The biology and biochemistry of PhoH2 proteins

A thesis

submitted in fulfilment

of the requirements for the degree

of

Doctor of Philosophy in Biological Sciences

at

The University of Waikato

by

EMMA SOPHIE VOUT ANDREWS

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2013

Abstract

PhoH2 proteins are found in a diverse range of organisms. To date little is known about these proteins and the role they play in the organisms in which they reside. PhoH2 is a PIN-PhoH domain fusion, and these proteins are currently annotated as having unknown function and are described as PhoH-like. PhoH-domains are thought to be ATPases and all characterised PIN-domain proteins are RNases. Most efforts have focussed on determining the role of PIN-domain proteins that comprise the toxic component (VapC) of VapBC toxin-antitoxin systems, in which the PIN-domain is coexpressed as part of an operon with an inhibitor (VapB). In the remaining cases where PIN-domain proteins can be found such as PIN-PhoH domain fusions, these cases remain unexamined. This thesis describes the biological and biochemical characterisation of the PIN-PhoH protein, PhoH2 from *Mycobacterium tuberculosis* and *Mycobacterium smegmatis*, along with a preliminary structural characterisation of a thermophilic PhoH2 protein homologue.

The *phoH2* gene from both mycobacterial organisms was found to be expressed as part of a long mRNA transcript. Examination of these transcripts revealed possible alternative 5' start sites out of frame with the *phoH2* gene. For protein overexpression, and 'normal' growth and colony formation with conditional overexpression, *phoH2* from *M. tuberculosis* required 152 bp of the 5' sequence directly upstream of the annotated *phoH2* gene (annotated here as $phoH2_{alt}$). PhoH2 proteins: $PhoH2_{alt\ MTB}$, $PhoH2_{alt\ MSMEG}$ and $PhoH2_{MSMEG}$ show ATP/Mg^{2+}-dependent, sequence-specific RNA unwinding and cleavage. The sequence (A C) (A/U) (A/U) (G/C) U was deduced as a substrate for PhoH2, and $PhoH2_{alt\ MTB}$ also demonstrated unwinding and cleavage activity on its upstream ~152 base RNA transcript, suggesting a potential autoregulatory mechanism. Structural analysis of a thermophilic PhoH2 protein homologue has provided preliminary crystallographic data which along with electron microscopy suggest a ring-like hexameric PhoH2 oligomer.

Acknowledgements

I would firstly like to thank my primary supervisor Professor Vic Arcus. Thank you for your ongoing support, expertise and guidance throughout this project. I am indebted for the opportunities you have given me not only to work on this project, but throughout my PhD studies. I would also like to thank Dr Ray Cursons for encouraging me to stay on and do a PhD and for your support during this project.

To everyