer to remove ammonium sulphate. This caused a significant amount of protein to precipitate during dialysis, and analysis by SDS-PAGE showed that His-SUMO-M2-1 was insoluble as a result of this process (figure 3.16). Another technique used to increase His-SUMO-M2-1 binding to the Ni column was partial unfolding with urea. 200 mM urea was added to clarified lysates and incubated at 4 °C for 30 mins before being applied to a Ni column. This did not increase binding of His-SUMO-M2-1 to the column, however.

Purity and yield of cleaved, eluted M2-1 protein were poor after Ni-affinity chromatography. The addition of 5 mM imidazole to lysates before binding reduced some non-specific binding of bacterial proteins, as did additional washes of the column, however some contaminants remained that could not be removed even with further purification by size exclusion chromatography (figure 3.10 B). Typical yields of cleaved, SEC purified M2-1 expressed from pETSUMO-M2-1 in BL21 Gold cells were 0.5-1 mg/L bacterial culture, which was insufficient for structural analysis.

Another considerable barrier to further use of His-SUMO-M2-1 was the stability of the eluted protein when imidazole was removed by dialysis. Both the fusion and cleaved-M2-1 precipitated upon removal of imidazole, when lysis and Ni-affinity purification were performed in Tris/NaCl-based buffers. The proportion of M2-1 protein that precipitated was decreased when using phosphate-based buffers [1x phosphate-buffered saline, 10 % (v/v) glycerol, 0.1 % Triton X-100], however this buffer system is not ideal for crystallographic studies due to the formation of salt crystals.

It was decided at this point that no further optimisation was reasonably possible with M2-1 in pET28aSUMO, and so the ORF was subcloned into an alternative vector system.

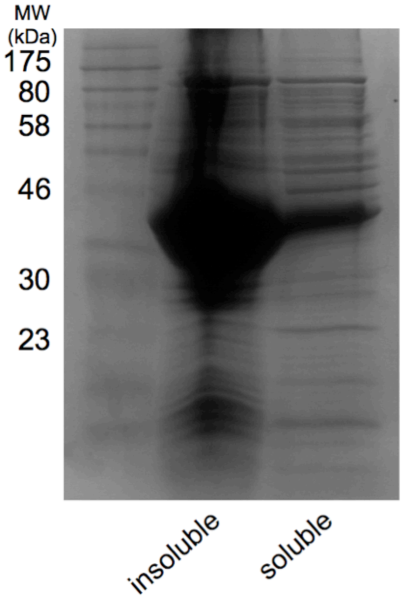


Figure 3.16 Solubility of His-SUMO-M2-1 post-ammonium sulphate precipitation.

Dialysis of ammonium sulphate from precipitated and resuspended protein caused His-SUMO-M2-1 to be predominantly insoluble (43 kDa).

3.6.7 Subcloning of M2-1 ORF into GST-tag vector

To test expression of M2-1 with an alternative affinity tag, the M2-1 ORF was subcloned into a pGEX vector that would express M2-1 as an N-terminal fusion with glutathione S-transferase (GST) – an alternative tag known to aid in the solubility of recombinant proteins that allows purification by glutathione affinity chromatography.

M2-1 ORF was subcloned by directly digesting the fragment from pETSUMO-M2-1 and ligating into pGEX4T2 and pGEX6P2, vectors with thrombin and HRV 3C protease cleavage sites, respectively. Double restriction digest with *BamHI* and *XhoI* of both pETSUMO-M2-1 and empty pGEX6P2/4T2 vectors was performed and the appropriate fragments purified by agarose gel electrophoresis and gel extraction. M2-1 ORF was ligated into the empty pGEX vectors and transformed into DH5α cells. 6 colonies for each vector were minipreped and diagnostically digested with *BamHI* and *XhoI* to confirm the presence of the M2-1 ORF. Results for pGEX6P2 are shown in figure 3.17; 5 of the 6 colonies were positive for the insert, and orientation was confirmed by DNA sequencing. This construct is referred to as pGEX-M2-1 and resulted in N-terminally GST-tagged M2-1, with 5 non-native residues (Gly-Pro-Leu-Gly-Ser) remaining after GST cleavage with HRV 3C protease. pGEX4T2 cloning was also successful, but was deemed not cost-effective and expression was not optimised with this construct.

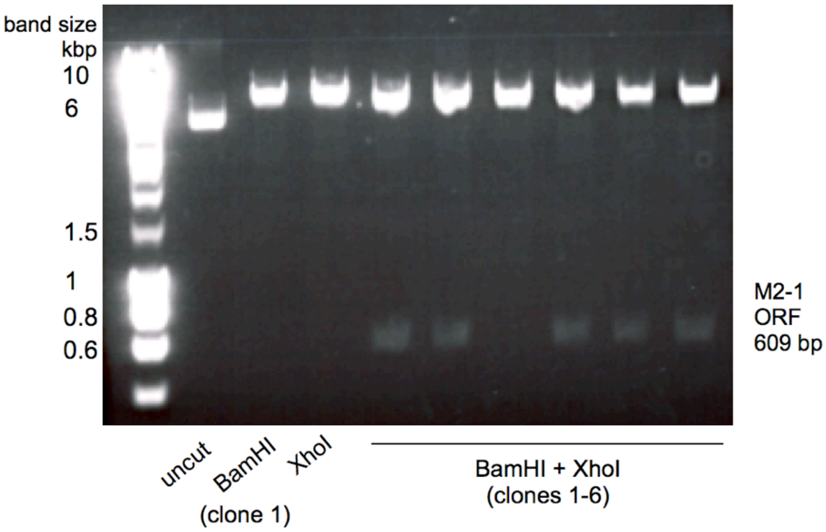


Figure 3.17 Identification of pGEX 6P2 clones containing M2-1 ORF.
6 clones were minipreped and diagnostically digested with *BamHI* and *XhoI*, with single digests and uncut vector shown here for clone 1 as controls for enzyme activity. 5 of 6 clones contained the M2-1 ORF (609 bp).

3.6.8 Optimising GST-M2-1 expression and purification

Initial expression of GST-M2-1 from pGEX-M2-1 was tested using the best conditions achieved for His-SUMO-M2-1 expression: overnight induction at 30 °C in BL21 Gold cells with 340 μM IPTG and 50 μM $ZnSO_4$; cell recovery by centrifugation and resuspension in 1x PBS, 0.1 % (v/v) Triton-X100, 10 % (v/v) glycerol, 1 mM DTT, 1 mg/mL lysozyme; incubation on ice for 30 mins followed by final lysis by sonication and clarification of lystes by centrifugation. Cells were also resuspended in a Tris buffer reported in the literature for comparison; this buffer contained many of the components found to improve His-SUMO-M2-1 solubility and stability, with the advantage of not containing phosphate buffer [50 mM Tris pH 7.4, 1 M NaCl, 1 mM DTT, 2 % (v/v) Triton X-100, 10 mM $MgSO_4$, 1 mM $CaCl_2$] (Tran et al., 2009).

SDS-PAGE analysis revealed that GST-M2-1 protein expressed well and was predominantly soluble irrespective of the buffer used to resuspend cells, and has an apparent molecular weight of 48 kDa as predicted (figure 3.18). The Tris buffer, supplemented with 1 mg/mL lysozyme and 1x complete protease inhibitors EDTA-free (Roche), was adopted from this point onwards to avoid incompatibility with downstream crystal trials.

Purification of GST-M2-1 was by glutathione sepharose affinity chromatography. Briefly, clarified lysate was applied to resin in batch and incubated at 4 °C with rotation. It was found that due to the slow binding kinetics of GST and glutathione, and the additional difficulty of possible higher-order oligomers forming through dimers of tetramers via GST, overnight binding was required – although this still resulted in a considerable amount of GST-M2-1 remaining unbound (figure 3.19). Resin was washed thoroughly in lysis buffer, then equilibrated in HRC 3C protease cleavage buffer and incubated with 3C protease. This on-resin cleavage resulted in the highly concentrated band of protein at ~26 kDa when visualised by SDS-PAGE, containing M2-1. Expression and purification by this method produced 10-15 mg of M2-1 per litre of culture, a >15-fold increase over the highest yields achieved with His-SUMO-tagged M2-1. No further optimisation of expression or affinity purification was required.

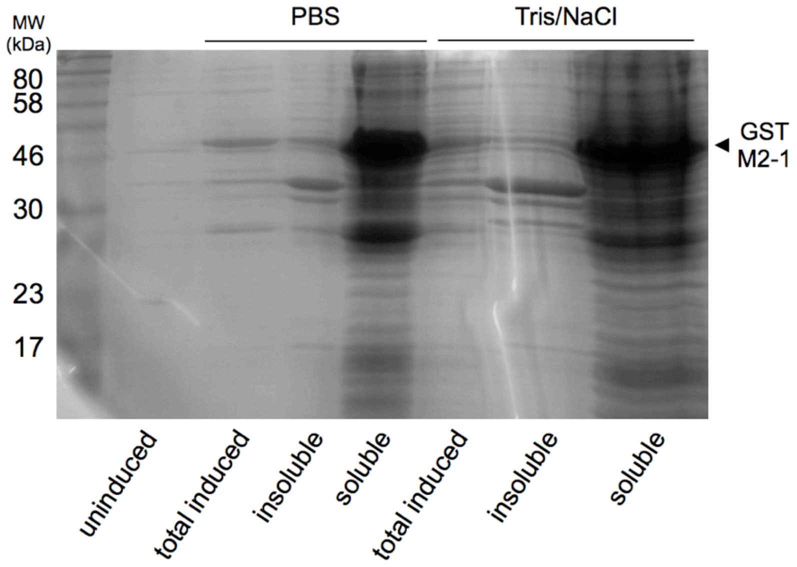


Figure 3.18 Expression of GST-M2-1 and solubility in two buffer systems.
Expression from pGEXM2-1 in BL21 Gold cells overnight at 30 °C with 340 μM IPTG and 50 μM $ZnSO_4$, followed by pellet resuspension in PBS buffer [1x PBS, 0.1 % (v/v) Triton-X100, 10 % (v/v) glycerol, 1 mM DTT, 1 mg/mL lysozyme] or Tris/NaCl buffer [50 mM Tris pH 7.4, 1 M NaCl, 1 mM DTT, 2 % (v/v) Triton X-100, 10 mM $MgSO_4$, 1 mM $CaCl_2$, 1 mg/mL lysozyme]. GST-M2-1 has a molecular weight of 48 kDa.

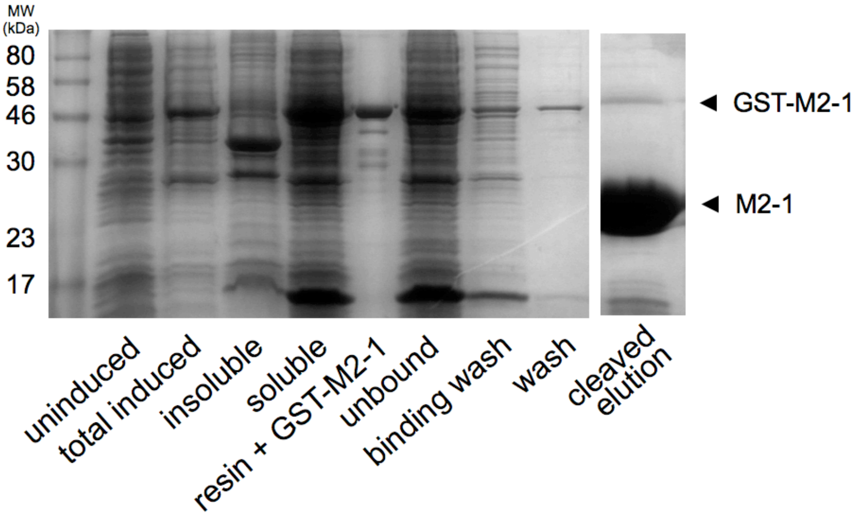


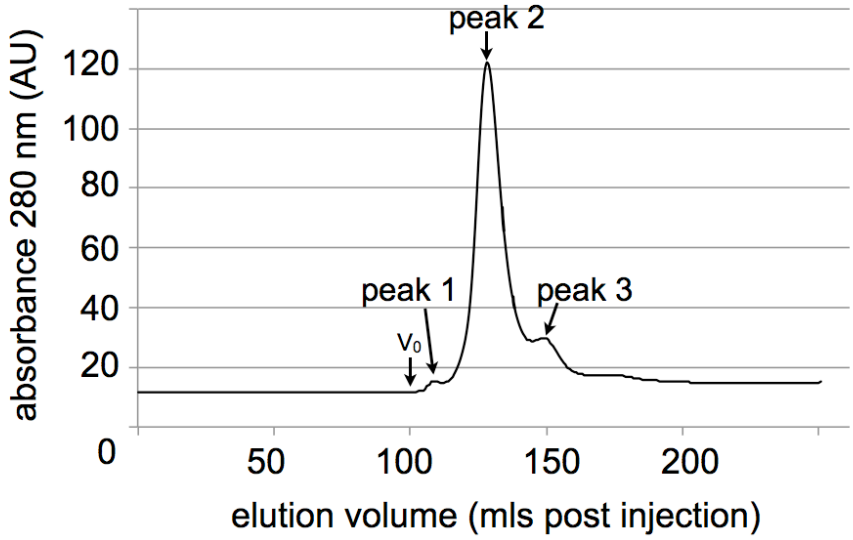
Figure 3.19 Purification of GST-M2-1.
GST-M2-1 appears as a 48 kDa, predominantly soluble species. GST-M2-1 was bound to resin overnight, however a significant proportion of protein remains in the unbound fraction. Resin was washed and GST cleaved from M2-1 on the resin by HRC 3C protease, resulting elution of M2-1 at ~26 kDa.

3.6.9 Size exclusion chromatography

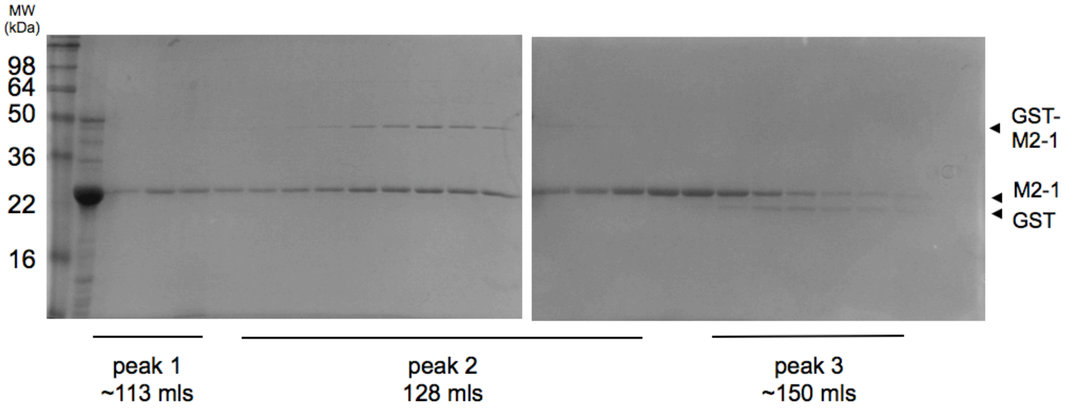
Further purification of M2-1 was required after proteolytic GST removal, to isolate the tetramer with GST fully cleaved from each monomer. Size exclusion chromatography was used as in Methods 2.2.4. Briefly, the column was equilibrated and run in 50 mM Tris pH 7.5, 150 mM NaCl, 1 mM DTT (the 3C protease cleavage buffer), the protein injected with a 5 mL loop, and the column run at 2 mL/min until a full column-volume of buffer had eluted (320 mL). Fractions were collected from the void volume (~100 mL) onwards. A typical A_{280} trace is shown in figure 3.20 A and SDS-PAGE analysis of the corresponding fractions is shown in figure 3.20 B. M2-1 appeared to elute in all fractions analysed between 113 and 160 mLs, although predominantly in peak 2 at 128 mLs, which corresponds to a 90 kDa species as the molecular weight predicts. The presence of M2-1 across multiple peaks is likely as a result of incomplete cleavage of GST from all 4 monomers of the tetramer, creating a heterogeneous population of GST-M2-1 oligomers. There is also a 48 kDa species co-eluting in peak 2, which is likely uncleaved GST-M2-1. Some cleaved GST that has eluted from the resin can be seen in peak 3 at 150 mLs, corresponding to 52 kDa as expected for a GST dimer.

The ratio of HRV 3C protease to GST-M2-1 was increased from 1:100 to 1:25, volume to volume of a 4 mg/mL solution of protease, in an attempt to reduce the heterogeneity of species present in the size exclusion chromatography. Increasing incubation to 24 hours at 4 °C was also tried. Whilst complete cleavage could not be achieved, the proportion of protein present in peak 2 of the size exclusion was increased to >95 % (figure 3.20 C).

A


peak 2
peak 1
peak 3
V0
absorbance 280 nm (AU)
0
20
40
60
80
100
120
50
100
150
200
elution volume (mls post injection)

B


MW (kDa)
98
64
50
36
22
16
GST-M2-1
M2-1
GST
peak 1 ~113 mls
peak 2 128 mls
peak 3 ~150 mls

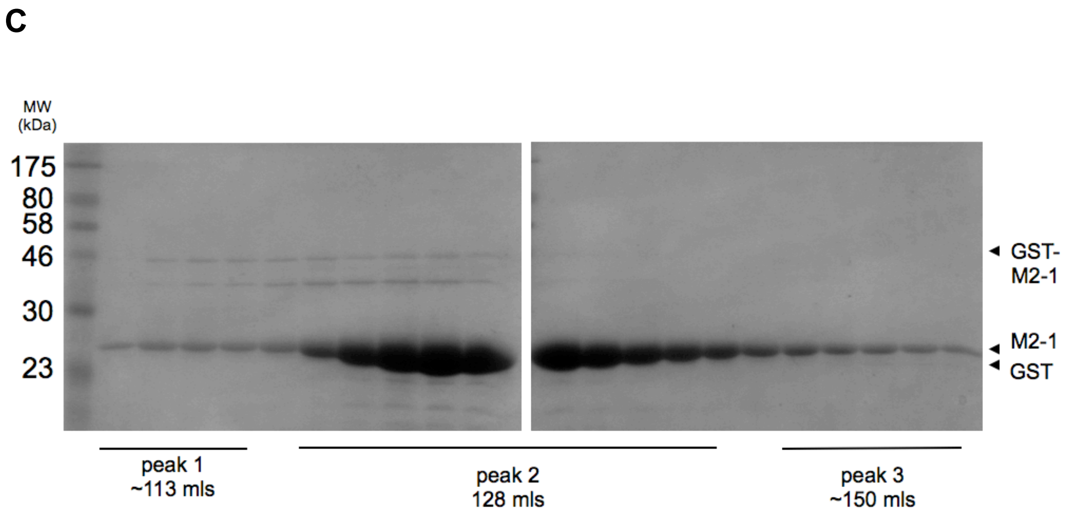


Figure 3.20 Size exclusion chromatography of pGEX-M2-1 expressed, cleaved M2-1.

A) A_{280} absorbance trace shows protein eluting in 3 peaks, with most protein in peak 2 at 128 mLs, 90 kDa. B) SDS-PAGE analysis of fractions from SEC with the corresponding peaks labelled. M2-1, 26 kDa, appears in all fractions. C) The proportion of peak 2 protein attributed to cleaved M2-1 was increased by optimising HRV 3C protease cleavage.

3.6.10 Ion exchange chromatography

GST and M2-1 have significantly different pIs, ~6 and 9.1 respectively, and so should be separated by ion exchange chromatography. A cation exchange column was used with all buffers at pH 7.5 so that negatively charged M2-1 would bind the column, whilst positively charged GST would pass through. However, comparison of M2-1 before and after ion exchange revealed that very few contaminants had been removed (figure 3.21), despite the fact that M2-1 clearly bound the column and eluted at ~300 mM NaCl. In addition to providing no purification benefit, approximately 30 % of M2-1 was lost during this process, possibly due to column saturation.

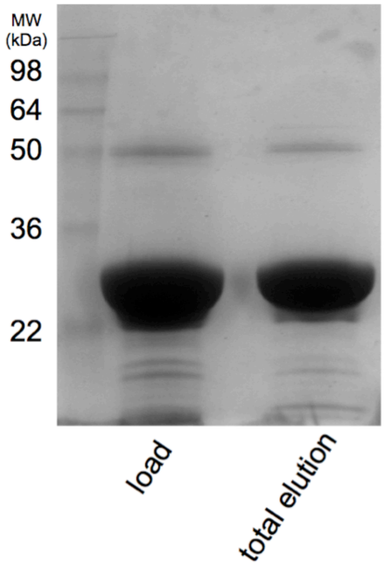


Figure 3.21 Further purification of M2-1 by ion exchange chromatography.

Comparing the IEX column load and pooled elution fractions, shows little difference i.e. very few contaminants have been removed. Cleaved M2-1 at 26 kDa.

3.6.11 Characterisation of pGEX-expressed M2-1 protein

Expression and purification of M2-1 by the above method (pGEXM2-1 expression in BL21 Gold cells, glutathione sepharose affinity chromatography followed by size exclusion chromatography) yielded ~5 mg of >95 % pure tetrameric M2-1 per litre of bacterial culture. Confirmation of accurate mass of monomer and tetramer were acquired by mass spectrometry, and the removal of bacterially-derived RNA bound to M2-1 confirmed by agarose gel electrophoresis.

3.6.11.1 Mass spectrometry

Accurate mass determination of M2-1 eluting in 'peak 2' during size exclusion chromatography (i.e. the 90 kDa species assumed to be tetrameric M2-1) was performed using nano-electrospray-ionization mass spectrometry in positive ionization mode, by Helen Beeston (Alison Ashcroft group, University of Leeds) as in Methods 2.2.7. Denaturing MS of M2-1 in acetonitrile was used to determine the mass of the monomer, whilst the intact mass of the tetramer was determined in ammonium acetate at pH 7.5. Examples of the spectra obtained are shown below (figure 3.22). The denaturing mass spectrum is presented in green and shows peaks for protonated molecular ions $(M+nH)^{n+}$, labelled up to n = 21, giving a monomeric mass of 22,565.5 ±1 Da; this is within 0.2 Da of the mass predicted for Gly-Pro-Leu-Gly-Ser-M2-1. The native, intact mass spectrum is presented in purple and shows peaks for protonated molecular ions $(M+nH)^{n+}$, labelled up to n = 21, giving a tetrameric mass of 90,528 ±5 Da. There is a 266 Da disparity between the observed native tetrameric mass and four times the denatured monomeric mass (4 x 22,565.5 = 90,262 Da), which is likely attributable to 4 atoms of zinc coordinated by the natively folded zinc finger motifs of M2-1; 266 divided by the atomic mass of zinc (65.39 Da) is 4.07.

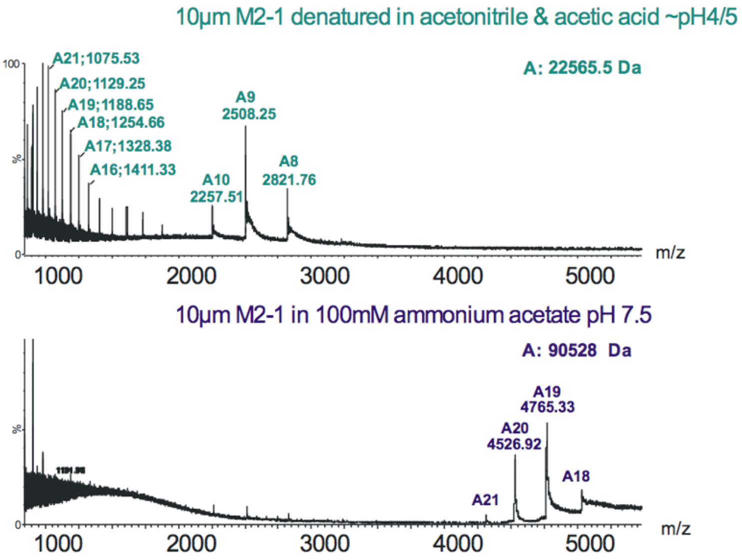


Figure 3.22 Mass spectrometric analysis of M2-1.

Accurate mass spectra are shown for denaturing MS (monomeric mass; green) and native MS (intact tetrameric mass; purple). Spectra show the relative abundance (%) of peaks for a given mass to charge ration (m/z). Protonated molecular ions $(M+nH)^{n+}$ are labelled with their m/z and the number of charges.

3.6.11.2 Binding of bacterially-derived nucleic acid

Both crystal trials and RNA binding assays require a source of RNA-free purified protein. Given that M2-1 is an established RNA binding protein, phenol-chloroform extraction and agarose gel electrophoresis were used to identify any bacterially-derived nucleic acid that had remained bound to M2-1 during the purification process. Tetrameric M2-1 from peak 2 size exclusion chromatography fractions, and purified tomato spotted wilt virus nucleocapsid protein as a known RNA-binder, were treated as in Methods 2.2.14.1. Nucleic acid was incubated with or without RNase A for 1 hour at 37 °C and samples analysed by 1.5 % agarose-Tris-borate-EDTA buffer (TBE) gel electrophoresis. RNase A-sensitive nucleic acid was present in both samples, showing that tetrameric M2-1 was still bound to bacterial RNA after both affinity chromatography and size exclusion chromatography (figure 3.23). The 260nm:280nm ratio for RNA-bound M2-1 was >1, as determined by spectrophotometry, and future assessment of RNA content was made by comparing 260nm:280nm measurements to this value.

To remove the RNA present in purified M2-1, two methods were used: RNase A treatment of all purified protein successfully removed RNA, however it caused some precipitation of M2-1 and reduced its long-term stability at 4 °C; introducing additional high salt washes during glutathione affinity purification, between washing with lysis buffer and equilibrating in protease cleavage buffer, removed RNA contamination, as determined by a 260nm:280nm ratio of <0.6. Future purifications all contained thorough washes with 20 mM Tris pH 7.4, 1 M NaCl.

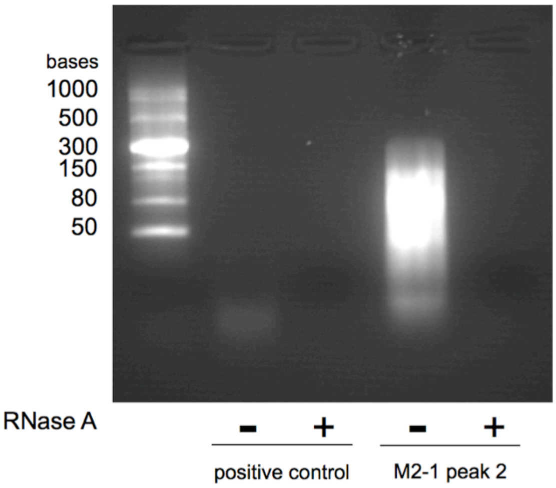


Figure 3.23 Visualising RNA bound to purified M2-1.

Phenol-chloroform was used to extract total nucleic acid from 'peak 2' fractions of pGEXM2-1-expressed M2-1. Nucleic acid was treated with RNase A to distinguish DNA from RNA and samples electrophoresed in a 1.5 % agarose-TBE gel. Tomato spotted wilt virus nucleocapsid protein, was used as a positive control. A low range single stranded RNA ladder is shown.

3.6.12 Final optimal expression and purification of RSV M2-1

Final optimal expression and purification conditions of RSV M2-1 were as follows (figure 3.24):

BL21 Gold cells transformed with pGEXM2-1 were induced at OD_{600} 0.6-0.8 with 340 μM IPTG and 50 μM $ZnSO_4$ overnight at 30 °C. Cells were lysed in 45 mL lysis buffer per 1 L pellet [50 mM Tris pH 7.4, 1 M NaCl, 1 mM DTT, 2 % (v/v) Triton X-100, 10 mM $MgSO_4$, 1 mM $CaCl_2$, 1 mg/mL lysozyme, 1x complete protease inhibitors EDTA-free (Roche)] for 30 minutes on ice, with further lysis by sonication. Lysates were clarified by centrifugation and applied in batch to glutathione sepharose overnight at 4 °C. Resin was pelleted and washed in lysis buffer, followed by extensive washing in high salt buffer to remove RNA. Resin was then equilibrated in 1 volume of protease cleavage buffer and incubated overnight at 4 °C with a 1:25 HRV 3C protease to equilibrated resin ratio. Cleaved M2-1 was eluted from resin by washing 3 times with 1 volume of protease cleavage buffer and elutions were pooled for size exclusion chromatography purification. Tetrameric M2-1 eluted at 128 mL as a ~90 kDa species.

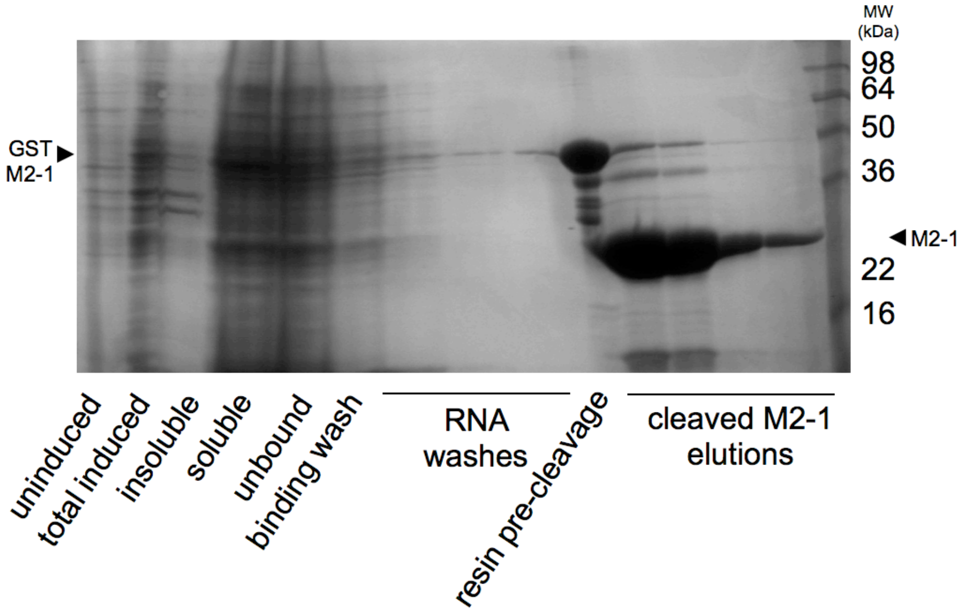


Figure 3.24 Final optimised GST-M2-1 expression and purification.

3.7 SECTION SUMMARY

This section has presented the optimisation of RSV M2-1 bacterial expression, providing a protocol for expression and purification of ~5 mg >95 % pure tetrameric M2-1 protein per litre of bacterial culture.

Initially, to increase the likelihood of optimally expressing recombinant M2-1 in *E.coli*, a N-terminal SUMO fusion was made as this system had been successfully used both within our group (Jenkins et al., 2011, Ariza et al., 2013) and for many published difficult-to-express proteins (Butt et al., 2005). SUMO-tags have been shown to increase solubility, improve folding, and decrease proteolytic degradation of recombinant proteins. SUMO proteins, or small ubiquitin-like modifier, are a family ubiquitous to eukaryotes that covalently modify cellular proteins to affect their role in a variety of cellular pathways, including prevention of ubiquitin-mediated degradation and nuclear transport. The SUMO encoded in the pET-28a vector is *SMT3* from *Saccharomyces cerevisiae*. Despite the recent emergence of SUMO fusion as one of the most successful ways to express recombinant protein in *E. coli*, robust expression and purification of M2-1 in pET28aSUMO was not achieved; extensive optimisation of induction conditions, buffer components, and the purification process, in addition to a free-cysteine mutant to reduce possible aggregation by disulphide bonds, all failed to improve M2-1 expression.

At this point, M2-1 was subcloned and expressed as a GST-fusion, an older but well-documented solubility aid for recombinant proteins. GST proteins are a family of enzymes that covalently modify toxic molecules with their substrate glutathione, a tripeptide of glutamine-cysteine-glycine with a non-canonical peptide linkage between the cysteine amine group and the side chain carboxyl group of the glutamine. It is the high affinity, specificity, and reversibility of this reaction that makes GST ideal for affinity chromatography, and gives glutathione affinity chromatography its reputation for yielding high purity protein. GST also aids in correct translation and folding of recombinant protein, and offers some protection from enzymatic degradation in cells (Walls and Loughran, 2010). pGEX vectors encode a 26 kDa GST derived from *Schistosoma japonicum* (Smith and Johnson, 1988), which forms homodimers in solution – one of the downsides of using a GST-tag for purification of oligomeric proteins, as GST dimerization can cause higher order oligomers of fusion protein to form. Due to the similarity in molecular weight between M2-1 (22.5 kDa) and GST (26 kDa), and their virtually identical electrophoretic mobilities, removal of the GST-tag by proteolysis was performed on resin-bound GST-M2-1. Cleavage by this method was inefficient, however, as the protease was sterically hindered by the orientation of the fusion protein on the resin. Additional inefficiency was introduced by the oligomeric nature of M2-1; as a tetramer, all four GST-M2-1 monomers must be successfully cleaved for the resulting M2-1 tetramer to be isolated as the correct species by size exclusion chromatography.

Although there appear to be several advantages to His-SUMO fusion over GST fusion, including a lack of oligomerisation and the use of a tertiary structure-specific protease, significantly higher levels of M2-1 expression were achieved with the GST system and the stability in solution of GST-expressed protein was also improved over the His-SUMO protein. When the size exclusion chromatography profiles of M2-1 produced by both methods were compared, a difference in elution volumes was noticed. His-SUMO-expressed M2-1 eluted at ~113 mLs, which corresponded to a molecular weight of 150 kDa, and GST-expressed M2-1 eluted at ~128 mLs, ~90 kDa. Although there is precedent for M2-1 protein appearing to have a hydrodynamic volume larger than its molecular weight would suggest (Esperante et al. 2011), the elution of His-SUMO-expressed M2-1 so close to the void volume of the column, added to its instability in solution upon dialysis of imidazole and significant losses during concentration, suggested that this protein may have been predominantly aggregated.

3.8 CHAPTER SUMMARY

This chapter has outlined the optimisation of RSV M2-1 expression in two systems: recombinant baculovirus-assisted insect cells and bacterial cells.

Prior to this work there was published precedent for expression of M2-1 by both methods, however neither had been optimised for the kind of large scale required for structural studies. During this project, a number of publications also presented methods for high-yield bacterial expression of M2-1: Tran et al (2009) published an expression method also using a GST fusion and a similar buffer to that arrived at by the optimisation presented here – elements of their method were used to improve purification in this work, including the addition of 1 M NaCl washes to remove bacterially-derived RNA; Esperante et al (2011) published a method of expressing M2-1 using C41 (DE3) cells without a fusion tag, followed by heparin affinity chromatography as a form of cation exchange – although we did not replicate this method exactly, as we did not have access to the unusual bicistronic vector they used, we did not see an increase in expression using the C41 strain.

Whilst both expression systems have previously been extensively used for producing crystallography-quality protein, the expression of M2-1 as a doublet in insect cells as a result of phosphorylation, now confirmed to involve serines 58 and 61, and the difficulty of separating the two phospho-species on a large scale, meant that this protein was unsuitable for structural studies. Successful optimisation of the bacterial expression of M2-1 as a GST fusion provided a high-yield source of homogeneous protein appropriate for crystal trials and RNA binding studies, and these experiments are presented in the remaining chapters of this thesis.

4 CHAPTER 4: DETERMINING THE X-RAY CRYSTAL STRUCTURE OF M2-1

4.1 CHAPTER INTRODUCTION

This chapter describes the process of determining the X-ray crystal structure of both wild type (WT) and phosphomimetic (S58DS61D) M2-1 protein; outlining the initial screening and optimisation of crystallisation conditions, X-ray data collection, multiple strategies used to address the 'phase problem', and the data reduction and structure solution. The structure itself is then described, and the new and unique insights into M2-1 structure and function it provides are summarised.

4.2 RESULTS

4.2.1 Initial crystallisation screens

Purified RSV M2-1 protein was produced as described in Chapter 3 and screened for initial crystallisation conditions using commercially available sparse matrix screens. These were designed to sample a wide range of precipitants, salts, buffers, pH, polymers and organic molecules, in a high-throughput 96-well format, and were biased towards previously successful crystallisation conditions for macromolecules. The screens were compatible with an automated robot system, which was used in these initial trials to quickly and reproducibly set up sitting-drops with various protein:mother liquor ratios across all conditions and at a range of temperatures, amounting to 2,880 combinations (see Methods 2.2.13.1).

Several conditions resulted in initial 'hits' that appeared within 2-3 days and ranged from 10-20 small crystals (20-40 μm) to 2-5 larger crystals (50-

100 μm) within a drop. A number of other conditions showed signs of the potential to be optimised for crystal growth: birefringent precipitate, a crystalline precipitate that is able to polarize light due to the anisotropy inherent in an ordered, crystalline material; phase separation, the formation of liquid-liquid interface droplets that are metastable, i.e. unable to nucleate; showers of microcrystals due to excessive nucleation. However, the growth of crystals in multiple conditions meant optimisation of these potential conditions was not required.

Initial hits, although from different screens, all centred around a common theme: 20-30 % (v/v) low molecular weight PEG (PEG 400, PEG MME 550, PEG MME 2000), ±0.05-0.2 M salt (NaCl, $MgCl_2$, $CaCl_2$), and a wide range of pHs from 7-10.5. Crystals of sufficient size (>20 μm) and quality (few splits, single crystals, no plates) were picked from drops using nylon loops and cryo-cooled directly from drops in liquid nitrogen; the low molecular weight PEG provided cryo-protection. Crystals were screened for diffraction at the Diamond Light Source synchrotron, by taking test images at $\phi = 0°$ and 90° using 12,658 eV X-rays (0.9795 Å). All crystals were found to be protein and diffracted to 3.5-6 Å. Two conditions were selected for further optimisation, based on diffraction quality, visual appearance of crystals, and reproducibility of crystals across multiple pHs and temperatures: Index G7 - 0.05 M magnesium chloride hexahydrate, 0.1 M HEPES pH 7.5, 30 % (v/v) PEG MME 550; Wizard E1 - 0.1 M CAPS pH 10.5, 30 % (v/v) PEG 400.

4.2.2 Crystal optimisation

4.2.2.1 Index G7 optimisation

Crystals directly from Index G7 (IG7) in the initial screen diffracted to ~3.5 Å, space group P2 with unit cell dimensions of a = 89.6 Å, b = 140.8 Å, c = 140.6 Å, $\alpha = \gamma = 90°$, $\beta = 96.3°$. However, these crystals were highly anisotropic, diffracting as well as 2.8 Å at some angles, but as poorly as 5 Å in others - a characteristic that was highlighted visually by the wedge-like shape of the crystals (figure 4.1). These crystals were also split or formed of stacks of thin plates. Matrices were set up manually in 24-well hanging-drop format to optimise % PEG and pH, at 12 °C, 16 °C and 25 °C. Crystals were grown at all temperatures and at several pHs and concentrations of PEG, however there was no discernable pattern of improvement when altering any of these parameters, either visually (figure 4.1, right) or when screened for diffraction.

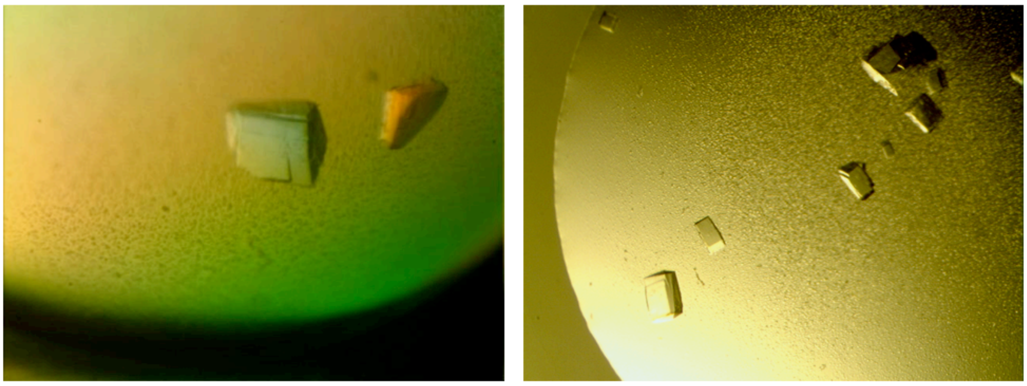


Figure 4.1 Index G7 crystals.
Crystals from Index G7 condition in initial screen were wedge-like and split, polarised light (left) and diffracted to 3.5 Å although with severe anisotropy. 24-well hanging drop optimisation of pH and % PEG did not reproducibly improve crystals (right).

A problem encountered during optimisation was batch-to-batch variability. Relatively low yields, in terms of crystallographic studies, were achieved during optimisation of GST-M2-1 expression and purification (~5 mg/L) and so multiple preparations of protein were needed. Small, unavoidable differences between preparations, such as contaminant levels, or small differences in pH or protein concentration, caused difficulties in reproducing crystals and occasionally resulted in the growth of atypical crystal forms that were poorly diffracting (figure 4.2). To improve consistency during crystallisation optimisation, a number of freezing conditions for M2-1 were tested. M2-1 was supplemented with 5 or 10 % (v/v) DMSO, glycerol or MPD, or left unsupplemented, and snap-frozen in liquid nitrogen before storing at -80 °C. After a few days, protein was thawed on ice before setting up known crystal growth conditions. Unfortunately it was found that M2-1 did not crystallise after freezing with or without cryo-protection, and so batch-to-batch variability remained a significant problems for crystallisation optimisation.

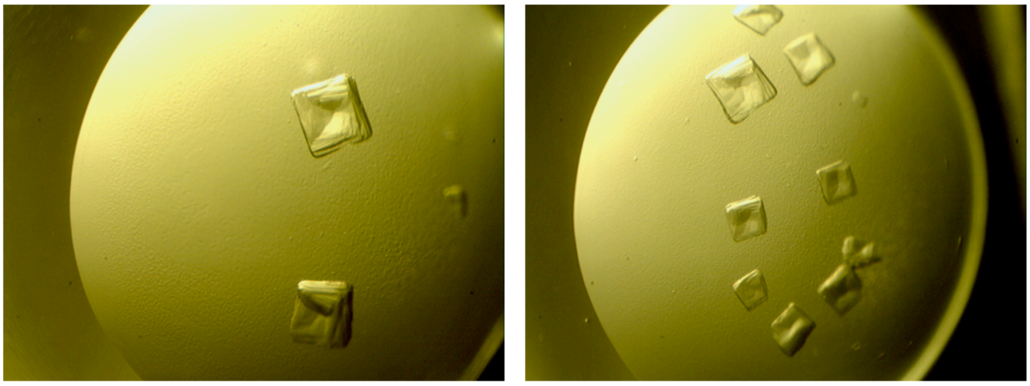


Figure 4.2 Atypical crystal forms.
Batch-to-batch variation caused different, poorly diffracting crystal forms in previously successful conditions.

A commercially available Additive Screen (Hampton Research) was used in 96-well sitting-drop format at 16 °C, based on the best-diffracting condition so far: 0.1 M Bis-Tris propane pH 7, 0.05 M $MgCl_2$, 30 % (v/v) PEG MME 550. This is a screen consisting of small molecules known to affect crystallisation by altering protein solubility, manipulating protein-solvent interactions, or perturbing the structure that water molecules form around the target protein. A 110 % stock solution of the mother liquor was made, and additives added to a 10 % concentration. Crystals appeared after 2-3 days in almost all conditions and several additives appeared to visually improve crystal quality, including 10 mM nickel chloride (figure 4.3, left) and 10 mM TCEP hydrochloride, a reducing agent (figure 4.3, right). Both of these additives improved diffraction quality, but resolution was not improved beyond 3.1 Å.

Other methods of improving crystal quality, such as micro- and streak-seeding, and increasing crystal size by using larger volume drops, did not improve diffraction quality or resolution and so no further optimisation of this condition was attempted.

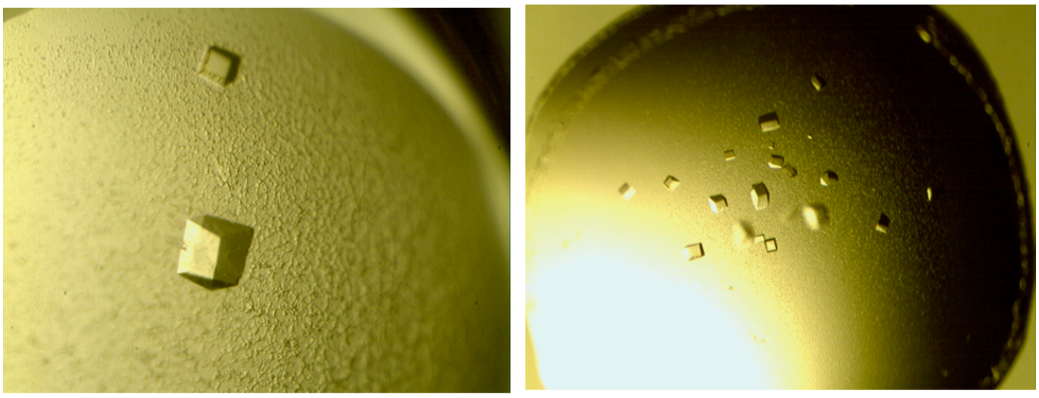


Figure 4.3 Additive screening.

Hampton Research Additive Screen was used with crystallisation condition: 0.1 M Bis-Tris propane pH 7, 0.05 M $MgCl_2$, 30 % (v/v) PEG MME 550. Left, 10 mM nickel chloride, right, 10 mM TCEP hydrochloride.

4.2.2.2 Wizard E1 Optimisation

An alternative diffracting condition from the initial screening was optimised. Crystals grown in Wizard E1 (WE1) [0.1 M CAPS pH 10.5, 30 % (v/v) PEG 400] were cryo-cooled directly from 96-well drops and screened for diffraction at Diamond. Although the crystals were smaller than the previous condition (IG7) and only diffracted to 3.69 Å in a new space group P422, WE1 crystals were cubic and less split (figure 4.4, left) and displayed higher quality diffraction than any IG7 crystal, i.e. sharper spot profiles, high intensity spots, and low background noise (figure 4.5). Extensive matrix optimisation of %

PEG, pH, and the addition of salt (based on its presence in the previous IG7 condition) was attempted, but it was found that the condition could not be translated from 96-well sitting drop to 24-well hanging drop. This is not uncommon, as factors such as nucleation from the plate in sitting drops or the base of a hanging drop, and differences in surface tension due to the shape of hanging vs. sitting drops, can be parameters that affect crystallisation. The problems that using 96-well sitting drops introduces are difficuty in picking crystals due to the curvature of the surface in which the drops sit, and a limitation on the size of drops that can be set up (2 μL) which can, in turn, limit the size of the crystals grown. The remaining crystal trials were set up in 96-well format, both with the robot and manually. Some increase in size of crystals and resolution of diffraction was achieved with optimisation of pH; a single crystal grown in 0.1 M glycine pH 9.5, 30 % (v/v) PEG 400, diffracted at 12.957 keV to a maximum of 2.95 Å in space group $P2_1$, different to the initial hit in P422. This was the highest resolution native WT dataset collected and was named 'WT-$P2_1$' (figure 4.4, right). Data collection statistics are shown in table 2, section 4.2.5.

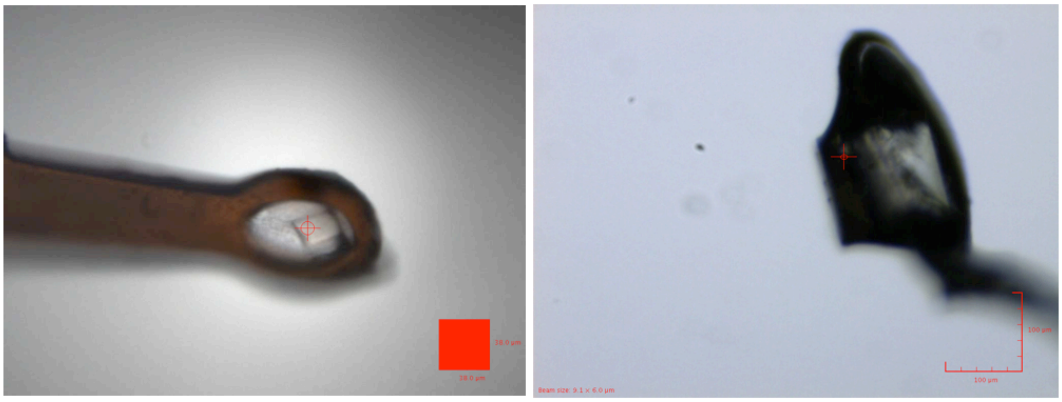
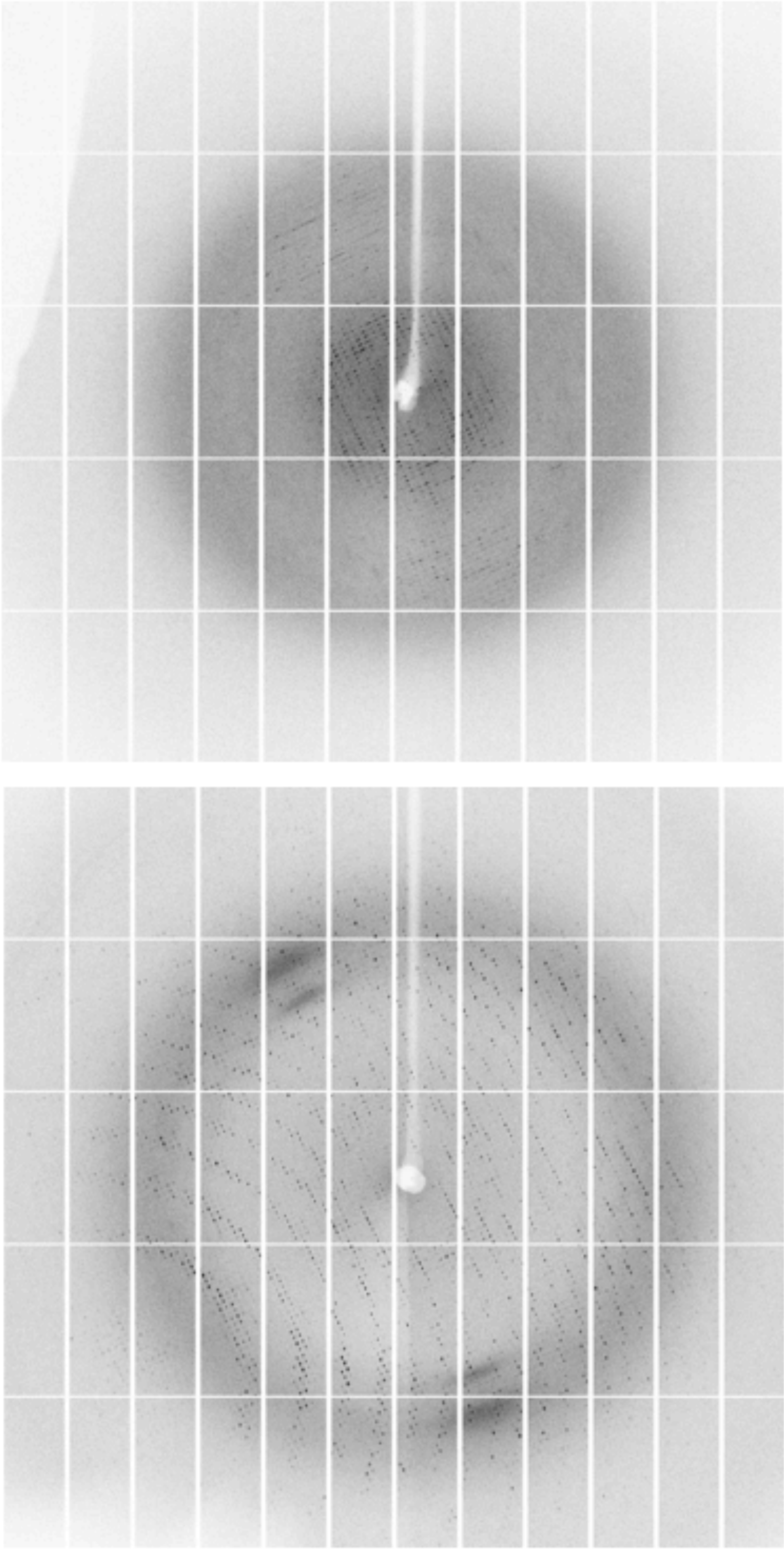


Figure 4.4 Wizard E1 crystals.

Images taken of crystals in loops on the beamline at Diamond. Left, crystal from initial Wizard screen was small (~30 μm^3) and cubic, and diffracted to 3.69 Å in P422. Right, crystal from matrix optimisation of pH and % PEG was larger (~120^3 μm), and diffracted to 2.95 Å in $P2_1$.

Figure 4.5 Comparing diffraction patterns.

(below) Comparison of low quality diffraction from Index G7, top image, (0.1 M Bis-Tris propane pH 7, 0.05 M $MgCl_2$, 30 % (v/v) PEG MME 550) with good quality diffraction from Wizard E1, bottom image, (0.1 M glycine pH 9.5, 30 % (v/v) PEG 400).



4.2.3 S58DS61D – cloning, expression and crystallisation

To assess the possible structural impact of phosphorylation at serines 58 and 61, a phosphomimetic mutant 'S58DS61D' was synthesised to replace both serines with aspartates, thus mimicking the negative charge of this post-translational modification. Site directed mutagenesis of pGEXM2-1 was used, with one set of complementary mutagenic primers encoding both serine to aspartate mutations. Success was confirmed by DNA sequencing and the construct named pGEXM2-1DD. This was transformed into BL21 (DE3) Gold cells and expressed with the same protocol as wild-type. S58DS61D had a similar expression and purification profile to WT with comparable yields and purity, although S58DS61D protein appeared to be more stable in solution as it was amenable to manipulation of buffer conditions and concentration with less precipitation than WT.

Crystallisation trials with S58DS61D were performed as for WT protein: initial screening was performed with commercially available sparse matrix screens in 96-well sitting drops with the automated robot system, at several temperatures. Hits were seen in many of the same conditions as for WT, however the best diffracting crystals were from a unique condition: 0.1 M Tris pH 8.5, 0.2 M trimethylamine N-oxide (TMANO), 20 % (w/v) PEG MME 2000. Crystals from this condition were picked from drops and transferred into a fresh drop containing 1:1 ratio of mother liquor to cryo-protectant solution [5 % (v/v) of each: glycerol, PEG-400, 2-Methyl-2, 4-pentanediol (MPD), ethylene glycol], then cryo-cooled in liquid nitrogen. These crystals, directly from the initial screens, diffracted to 2.4 Å in $P2_1$ with a higher quality diffraction pattern than WT crystals. No further optimisation of S58DS61D crystals was required and the best dataset collected from this condition was named 'DD-$P2_1$'.

4.2.4 Initial phasing attempts

4.2.4.1 Molecular replacement

There are no X-ray crystal structures available of M2-1 protein from any of the pneumoviruses, nor are there predicted to be any homologous proteins based on amino acid sequence, and so there was no appropriate structure in the PDB to use as a search model for molecular replacement when this work began. However, during crystal trials for the WT M2-1, a partial solution NMR structure of M2-1 residues 58-177 was published (Blondot et al., 2012, Dubosclard et al., 2011). This structure was monomeric, as it lacked the proposed tetramerisation domain in the N-terminus. Attempts were made to use each of the 20 structures from the PDB, in turn, as search models using the WT-P2_1 dataset with PHASER (McCoy et al., 2007). The structure could not be solved by this method.

An increasingly powerful option for structure solution in the absence of a suitable search model in the PDB is the generation of a predicted structure. Robetta is an online protein structure prediction server that takes a protein sequence and divides it into putative domains, which it then models based on homologues in the PDB or *ab initio* using a combination of basic protein structure principles (permissible peptide bond angles, helix formation etc.) and common folds found in the PDB. Robetta divided M2-1 into a domain corresponding to the partial NMR structure, which it modelled based on that PDB entry, and a domain comprising the remaining N-terminal residues, which was *ab initio* modelled but failed to produce a zinc finger-like motif (figure 4.6). None of the Robetta models were capable of phasing our X-ray data by molecular replacement.

Figure 4.6 Robetta structure predictions.

Two examples of structure predictions made by the online server Robetta. These are based on a combination of homology modelling with the partial NMR structure and *ab initio* modelling with basic principles of protein structure.

4.2.4.2 Experimental phasing – single wavelength anomalous dispersion (SAD)

4.2.4.2.1 Selenomethionine

The naturally occurring sulphur atoms in methionine residues were replaced with selenium, a heavier atom that provides an anomalous signal when selenium-edge X-rays are used. pGEXM2-1 was retransformed into B834(DE3) cells, a methionine auxotrophic strain of *E. coli*, and 50 mL LB medium starter cultures grown overnight as in Methods 2.2.1.7. Cells were pelleted and washed 3 times in sterile water, resuspended in 1 mL water and used to innoculate 1 L SelenoMet Medium (a premade M9 minimal media supplemented with all amino acids except methionine; Molecular Dimensions Limited) containing 30 mg/L L-selenomethionine. Cultures were grown to OD_{600} 0.6-0.8 and induced with 1 mM IPTG and 50 μM $ZnSO_4$ for 6 hours at 37 °C, as per the manufacturer's protocol. 'Semet' M2-1 was purified as per wild type and crystal trials were attempted. This protein crystallised poorly, and all crystals that were obtained either did not diffract or diffracted to worse than 5 Å.

4.2.4.2.2 Naturally-bound metal atoms

M2-1 protein contains a predicted Cys_3-His_1 zinc finger motif between residues 7 and 25 that has been shown to coordinate one atom of zinc per monomer. This zinc could also be used to provide the anomalous signal for phasing by SAD. A fluorescence scan was used to determine if zinc was present in the crystals and the exact wavelength of the zinc absorption edge (figure 4.7), which can vary from the theoretical value based on the chemical environment of the heavy atom. Zinc was found to be present in all crystals, however one zinc atom per 199 residues did not provide sufficient anomalous signal to begin an initial estimate of phases.

An additive screen used to optimise Wizard E1 crystals [0.1 M CAPS pH 10, 30 % (v/v) PEG 400] yielded well-diffracting crystals in the presence of 0.01 M cadmium chloride; diffracting to 2.52 Å in space group P422 (figure 4.8). The data were of poorer quality than both the WT-$P2_1$ and the DD-$P2_1$

datasets collected previously. The absorption edge for cadmium was outside the range of possible fluorescence scans at Diamond, however there remained a significant contribution at the zinc edge wavelength, as predicted, which increased the magnitude of the anomalous signal and allowed phasing by SAD as described below.

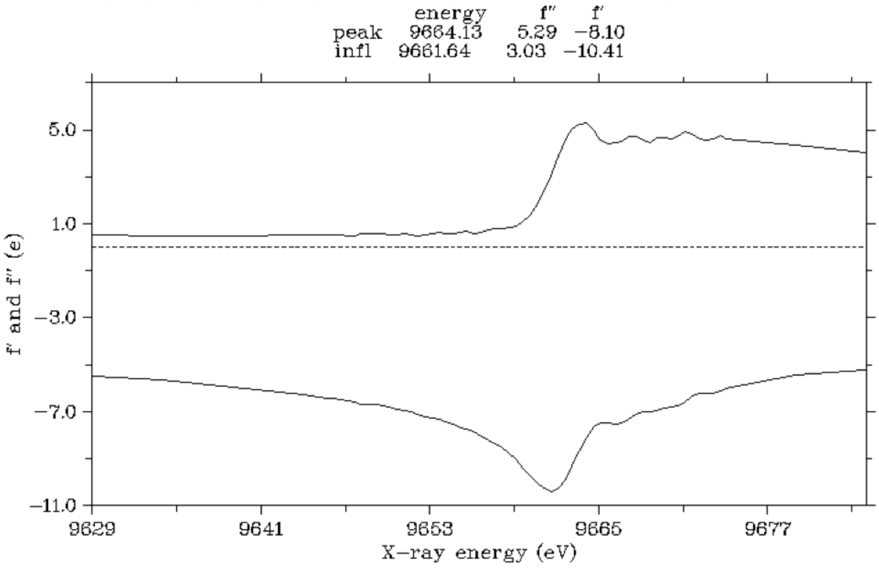


Figure 4.7 Fluorescence scan of zinc absorption edge.

Spectrum displaying the magnitudes of the anomalous (f'') and dispersive (f') components of scattering at a given X-ray energy (eV). The anomalous peak is at 9664.13 eV and is close to the theoretical value for the zinc K-edge, 9658.6 eV.

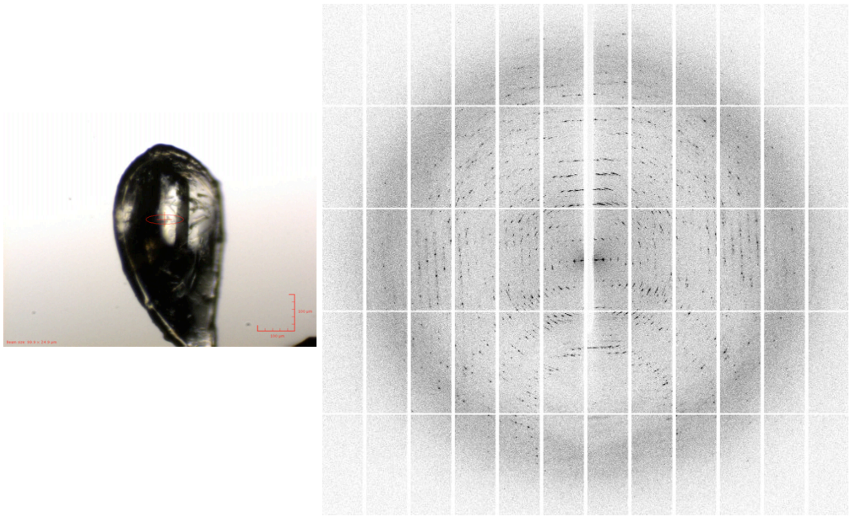


Figure 4.8. Wizard E1 with 0.01 M cadmium chloride, crystal and diffraction pattern.

The crystal used to collect dataset WT-P422 at the zinc absorption edge diffracted to 2.52 Å in space group P422, with significant anomalous signal.

4.2.5 Data processing

X-ray data for the M2-1 WT-P422 crystal were indexed and integrated using iMOSFLM (Battye et al., 2011, Leslie, 2006). Indexing is the process of assigning Miller indices (hkl) to each reflection in the diffraction pattern and determining information such as unit cell size and mosaicity. This assignment is based on the Bravais lattice class observed from the geometric location of reflections, and must be manually checked to confirm that the reflections predicted by the lattice group match those observed in the diffraction images. Integration is the process of recording an intensity and associated error for each reflection. These two processes convert the dataset from a series of images to a numerical list of reflection intensities. A primitive tetragonal lattice was assigned with unit cell parameters of a = b = 102.28 Å, c = 90.99 Å, α = b = γ = 90.00 °, and POINTLESS used to confirm the Laue group (based on the symmetry of the reflections) and the space group (using systematic absences in the diffraction pattern to identify screw axes). P422 was assigned and scored as the only probable space group. This reduced data was then scaled to make the dataset internally consistent by accounting for variations such as beam intensity, radiation damage, and detector inconsistencies, using scale-factors determined by comparing symmetry related reflections. Multiple observations of each reflection were merged to create an average intensity (I_{hkl}), and a list of normalised structure factors amplitudes ($|F_{hkl}|$) created using the following relationship:

$$I_{hkl} \propto |F_{hkl}|^2$$

Scaling and merging were performed with AIMLESS, and this programme also outputs a number of statistics to objectively assess data quality:

> R_{merge} – the basic measure of internal consistency that assesses the agreement between the average intensity of a reflection, and each individual observation of that reflection. A lower value is an indication of higher quality, however R_{merge} increases with multiplicity and so is not always an accurate indicator.

R_{pim} – the 'precision-indicating R-factor' is like the R_{merge} but accounts for multiplicity, thus reflecting the increase in accuracy as more observations are merged.

Completeness – the number of unique reflections as a percentage of the total reflections.

Multiplicity – the number of total reflections divided by the unique reflections; it is the average number of independent measurements of a reflection.

$I/\sigma(I)$ – the average signal to noise ratio of the intensities.

These statistics for WT-P422 dataset (and the two datasets solved by molecular replacement, below) are presented in table 2, and were all acceptable at the 2.52 Å resolution cut-off that was chosen. Approximately 5 % of the data were randomly selected in AIMLESS to be omitted from refinement, allowing the R_{free} to be monitored during refinement (see 4.2.7). Tests for twinning of the data, where two or more lattices grow together so that their unit cells are related by symmetry, were also performed here. All data presented in this thesis were not twinned.

Data set	WT-P422	WT-P21	DD-P21
Wavelength (Å)	1.2823	1.2819	0.9889
Space group	$P422$	$P2_1$	$P2_1$
Cell parameters (Å,º)	a = 102.28 b = 102.28 c = 90.99 α= 90.00 β = 90.00 γ = 90.00	a = 91.85 b = 141.6 c = 141.7 α= 90.00 β = 96.14 γ = 90.00	a = 89.68 b = 141.60 c = 142.16 α = 90.00 β = 93.92 γ = 90.00
Total reflections	884975 (49270)	1135690 (62535)	840458 (34707)
Unique reflections	16695(1565)	75535 (4385)	138151 (6792)
Resolution shells (Å)			
Low	45.74-2.61	140.88-3.01	78.12-2.44
High	2.61-2.52	3.01-2.95	2.44–2.40
R_{merge}, %	9.6 (64.7)	19.1 (138.4)	9.8 (60.1)
R_{pim}, %	1.8(16.1)	7.4 (55.2)	6.5 (44.1)
Completeness, %	98.9 (91.2)	99.5 (98.2)	100 (99.5)
Multiplicity	30 (6.0)	15.0 (14.3)	6.1 (5.1)
$I/\sigma(I)$	14.8 (3.8)	14.4 (2.3)	11.1 (2.4)

Table 2 Data collection and scaling statistics.

Three datasets are presented here: WT-P422, the SAD dataset described above; WT-P2_1, a wild type dataset solved by molecular replacement (below); DD-P2_1, the phosphomimetic S58DS61D mutant dataset, also solved by molecular replacement (below). Parentheses indicate highest resolution shell values.

4.2.6 SAD Phasing

Phasing of the M2-1 WT-P422 data was carried out with PHENIX AutoSol (Adams et al., 2010) using the single-wavelength anomalous dispersion (SAD) method. AutoSol is an automated structure solution pipeline that combines a series of programs to identify potential heavy-atom sites, calculate experimental phases, and build an initial density-modified electron density map. It is a fast way to try multiple solutions, such as different possible heavy atom sites, related space groups, or opposing ‘handed’ structures. For the WT-P422 dataset, AutoSol first used HySS (Hybrid Substructure Search) to identify all possible locations of anomalous scatterers through a combination of Patterson methods (see Introduction 1.5.4). Then for each potential heavy-atom substructure a set of reference phases (φ_A) for the heavy atoms was calculated, as their location, structure factor amplitude and phase are related through the electron-density equation. Initial phases for the rest of the protein (φ_P) were estimated with PHASER, using the relationship:

$$\varphi_P = \varphi_A + \alpha$$

where α is the shift in phase from heavy atom phase to protein phase, and is related to the anomalous difference and heavy atom structure factor magnitude (Dauter et al., 2002). Using these initial protein phases electron density maps were built and statistically density modified with RESOLVE, which works to increase the probability of an electron density map being correct by incorporating prior knowledge, such as: non-crystallographic symmetry averaging from symmetry of heavy atoms in the substructure; solvent flattening, by subtracting the mean solvent density from the whole map and masking the protein region based on knowledge of the unit cell solvent content; histogram matching, where the density distribution of the map is modified until it fits the pattern expected for a protein structure.

Electron density maps built using phase estimates from different substructure solutions were scored based on a number of quality-assessment criteria: skew of the electron density (i.e. density next to density and solvent next to solvent), regions of connected density, contiguous flat solvent, R-factor for density modification (difference between calculated and observed structure factors). The most likely solution resulted in a map with two monomers per asymmetric unit (AU) that combined with their symmetry mates to form two tetramers in the unit cell. This was in line with the Matthews coefficient (V_M), which predicts the most probable content of the asymmetric unit based on protein molecular weight and estimated solvent content; 50 ±10 % for most proteins in the PDB. Four heavy atoms were identified in the substructure, which were later confirmed as two cadmiums coordinated in the Cys_3-His_1 zinc finger motifs, and two zinc atoms bound at a surface exposed site on M2-1 as an artefact of crystallisation.

4.2.7 Model building and refinement

A preliminary structure was auto-built into the density-modified map by BUCCANEER (Cowtan, 2006). This programme finds possible Cα positions and then extends these into chain fragments by laterally adding more Cα candidates whilst obeying Ramachandran constraints. These fragments are gradually joined and side chains assigned according to the sequence provided. This model was sufficient to see that the solution was correct, as some major features such as α-helices had been auto-built, however there was no density apparent for the C-terminal residues 175-194, or residues 52-67, resulting in an incomplete chain.

The preliminary model was improved by manual building in COOT (Emsley and Cowtan, 2004), where the protein chain was modified residue by residue to improve both its fit in the electron density, and properties such as bond lengths and torsion angles. This process used two types of electron density maps: a maximum likelihood map, which shows the current best estimate of electron density for the structure whilst accounting for errors in the phases, and attempts to limit any potential model bias; a difference map, which estimates the difference between the true structure and the current model, and highlights areas of positive or negative density where atoms should be added or removed, respectively.

Manual building was followed by model refinement in REFMAC5 (Murshudov et al., 2011), where the difference between observed and calculated structure factor amplitudes (R_{cryst}) was minimised by improving the atomic positions in the model. This was achieved by providing additional information in the form of chemical restraints, such as preferred bond lengths, angles and torsions, towards which refinement is weighted at lower resolutions, such as those presented in this thesis. Refinement also improves the isotropic B-factor of atoms, a measurement of the degree of movement around the atomic position in the model. This process of minimising R_{cryst} improved the calculated phases, which were then used to build an improved electron density map (and the model that fits it) in an iterative cycle.

The progress of refinement was assessed by calculating the R_{cryst} for the 5 % of data omitted after scaling: the R_{free}. This measures how well the current model can predict the measured intensities in the 5 % of unrefined data. General 'rules of thumb' aim to have R_{crys} below 10 % of the resolution, i.e. <0.25 or 25 % at 2.5 Å, and R_{free} and R_{cryst} should not deviate by more than 5-7 %. However these statistics are often poorly representative of the true quality of the structure, which should be determined by examining the electron density itself and biophysical factors such as: root-mean-square deviation (rmsd) bond lengths (<0.02 Å) and rmsd bond angles (1-2 °), the number of amino acids residing in the 'preferred' region of a Ramachandran plot, and steric clashes between residues and/or solvent molecules.

Refinement of the WT-P422 dataset become unstable after several cycles, i.e. R_{cryst} and R_{free} began to diverge, and no further improvement of R-factors could be achieved, likely due to the relatively poor quality diffraction images that lead to poor electron density. Refinement of this dataset was not completed and statistics are presented in table 3. Despite this, the model was assessed using the MolProbity server (Chen et al., 2010), which outputs a series of scores that assess the quality of the model based on the biophysical criteria described above. The WT-P422 model had a MolProbity score in the

90th percentile of PDB structures; this takes into account sidechain clashes, whether sidechains are in preferable rotomers, and analysis of Ramachandran outliers, and normalises these scores to the resolution of the structure. Figure 4.9 shows Ramachandran plots for glycines (right) and all other amino acids (left), with favoured regions in pale blue and allowed regions in dark blue. The two outliers, glycine 18 on both monomers in the AU, are shown in green and their location within the zinc finger motif (resides 7-25) likely explains their unfavourable orientation, as this region will be constrained by the zinc-coordinating residues.

The MolProbity analysis suggested that although the refinement of WT-P422 was incomplete, it was a suitable model for molecular replacement of higher quality datasets.

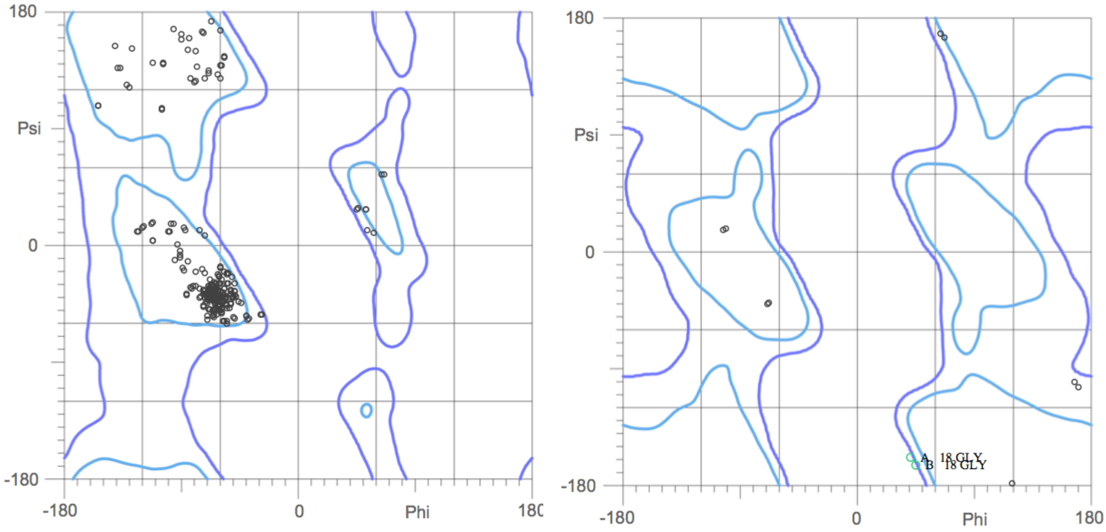


Figure 4.9 Ramachandran plots for WT-P422 model.

Plots of phi and psi Cα backbone angles are the output from the MolProbity server analysis. Right, glycine, Left, all other amino acids. Light blue, favoured region, Dark blue, allowed region. Outliers are shown in green and labelled.

Data set	**WT-P422**	**WT-P21**	**DD-P21**
Resolution shells (Å)			
Low	45.74-2.61	140.88-3.01	78.12-2.44
High	2.61-2.52	3.01-2.95	2.44–2.40
R_{cryst}, %	21.12	20.54	20.36
R_{free}, %	26.15	25.46	23.78
V_M (Å^3/Da)	2.69	2.59	2.53
Mol. per AU	2	16	16
Reflections working set	15849	71711	131263
Free R-value set (no. of reflections)	5.1% (845)	5.0% (3795)	5.0% (6851)
RMSD bond lengths, Å	0.013	0.013	0.015
RMSD bond angles, º	1.459	1.523	1.579
No. of atoms used in refinement			
Non-hydrogen atoms	2669	21688	21997
Protein atoms	2604	21494	21586
Metal atoms	4	16	16
Water molecules	61	178	395
Mean B-factor, Å^2			
Total	71.8	69.4	50.2
Protein atoms	72.3	69.8	50.5
Metal atoms	49.5	44.4	27.61
Water molecules	53.7	38.5	39.3
Ramachandran plot statistics, %			
Preferred region	96.49	94.42	96.21
Allowed region	2.88	4.54	3.15
Outliers	0.64	1.04	0.63

Table 3 Refinement statistics.

4.2.8 Molecular replacement of other datasets

Two previously described datasets were indexed, integrated and scaled as for the WT-P422 data (see 4.2.5): WT-$P2_1$ data (see 4.2.2.2) 0.1 M glycine pH 9.5, 30 % (v/v) PEG 400, diffracted to maximum of 2.95 Å in space group $P2_1$, and phosphomimetic DD-$P2_1$ (see 4.2.3) 0.1 M Tris pH 8.5, 0.2 M trimethylamine N-oxide (TMANO), 20 % (w/v) PEG MME 2000, diffracted to 2.4 Å in space group $P2_1$. For both datasets, 16 monomers (corresponding to four tetramers) were found in the AU. Data collection statistics are presented in table 2.

A monomer from the M2-1 WT-P422 model was used to solve the M2-1 WT-$P2_1$ and M2-1 DD-$P2_1$ X-ray datasets via molecular replacement (MR) using PHASER (McCoy et al., 2007). Molecular replacement works by manipulating the model molecule, an M2-1 monomer from WT-P422 in this instance, into the same orientation and location in the unit cell as the unknown molecule. This is achieved by correlating intra- and inter-molecular vectors in the Patterson maps of both molecules. Translation function Z-scores (TFZ) for both datasets, a statistic used to assess the probability of a molecular replacement solution being correct, were considerably above the acceptable cut-off of 8. Once the orientation and position of the unknown molecule are known, the phases from the model can be applied to the structure factor amplitudes of the unknown protein to build an electron density map.

Maps were built in this way for WT-$P2_1$ and DD-$P2_1$, and the structures were fully refined as described in 4.2.7, with cycles of manual building in COOT and refinement in REFMAC5. 16 zinc atoms were identified per AU, as expected, and no cadmium as none was present in the mother liquor for either of these datasets. Automatically generated non-crystallographic symmetry restraints were used during refinement, to restrain all 16 copies of M2-1 in the AU to the average atomic position.

Refinement statistics are presented in table 3. Although the Ramachandran statistics for WT-$P2_1$ did not suggest that this model of M2-1 was significantly better than the WT-P422 model, the R_{cryst} and R_{free} were slightly improved and the electron density map was deemed to be visibly better defined (figure 4.10). Density for residues 52-67, which had been unresolved in the WT-P422 model, was seen in 2 of 16 molecules in the AU. Statistics for DD-P2$_1$, the phosphomimetic mutant S58DS61D, were significantly better than either WT dataset, with a smaller deviation between R_{cryst} and R_{free}, a lower R_{free} of 23.78 %, lower mean B-factors, and fewer Ramachandran outliers.

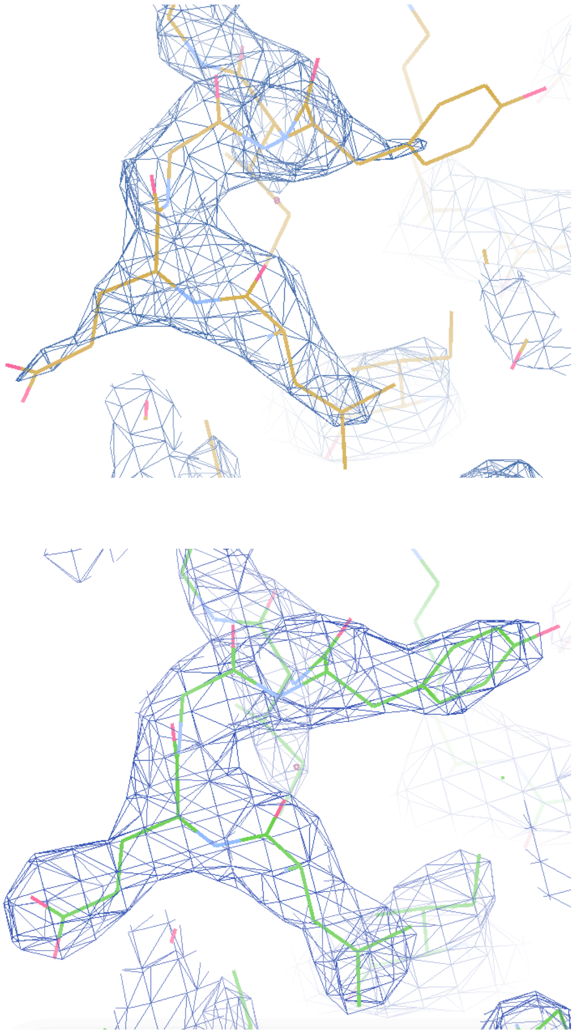


Figure 4.10 Comparison of electron density maps.

Electron density (blue; contoured to 2.0 rmsd) and structure models are shown for the original SAD-phased WT-P422 dataset (top; yellow model) and for the molecular replaced WT-P2$_1$ dataset (bottom; green). The WT-P2$_1$ density and model are in closer agreement than for WT-P422, which has poorly resolved sidechains.

4.2.9 The Structure

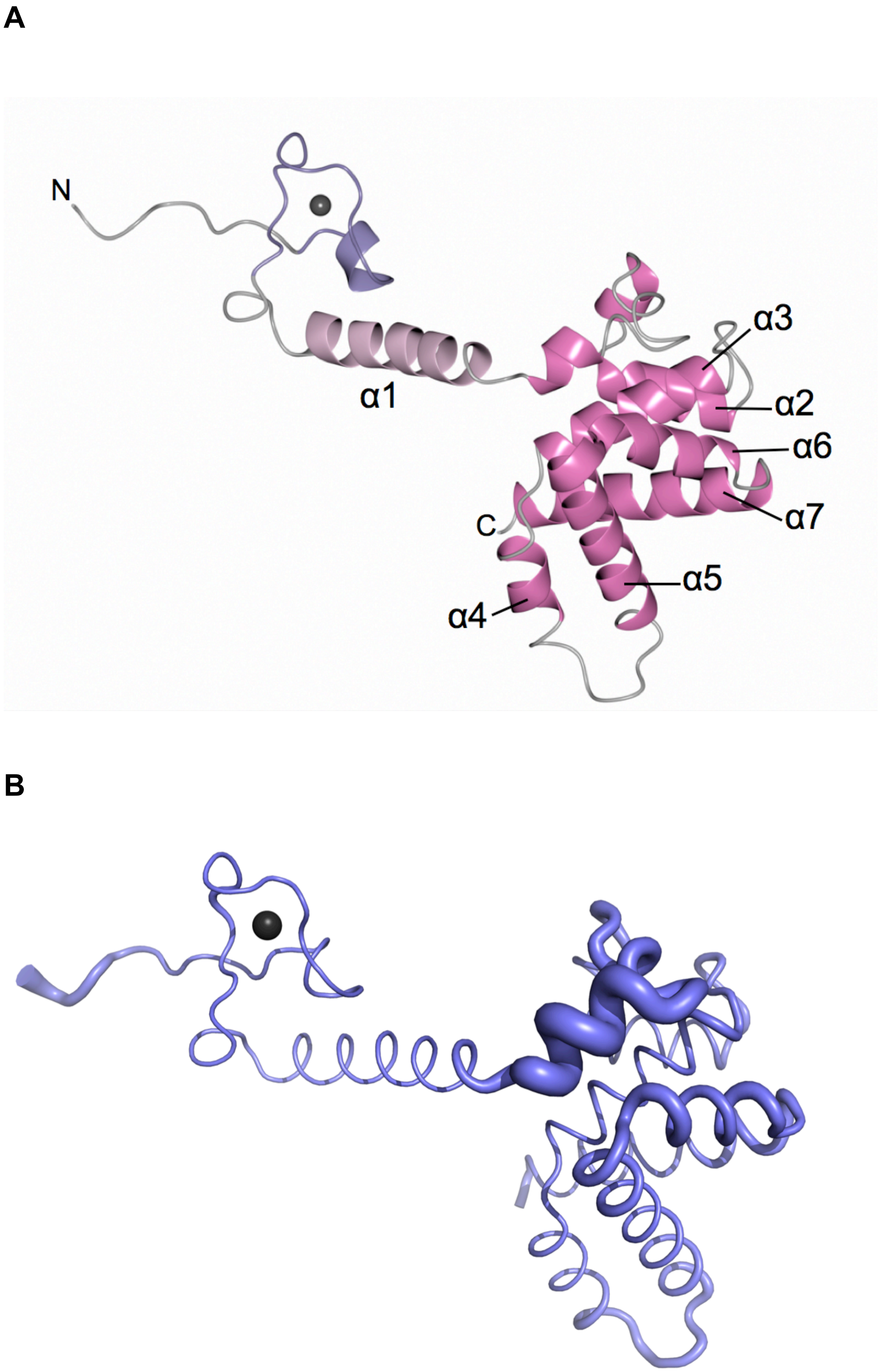
PDB codes are as follows: WT-P422, 4c3d; WT-P21, 4c3b, DD-P21, 4c3e.

4.2.9.1 M2-1 monomer

The X-ray crystal structure of RSV WT M2-1 was solved by the SAD method followed by molecular replacement into a higher quality dataset, to a maximum resolution of 2.52 Å, and is the first example of a full-length M2-1 structure from any pneumovirus. M2-1 crystallised as a tetramer in all space groups, however a monomeric view of the crystal structure is shown in figure 4.11 for simplicity. M2-1 forms four distinct regions linked by unstructured or flexible sequences, these are (N to C terminus): the zinc finger region (residues 7-25) and tetramerisation helix (residues 32-49), which extend on an N-terminal arm and are linked via a flexible loop (residues 52-67) to the core domain (residues 69-172), a globular α-helical fold. An additional region is comprised of the C-terminal 20 residues, which were not resolved in any dataset as they are unstructured (Tran et al., 2009). Figure 4.11 A) shows the structure of monomeric M2-1 with the above regions in different colours, and also numbers the helices α1 to α7. The flexible loop region linking the tetramerisation helix and the core domain was poorly resolved, only showing density in 2 out of 16 molecules in the WT-P2_1 dataset, and this is reflected in the high B-factors for this region (figure 4.11 B). All monomer figures shown here are of a monomer from the WT-P2_1 dataset, containing well-defined density for this region.

Figure 4.11 The monomer representation of M2-1 crystal structure.

(below) One monomer is shown from the crystallised tetramer for simplicity. **A)** N and C termini are labelled (N and C, respectively), and the three helical regions are coloured as follows: zinc finger, purple (residues 7-25); the tetramerisation helix, pale pink (residues 32-49); the core domain, pink (residues 69-172). Unstructured regions are shown in grey. Helices are numbered α1 (tetramerisation helix) through to α7. **B)** 'Worm' representation of monomer where width is proportional to B-factor.


A
N
α3
α1
α2
α6
α7
C
α5
α4
B

4.2.9.2 Tetrameric organisation of M2-1

M2-1 crystallised as a tetramer – the proposed physiologically active form of M2-1 and the only oligomer purified by size exclusion chromatography in this work (figure 4.12). All tetramer figures shown here were built in PISA with the flexible loop-containing monomer described above, in order to present a full-length tetramer and ensure that surface representations of the protein were more appropriate. PISA is an online server that uses criteria such as Gibbs free energy of formation, interface size, and hydrogen bonding to predict the buried surface area and probability of stability for higher order assemblies, allowing inferences about biological relevance to be made. It is imperative that analyses like these are performed on crystal structures to help to distinguish between artefactual contacts as a result of crystallisation, and biologically relevant interactions. The agreement between PISA and the oligomeric status of M2-1 in solution, both in this work and in previous publications, strongly suggests that the tetrameric form crystallised here is physiologically relevant.

The packing of the tetramerisation helix (α1) appears to drive oligomerisation by burying a set of hydrophobic residues (L36, L43, I46, M50 are all on one face of the helix) within a four-helix bundle (figure 4.13). Residues from a single protomer contact all neighbouring protomers forming a very stable tetrameric structure. Further stabilisation of the tetrameric arrangement of M2-1 is provided by the Cys_3-His_1 zinc finger motif (Fig. 4.12 A, inset), which interacts mainly with the N-terminal face of the core domain of an adjacent protomer, increasing the buried surface. PISA analysis estimated the total buried surface area in the tetramer as 12,320 $Å^2$ with a predicted change in Gibbs free energy of formation (ΔG_{form}) of -135 kcal/mol upon tetramer formation, implying spontaneous, exothermic formation of an extremely stable tetramer. A homotetramer of M2-1 was predicted to be its only probable form and explains why no other oligomers were ever seen in solution during protein purification.

The zinc finger motif is visualised here for the first time and appears to be non-canonical, with no structurally similar folds found using the DALI server. However, a sequence search of the PDB suggests that the zinc finger of tristetraprolin (TTP)/Nup475, (a mammalian RNA processing factor) is related and appears to recognise zinc in a similar fashion. The M2-1 zinc finger is able to bind cadmium in addition to zinc, as evidenced by the presence of cadmium at this location within WT-P422 crystals, which contributed to the anomalous signal. In the context of the tetramer, the four zinc fingers lie on the surface, and this raises the possibility that they may also participate in RNA binding.

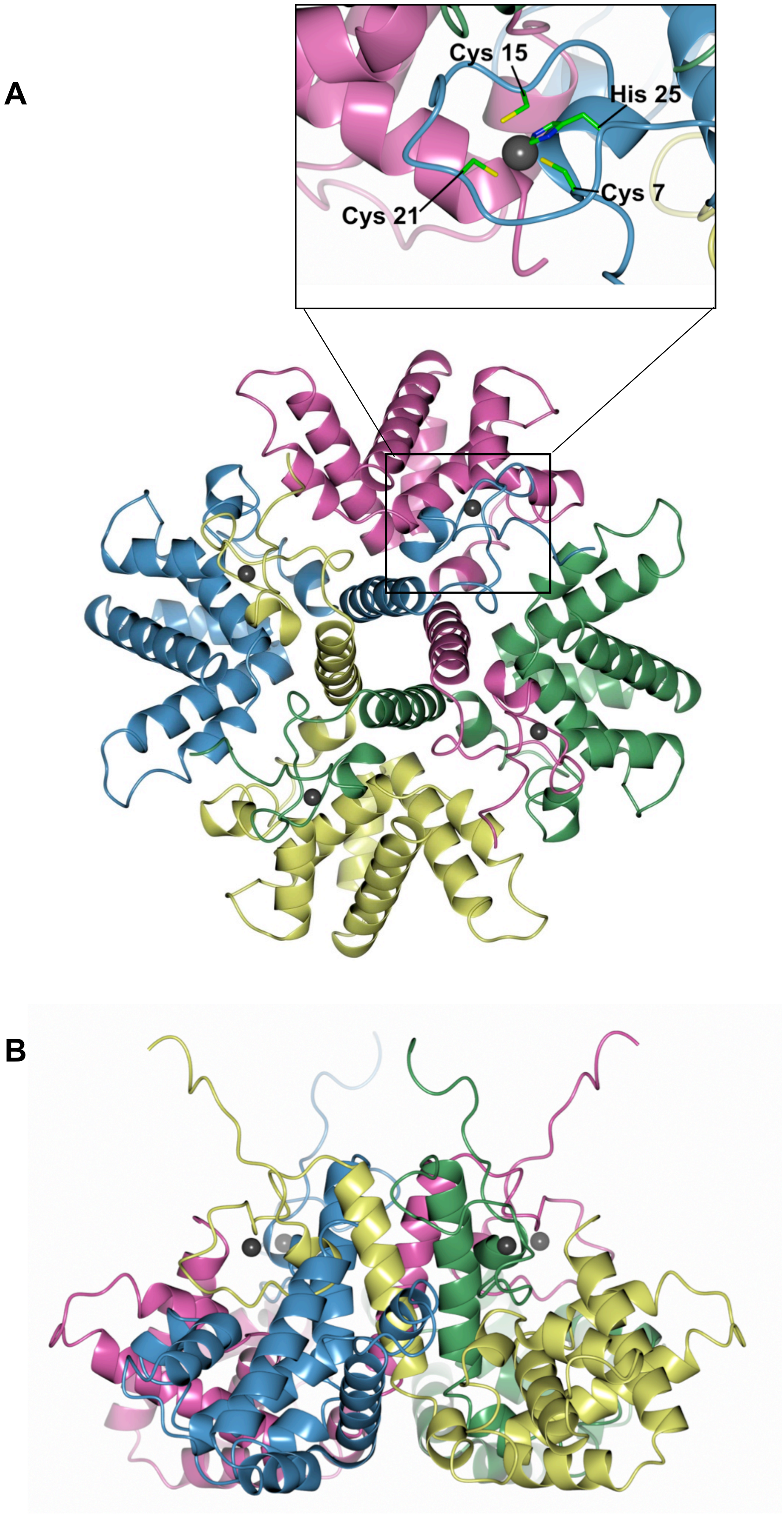
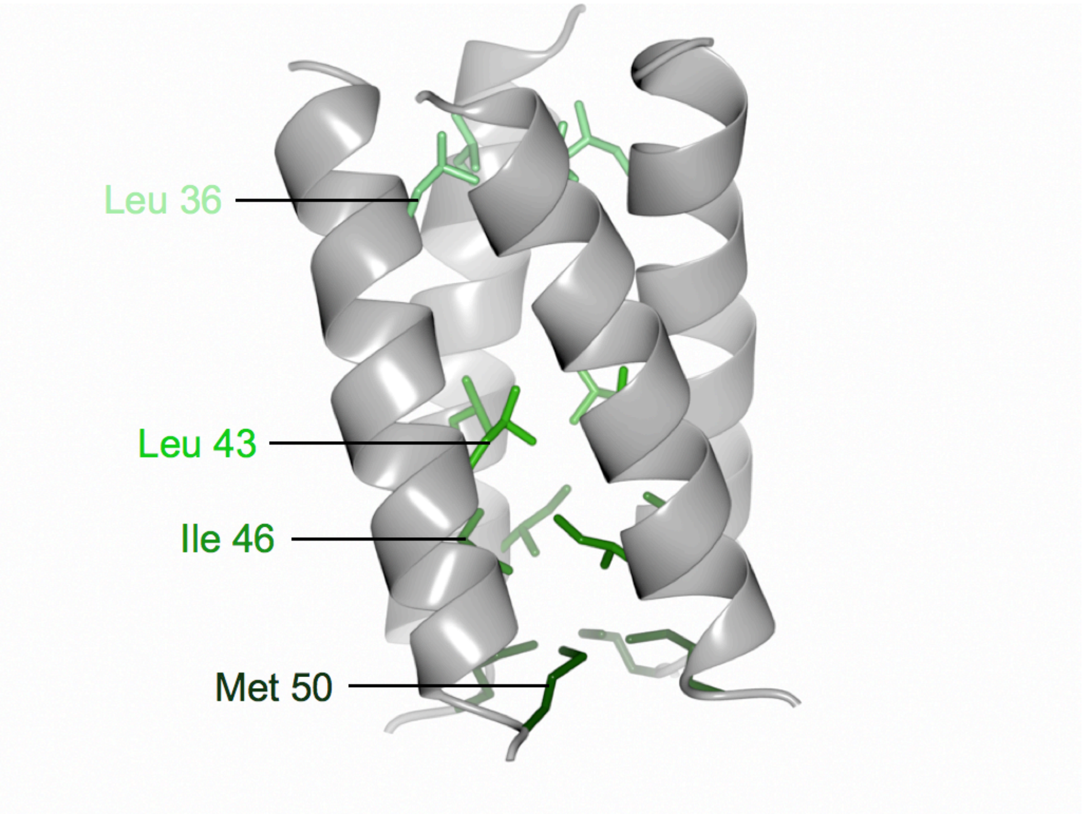

A
Cys 15
His 25
Cys 7
Cys 21
B

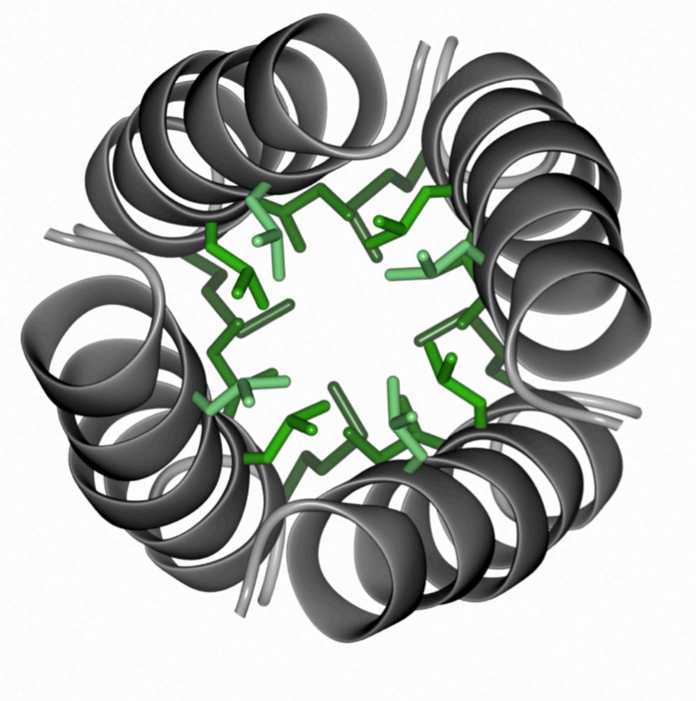
Figure 4.12 Tetrameric crystal structure of RSV M2-1.

(above) **A)** N-terminal and **B)** side views of the M2-1 tetramer are shown with monomers highlighted in different colours. The zinc finger is shown inset in A) with zinc atoms in grey. The side view highlights the interlocking protomers as the zinc finger motif extends through the tetramer to contact the N-terminal face of the adjacent protomer.

Figure 4.13 Tetramerisation via a four-helix bundle.

(below) Helix α1 mediates tetramerisation by forming a four-helix bundle that buries hydrophobic residues L36, L43, I46, M50 on all four helices. Top, side view, Bottom, top-down view. Hydrophobic residues are coloured green, graduating from light to dark from N- to C-terminus of helix α1.


Leu 36
Leu 43
Ile 46
Met 50



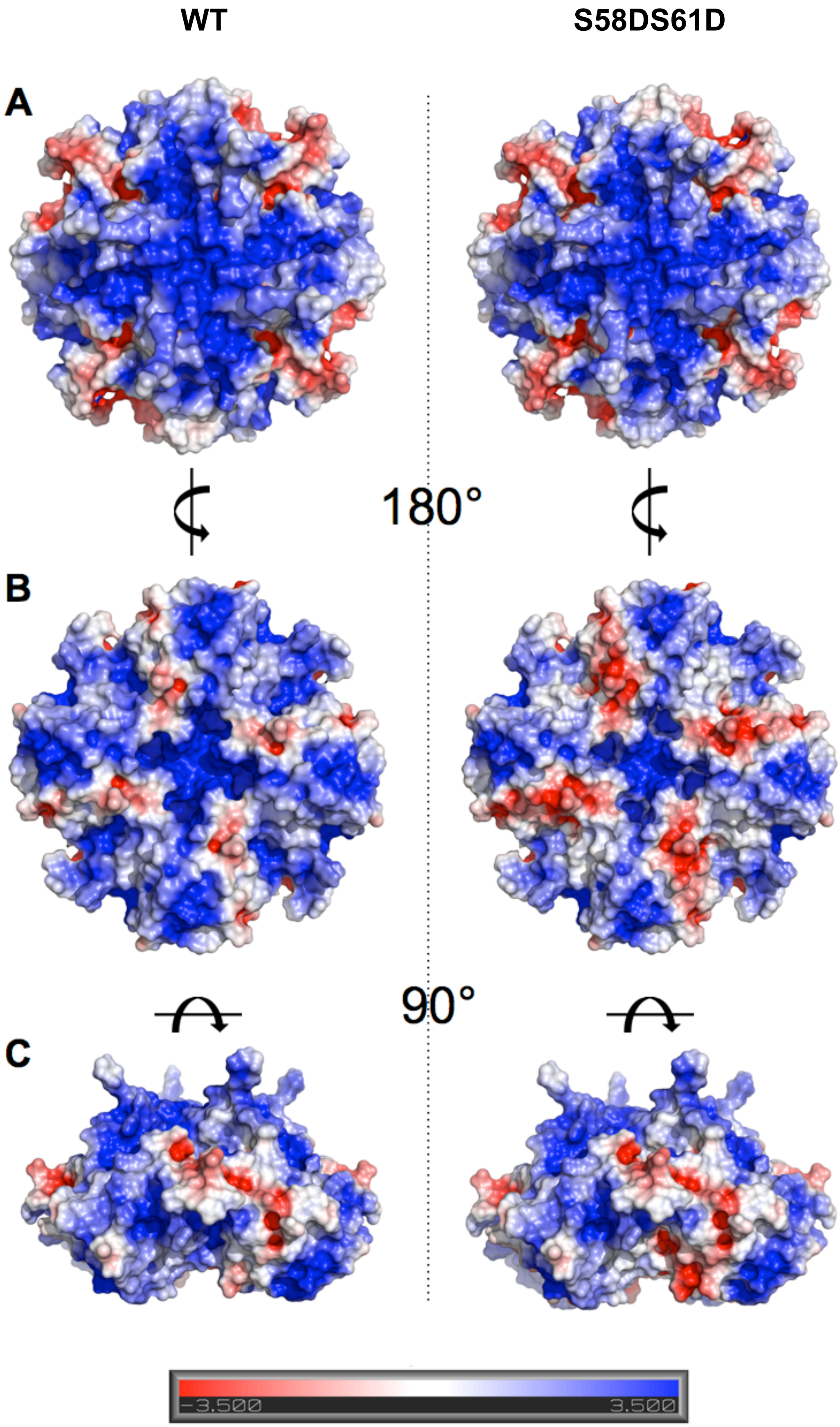
4.2.10 Electrostatics

Given the RNA-binding nature of RSV M2-1 protein, the electrostatic surface potential of the tetramer was analysed. Figure 4.14 (left, A, B and C) shows surface representations of the WT protein colour-coded blue to red according to charge (positive to negative, respectively), which was calculated using a programme called Adaptive Poisson-Boltzmann Solver (APBS). Areas of positive charge are traditionally linked to electrostatic interactions with the negative phosphate groups of nucleic acids. This analysis shows extensive, contiguous surfaces of blue positive charge that cover the N-terminal (A) and C-terminal (B) faces, and additionally wrap around the side face (C), suggesting that the region of M2-1 involved in RNA binding may encompass a large number of residues. The sideview of the tetramer also shows areas of negative charge (red). The M2-1 tetramer does not possess an obvious RNA-binding cleft or channel; the N- and C-terminal surfaces are relatively flat, which may explain why the bacterially-derived RNA was removed from M2-1 during purification with relative ease. The extensive positive charge of the tetramer is also consistent with the high pI of M2-1, 9.1. Tracts of negative charge, in red, crossed over the C-terminal surface (figure 4.14 B, left) and corresponded to the surface-exposed flexible loop region resolved in the WT-$P2_1$ structure.

4.2.11 S58DS61D

M2-1 protein is phosphorylated at serines 58 and 61 in the context of viral infection. The WT-$P2_1$ M2-1 crystal structure revealed that these sites lie on the flexible loop connecting the tetramerisation helix to the core domain. The location of these residues on a mobile, surface-exposed region suggested they might be involved in regulating the conformation of the loop, or of global M2-1 structure. The X-ray crystal structure of the M2-1 mutant S58DS61D was solved from DD-$P2_1$ data, by molecular replacement with a monomer from the WT-P422 structure (see 4.2.8). However, the structure of this mutant was identical to WT M2-1 in all respects (figure 4.14 right, A, B and C), except for the increase in negative charge at residues 58 and 61 expected from a serine to aspartate mutation (figure 4.14 B), suggesting that phosphorylation at these sites does not induce conformational change in M2-1. The flexible loop (residues 52-67) was poorly resolved in the DD-P2$_1$ dataset, suggesting that the position of the loop was not stabilised by mimicking phosphorylation at serines 58 and 61.

Figure 4.14 Surface representation of electrostatic potential.
(below) Comparing the surface charge of WT (left) and phosphomimetic S58DS61D (right) M2-1 protein. **A)** N-terminal face, **B)** C-terminal face highlighting the increased negative charge in S58DS61D (right), **C)** side view. Electrostatic potential is shown colour coded from red (negative) to blue (positive), in dimensionless units of k_b T e_c^{-1}, where k_b is Boltzmann's constant, T is the temperature, e_c is the charge of an electron.


WT
S58DS61D
A
180°
B
90°
C
-3.500
3.500

4.2.12 Comparison with other structures

4.2.12.1 M2-1 NMR structure

Blondot et al (2012) expressed truncated monomeric M2-1 and determined the solution NMR structure, comprising residues 58-177 ($M2\text{-}1_{58\text{-}177}$), which consequently omitted the N-terminal arm and zinc finger motif, the tetramerisation helix, part of the flexible loop containing phosphorylatable residues S58 and S61, and also the unstructured C-terminus (Introduction figure 1.9) (Blondot et al., 2012, Dubosclard et al., 2011). Superposition of the partial NMR structure with the full-length crystal structure described here (figure 4.15, left) shows highly similar arrangement of α-helices 2 to 7 (rmsd of 1.26 Å over the 93 residues of the core α-helices), demonstrating that the core domain of M2-1 is capable of independently folding into an RNA- and P-binding domain in the absence of the tetramerisation helix or the omitted N and C-terminal regions. The superposition also shows that N-terminus of the NMR model is unstructured, whereas this region forms part of the flexible loop in our crystal structure and contains some helical elements. Surprisingly, we were unable to use the NMR core structure as a search model for molecular replacement of our X-ray diffraction data.

4.2.12.2 Ebola virus VP30 protein

Functional similarities have been drawn between the RSV M2-1 protein and the VP30 protein of Ebola virus (EBOV) from the family *Filoviridae*, which are also members of the order *Mononegavirales*. Both are Cys_3-His_1 zinc finger-containing polymerase co-factors essential for viral transcription but dispensable for replication, and both are functionally regulated by dynamic phosphorylation at conserved N-terminal sites. However, they are proposed to function at different stages of transcription: VP30 is required for transcription initiation, a role that has been discounted for M2-1 (Collins et al., 1996), whereas M2-1 is required for antitermination, promoting both inter and intragenic polymerase processivity. Additionally, VP30 is hexameric in solution. Comparison between the VP30 C-terminal domain (only residues

142-272 out of 288 have been resolved) and the M2-$1_{58\text{-}177}$ NMR structure have previously demonstrated similarities (Blondot et al., 2012), however here we show superposition of EBOV VP30 with our full-length M2-1 X-ray crystal structure (rmsd 2.75 Å over the central 61 Cαs), which allows a more complete alignment due to the inclusion of the full M2-1 protein (figure 4.15, right). The superposition shows that, in particular, helices 3, 6 and 7 align well with M2-1, whilst there are minor shifts in orientation of helices 4 and 5. Helix 2 is considerably shifted in VP30, however this may align better with M2-1 in the presence of the VP30 N-terminus, including its zinc finger domain. VP30 also contains an additional helix at its C-terminus that has no homologue in M2-1.

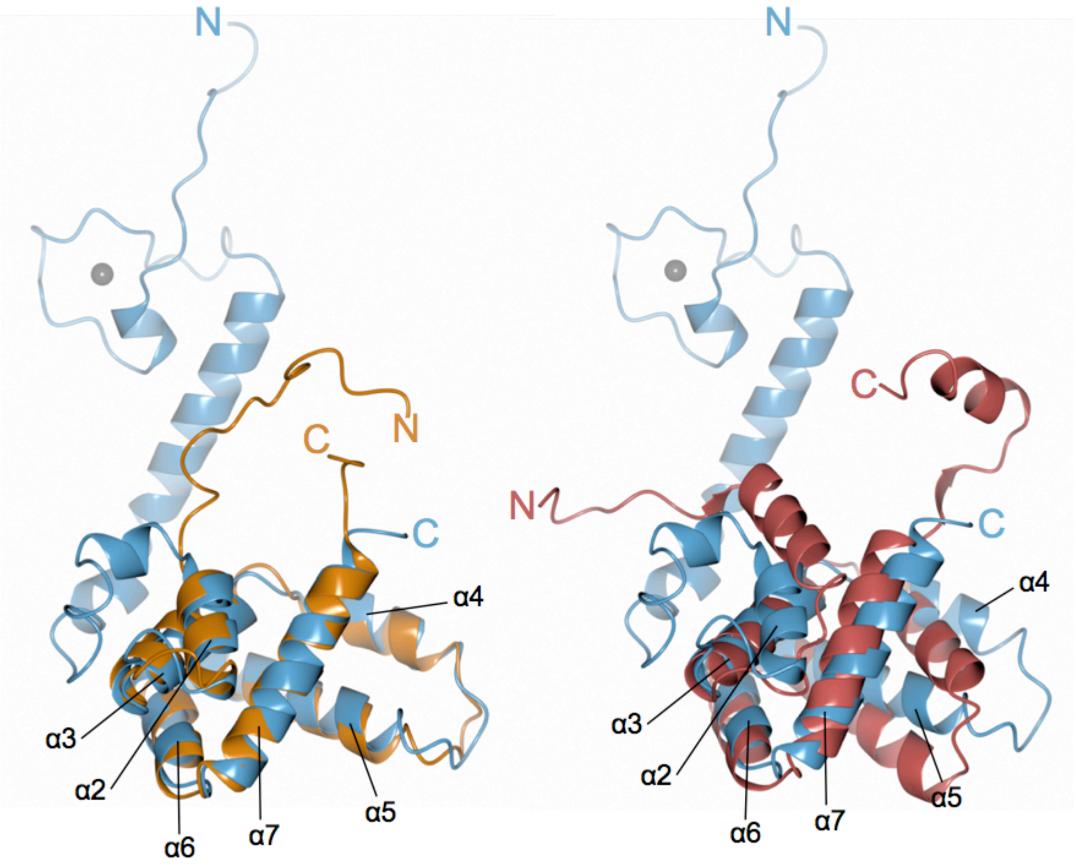


Figure 4.15 Superposition of M2-1$_{58\text{-}177}$ NMR structure and EBOV VP30 C-terminal domain with full length RSV M2-1.

Left, superposition of full length M2-1 crystal structure (blue) and partial M2-1 NMR (orange). Full length M2-1 helices are labelled and show good agreement between structure for helices 2 to 7. Right, superposition of full length M2-1 crystal structure (blue) and C-terminal VP30 (red). M2-1 helices are labelled and show good agreement for helices 3, 6 and 7, with small displacements for helices 2, 4 and 5.

4.2.12.3 Tristetraprolin/Nup475

Comparisons between the non-canonical zinc finger motifs of RSV M2-1 and TTP/Nup475 have been made in several previous publications, based on the assumption that their similar sequences would cause them to coordinate zinc in a similar manner: by 3 cysteines and 1 histidine. Nup475 has tandem Cys-X_8-Cys-X_5-Cys-X_3-His type zinc finger motifs, and M2-1 has a single Cys-X_7-Cys-X_5-Cys-X_3-His type. Whilst no structurally similar folds were found using the DALI server, including that of Nup475, a least-squares superposition of the two motifs reveals that although there is minimal secondary structure similarity, the overall disk-like fold of the zinc finger is very similar (figure 4.16). Nup475 has also been purified as a tetramer (Cao, 2004), and the short helix-turn at the C-terminus of the partial NMR structure may indicate that this region forms a longer helix analogous to the tetramerisation helix α1 of M2-1, in the full length Nup475 protein.

In the original WT-P422 data phased using SAD, the M2-1 zinc finger coordinated cadmium from the crystallisation condition, not zinc as in the other two datasets. Cadmium has no known physiological role in the human body, and its presence at anything other than trace concentration is considered toxic. This is likely due to the established ability of Cd ions to displace zinc in zinc finger motifs: in the context of cellular transcription factors such as Nup475, this has been proposed to interfere with gene expression by altering nucleic acid binding affinity (Michalek et al., 2012). Whilst we observed no structural difference between coordinating zinc, atomic radius 0.74 Å, and cadmium, 0.95 Å, and have not attempted to investigate the functional implications of this displacement here, it is of note that the coordination of cadmium by zinc fingers has been shown in a physiological context for other proteins, and may not represent a solely artefactual event.

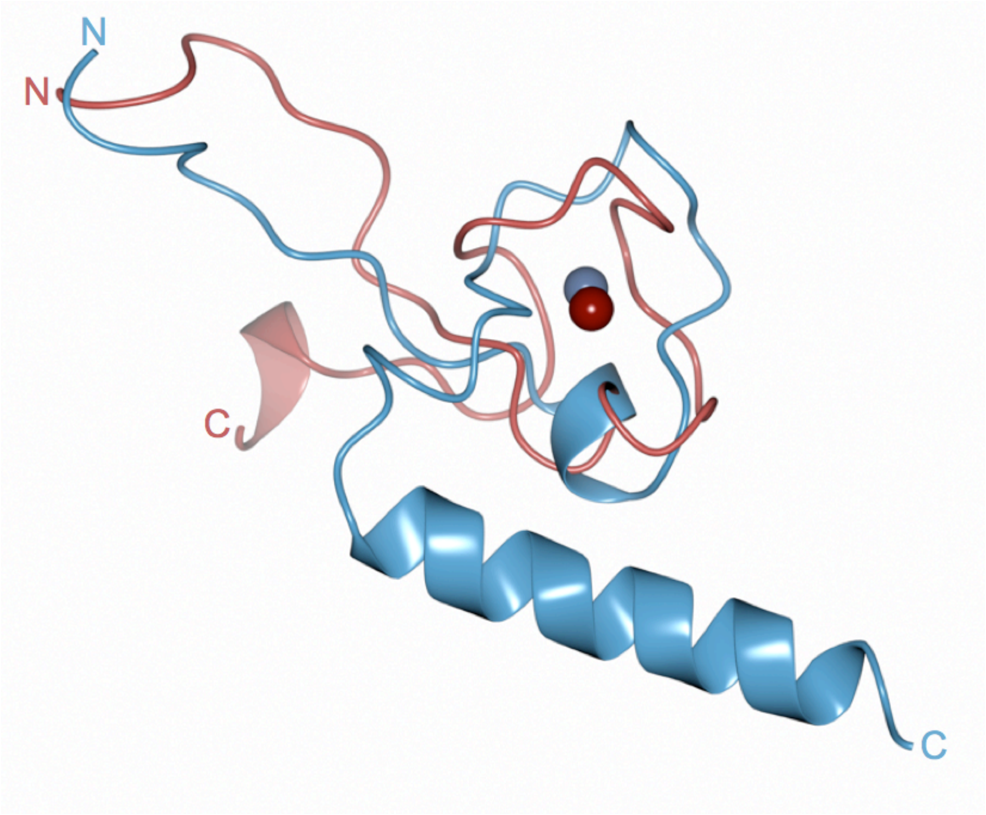


Figure 4.16 Superposition of zinc finger domains from M2-1 and Nup475. Superposition of M2-1 crystal structure (blue) and Nup475 NMR structure (red) was by a least squares fit of the Cαs from the three cysteines and 1 histidine of the motif. Both proteins have an unstructured N-terminus, with a disk-like Cys_3-His_1 zinc finger coordinating one atom of zinc per monomer. The Nup475 structure terminates before a possible helix equivalent to α1 in M2-1 at the C-terminus of this motif.

4.3 CHAPTER SUMMARY

This chapter presents the crystallisation and structure solution of RSV M2-1 protein – the first X-ray crystal structure of M2-1 from any pneumovirus. Wild type M2-1 was experimentally phased using cadmium and zinc atoms to a maximum resolution of 2.52 Å. Phosphomimetic M2-1, S58DS61D, was solved by molecular replacement to 2.4 Å. All three structures determined in this work, outlined in tables 1 and 2, were the same with the exception of their surface charge distribution and small differences in side chain orientation. During this project, a partial solution NMR structure of M2-1 core domain (residues 58-177) was published. The full-length M2-1 structure we present here is largely in agreement with the NMR structure, which suggests the core is able to independently fold into a compact domain without influence of either N-terminal or C-terminal regions.

The crystal structure confirms that M2-1 forms many of the features that had been predicted in the literature, such as a zinc finger motif coordinating one atom of zinc/cadmium per monomer, an oligomerisation region, and a globular α-helical domain. What this crystal structure reveals for the first time is the extended nature of the M2-1 monomer, with the zinc finger projecting away from the core on an N-terminal arm, and the structural basis behind the 'tight' tetramer formation seen in solution: the extensive oligomerisation interface mediated by the four-helix bundle and the zinc finger, which contacts the adjacent protomer. The extensive contact made by the zinc finger is consistent with the essential role of this motif in M2-1 function, and suggests it plays an important role in preserving the structural integrity of the tetramer. Other virus-encoded tetramers are known to use similar four-helix bundles to drive oligomerisation, with prominent examples being the influenza A virus M2 ion channel and the P protein of the paramyxovirus Sendai virus.

This is only the second structure of a zinc finger motif coordinating cadmium in the PDB, and is the first crystal structure: a solution NMR structure of the 'classic' Cys_2-His_2 zinc finger from *Arabidopsis thaliana*

SUPERMAN protein, shows a conformational change when zinc is displaced by cadmium (Malgieri et al., 2011). We saw no such change in our M2-1 crystal structures, however the two zinc fingers are from different classes and may respond differently to alternative cations. Equally, it may be that outside of the physical constraints that crystallisation can impose, the M2-1 zinc finger motif may also undergo a conformational change when coordinating cadmium. In either case, the structure presented here provides new and interesting information to the growing field of zinc finger-mediated cadmium toxicity.

The tight tetramer formation displayed in our structure suggests that M2-1 is not capable of forming alternative higher order oligomers, and this is consistent with our observation that no higher order species can be detected in solution by either gel filtration or mass spectrometry. M2-1 has recently been proposed to perform a role in virion assembly, forming a matrix-like layer surrounding the RNP within the mature virus particle (Liljeroos, 2013). However, our results are not consistent with this role, as M2-1 does not possess the necessary architecture needed to generate a large two-dimensional planar lattice. Also inconsistent with a role for M2-1 as a matrix layer, is its relatively low abundance within purified virus particles (Radhakrishnan et al, 2010), which suggests it is a minor component of the virion. However, we cannot rule out the possibility that alternative oligomeric forms of M2-1 exist in the context of viral infection, with radically different functions as has recently been shown for the Ebola VP40 matrix protein (Bornholdt et al, 2013).

The electrostatic surface potential of the M2-1 tetramer allows speculation over regions involved in RNA binding, a known function of this protein. Extensive, continuous, positively charged regions cover all three faces of the structure. The lack of an apparent RNA-binding cleft supports the theory that RNA binding by M2-1 is electrostatically, not sequence, driven and that this is reflected in a lack of sequence specificity. The X-ray crystal structure of a phosphomimetic M2-1 mutant did not reveal a global, or local, conformational change in response to increased negative charge at serines 58 and 61, although the poorly defined electron density for the flexible loop means we cannot rule out a shift in equilibrium between multiple conformations of this region. The discovery that negatively charged phospho-serines 58 and 61 reside on a flexible loop that lies over the positively charged surface of the C-terminal face, suggests that regulation of binding (RNA or otherwise) to this surface could be regulated by disrupting the positive charge with phosphorylation at these sites. Both the concept of RNA-binding specificity, and the role phosphorylation might play in this, are investigated in Chapter 5 of this thesis.

5 CHAPTER 5: UNDERSTANDING M2-1 RNA BINDING AND THE ROLE OF PHOSPHORYLATION

5.1 CHAPTER INTRODUCTION

RSV M2-1 was first described as an RNA-binding protein after a reduction in electrophoretic mobility upon addition of RNA was observed in gel-retardation assays (Cuesta et al., 2000), and after the coimmunoprecipitation of a diffusely migrating RNase A-sensitive species at the top of gels analysing inorganic [^{33}P]phosphate-labelled M2-1 (Cartee and Wertz, 2001). The specificity of this RNA binding, and its relevance to the RSV lifecycle, has been continually debated since these original publications: M2-1 has been proposed to bind to long RNAs (>700 nts) with no sequence specificity, short RNAs (<80 nts) containing the antigenomic leader sequence, RSV-specific mRNAs, A-rich RNAs, and with no sequence specificity at all. Each of these suggestions carries with it implications for the RNA species targeted and thus the mechanism of M2-1 antitermination activity, e.g. association with the genome vs. mRNA. This chapter presents the interrogation of M2-1 RNA-binding specificity, and the residues of the protein that are involved, using a fluorescence polarisation anisotropy (FPA) assay and a range of both RNAs and mutant M2-1 proteins.

The principle of FPA is that when a fluorescently labelled molecule (e.g. RNA) is excited by polarised light, the fluorophore emits light with a degree of polarisation that is inversely proportional to the ‘tumbling rate’ of the labelled molecule. For example, a small labelled RNA has a high tumbling rate due to rapid Brownian motion in solution, which will result in depolarisation of the emitted light as the fluorophore reorients during the lifetime of its excitation (~4 nanoseconds for fluorescein, used here); if the small labelled RNA is bound to a larger protein molecule, this complex will have a slower rate of rotation and the emitted light will still be polarised to an extent (figure 5.1). Binding of M2-1 to a range of 3’-fluorescein-labelled RNAs was assessed in this way, with higher polarisation values indicating binding.

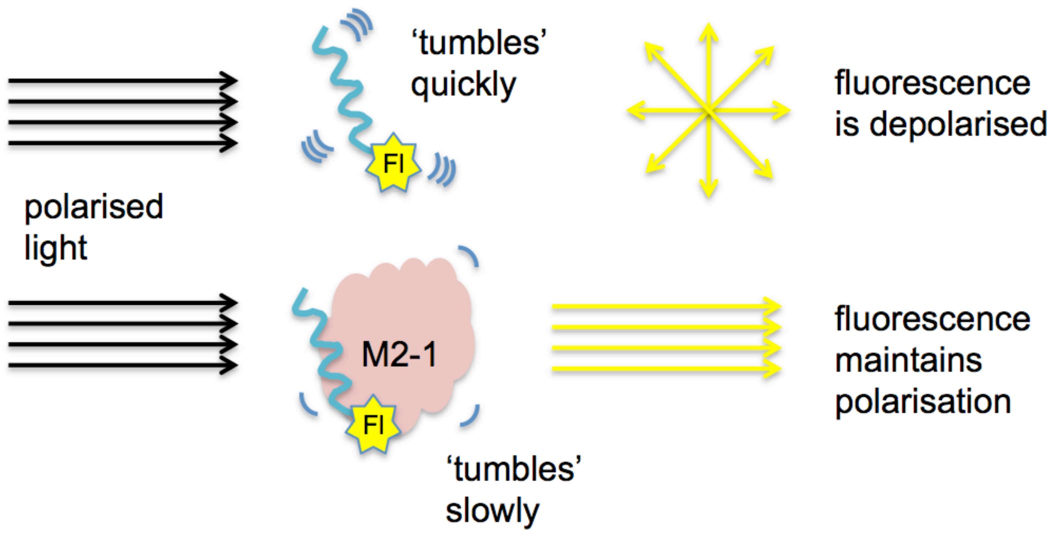


Figure 5.1 Fluorescence polarisation anisotropy.

(Top) Small 3'-fluorescently-tagged RNAs have quick rotation, or tumbling, in solution and so emitted fluorescence is depolarised by the rapid motion. (Bottom) Binding of a larger protein (e.g. M2-1) to the RNA slows the rate of tumbling, and so emitted fluorescence is depolarised less i.e. it maintains more polarisation.

This chapter also investigates the effect of M2-1 phosphorylation on its RNA-binding ability. Whilst data suggests that ablation of M2-1 phosphorylation reduces its transcription antitermination efficiency in the context of the established minigenome system (Cartee and Wertz, 2001), there is little concrete evidence for phosphorylation having either a positive or negative effect on M2-1 RNA binding: Cartee et al found that mutating either serine 58 or 61 to alanine had no effect on coimmunoprecipitation of RNA from mammalian tissue culture; Cuesta et al suggest that phosphorylation reduces the RNA-binding capacity of M2-1, however they present no primary data and incorrectly reported the phosphorylation sites as threonine 56 and serine 58 (Cuesta et al., 2000); Tran et al saw no effect on binding of yeast tRNA to M2-1 protein that had been *in vitro* phosphorylated whilst GST-tagged and bound to glutathione sepharose (Tran et al., 2009). These publications either focus on ablation of phosphorylation, or use techniques that are poorly representative of M2-1 in the virus life cycle. Here, we demonstrate that mimicking phosphorylation reduces the RNA-binding affinity of M2-1 for multiple RNA sequences, and discuss how this links to the location of serines 58 and 61 as revealed by our X-ray crystal structure in Chapter 4. This chapter also summarises data produced by our collaborators in Jean-François Eléouët's group for the publication "Crystal structure of the essential M2-1 antiterminator of HRSV and consequences of dynamic phosphorylation" – under review, where it was shown that both mimicking and ablating phosphorylation of M2-1 in the minigenome system has a deleterious effect on M2-1-dependent expression.

5.2 RESULTS

5.2.1 Confirmation of M2-1 RNA binding

To confirm that bacterially-expressed, RNA-free M2-1 produced as in Chapter 3 was capable of binding synthetic RNAs, 10 μM M2-1 protein was incubated with 10 μM leader RNA [a 3'-fluorescein 44 nucleotide RNA representing the 3' genomic leader sequence, see Appendix II] and analysed by nanoESI time-of-flight mass spectrometry. Figure 5.2 shows a comparison of a protein-only spectrum (top) with that of the M2-1 tetramer bound to RNA (bottom), demonstrating an increase in mass:charge ratio (m/z) that suggests M2-1 was bound to 2 RNAs (44mer RNA ~14,960 Da). M2-1 protein was not observed binding to any other number of RNAs in these analyses, although a small amount of an undecipherable higher-order species (m/z >8000) was seen (data not shown) which could be RNA-mediated aggregation. A low signal:noise ratio, shown by the low m/z background peaks and the poorly defined peaks of interest, indicated that a relatively small proportion of the protein injected was successfully being detected (figure 5.2, bottom). In addition, no free M2-1 was detected in the presence of RNA. Taken together, these data indicated that the majority of M2-1 may have bound more/fewer than 2 RNAs per tetramer, but these complexes were not identified by MS - possibly due to the long RNAs 'mopping up' free protein by binding M2-1 at multiple points along their length, causing aggregation that resulted in poor signal:noise.

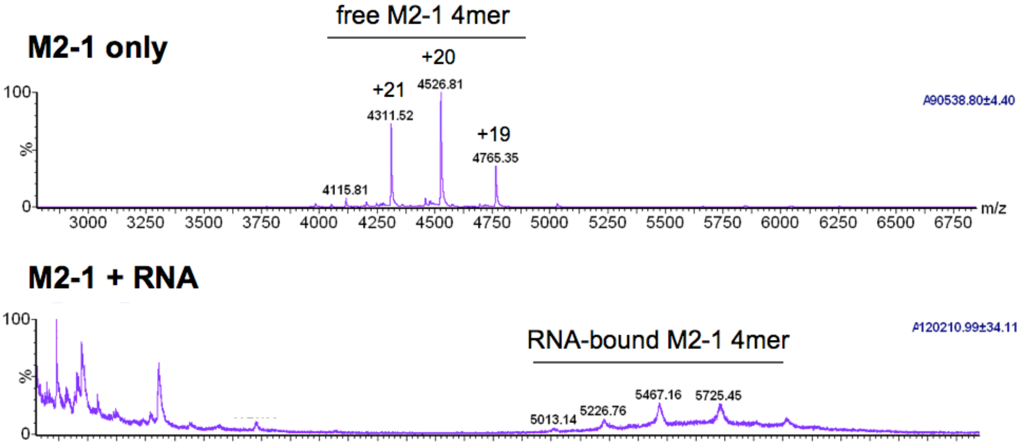


Figure 5.2 Mass spectrometric analysis of M2-1 RNA binding.

NanoESI-TOF MS was used to confirm that bacterially-expressed, RNA-free M2-1 was capable of binding 3'-fluorescein-tagged synthetic RNAs. **Top,** M2-1 only spectrum shows an RNA-free M2-1 tetramer; peaks with 19, 20 and 21 charges are labelled and give a mass of 90,538.80 Da. **Bottom,** M2-1 + RNA spectrum shows an RNA-bound tetramer with a mass of 120,210.99 Da; 44mer RNA ~14,960 Da, therefore this species is M2-1 bound to 2 RNAs. Poor signal:noise is evidenced by background peaks at the low m/z end of the spectrum and broad peaks of interest.

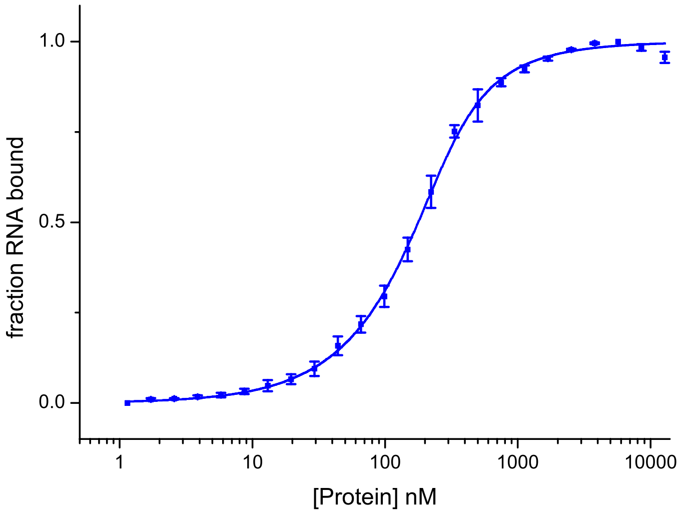


Figure 5.3 RNA binding of Bunyamwera nucleocapsid protein.

A typical binding isotherm for BUNV N protein, used as a known RNA-binder to establish the FPA assay. Full details of this work are outlined in Ariza et al. 2013.

5.2.2 Assay development

A fluorescence polarisation anisotropy assay was developed in 96- and 384-well plate format for determining RNA binding affinity of M2-1 to various sequences. To establish the technique, a known RNA-binder from previous work in our group, the nucleocapsid protein (N) of Bunyamwera virus (BUNV), was used with its cognate RNA [a 3'Fl 48 nucleotide RNA representing the 5' terminus of the BUNV S segment antigenome] (Ariza et al., 2013). Binding of BUNV N protein was achieved using the method outlined in Methods 2.2.14.3 (figure 5.3, above) and so this method was applied to M2-1 protein. 3'-Fl RNA had low intrinsic polarisation in the absence of M2-1 protein, and a titration of M2-1 with no RNA showed a consistent level of polarisation equivalent to that of buffer only controls. For initial M2-1 RNA binding tests, plates were read after equilibrating at room temperature for 30 minutes, and then again the following day. Values at both timepoints were the same, and so M2-1-RNA interactions were deemed to have equilibrated after 30 minutes.

5.2.3 Assessing sequence specificity of RNA binding

5.2.3.1 Longer RNAs (44 nucleotides)

The 44mer leader RNA, used in the mass spectrometric analysis described above, was tested for binding to M2-1 in the FPA assay in addition to a 44 nucleotide RNA representing the M/SH gene junction, including M gene end and SH gene start sequences (see Appendix II). Neither of these RNAs bound to M2-1 in this assay, possibly due to the aggregation effect suggested by the mass spectrometry, or due to the 'propeller effect' where the rotation of the fluorophore and the rotation of the RNA-protein complex become uncoupled, as a considerable proportion of the RNA is not in contact with the protein.

5.2.3.2 Shorter RNAs ≤13 nucleotides

A panel of short RNAs, 13 nucleotides or less, were selected to analyse the specificity of M2-1 RNA binding in light of our results with the 44mer RNAs. These included *cis*-acting signals from the RSV genome, sequences present in RSV mRNAs, polyG, A, U or C homo-oligomeric sequences, and other non-viral RNAs of unrelated sequence, all of which were tagged at their 3' ends with fluorescein (Fl) (see figure 5.5 A for sequences). Figure 5.4 A shows binding isotherms typical of the data collected and represents the highest and lowest affinity RNAs tested here. All data were fitted using logistic regression to an R-square value > 0.98 with Hill coefficients in the range of 0.5-0.9, which may suggest some negative cooperativity of RNA binding, however further analysis of this was beyond the scope of this thesis.

Of the four homo-oligomeric RNAs, polyA RNA was bound with the highest affinity (figure 5.4 B). Binding of polyU and C RNAs could not be saturated at available protein concentrations and consequently data could not be accurately fitted; suggested Kds of > 1 mM are based on an extrapolated estimate of the 100 % RNA bound protein concentration. In contrast, polyG RNA showed no binding to M2-1 (data not shown), whereas a polyU RNA with an inserted single G nucleotide, 'g8mer', showed significantly higher binding affinity than the unmodified polyU.

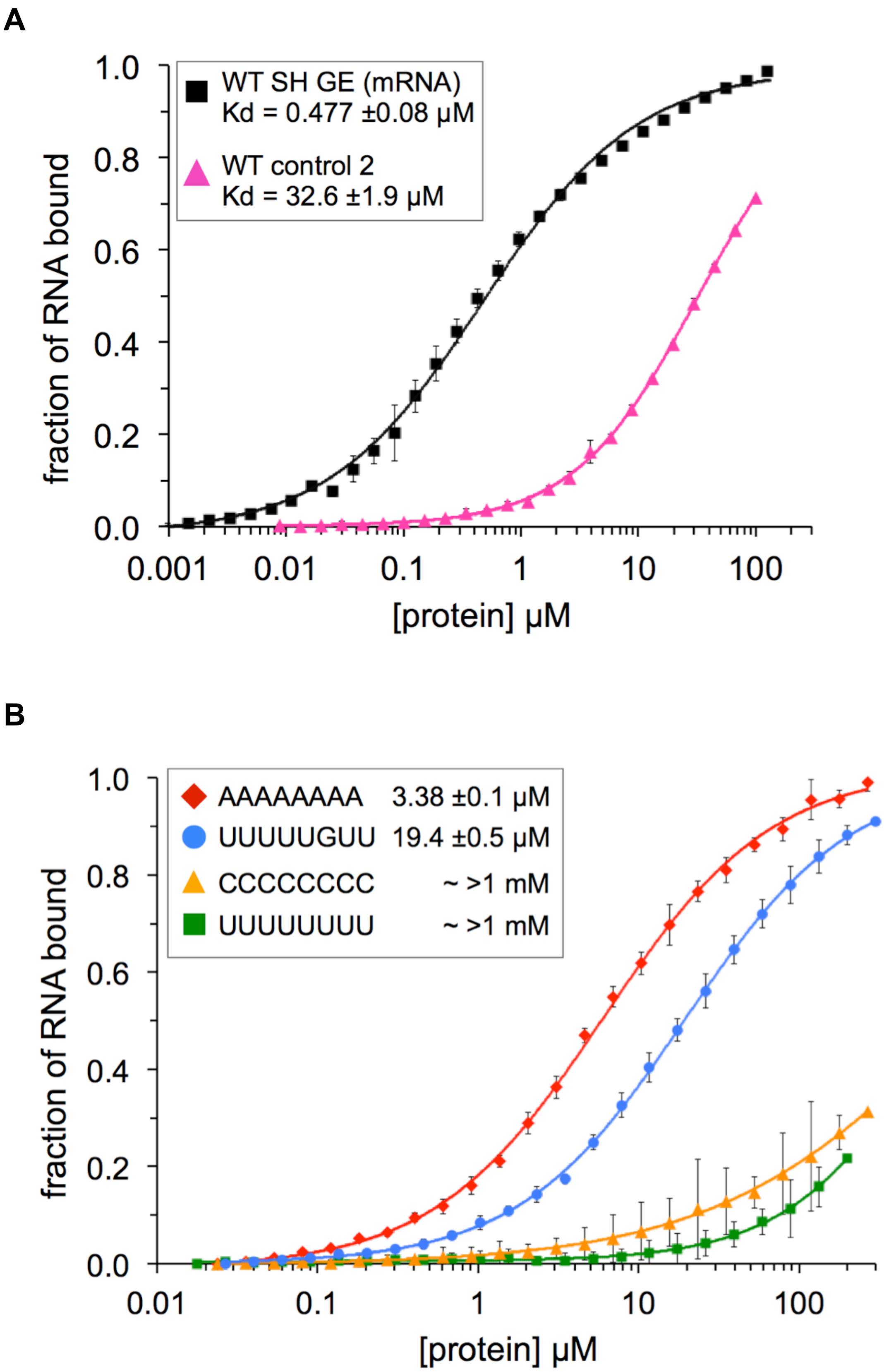
We next wanted to examine M2-1 binding to RSV-specific sequences. M2-1 possesses transcription antitermination activity at gene ends, and as the termination event is signalled by the conserved gene ends alone, we chose to focus our analysis on the binding of M2-1 to RNAs bearing these sequences. Two gene end signals were selected, SH and F, and corresponding genome and mRNA sense oligonucleotides representing these sequences were synthesized and tested in binding assays. For both gene ends, significantly higher binding affinities were recorded for the mRNA sense RNAs (figure 5.5 B) compared to their genomic sense reverse complement, consistent with a preference for A-rich sequences. Although this is difficult to see in the bar chart of figure 5.5, the table in 5.5 B presents the statistical analysis that confirmed this significance. In addition, we observed a higher binding affinity for SH than for F gene end sequences, in both senses, despite their similarity in 'A' content (figure 5.5 B). One possible explanation is that M2-1 not only has a preference for A-rich sequences, but that the spacing of adenosine nucleotides is important: the genomic sense A-rich region of the gene ends for F (3'-AAUAUA-5') and SH (3'-AAUUAA-5') vary in their arrangement, but not their 'A' content.

Following on from this finding, we examined the binding of M2-1 to the conserved RSV gene start sequence responsible for transcription initiation. This sequence in genome sense is highly pyrimidine rich, possessing only a single A-residue in its entire length, and the binding affinity of this RNA was the lowest of any RSV-specific RNAs tested (figure 5.5 A). Other non-viral RNAs tested included oligonucleotides control 1 and control 2 (figure 5.5 A), both of which are low in 'A' nucleotide content, and exhibited correspondingly low binding affinities.

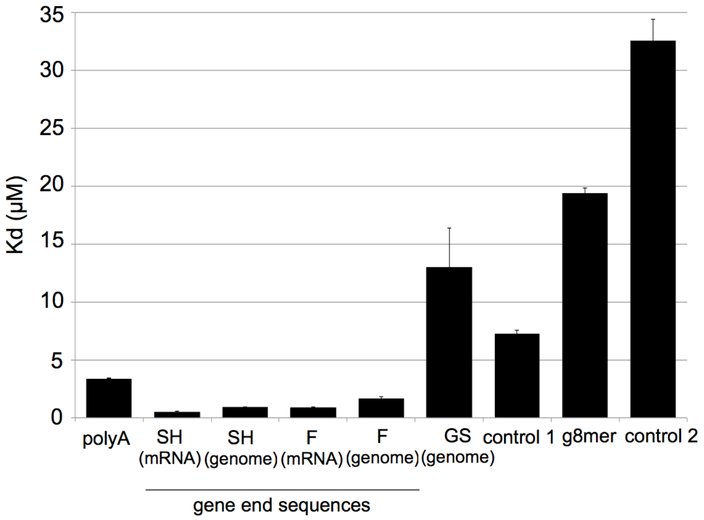
All sequence analysis of RNA binding presented in this chapter is summarised in figure 5.5 A. Taken together, these data indicate that tetrameric, full-length M2-1 exhibits a consistent preference for A-rich sequences, particularly in the context of mRNA sense (positive) gene ends, as has been previously shown for monomeric M2-1$_{58-177}$ by NMR (Blondot et al., 2012). In contrast to this work, however, full-length tetrameric M2-1 protein displays a significant difference in RNA binding affinity to gene ends from different junctions.

Figure 5.4 RNA binding isotherms for various sequences.

(below) A) Binding isotherms for the highest (SH gene end, mRNA sense) and lowest (control 2) affinity RNAs. **B)** Isotherms for homo-oligomeric sequences and g8mer (blue), with corresponding Kds. Sequences are shown 3' to 5'. Data were fitted by logistic regression and error bars represent the standard deviation from the mean of triplicate data. Kd was calculated from the concentration of protein required for 50 % RNA to be bound. PolyU and polyC RNAs did not saturate at available protein concentrations, and so Kds are estimates based on the projected 100 % RNA bound concentration.


A
WT SH GE (mRNA)
Kd = 0.477 ±0.08 µM
WT control 2
Kd = 32.6 ±1.9 µM
fraction of RNA bound
1.0
0.8
0.6
0.4
0.2
0.0
0.001
0.01
0.1
1
10
100
[protein] µM
B
AAAAAAAA 3.38 ±0.1 µM
UUUUUGUU 19.4 ±0.5 µM
CCCCCCCC ~ >1 mM
UUUUUUUU ~ >1 mM
fraction of RNA bound
1.0
0.8
0.6
0.4
0.2
0.0
0.01
0.1
1
10
100
[protein] µM

A



RNA	sequence (Fl-3’-5’)
polyA	AAAAAAAA
SH GE (mRNA)	AAAAAUUAAUUGA
SH GE (genome)	UCAAUUAAUUUUU
F GE (mRNA)	CAAAAUAUAUUGA
F GE (genome)	UCAAUAUAUUUUG
GS (genome)	CCCCGUUUAU
control 1	AAUAGAUGUUC
g8mer	UUUUUGUU
control 2	CCUGUUGCU

B

RNA (Fl-3’-5’)	Kd (µM)	T-Test P-value, mRNA vs. genome	T-Test P-value, F vs. SH
F (genome) UCAAUAUAUUUUG	1.68 ±0.03	0.000787	genome = 0.000476 mRNA = 0.005626
F (mRNA) CAAAAUAUAUUGA	0.86 ±0.04		
SH (genome) UCAAUUAAUUUUU	0.92 ±0.04	0.000829	
SH (mRNA) AAAAAUUAAUUGA	0.48 ±0.02		

Figure 5.5 Sequence specificity of M2-1 RNA binding.

A) Chart of Kds for all RNAs tested and a table of their sequences. Genome and mRNA labels indicate the sense of the RNA in the context of viral life cycle. GS, gene start. **B)** Table comparing affinities for SH and F gene end sequence in both mRNA and genome sense. Unpaired, 2-tailed Student’s T-tests were performed for mRNA vs genome sequences, and for F vs. SH gene end sequences. All p-values were significant to $p < 0.01$.

5.2.4 Residues involved in RNA binding

To examine the contribution of specific M2-1 residues towards RNA binding, a panel of M2-1 mutants was created using site directed mutagenesis (see Appendix I for primer sequences), and expressed as full length tetrameric GST-M2-1 in bacterial cells. Mutants were expressed, purified and cleaved from GST as per WT M2-1, although size exclusion chromatography (SEC) was not used to allow for higher throughput: binding of pre- and post-SEC M2-1 was found to be comparable, however all affinities presented here are relative to WT produced in parallel, in the same manner as the mutants.

Single and double mutants were selected based on the full-length M2-1 tetramer electrostatic surface model (Chapter 4, figure 4.14) and consideration of previously examined M2-1 residues within the monomer core region (Blondot et al, 2012). Residues directly contributing to the zinc finger motif were not selected as these mutants have been shown to misfold and aggregate (Jean-François Eléouët, personal communication). The positively charged residues selected for alteration were located in both the core and N-terminal regions, and were mutated to either alanine (to negate charge) or aspartate (to reverse the charge). Figure 5.6 A shows a surface representation of M2-1 with all mutants tested in the FPA assay highlighted in yellow on one monomer (dark grey).

Mutants were tested for their binding to our highest affinity sequence described above, the SH mRNA sense gene end. The data were expressed as the fold change in Kd relative to WT (figure 5.6 B), and asterisks indicate statistical significance. Of the panel of mutants tested, only K19AR20A exhibited essentially unchanged RNA binding affinity. All others resulted in significant loss of RNA binding affinity, including K92A, K92D, K150A, R150D, R151D, K150AR151A and K159A.

These findings confirm Blondot et al (2012), which found that, qualitatively, K92, K150 and R151 were the residues essential for the binding of GST-M2-1$_{58\text{-}177}$ to yeast tRNA. Our structure revealed that N-terminal arginine residues R3 and R4 also contributed to the positively charged N-terminal face, and FPA analysis of the R3AR4A double mutant shows these residues also significantly influence RNA binding, implying the M2-1 residues that participate in RNA binding are not solely confined to the main basic cluster of the core. We saw no significant difference between mutating selected positively charged residues to alanine (uncharged) or aspartate (negatively charged), unlike in previous studies where introducing negative charge increased the loss of RNA binding affinity relative to neutralising positive charge (Blondot et al., 2012).

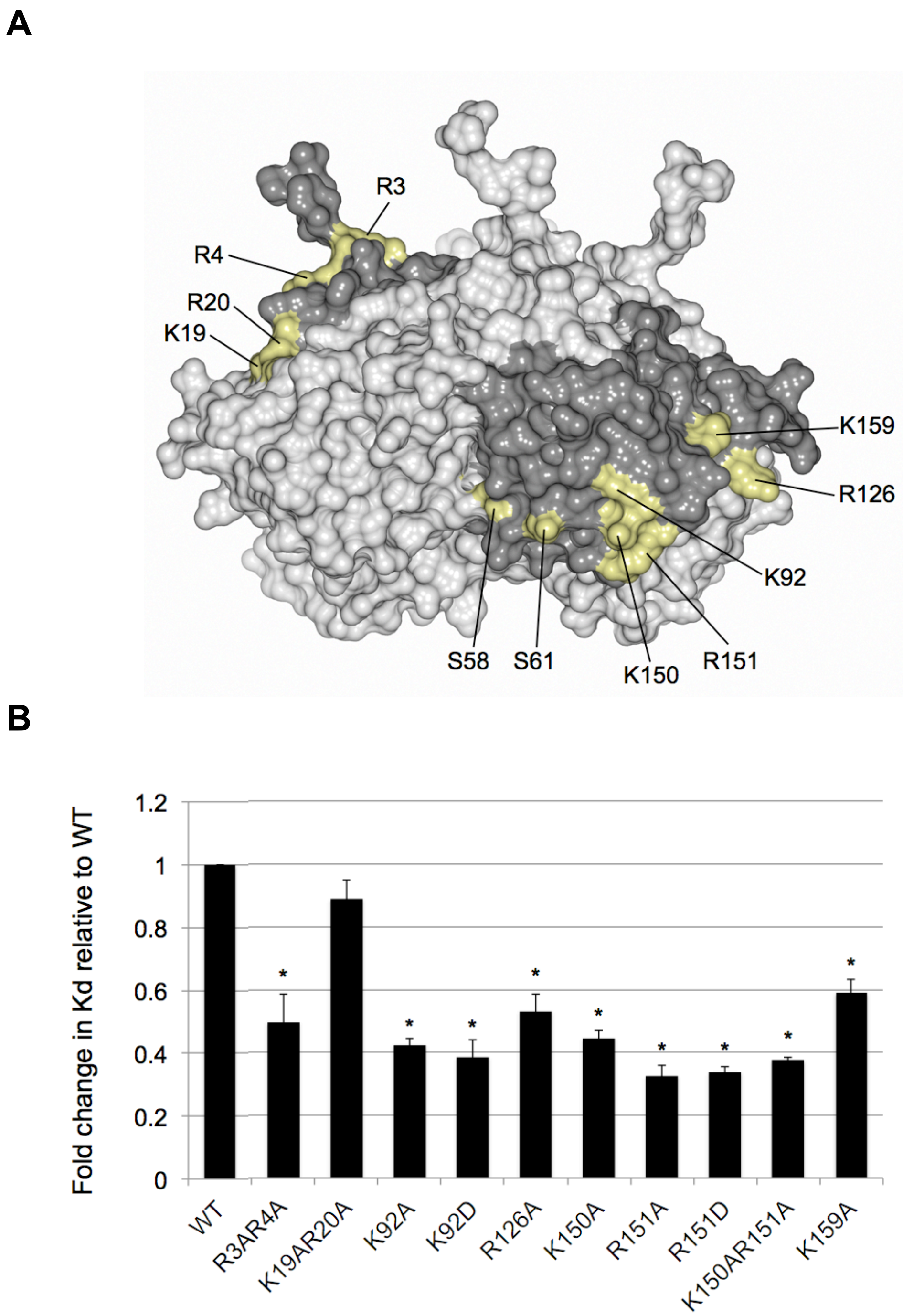


Figure 5.6 Identification of M2-1 residues involved in RNA binding.

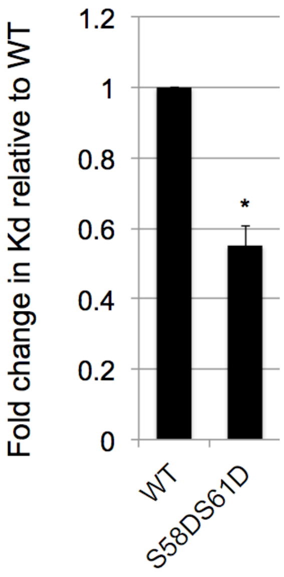
A) A tilted side view surface representation of M2-1 tetramer highlighting all residues tested in the FPA RNA binding assay, including the phosphomimetic mutations described below. **B)** Bar chart comparing the binding affinity of M2-1 mutants, which is expressed as the fold change in Kd relative to WT. Asterisks indicate statistical significance, where $p < 0.01$ with an unpaired, 2-tailed Student's T-test.

5.2.5 Phosphorylation and RNA binding

Figure 5.6 A also highlights the phosphomimetic mutant, S58DS61D, that was crystallised and the structure solved in Chapter 4. The location of S58 and S61 adjacent to the positively charged surface of the C-terminal core, raised the possibility that increased negative charge would reduce RNA binding due to electrostatic repulsion of the RNA phosphate backbone. The role of phosphorylation in M2-1 RNA binding was investigated by comparing the binding of this mutant and WT M2-1 (unphosphorylated in this system) to various RNA sequences in the FPA assay.

Analysis of this mutant with the SH mRNA sense gene end RNA, revealed a significantly reduced RNA binding ability compared to the unphosphorylated WT M2-1 (figure 5.7 A), confirming that the altered negative charge density conferred by mimicking phosphorylation significantly affected the ability of S58DS61D to bind RNA. We also tested the RNA binding ability of this mutant with other RNAs of differing lengths and nucleotide sequences. These results showed that phosphomimetic mutant S58DS61D consistently demonstrated a statistically significant reduction in affinity of approximately 2-fold across all RNAs tested (figure 5.7 B). This finding supports the conclusion that the negative charge density at residues 58 and 61 influences the ability of M2-1 to bind RNA, and suggests that phosphorylation of M2-1 during RSV infection would similarly reduce binding affinity to target RNAs, which may consequently affect its processivity function.

A



B

RNA (Fl-3'-5')	WT Kd (µM)	S58DS61D Kd (µM)	T-Test P-value, WT vs. S58DS61D
polyA AAAAAAAA	3.38 ±0.1	10.95 ±0.2	0.0000009
GS (genome) CCCCGUUUAU	12.59 ±0.7	28.59 ±0.9	0.0033492
control 1 AAUAGAUGUUC	7.29 ±0.3	14.47 ±0.8	0.0000245
SH (mRNA) AAAAAUUAAUUGA	0.48 ±0.02	0.776 ±0.02	0.0096107

Figure 5.7 RNA binding affinity of the phosphomimetic mutant S58DS61D.

A) Comparison of the affinity of S58DS61D for SH gene end (mRNA) relative to WT. **B)** Binding of S58DS61D to a number of RNA sequences, tested with WT M2-1 in section 5.2.3.2. The Kds for WT and S58DS61D are compared here. P-values (asterisk in A) are from unpaired, two-tailed Student's T-tests, and all indicate statistical significance, where $p < 0.01$.

5.2.6 M2-1 antitermination in minigenome system

To analyse the effect of both RNA binding mutants and phosphorylation on transcriptional antitermination activity of M2-1, the phenotypes of our mutants were assessed in the context of the well-established RSV minigenome system (Yu et al., 1995, Hardy and Wertz, 1998). This work was performed by our collaborators in Jean-François Eléouët's group (Institut National de la Recherche Agronomique, Jouy-en-Josas, France), as featured in their previous M2-1 publications (Tran et al., 2009, Blondot et al., 2012). Briefly, this reverse genetics system requires T7 RNA polymerase-expressing mammalian cells, either by recombinant vaccinia infection or by using the stable cell line, BSR-T7/5. These are co-transfected with a truncated substitute genome (containing two transcriptional units separated by a native gene junction, and flanked by the leader and trailer regions required for replication and transcription) and the support plasmids encoding the N, P, L and M2-1 proteins under T7 promoter control – the four proteins that are necessary and sufficient for transcription and replication of the minigenome.

For this work, the downstream transcriptional unit was replaced with the firefly luciferase gene (Tran et al., 2009), the expression of which is absolutely dependent on functional M2-1 expression. Site directed mutagenesis of the M2-1 support plasmid was used to assess the impact of our panel of RNA binding mutants, discussed above.

5.2.6.1 RNA binding and antitermination

The panel of mutants that were incorporated in our FPA assay, described above, were tested to compare their M2-1 antitermination and RNA binding abilities (figure 5.8). Mutant R3AR4A, which possessed reduced RNA binding activity, also exhibited a significant loss of luciferase expression (<20 % of WT), thus identifying the corresponding residues as critical for M2-1 transcriptional antitermination. Interestingly, single mutants R3A and R4A showed similar minigenome activity to that of WT M2-1, possibly because altering the charge on these single residues was insufficient to significantly disrupt the ligand binding capability of the extensive positive charge on the M2-1 N-terminal face. Alteration of positively charged residues K19AR20A, which resulted in no significant change in RNA binding activity, correspondingly resulted in essentially unchanged luciferase expression in the replicon assay, suggestive of no role in antitermination. M2-1 mutants with alterations at core residues K150, R151 and K159 also showed minigenome activities that reflected their RNA binding abilities: double mutant K150AR151A that exhibited approximately 40 % RNA binding ability of WT, showed a comparable reduction in luciferase expression, while mutant K159A exhibited both RNA binding and luciferase expression of between 60-80 % compared to WT M2-1.

Taken together, our results showed a consistent trend; mutants with reduced RNA binding ability showed a corresponding reduction in minigenome activity, further emphasising the importance of RNA binding for M2-1 antitermination function.

5.2.6.2 Phosphorylation and antitermination

To better understand the role of phosphorylation in M2-1 antitermination, the minigenome assay was used to compare the antitermination activity of previously described phosphomimetic mutant S58DS61D and a phosphoablatant mutant, S58AS61A – electrostatically equivalent to bacterially-expressed WT M2-1, which was not phosphorylated. These mutants represent forms of M2-1 that are either maximally charged (DD) or uncharged (AA) at residues 58 and 61, and in contrast to the WT M2-1 in this system these charge states are not interchangeable. Consequently, these mutants are not directly comparable to the panel of RNA binding mutants described above, as the RNA binding mutants were still capable of dynamic phosphorylation.

Both mutants AA and DD resulted in significantly diminished antitermination activity (figure 5.8): phosphoablatant S58AS61A exhibiting

<20 % WT activity; phosphomimetic S58DS61D exhibiting ~45 % WT activity. These results imply that during RSV infection unphosphorylated (ie. uncharged) and phosphorylated (i.e. charged) M2-1 variants would possess different antitermination activities, and that a homogenous population of either state is suboptimal. This suggests that the increased antitermination function exhibited by WT M2-1 in the minigenome system depends on its switchable charge state, which can be achieved by reversible phosphorylation of the WT serine residues.

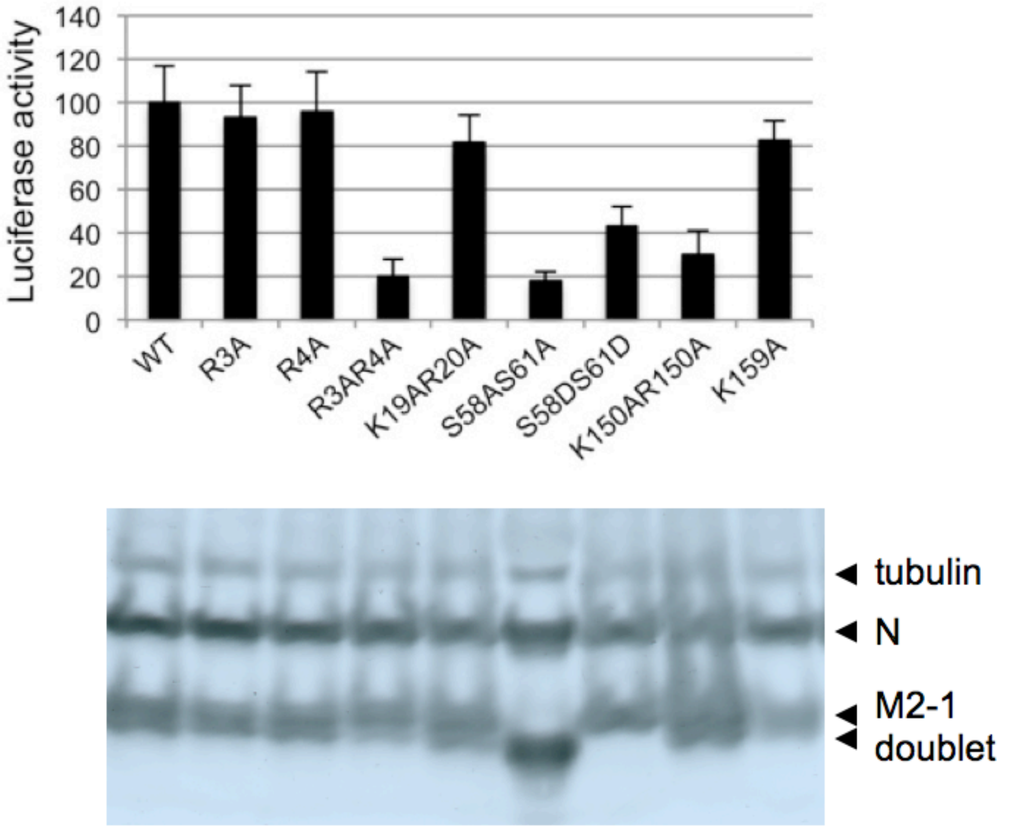


Figure 5.8 Minigenome activity of RNA and phosphorylation mutants.

Top, a histogram of M2-1-dependent luciferase activity relative to WT (dynamically phosphorylated in this system). Bottom, anti-M2-1 western blot confirming expression of all mutants. Lane order is the same as the histogram above. Non-specific binding resulted in two higher molecular weight bands suspected to be tubulin and RSV N protein. Work performed, and figure provided, by Jean-François Eléouët and his group.

5.3 CHAPTER SUMMARY

M2-1 is an RNA binding protein and abrogation of this ability coincides with loss of viral polymerase processivity, indicating RNA binding is critical to M2-1 transcriptional antitermination function (Blondot et al., 2012). To better understand how RNA binding affinity and specificity may influence the mode of action of M2-1 in promoting polymerase processivity, we quantitatively determined the binding affinity of bacterially-expressed full-length tetrameric RSV M2-1 protein to a variety of viral and non-viral RNA oligonucleotides using fluorescence polarisation anisotropy (FPA).

This work presents further evidence for the suggestion that M2-1 protein has a preference for A-rich RNA sequences and, in particular, the gene end sequences of RSV. Given the high adenosine content of 3' mRNA sense gene end sequences, and the exposed nature of nascent viral mRNA compared with the encapsidated viral genome, it is tempting to speculate that viral mRNA is also the more likely target for M2-1. This idea is notionally supported by the observation that transcription termination signals are ignored by the RSV polymerase during replication as it moves along the genomic RNP template, likely as a result of the concurrent encapsidation of the newly synthesized replication product by the N protein (Cowton, McGivern and Fearns, 2006). This role of N may be partially recapitulated by M2-1 during transcription through its interaction with specific sequences within the nascent mRNA. Our data also present a difference in RNA binding preference of M2-1 for the various gene end sequences of RSV. Although only two such sequences have been studied here and a full panel of gene ends is required for further analysis, our data suggest that differential binding to gene end *cis*-acting signals may be responsible for the previously observed variation in sensitivity to M2-1 of the gene junctions (Hardy et al., 1999). This is in contrast to the findings of Blondot et al, who saw no distinction between the binding affinity of M2-$1_{58\text{-}177}$ for either F or SH gene end sequences, however the truncated monomeric nature of their protein leads us to believe that our results are more representative of M2-1 in the context of viral infection.

The position of the surface responsible for the critical M2-1:RNA interaction that was defined on the core domain by Blondot et al, can now be visualised in the context of the complete tetramer (figure 5.9). This previously assigned surface is shown in blue and many of its residues (K92, K150, R151) have been confirmed in this chapter to be involved in RNA binding in the context of full length M2-1 tetramer. The surface corresponds to an area of positive charge on the side face of the tetramer that can now be seen to extend across both N- and C-terminal faces; the extension of positive charge combined with the identification of residues outside of this previously defined surface, such as R3, R4 and R126 (shown in pink in figure 5.9), points to the involvement of considerably larger surface of the M2-1 tetramer in its RNA binding capability, including a critical contribution to the N-terminal binding surface made by each adjacent monomer within the tetramer. Residues that were identified as playing important roles in RNA binding also exhibited correspondingly reduced levels of transcriptional antitermination in the context of the minigenome system when mutated to alanine, further confirming the critical role of RNA binding in M2-1 function.

Figure 5.9 also highlights an interesting and novel finding revealed by our crystal structures: The sites of phosphorylation, serines 58 and 61, are in close proximity to the proposed RNA binding surface, leading to the possibility of their involvement in the regulation of RNA binding and, consequently, antitermination activity of M2-1. A combination of FPA and minigenome analysis allowed interrogation of the possible effects of M2-1 phosphorylation on its RNA binding and antitermination activity, respectively, using phosphomimetic and phosphoablatant mutants. The consequence of preventing phosphorylation at either S58 or S61 was previously shown to be a significant reduction in transcription antitermination efficiency (Cartee and Wertz, 2001), however neither this publication nor any other published study has investigated the phenotype of fully phosphorylated M2-1. Here we not only confirm that complete abrogation of phosphorylation at both these sites, as exhibited by mutant S58AS61A, profoundly inhibits antitermination, but we also show that mimicking phosphorylation at these residues by using the phosphomimetic S58DS61D M2-1 mutant also significantly reduces antitermination activity. These results show that the antitermination ability of M2-1 protein is highest for the WT, suggesting that dynamic phosphorylation is essential for efficient M2-1 antitermination function.

To correlate these antitermination phenotypes to a loss of RNA binding affinity, bacterially-expressed phosphomimetic and WT (intrinsically unphosphorylated in this system) proteins were compared in the FPA assay. The data demonstrated that although unphosphorylated M2-1 had lower antitermination activity, it was capable of binding RNA with higher affinity than phosphomimetic M2-1. Whilst a comparison with a heterogeneous population of M2-1 containing both charge states is required to truly understand the involvement of phosphorylation in RNA binding, data in this chapter suggest that phosphorylation may not simply regulate RNA binding of M2-1; rather phosphorylation may also be involved in regulating M2-1 interactions with other ligands, such as the viral P protein or as yet undiscovered host factors.

In summary, this chapter has confirmed the RNA binding preference of full length tetrameric M2-1 for A-rich sequences, which are strongly suggested to be the gene end sequences at the 3’ end of nascent viral mRNAs. We have defined the RNA binding surface in the context of tetrameric M2-1 and suggest the involvement of additional residues, including several in the N-terminus of the protein. Phosphomimetic M2-1 has decreased RNA binding affinity across a range of sequences, relative to unphosphorylated wild type, but both charge states show lower antitermination activity than WT; dynamic phosphorylation of M2-1 protein is required for efficient antitermination in the context of viral transcription and replication.

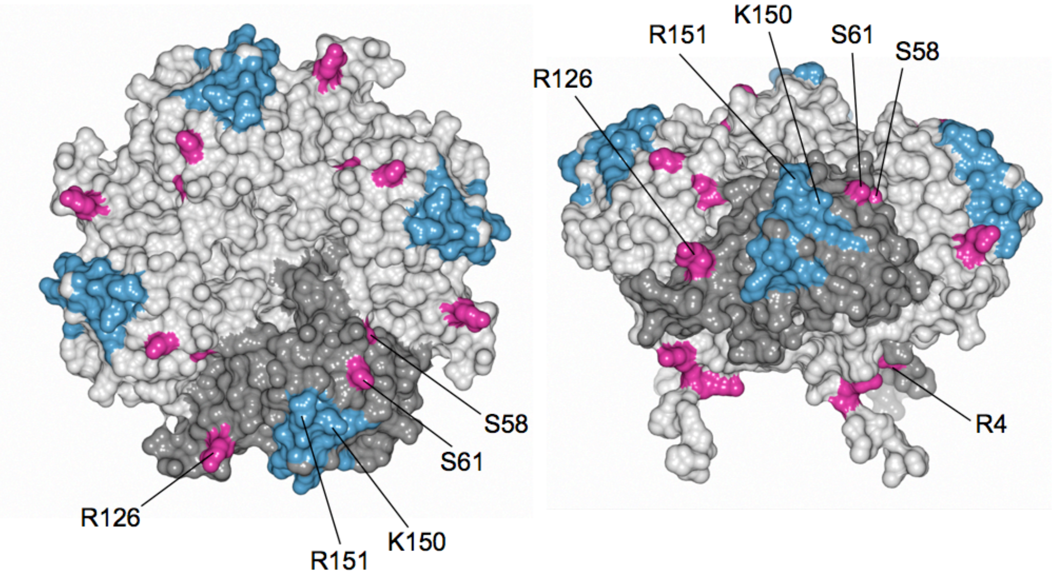


Figure 5.9 Defining the RNA binding surface of M2-1.

A surface representation of the M2-1 tetramer, shown as a C-terminal (left) and tilted side view (right), with one monomer coloured grey for clarity. The previously outlined RNA binding surface, as defined by NMR using M2-1$_{58\text{-}177}$, is shown in blue and includes residues K92, K150, R151 and K159 tested in this chapter. Additional residues identified in this study (R3, R4, S58, S61, R126) are highlighted in pink. Some residues are labelled for orientation purposes.

6 CHAPTER 6: DISCUSSION

6.1 THE X-RAY CRYSTAL STRUCTURE OF RSV M2-1 PROVIDES NEW UNDERSTANDING OF THE FUNCTIONAL IMPLICATIONS OF ITS ORGANISATION

The high-resolution crystal structures we present here reveal new and important insights into the structure and function of the full length RSV M2-1 in its biologically relevant tetrameric form. These structures identify multiple elements that drive oligomerisation, including a four-helix bundle and non-canonical zinc finger that interacts with adjacent monomers within the assembled tetramer. We pinpoint the location of residues essential for M2-1 mediated antitermination, and define the location of regions responsible for critical M2-1:RNA interactions on the tetramer surface. We also provide a structural and functional interrogation of critical phosphorylated residues S58 and S61, known to play essential roles in M2-1 polymerase processivity activity.

Whilst the crystal structure outlined in this thesis is in close agreement with the partial solution NMR structure of monomeric M2-1$_{58\text{-}177}$ that was published during this project (Blondot et al., 2012), the full-length nature of our protein provides previously unknown information about the quaternary organisation of M2-1, which is essential for understanding functional interrogation in a physiologically relevant context. Equally, it is reassuring to note the similarity between NMR and crystal structures, as the possible irrelevance of crystalline protein compared to protein in solution (as with NMR) is often cited as a disadvantage of X-ray crystallography.

Oligomerisation of M2-1 to form a stable tetramer has previously been inferred from cross-linking experiments and measurements of hydrodynamic volume. We have confirmed that M2-1 is indeed a homo-tetramer and that this organisation is strongly favoured; it is driven primarily by the long tetramerisation helix α1 forming a four-helix bundle stabilised by clustered hydrophobic residues. The formation of a zinc finger domain, critical for M2-1 antitermination and correct folding, is confirmed here for the first time and the structure shows it extends on the N-terminal arm and interacts with an adjacent protomer, further stabilising the tetramer. The involvement of the zinc finger in conformational stability explains its indispensability for M2-1 function, which was not previously understood given its apparent lack of involvement in RNA binding or phosphorylation (Hardy and Wertz, 2000, Cartee and Wertz, 2001).

The only other mononegavirus protein with known similar function to M2-1 is VP30 from the Ebola virus (EBOV). Both M2-1 and VP30 influence transcription, interact with multiple components of the viral RNA synthesis machinery, possess related zinc finger motifs, and are modified by dynamic phosphorylation. It was hoped that revealing the structural similarities between VP30 and full-length M2-1 might point to similar mechanisms of regulation, however the data presented here serve more to highlight the functional differences between M2-1 and VP30, than to confirm their similarities; the loss of negative charge through phosphoablation at M2-1 S58 and S61 dramatically reduces productive transcription, whereas for VP30 phosphoablation increases transcription in a minigenome system (Biedenkopf et al., 2013). Comparison of the VP30 model with the known functions of M2-1 is complicated by the multiple binding partners of VP30 within the polymerase complex (VP35, N, L and RNA), in contrast to M2-1, which has only 2 known associates (RNA and P). Furthermore, here we show that the ability of M2-1 to bind RNA is both modulated by phosphorylation, and is a critical determinant of antitermination function, whereas for VP30 no such association has been made. Although it may seem unlikely that two proteins sharing such considerable structural similarity would perform different functions in what are, otherwise, two highly conserved lifecycle-strategies, M2-1 protein already has precedent for such an occurrence: in contrast to the M2-1 protein of RSV, human metapneumovirus M2-1 protein is not essential to the lifecycle of the virus nor is it required for the complete and efficient transcription of viral mRNAs despite considerable sequence identity (Buchholz et al., 2005). Thus the differences in RSV M2-1 and EBOV VP30 function, despite their structural similarities, may be as a result of the binding partners they mediate their function through, rather than a fundamental difference in their modes of action.

6.2 AN EXTENSIVE AREA OF M2-1 SURFACE MEDIATES ELECTROSTATICALLY-DRIVEN RNA BINDING OF A-RICH SEQUENCES

The surface of the M2-1 tetramer is covered in extensive regions of positive charge, classically associated with binding to nucleic acid. Our data describe a number of positively charged residues whose mutation to an uncharged residue results in a reduction in RNA binding affinity of M2-1. These residues not only feature in the RNA binding region previously assigned by NMR chemical shifts, a region which can now be mapped to the side face of the M2-1 tetramer, but are also present on the N- and C-terminal faces on the protein.

Binding of RNA to M2-1 was recently found to be mutually exclusive from the M2-1:P interaction, due to partially overlapping interaction surfaces (Blondot et al., 2012, Tran et al., 2009). Our structure of the full length M2-1 tetramer, onto which the P and RNA binding surfaces can be mapped, further supports this exclusivity as the regions in question remain sufficiently close that binding of a ligand to one would sterically hinder the other. M2-1 has also recently been shown to bind P with a 1:1 stoichiometry at high affinity (Kd in the low nM range) with each M2-1 tetramer binding a P tetramer (Esperante et al., 2012). Although no high-resolution structure of RSV P is currently available, the large hydrodynamic volume of the P tetramer (elutes as 500 kDa tetramer despite mass spectrometry confirmation of mass as 125 kDa) and the overlap of P and RNA binding sites would further support the exclusivity of P and RNA binding. The interaction between P and M2-1 has been proposed to recruit M2-1 to the polymerase complexes during transcription, and abrogation of P-binding causes relocalisation (by immunofluorescence) of M2-1 from the locations of viral transcription and replication in transfected cells (Blondot et al., 2012). Given the low affinity RNA binding presented here, it is interesting to speculate how the strong M2-1:P interaction could be disrupted by RNA, as would be necessary to permit interaction with RNA for transcriptional antitermination. An alternative possibility is that the M2-1:P interaction is disrupted by the phosphorylation of P at threonine 108, which has been shown to prevent the P:M2-1 interaction (Asenjo et al, 2003).

Considering the lack of an obvious RNA binding cleft or solvent-exposed aromatic residues that could stack against nucleotide bases, and the extensive surfaces of positive charge described above, it seems likely that RNA binding of M2-1 is electrostatically driven rather than by the recognition of a specific sequence. This is supported by the relatively low affinities presented in this work (0.48-33 μM) and in previous publications (Blondot et al, 2.5 to >600 μM), which point to a preference for A-rich RNAs but no obvious preference for a particular viral sequence. Our RNA binding data suggested a slight preference for the mRNA sense gene end *cis*-acting signals, in agreement with the high 'A' content of these sequences within the nascent viral mRNAs. In addition, the variability of these sequences between genes may account for the apparent lack of sequence specificity of M2-1; it must be capable of recognising all gene end signals despite their conservation amounting only to a A-G-U-U/A-A sequence, followed by a 1-4 nucleotide variable region and a 4-6 polyA tract. The conservation of only small elements of the gene end signals may also support the idea that gene end-like sequences, only containing some characteristics of these sequences such as a polyA tract, may be the cause of premature polymerase transcription termination within genes and that M2-1 must be flexible enough to recognise these too, for its intragenic antitermination function.

6.3 THE ROLE OF PHOSPHORYLATION

The role of M2-1 phosphorylation has long been debated, although it is known to be required for antitermination function (Cartee and Wertz, 2001). In this work, we have not only confirmed the identification of the phosphorylation sites as serines 58 and 61, but we suggest that it is dynamic phosphorylation that is required by M2-1; homogenous populations of either phosphomimetic or phosphoablatant M2-1 protein have suboptimal antitermination activity relative to wild type, with phosphoablatant M2-1 being the most deleterious to transcription of a minigenome. In contrast, analysis of the RNA binding abilities of these charge states (S58DS61D phosphomimetic, and bacterially-expressed WT phosphoablatant) revealed that unphosphorylated M2-1 had higher RNA binding affinity across a range of sequences than the phosphomimic. This suggests that phosphorylation is not simply regulating the binding of RNA to M2-1, although it is clearly impacting on this interaction, nor is it affecting the sequence specificity of RNA binding, as we saw the same pattern of affinity for different sequences with phosphomimetic and unphosphorylated protein. The consistent two-fold reduction in binding affinity of mutant S58DS61D across all RNAs tested is likely due to electrostatic effects, as residues S58 and S61 are in close proximity to the RNA binding cluster on the tetramer surface, and increased negative charge would be expected to reduce RNA binding affinity – although the less deleterious effect on antitermination of mimicking charge compared with a lack of charge, suggests that an effect on RNA binding is not the primary role of phosphorylation in regulating antitermination of M2-1.

One possibility is that phosphorylation of M2-1 could be regulating a currently undiscovered interaction with a viral protein or host cell factor. To determine if the regulation of such an interaction is through the induction of structural rearrangement that exposes a new binding surface, the crystal structures of WT and S58DS61D M2-1 protein were compared. The full length M2-1 tetramer structure revealed the position of S58 and S61 on a negatively charged flexible loop that lies across the positively charged C-terminal face of each monomer, raising the possibility that altered charge resulting from phosphorylation could shift its position, exposing an alternative binding surface. However, our data showed that the altered negative charge density of the loop had no detectable effect on its position or any other M2-1 structural element. This finding does not rule out the regulation of such an interaction via electrostatic effects in the existing surfaces presented by our structure, but more work is required to fully identify the binding partners of M2-1; RSV is known to require a number of host cell proteins for poorly-described roles in its virus lifecycle, such as actin and HSP90 (Radhakrishnan et al., 2010), and an interaction of these proteins with M2-1 itself cannot be discounted.

Although there is extensive precedent for the use of aspartate mutation to mimic the negative charge imparted by a phosphorylation event, there is the possibility that our mutants do not adequately reflect the post-translational modification *in vivo*. There have been published incidences where two charged residues were required to mimic phosphorylation, or where aspartate mutants have activity more like the unphosphorylated form (reviewed in Pearlman et al., 2011). However the substitution of serines for aspartates successfully mimics phosphorylation in the majority of cases, and Pearlman et al even suggest that the mutation of aspartate/glutamate residues to phospho-serines/tyrosines is how many regulatory phosphorylation sites have evolved.

6.4 A MODEL OF M2-1 FUNCTION

Taking the available evidence together, we propose that M2-1 may preferentially bind nascent mRNAs as they are being transcribed: the 3' viral mRNA contains the gene end consensus transcription termination signal, which is A-rich by virtue of the U-tract that is a conserved genomic gene end feature. Recognition of this region of RNA by M2-1 may be what allows the polymerase to read-through these *cis*-acting signals, a function that could be performed by the nucleocapsid protein during the concurrent encapsidation of replication. How RNA binding of M2-1 translates to an effect on the viral polymerase is unknown, but it could be mediated through the M2-1:P interaction or the predicted (although not proven) interaction of M2-1 with the polymerase itself. In addition, we propose that RNA sequences that possess conserved gene end characteristics may also arise sporadically within the body of a transcript, and the binding of M2-1 to these sequences in nascent mRNA may account for intragenic antitermination. It is of note that the majority of transcription antiterminators in bacterial systems bind sequences within nascent mRNAs, and act by preventing formation of *cis*-acting secondary structure termination signals that are consequently ignored by the transcribing polymerase (Greenblatt et al., 1993). Possibly the best understood virus-encoded processivity factor is HIV Tat, which through binding to the TAR element in nascent mRNA, serves to recruit kinases that modify the C-terminal domain of RNA pol II to modify its processivity. In light of the encapsidated state of the RSV genome in contrast to naked RSV mRNAs, it seems likely that M2-1 may mediate its antitermination through a similar mechanism.

The discovery presented here, that dynamic phosphorylation is required for M2-1 transcriptional antitermination function, poses an important question: Why should the ability to dynamically switch between different charged states be critical to overall M2-1 function? One possible explanation for this requirement is that M2-1 function is a multi-step process, at least one of which requires the uncharged state and, conversely, at least one other step requires M2-1 in the charged, phosphorylated state. In the context of viral infection, this could equate to the reversible binding of RNA or P protein ligands, which consequently allow recycling of various forms of M2-1 into or away from the transcribing polymerase complex. Transient binding of M2-1 to a host factor is another possibility, as several other viruses of the order *Mononegavirales* require cellular binding partners for transcription and replication, such as the paramyxovirus Sendai virus whose polymerase interacts with tubulin to promote processivity, and human parainfluenza virus type 3 whose ribonucleocapsids bind beta-catenin (an actin-binding protein) as a transcriptional activator (Collins and Crowe Jr, 2007).

6.5 FUTURE DIRECTIONS

The crystal structure presented here provides the opportunity for the phasing of future X-ray diffraction data by molecular replacement, such as the structures of other M2-1 proteins within the subfamily *Pneumovirinae* or RSV M2-1 co-crystallised with some of the ligands discussed in this thesis. In particular, a crystal structure of M2-1 bound to an mRNA sense gene end RNA sequence would provide essential confirmation of the residues involved in the interaction and would allow further inferences to be made about how M2-1 performs its functions. Equally useful would be a M2-1:P co-crystallisation experiment, however the large size of such a complex will make this technically challenging, and truncated domains predicted to interact may be more easily achievable.

The role of phosphorylation in M2-1 function suggested here is speculative, and further investigation is required to understand this essential process. Structural studies with *in vitro* phosphorylated M2-1 protein are necessary to discount a conformational role for this modification. The possible regulation of RNA binding by phosphorylation should be examined more closely, possibly by repeating the mutational analysis performed in this thesis but with both S58AS61A and S58DS61D mutations coupled to possible RNA binding mutants, to determine if the effect of negating positive charge at these residues is phosphorylation dependent. Another possible avenue of investigation is a proteomic study comparing the potential host factor binding partners of phosphorylated vs. unphosphorylated M2-1, using GST pull-down assays coupled to mass spectrometric analysis.

There remains a considerable amount of work to be done with this exciting RSV protein, as much remains poorly understood about the role of M2-1 in the virus life cycle, how it performs its functions, and how these functions are regulated. We believe that the research presented here makes a significant contribution to the fundamental knowledge of M2-1 molecular biology, and not only paves the way for further structure-guided investigation of M2-1 function, but allows reinterpretation of previous data in a new and informative light.

REFERENCES

ABRAHAM, G. & BANERJEE, A. K. 1976. Sequential Transcription Of Genes Of Vesicular Stomatitis-Virus. *Proceedings of the National Academy of Sciences of the United States of America,* 73**,** 1504-1508.

ADAMS, P., AFONINE, P., BUNKOCZI, G., CHEN, V., DAVIS, I., ECHOLS, N., HEADD, J., HUNG, L., KAPRAL, G., GROSSE-KUNSTLEVE, R., MCCOY, A., MORIARTY, N., OEFFNER, R., READ, R., RICHARDSON, D., RICHARDSON, J., TERWILLIGER, T. & ZWART, P. 2010. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D-Biological Crystallography,* 66**,** 213-221.

AHMADIAN, G., RANDHAWA, J. S. & EASTON, A. J. 2000. Expression of the ORF-2 protein of the human respiratory syncytial virus M2 gene is initiated by a ribosomal termination-dependent reinitiation mechanism. *Embo Journal,* 19**,** 2681-2689.

ARIZA, A., TANNER, S. J., WALTER, C. T., DENT, K. C., SHEPHERD, D. A., WU, W., MATTHEWS, S. V., HISCOX, J. A., GREEN, T. J., LUO, M., ELLIOTT, R. M., FOOKS, A. R., ASHCROFT, A. E., STONEHOUSE, N. J., RANSON, N. A., BARR, J. N. & EDWARDS, T. A. 2013. Nucleocapsid protein structures from orthobunyaviruses reveal insight into ribonucleoprotein architecture and RNA polymerization. *Nucleic Acids Res,* 41**,** 5912-5926.

ASENJO, A., CALVO, E. & VILLANUEVA, N. 2006. Phosphorylation of human respiratory syncytial virus P protein at threonine 108 controls its interaction with the M2-1 protein in the viral RNA polymerase complex. *Journal of General Virology,* 87**,** 3637-3642.

ATREYA, P. L., PEEPLES, M. E. & COLLINS, P. L. 1998. The NS1 protein of human respiratory syncytial virus is a potent inhibitor of minigenome transcription and RNA replication. *Journal of Virology,* 72**,** 1452-1461.

BANERJEE, A. K. 2008. Non-segmented negative-strand RNA virus RNA synthesis in vivo - Response. *Virology,* 371**,** 231-233.

BARR, J. N. & WERTZ, G. W. 2001. Polymerase slippage at vesicular stomatitis virus gene junctions to generate poly(A) is regulated by the upstream 3 '-AUAC-5 ' tetranucleotide: Implications for the mechanism of transcription termination. *Journal of Virology,* 75**,** 6901-6913.

BARR, J. N., WHELAN, S. P. & WERTZ, G. W. 2002. Transcriptional control of the RNA-dependent RNA polymerase of vesicular stomatitis virus. *Biochim Biophys Acta,* 1577**,** 337-53.

BATTYE, T. G., KONTOGIANNIS, L., JOHNSON, O., POWELL, H. R. & LESLIE, A. G. 2011. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr D Biol Crystallogr,* 67**,** 271-81.

BERMINGHAM, A. & COLLINS, P. L. 1999. The M2-2 protein of human respiratory syncytial virus is a regulatory factor involved in the balance between RNA replication and transcription. *Proceedings of the National Academy of Sciences of the United States of America,* 96**,** 11259-11264.

BIACCHESI, S., SKIADOPOULOS, M. H., BOIVIN, G., HANSON, C. T., MURPHY, B. R., COLLINS, P. L. & BUCHHOLZ, U. J. 2003. Genetic diversity between human metapneumovirus subgroups. *Virology,* 315**,** 1-9.

BIEDENKOPF, N., HARTLIEB, B., HOENEN, T. & BECKER, S. 2013. Phosphorylation of Ebola virus VP30 influences the composition of the viral nucleocapsid complex: impact on viral transcription and replication. *J Biol Chem,* 288**,** 11165-74.

BLONDOT, M. L., DUBOSCLARD, V., FIX, J., LASSOUED, S., AUMONT-NICAISE, M., BONTEMS, F., ELEOUET, J. F. & SIZUN, C. 2012. Structure and functional analysis of the RNA- and viral phosphoprotein-binding domain of respiratory syncytial virus M2-1 protein. *PLoS Pathog,* 8**,** e1002734.

BUCHHOLZ, U. J., BIACCHESI, S., PHAM, Q. N., TRAN, K. C., YANG, L., LUONGO, C. L., SKIADOPOULOS, M. H., MURPHY, B. R. & COLLINS, P. L. 2005. Deletion of M2 gene open reading frames 1 and 2 of human metapneumovirus: effects on RNA synthesis, attenuation, and immunogenicity. *J Virol,* 79**,** 6588-97.

BUKREYEV, A., YANG, L. & COLLINS, P. L. 2012. The secreted G protein of human respiratory syncytial virus antagonizes antibody-mediated restriction of replication involving macrophages and complement. *J Virol,* 86**,** 10880-4.

BUTT, T. R., EDAVETTAL, S. C., HALL, J. P. & MATTERN, M. R. 2005. SUMO fusion technology for difficult-to-express proteins. *Protein Expr Purif,* 43**,** 1-9.

CAO, H. 2004. Expression, purification, and biochemical characterization of the antiinflammatory tristetraprolin: a zinc-dependent mRNA binding protein affected by posttranslational modifications. *Biochemistry,* 43**,** 13724-38.

CARROMEU, C., SIMABUCO, F. M., TAMURA, R. E., ARCIERI, L. E. F. & VENTURA, A. M. 2007. Intracellular localization of human respiratory syncytial virus L protein. *Archives of Virology,* 152**,** 2259-2263.

CARTEE, T. L. & WERTZ, G. W. 2001. Respiratory syncytial virus M2-1 protein requires phosphorylation for efficient function and binds viral RNA during infection. *Journal of Virology,* 75**,** 12188-12197.

CHAMBERS, S. P., AUSTEN, D. A., FULGHUM, J. R. & KIM, W. M. 2004. High-throughput screening for soluble recombinant expressed kinases in Escherichia coli and insect cells. *Protein Expr Purif,* 36**,** 40-7.

CHANOCK, R. & FINBERG, L. 1957. Recovery from infants with respiratory illness of a virus related to chimpanzee coryza agent (CCA). II. Epidemiologic aspects of infection in infants and young children. *Am J Hyg,* 66**,** 291-300.

CHANOCK, R., ROIZMAN, B. & MYERS, R. 1957. Recovery from infants with respiratory illness of a virus related to chimpanzee coryza agent (CCA). I. Isolation, properties and characterization. *Am J Hyg,* 66**,** 281-90.

CHEN, V. B., ARENDALL, W. B., HEADD, J. J., KEEDY, D. A., IMMORMINO, R. M., KAPRAL, G. J., MURRAY, L. W., RICHARDSON, J. S. & RICHARDSON, D. C. 2010. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr,* 66**,** 12-21.

COLLINS, P. L. & CROWE JR, J. E. (eds.) 2007. *Fields Virology, 5th Edition*: Lippincott Williams & Wilkins.

COLLINS, P. L., CROWE, J. E. 2007. *Respiratory syncytial virus and metapneumovirus*, Lippincott Williams & Wilkins, Philadelphia, PA.

COLLINS, P. L. & GRAHAM, B. S. 2008. Viral and host factors in human respiratory syncytial virus pathogenesis. *Journal of Virology,* 82**,** 2040-2055.

COLLINS, P. L., HILL, M. G., CAMARGO, E., GROSFELD, H., CHANOCK, R. M. & MURPHY, B. R. 1995. Production Of Infectious Human Respiratory Syncytial Virus From Cloned Cdna Confirms An Essential Role For The Transcription Elongation-Factor From The 5'-Proximal Open Reading Frame Of The M2 Messenger-Rna In Gene-Expression And Provides A Capability For Vaccine Development. *Proceedings of the National Academy of Sciences of the United States of America,* 92**,** 11563-11567.

COLLINS, P. L., HILL, M. G., CRISTINA, J. & GROSFELD, H. 1996. Transcription elongation factor of respiratory syncytial virus, a nonsegmented negative-strand RNA virus. *Proceedings of the National Academy of Sciences of the United States of America,* 93**,** 81-85.

COLLINS, P. L., HILL, M. G. & JOHNSON, P. R. 1990. The 2 Open Reading Frames of The 22K Messenger-RNA Of Human Respiratory Syncytial Virus - Sequence Comparison of Antigenic Subgroup-A and Subgroup-B and Expression in vitro. *Journal of General Virology,* 71**,** 3015-3020.

COLLINS, P. L. & MELERO, J. A. 2011. Progress in understanding and controlling respiratory syncytial virus: Still crazy after all these years. *Virus Res*.

COLLINS, P. L., OLMSTED, R. A., SPRIGGS, M. K., JOHNSON, P. R. & BUCKLER-WHITE, A. J. 1987. Gene overlap and site-specific attenuation of transcription of the viral polymerase L gene of human respiratory syncytial virus. *Proc Natl Acad Sci U S A,* 84**,** 5134-8.

COLLINS, P. L. & WERTZ, G. W. 1983. cDNA cloning and transcriptional mapping of nine polyadenylylated RNAs encoded by the genome of human respiratory syncytial virus. *Proc Natl Acad Sci U S A,* 80**,** 3208-12.

COLLINS, P. L. & WERTZ, G. W. 1985. The envelope-associated 22K protein of human respiratory syncytial virus: nucleotide sequence of the mRNA and a related polytranscript. *J Virol,* 54**,** 65-71.

COWTAN, K. 2006. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr D Biol Crystallogr,* 62**,** 1002-11.

COWTON, V. M. & FEARNS, R. 2005. Evidence that the respiratory syncytial virus polymerase is recruited to nucleotides 1 to 11 at the 3 ' end of the nucleocapsid and can scan to access internal signals. *Journal of Virology,* 79**,** 11311-11322.

COWTON, V. M., MCGIVERN, D. R. & FEARNS, R. 2006. Unravelling the complexities of respiratory syncytial virus RNA synthesis. *Journal of General Virology,* 87**,** 1805-1821.

CUBITT, B., OLDSTONE, C., VALCARCEL, J. & CARLOS DE LA TORRE, J. 1994. RNA splicing contributes to the generation of mature mRNAs of Borna disease virus, a non-segmented negative strand RNA virus. *Virus Res,* 34**,** 69-79.

CUESTA, I., GENG, X. H., ASENJO, A. & VILLANUEVA, N. 2000. Structural phosphoprotein M2-1 of the human respiratory syncytial virus is an RNA binding protein. *Journal of Virology,* 74**,** 9858-9867.

CURRAN, J. & KOLAKOFSKY, D. 2008. Nonsegmented negative-strand RNA virus RNA synthesis in vivo. *Virology,* 371**,** 227-230.

DAUTER, Z., DAUTER, M. & DODSON, E. 2002. Jolly SAD. *Acta Crystallographica Section D-Biological Crystallography,* 58**,** 494-506.

DELGADO, M. F., COVIELLO, S., MONSALVO, A. C., MELENDI, G. A., HERNANDEZ, J. Z., BATALLE, J. P., DIAZ, L., TRENTO, A., CHANG, H. Y., MITZNER, W., RAVETCH, J., MELERO, J. A., IRUSTA, P. M. & POLACK, F. P. 2009. Lack of antibody affinity maturation due to poor Toll-like receptor stimulation leads to enhanced respiratory syncytial virus disease. *Nat Med,* 15**,** 34-41.

DEVINCENZO, J., LAMBKIN-WILLIAMS, R., WILKINSON, T., CEHELSKY, J., NOCHUR, S., WALSH, E., MEYERS, R., GOLLOB, J. & VAISHNAW, A. 2010. A randomized, double-blind, placebo-controlled study of an RNAi-based therapy directed against respiratory syncytial virus. *Proc Natl Acad Sci U S A,* 107**,** 8800-5.

DIBBEN, O., THORPE, L. C. & EASTON, A. J. 2008. Roles of the PVM M2-1, M2-2 and P gene ORF 2 (P-2) proteins in viral replication. *Virus Research,* 131**,** 47-53.

DUBOSCLARD, V., BLONDOT, M. L., ELEOUET, J. F., BONTEMS, F. & SIZUN, C. 2011. 1H, 13C, and 15N resonance assignment of the central domain of hRSV transcription antitermination factor M2-1. *Biomol NMR Assign,* 5**,** 237-9.

DUTCH, R. E. 2010. Entry and fusion of emerging paramyxoviruses. *PLoS Pathog,* 6**,** e1000881.

EGLI, M. 2010. Diffraction techniques in structural biology. *Curr Protoc Nucleic Acid Chem,* Chapter 7**,** Unit 7.13.

EMSLEY, P. & COWTAN, K. 2004. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr,* 60**,** 2126-32.

ESPERANTE, S. A., CHEMES, L. B., SANCHEZ, I. E. & PRAT-GAY, G. D. 2011. The respiratory syncytial virus transcription antiterminator M2-1 is highly stable, zinc binding tetramer with a strong pH dependent dissociation and a monomeric unfolding intermediate. *Biochemistry*.

ESPERANTE, S. A., PARIS, G. & PRAT-GAY, G. D. 2012. Modular unfolding and dissociation of the human respiratory syncytial virus phosphoprotein P and its interaction with the M2-1 antiterminator: a singular tetramer-tetramer interface arrangement. *Biochemistry*.

EVANS, P. 2006. Scaling and assessment of data quality. *Acta Crystallogr D Biol Crystallogr,* 62**,** 72-82.

EVANS, P. R. 2011. An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallogr D Biol Crystallogr,* 67**,** 282-92.

FALSEY, A. R., HENNESSEY, P. A., FORMICA, M. A., COX, C. & WALSH, E. E. 2005. Respiratory syncytial virus infection in elderly and high-risk adults. *New England Journal of Medicine,* 352**,** 1749-1759.

FEARNS, R. & COLLINS, P. L. 1999a. Model for polymerase access to the overlapped L gene of respiratory syncytial virus. *Journal of Virology,* 73**,** 388-397.

FEARNS, R. & COLLINS, P. L. 1999b. Role of the M2-1 transcription antitermination protein of respiratory syncytial virus in sequential transcription. *Journal of Virology,* 73**,** 5852-5864.

FEARNS, R., COLLINS, P. L. & PEEPLES, M. E. 2000. Functional analysis of the genomic and antigenomic promoters of human respiratory syncytial virus. *Journal of Virology,* 74**,** 6006-6014.

FEARNS, R., PEEPLES, M. E. & COLLINS, P. L. 2002. Mapping the transcription and replication promoters of respiratory syncytial virus. *J Virol,* 76**,** 1663-72.

FENTON, C., SCOTT, L. J. & PLOSKER, G. L. 2004. Palivizumab: a review of its use as prophylaxis for serious respiratory syncytial virus infection. *Paediatr Drugs,* 6**,** 177-97.

GAN, S. W., TAN, E., LIN, X., YU, D., WANG, J., TAN, G. M., VARARATTANAVECH, A., YEO, C. Y., SOON, C. H., SOONG, T. W., PERVUSHIN, K. & TORRES, J. 2012. The small hydrophobic protein of the human respiratory syncytial virus forms pentameric ion channels. *J Biol Chem,* 287**,** 24671-89.

GARCIA, J., GARCIABARRENO, B., VIVO, A. & MELERO, J. A. 1993. Cytoplasmic inclusions of respiratory syncytial virus-infected cells - formation of inclusion-bodies in transfected cells that coexpress the nucleoprotein, the phosphoprotein, and the 22k protein. *Virology,* 195**,** 243-247.

GHILDYAL, R., BAULCH-BROWN, C., MILLS, J. & MEANGER, J. 2003. The matrix protein of Human respiratory syncytial virus localises to the nucleus of infected cells and inhibits transcription. *Arch Virol,* 148**,** 1419-29.

GHILDYAL, R., HO, A., WAGSTAFF, K. M., DIAS, M. M., BARTON, C. L., JANS, P., BARDIN, P. & JANS, D. A. 2005a. Nuclear import of the respiratory syncytial virus matrix protein is mediated by importin beta1 independent of importin alpha. *Biochemistry,* 44**,** 12887-95.

GHILDYAL, R., LI, D. S., PEROULIS, I., SHIELDS, B., BARDIN, P. G., TENG, M. N., COLLINS, P. L., MEANGER, J. & MILLS, J. 2005b. Interaction between the respiratory syncytial virus G glycoprotein cytoplasmic domain and the matrix protein. *Journal of General Virology,* 86**,** 1879-1884.

GHILDYAL, R., MILLS, J., MURRAY, M., VARDAXIS, N. & MEANGER, J. 2002. Respiratory syncytial virus matrix protein associates with nucleocapsids in infected cells. *Journal of General Virology,* 83**,** 753-757.

GLENN, G. M., SMITH, G., FRIES, L., RAGHUNANDAN, R., LU, H., ZHOU, B., THOMAS, D. N., HICKMAN, S. P., KPAMEGAN, E., BODDAPATI, S. & PIEDRA, P. A. 2013. Safety and immunogenicity of a Sf9 insect cell-derived respiratory syncytial virus fusion protein nanoparticle vaccine. *Vaccine,* 31**,** 524-32.

GODDARD, N. L., COOKE, M. C., GUPTA, R. K. & NGUYEN-VAN-TAM, J. S. 2007. Timing of monoclonal antibody for seasonal RSV prophylaxis in the United Kingdom. *Epidemiology and Infection,* 135**,** 159-162.

GOLOVANOV, A., HAUBERGUE, G., WILSON, S. & LU-YUN, L. 2004. A simple method for improving protein solubility and long-term stability. *Journal of the American Chemical Society*, 8933-8939.

GONZALEZ-REYES, L., RUIZ-ARGUELLO, M. B., GARCIA-BARRENO, B., CALDER, L., LOPEZ, J. A., ALBAR, J. P., SKEHEL, J. J., WILEY, D. C. & MELERO, J. A. 2001. Cleavage of the human respiratory syncytial virus fusion protein at two distinct sites is required for activation of membrane fusion. *Proc Natl Acad Sci U S A,* 98**,** 9859-64.

GOULD, P. S. & EASTON, A. J. 2005. Coupled translation of the respiratory syncytial virus M2 open reading frames requires upstream sequences. *J Biol Chem,* 280**,** 21972-80.

GOULD, P. S. & EASTON, A. J. 2007. Coupled translation of the second open reading frame of M2 mRNA is sequence dependent and differs significantly within the subfamily Pneumovirinae. *Journal of Virology,* 81**,** 8488-8496.

GREENBLATT, J., NODWELL, J. R. & MASON, S. W. 1993. Transcriptional antitermination. *Nature,* 364**,** 401-6.

GROSFELD, H., HILL, M. G. & COLLINS, P. L. 1995. Rna Replication By Respiratory Syncytial Virus (Rsv) Is Directed By The N-Protein, P-Protein, And L Protein - Transcription Also Occurs Under These Conditions But Requires Rsv Superinfection For Efficient Synthesis Of Full-Length Messenger-Rna. *Journal of Virology,* 69**,** 5677-5686.

HALL, C. B. 2001. Medical progress - Respiratory syncytial virus and parainfluenza virus. *New England Journal of Medicine,* 344**,** 1917-1928.

HALL, C. B., WEINBERG, G. A., IWANE, M. K., BLUMKIN, A. K., EDWARDS, K. M., STAAT, M. A., AUINGER, P., GRIFFIN, M. R., POEHLING, K. A., ERDMAN, D., GRIJALVA, C. G., ZHU, Y. W. & SZILAGYI, P. 2009. The Burden of Respiratory Syncytial Virus Infection in Young Children. *New England Journal of Medicine,* 360**,** 588-598.

HARDY, R. W., HARMON, S. B. & WERTZ, G. W. 1999. Diverse gene junctions of respiratory syncytial virus modulate the efficiency of transcription termination and respond differently to M2-mediated antitermination. *Journal of Virology,* 73**,** 170-176.

HARDY, R. W. & WERTZ, G. W. 1998. The product of the respiratory syncytial virus M2 gene ORF1 enhances readthrough of intergenic junctions during viral transcription. *Journal of Virology,* 72**,** 520-526.

HARDY, R. W. & WERTZ, G. W. 2000. The Cys(3)-His(1) motif of the respiratory syncytial virus M2-1 protein is essential for protein function. *Journal of Virology,* 74**,** 5880-5885.

HARPER, D. M., FRANCO, E. L., WHEELER, C., FERRIS, D. G., JENKINS, D., SCHUIND, A., ZAHAF, T., INNIS, B., NAUD, P., DE CARVALHO, N. S., ROTELI-MARTINS, C. M., TEIXEIRA, J., BLATTER, M. M., KORN, A. P., QUINT, W., DUBIN, G. & GROUP, G. H. V. S. 2004. Efficacy of a bivalent L1 virus-like particle vaccine in prevention of infection with human papillomavirus types 16 and 18 in young women: a randomised controlled trial. *Lancet,* 364**,** 1757-65.

HARRISON, M. S., SAKAGUCHI, T. & SCHMITT, A. P. 2010. Paramyxovirus assembly and budding: building particles that transmit infections. *Int J Biochem Cell Biol,* 42**,** 1416-29.

HARTLIEB, B., MUZIOL, T., WEISSENHORN, W. & BECKER, S. 2007. Crystal structure of the C-terminal domain of Ebola virus VP30 reveals a role in transcription and nucleocapsid association. *Proc Natl Acad Sci U S A,* 104**,** 624-9.

ICTV. 2012. *Virus Taxonomy: 2012 release* [Online]. International Committee on Taxonomy of Viruses. Available: http://ictvonline.org/virusTaxonomy.asp.

JENKINS, H. T., MALKOVA, B. & EDWARDS, T. A. 2011. Kinked beta-strands mediate high-affinity recognition of mRNA targets by the germ-cell regulator DAZL. *Proc Natl Acad Sci U S A,* 108**,** 18266-71.

KABSCH, W. 2010. XDS. *Acta Crystallogr D Biol Crystallogr,* 66**,** 125-32.

KAPIKIAN, A. Z., MITCHELL, R. H., CHANOCK, R. M., SHVEDOFF, R. A. & STEWART, C. E. 1969. An epidemiologic study of altered clinical reactivity to respiratory syncytial (rs) virus infection in children previously vaccinated with an inactivated rs virus vaccine. *American Journal of Epidemiology,* 89**,** 405-&.

KOST, T. A., CONDREAY, J. P. & JARVIS, D. L. 2005. Baculovirus as versatile vectors for protein expression in insect and mammalian cells. *Nat Biotechnol,* 23**,** 567-75.

KRISHNAMOORTHY, N., KHARE, A., ORISS, T. B., RAUNDHAL, M., MORSE, C., YARLAGADDA, M., WENZEL, S. E., MOORE, M. L., PEEBLES, R. S., JR., RAY, A. & RAY, P. 2012. Early infection with respiratory syncytial virus impairs regulatory T cell function and increases susceptibility to allergic asthma. *Nat Med,* 18**,** 1525-30.

KRISSINEL, E. & HENRICK, K. 2007. Inference of macromolecular assemblies from crystalline state. *J Mol Biol,* 372**,** 774-97.

KRZYZANIAK, M. A., ZUMSTEIN, M. T., GEREZ, J. A., PICOTTI, P. & HELENIUS, A. 2013. Host cell entry of respiratory syncytial virus involves macropinocytosis followed by proteolytic activation of the f protein. *PLoS Pathog,* 9**,** e1003309.

KUEHNER, J. N., PEARSON, E. L. & MOORE, C. 2011. Unravelling the means to an end: RNA polymerase II transcription termination. *Nat Rev Mol Cell Biol,* 12**,** 283-94.

KUO, L., GROSFELD, H., CRISTINA, J., HILL, M. G. & COLLINS, P. L. 1996. Effect of mutations in the gene-start and gene-end sequence motifs on transcription of monocistronic and dicistronic minigenomes of respiratory syncytial virus. *Journal of Virology,* 70**,** 6892-6901.

KURT-JONES, E. A., POPOVA, L., KWINN, L., HAYNES, L. M., JONES, L. P., TRIPP, R. A., WALSH, E. E., FREEMAN, M. W., GOLENBOCK, D. T., ANDERSON, L. J. & FINBERG, R. W. 2000. Pattern recognition receptors TLR4 and CD14 mediate response to respiratory syncytial virus. *Nat Immunol,* 1**,** 398-401.

LAMBERT, D. M., HAMBOR, J., DIEBOLD, M. & GALINSKI, B. 1988. Kinetics of Synthesis and Phosphorylation of Respiratory Syncytial Virus Polypeptides. *Journal of General Virology,* 69**,** 313-323.

LESLIE, A. G. 2006. The integration of macromolecular diffraction data. *Acta Crystallogr D Biol Crystallogr,* 62**,** 48-57.

LI, D. S., JANS, D. A., BARDIN, P. G., MEANGER, J., MILLS, J. & GHILDYAL, R. 2008a. Association of respiratory syncytial virus M protein with viral nucleocapsids is mediated by the M2-1 protein. *Journal of Virology,* **82,** 8863-8870.

LI, J., RAHMEH, A., MORELLI, M. & WHELAN, S. P. 2008b. A conserved motif in region v of the large polymerase proteins of nonsegmented negative-sense RNA viruses that is essential for mRNA capping. *J Virol,* **82,** 775-84.

LILJEROOS, L., KRZYZANIAK, M. A., HELENIUS, A. & BUTCHER, S. J. 2013. Architecture of respiratory syncytial virus revealed by electron cryotomography. *Proc Natl Acad Sci U S A,* **110,** 11133-8.

LIUZZI, M., MASON, S. W., CARTIER, M., LAWETZ, C., MCCOLLUM, R. S., DANSEREAU, N., BOLGER, G., LAPEYRE, N., GAUDETTE, Y., LAGACÉ, L., MASSARIOL, M. J., DÔ, F., WHITEHEAD, P., LAMARRE, L., SCOUTEN, E., BORDELEAU, J., LANDRY, S., RANCOURT, J., FAZAL, G. & SIMONEAU, B. 2005. Inhibitors of respiratory syncytial virus replication target cotranscriptional mRNA guanylylation by viral RNA-dependent RNA polymerase. *J Virol,* **79,** 13105-15.

LU, B., MA, C. H., BRAZAS, R. & JIN, H. 2002. The major phosphorylation sites of the respiratory syncytial virus phosphoprotein are dispensable for virus replication in vitro. *J Virol,* **76,** 10776-84.

LUCKOW, V. A., LEE, S. C., BARRY, G. F. & OLINS, P. O. 1993. Efficient generation of infectious recombinant baculoviruses by site-specific transposon-mediated insertion of foreign genes into a baculovirus genome propagated in Escherichia coli. *J Virol,* **67,** 4566-79.

MADAN, V., CASTELLÓ, A. & CARRASCO, L. 2008. Viroporins from RNA viruses induce caspase-dependent apoptosis. *Cell Microbiol,* **10,** 437-51.

MALGIERI, G., ZACCARO, L., LEONE, M., BUCCI, E., ESPOSITO, S., BAGLIVO, I., DEL GATTO, A., RUSSO, L., SCANDURRA, R., PEDONE, P. V., FATTORUSSO, R. & ISERNIA, C. 2011. Zinc to cadmium replacement in the A. thaliana SUPERMAN Cys_2 His_2 zinc finger induces structural rearrangements of typical DNA base determinant positions. *Biopolymers,* **95,** 801-10.

MASON, S. W., ABERG, E., LAWETZ, C., DELONG, R., WHITEHEAD, P. & LIUZZI, M. 2003. Interaction between human respiratory syncytial virus (RSV) M2-1 and P proteins is required for reconstitution of M2-1-dependent RSV minigenome activity. *Journal of Virology,* **77,** 10670-10676.

MASTERS, P. S. & SAMUEL, C. E. 1984. Detection of in vivo synthesis of polycistronic mRNAs of vesicular stomatitis virus. *Virology,* **134,** 277-86.

MASTRANGELO, P. & HEGELE, R. G. 2013. RSV fusion: time for a new model. *Viruses,* **5,** 873-85.

MCCOY, A. J., GROSSE-KUNSTLEVE, R. W., ADAMS, P. D., WINN, M. D., STORONI, L. C. & READ, R. J. 2007. Phaser crystallographic software. *J Appl Crystallogr,* **40,** 658-674.

MCGIVERN, D. R., COLLINS, P. L. & FEARNS, R. 2005. Identification of internal sequences in the 3 ' leader region of human respiratory syncytial virus that enhance transcription and confer replication processivity. *Journal of Virology,* **79,** 2449-2460.

MCNAMARA, P. S. & SMYTH, R. L. 2002. The pathogenesis of respiratory syncytial virus disease in childhood. *Br Med Bull,* **61,** 13-28.

MCNICHOLAS, S., POTTERTON, E., WILSON, K. S. & NOBLE, M. E. 2011. Presenting your structures: the CCP4mg molecular-graphics software. *Acta Crystallogr D Biol Crystallogr,* **67,** 386-94.

MEYER, G., DEPLANCHE, M. & SCHELCHER, F. 2008. Human and bovine respiratory syncytial virus vaccine research and development. *Comparative Immunology Microbiology and Infectious Diseases,* **31,** 191-225.

MICHALEK, J. L., LEE, S. J. & MICHEL, S. L. 2012. Cadmium coordination to the zinc binding domains of the non-classical zinc finger protein Tristetraprolin affects RNA binding selectivity. *J Inorg Biochem,* 112, 32-8.

MONEY, V. A., MCPHEE, H. K., MOSELY, J. A., SANDERSON, J. M. & YEO, R. P. 2009. Surface features of a Mononegavirales matrix protein indicate sites of membrane interaction. *Proc Natl Acad Sci U S A,* 106, 4441-6.

MORIN, B., KRANZUSCH, P. J., RAHMEH, A. A. & WHELAN, S. P. 2013. The polymerase of negative-stranded RNA viruses. *Curr Opin Virol,* 3, 103-10.

MUNIR, S., HILLYER, P., LE NOUËN, C., BUCHHOLZ, U. J., RABIN, R. L., COLLINS, P. L. & BUKREYEV, A. 2011. Respiratory syncytial virus interferon antagonist NS1 protein suppresses and skews the human T lymphocyte response. *PLoS Pathog,* 7, e1001336.

MURSHUDOV, G. N., SKUBÁK, P., LEBEDEV, A. A., PANNU, N. S., STEINER, R. A., NICHOLLS, R. A., WINN, M. D., LONG, F. & VAGIN, A. A. 2011. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr,* 67, 355-67.

NAYLOR, C. J., BROWN, P. A., EDWORTHY, N., LING, R., JONES, R. C., SAVAGE, C. E. & EASTON, A. J. 2004. Development of a reverse-genetics system for Avian pneumovirus demonstrates that the small hydrophobic (SH) and attachment (G) genes are not essential for virus viability. *J Gen Virol,* 85, 3219-27.

NOTON, S. L., DEFLUBÉ, L. R., TREMAGLIO, C. Z. & FEARNS, R. 2012. The respiratory syncytial virus polymerase has multiple RNA synthesis activities at the promoter. *PLoS Pathog,* 8, e1002980.

OLSZEWSKA, W. & OPENSHAW, P. 2009. Emerging drugs for respiratory syncytial virus infection. *Expert Opinion on Emerging Drugs,* 14, 207-217.

PEARLMAN, S. M., SERBER, Z. & FERRELL, J. E. 2011. A mechanism for the evolution of phosphorylation sites. *Cell,* 147, 934-46.

POTTERTON, E., BRIGGS, P., TURKENBURG, M. & DODSON, E. 2003. A graphical user interface to the CCP4 program suite. *Acta Crystallogr D Biol Crystallogr,* 59, 1131-7.

QANUNGO, K. R., SHAJI, D., MATHUR, M. & BANERJEE, A. K. 2004. Two RNA polymerase complexes from vesicular stomatitis virus-infected cells that carry out transcription and replication of genome RNA. *Proceedings of the National Academy of Sciences of the United States of America,* 101, 5952-5957.

RADHAKRISHNAN, A., YEO, D., BROWN, G., MYAING, M. Z., IYER, L. R., FLECK, R., TAN, B. H., AITKEN, J., SANMUN, D., TANG, K., YARWOOD, A., BRINK, J. & SUGRUE, R. J. 2010. Protein analysis of purified respiratory syncytial virus particles reveals an important role for heat shock protein 90 in virus particle assembly. *Mol Cell Proteomics,* 9, 1829-48.

RODRÍGUEZ, L., CUESTA, I., ASENJO, A. & VILLANUEVA, N. 2004. Human respiratory syncytial virus matrix protein is an RNA-binding protein: binding properties, location and identity of the RNA contact residues. *J Gen Virol,* 85, 709-19.

ROUTLEDGE, E. G., WILLCOCKS, M. M., MORGAN, L., SAMSON, A. C., SCOTT, R. & TOMS, G. L. 1987. Heterogeneity of the respiratory syncytial virus 22K protein revealed by Western blotting with monoclonal antibodies. *J Gen Virol,* 68 (Pt 4), 1209-15.

SCHLICHTING, I. & MIAO, J. 2012. Emerging opportunities in structural biology with X-ray free-electron lasers. *Curr Opin Struct Biol,* 22, 613-26.

SCHRÖDINGER, L. The PyMOL Molecular Graphics System, Version 1.5.0.4.

SHAIKH, F. Y., UTLEY, T. J., CRAVEN, R. E., ROGERS, M. C., LAPIERRE, L. A., GOLDENRING, J. R. & CROWE, J. E. 2012. Respiratory syncytial virus assembles into structured filamentous virion particles independently of host cytoskeleton and related proteins. *PLoS One,* 7**,** e40826.

SMITH, D. B. & JOHNSON, K. S. 1988. Single-step purification of polypeptides expressed in Escherichia coli as fusions with glutathione S-transferase. *Gene,* 67**,** 31-40.

SOKOLENKO, S., GEORGE, S., WAGNER, A., TULADHAR, A., ANDRICH, J. M. & AUCOIN, M. G. 2012. Co-expression vs. co-infection using baculovirus expression vectors in insect cell culture: Benefits and drawbacks. *Biotechnol Adv,* 30**,** 766-81.

STOKES, H. L., EASTON, A. J. & MARRIOTT, A. C. 2003. Chimeric pneumovirus nucleocapsid (N) proteins allow identification of amino acids essential for the function of the respiratory syncytial virus N protein. *J Gen Virol,* 84**,** 2679-83.

SUTHERLAND, K. A., COLLINS, P. L. & PEEPLES, M. E. 2001. Synergistic effects of gene-end signal mutations and the M2-1 protein on transcription termination by respiratory syncytial virus. *Virology,* 288**,** 295-307.

SWEDAN, S., ANDREWS, J., MAJUMDAR, T., MUSIYENKO, A. & BARIK, S. 2011. Multiple functional domains and complexes of the two nonstructural proteins of human respiratory syncytial virus contribute to interferon suppression and cellular location. *J Virol,* 85**,** 10090-100.

TAWAR, R. G., DUQUERROY, S., VONRHEIN, C., VARELA, P. F., DAMIER-PIOLLE, L., CASTAGNE, N., MACLELLAN, K., BEDOUELLE, H., BRICOGNE, G., BHELLA, D., ELEOUET, J. F. & REY, F. A. 2009. Crystal Structure of a Nucleocapsid-Like Nucleoprotein-RNA Complex of Respiratory Syncytial Virus. *Science,* 326**,** 1279-1283.

TAYYARI, F., MARCHANT, D., MORAES, T. J., DUAN, W., MASTRANGELO, P. & HEGELE, R. G. 2011. Identification of nucleolin as a cellular receptor for human respiratory syncytial virus. *Nat Med,* 17**,** 1132-5.

TERWILLIGER, T. C., ADAMS, P. D., READ, R. J., MCCOY, A. J., MORIARTY, N. W., GROSSE-KUNSTLEVE, R. W., AFONINE, P. V., ZWART, P. H. & HUNG, L. W. 2009. Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard. *Acta Crystallogr D Biol Crystallogr,* 65**,** 582-601.

TERWILLIGER, T. C. & BERENDZEN, J. 1999. Automated MAD and MIR structure solution. *Acta Crystallogr D Biol Crystallogr,* 55**,** 849-61.

TRAN, T. L., CASTAGNE, N., DUBOSCLARD, V., NOINVILLE, S., KOCH, E., MOUDJOU, M., HENRY, C., BERNARD, J., YEO, R. P. & ELEOUET, J. F. 2009. The Respiratory Syncytial Virus M2-1 Protein Forms Tetramers and Interacts with RNA and P in a Competitive Manner. *Journal of Virology,* 83**,** 6363-6374.

TREMAGLIO, C. Z., NOTON, S. L., DEFLUBÉ, L. R. & FEARNS, R. 2013. Respiratory syncytial virus polymerase can initiate transcription from position 3 of the leader promoter. *J Virol,* 87**,** 3196-207.

WALLS, D. & LOUGHRAN, S. T. 2010. Tagging Recombinant Proteins to Enhance Solubility and Aid Purification. *Protein Chromatography: Methods and Protocols.* Springer Protocols.

WHELAN, S. P. J. 2008. Non-segmented negative-strand RNA virus RNA synthesis in vivo - Response. *Virology,* 371**,** 234-237.

WINN, M. D., BALLARD, C. C., COWTAN, K. D., DODSON, E. J., EMSLEY, P., EVANS, P. R., KEEGAN, R. M., KRISSINEL, E. B., LESLIE, A. G., MCCOY, A., MCNICHOLAS, S. J., MURSHUDOV, G. N., PANNU, N. S., POTTERTON, E. A., POWELL, H. R., READ, R. J., VAGIN, A. & WILSON, K. S. 2011. Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr,* 67**,** 235-42.

WU, H., PFARR, D. S., JOHNSON, S., BREWAH, Y. A., WOODS, R. M., PATEL, N. K., WHITE, W. I., YOUNG, J. F. & KIENER, P. A. 2007. Development of motavizumab, an ultra-potent antibody for the prevention of respiratory syncytial virus infection in the upper and lower respiratory tract. *J Mol Biol,* 368**,** 652-65.

YU, Q. Z., HARDY, R. W. & WERTZ, G. W. 1995. Functional cDNA clones of the human respiratory syncytial (rs) virus n-proteins, P-proteins, and L-proteins support replication of rs virus genomic RNA analogs and define minimal trans-acting requirements for RNA replication. *Journal of Virology,* 69**,** 2412-2419.

ZHAO, Y. G., CHAPMAN, D. A. G. & JONES, I. M. 2003. Improving baculovirus recombination. *Nucleic Acids Research,* 31.

APPENDICES

APPENDIX I - PRIMER SEQUENCES

Name	Sequence 5'-3'	Use
TRIEX2F	CGGAATTCCCATCTGGTGCCGCGCGGCAGCGGCAGCAGCCTCGAGCGG (48)	Creating ST01 in pTriEx1.1 Neo (Chapter 3.1.1.1)
TRIEX2R	CCGCTCGAGGCTGCTGCCGCTGCCGCGCGGCACCAGATGGGAATTCCG (48)	
SUMO F	GCATGGATCCATGTCACGAAGGAATCC (27)	pET28aSUMO cloning, *BamHI* (Chapter 3.2.1.1)
SUMO R	CGATCTCGAGTTAGGTAGTATCATTATTTTTGGC (34)	pET28aSUMO cloning, *XhoI* (Chapter 3.2.1.1)
C96S F	CTAAACAATCAGCAAGTGTTGCCATGAGC (29)	C96S mutant (Chapter 3.2.1.5.3)
C96S R	GCTCATGGCAACACTTGCTGATTGTTTAG (29)	
S58DS61D F	GTATAGATACCTTAGATGAAATAGATGGAGC (31)	S58DS61D mutant (Chapter 4.2.1.3)
S58DS61D R	GCTCCATCTATTTCATCTAAGGTATCTATAC (31)	
R3AR4A F	GGATCCATGTCAGCAGCGAATCCTTGC (27)	RNA binding mutants (Chapter 5.2.4)
R3AR4A R	GCAAGGATTCGCTGCTGACATGGATCC (27)	
K19AR20A F	CATTGCTTAAATGGTGCGGCGTGTCATTTTAGTC (34)	
K19AR20A R	GACTAAAATGACACGCCGCACCATTTAAGCAATG (34)	

Name	Sequence 5'-3'	Use
K92A F	CAATAAACAATATAACTGCACAATCAGCATGTGTTGC (37)	RNA binding mutants (Chapter 5.2.4)
K92A R	GCAACACATGCTGATTGTGCAGTTATATTGTTTATTG (37)	
K92D F	CAATAAACAATATAACTGATCAATCAGCATGTGTTGC (37)	
K92D R	GCAACACATGCTGATTGATCAGTTATATTGTTTATTG (37)	
R126A F	CACCCAAGATAGCAGTGTACAATACTGTC (29)	
R126A R	GACAGTATTGTACACTGCTATCTTGGGTG (29)	
K150A F	CTATCCATCTGTTAGCAAGATTGCCAGCAGAC (32)	
K150A R	GTCTGCTGGCAATCTTGCTAACAGATGGATAG (32)	
R151A F	CCATCTGTTAAAAGCATTGCCAGCAGACG (29)	
R151A R	CGTCTGCTGGCAATGCTTTTAACAGATGG (29)	
R151D F	CCATCTGTTAAAAGATTTGCCAGCAGACG (29)	
R151D R	CGTCTGCTGGCAAATCTTTTAACAGATGG (29)	
K150AR151A F	CTATCCATCTGTTAGCAGCATTGCCAGCAGACG (33)	
K150AR151A R	CGTCTGCTGGCAATGCTGCTAACAGATGGATAG (33)	
K159A F	GACGTATTGAAGGCAACCATCAAAAACAC (29)	
K159A R	GTGTTTTTGATGGTTGCCTTCAATACGTC (29)	

APPENDIX II - ADDITIONAL RNA SEQUENCES FOR FLUORESCENCE POLARISATION

Name	Sequence 5'-3'
Le	ACGCGAAAAAAUGCGUACAACAAACUUGCAUAAACCAAAAAAAU-FI (44)
M/SH	AAUGAUUAUUUGCCCCAUGUGUAUAUUUUUUAUUAACUUAUUUG-FI (44)

APPENDIX III – M2-1 SEQUENCES

RSV A2 strain M2-1 nucleotide sequence

Start codon, stop codon

ATGTCACGAAGGAATCCTTGCAAATTTGAAATTCGAGGTCATTGCTTAAATGGTAA
GAGGTGTCATTTTAGTCATAATTATTTTGAATGGCCACCCCATGCACTGCTTGTAA
GACAAAACTTTATGTTAAACAGAATACTTAAGTCTATGGATAAAAGTATAGATAC
CTTATCAGAAATAAGTGGAGCTGCAGAGTTGGACAGAACAGAAGAGTATGCTCTT
GGTGTAGTTGGAGTGCTAGAGAGTTATATAGGATCAATAAACAATATAACTAAAC
AATCAGCATGTGTTGCCATGAGCAAACTCCTCACTGAACTCAATAGTGATGATATC
AAAAAGCTGAGGGACAATGAAGAGCTAAATTCACCCAAGATAAGAGTGTACAATA
CTGTCATATCATATATTGAAAGCAACAGGAAAAACAATAAACAAACTATCCATCT
GTTAAAAAGATTGCCAGCAGACGTATTGAAGAAAACCATCAAAAACACATTGGAT
ATCCATAAGAGCATAACCATCAACAACCCAAAAGAATCAACTGTTAGTGATACAAA
TGACCATGCCAAAAATAATGATACTACCTAA

RSV A2 strain M2-1 amino acid sequence and secondary structure alignment

MSRRNPCKFEIRGHCLNGKRCHFSHNYFEWPPHALLVRQNFMLNRILKSMDKSIDT
LSEISGAAELDRTEEYALGVVGVLESYIGSINNITKQSACVAMSKLLTELNSDDIKKLR
DNEELNSPKIRVYNTVISYIESNRKNNKQTIHLLKRLPADVLKKTIKNTLDIHKSITIN
NPKESTVSDTNDHAKNNDTT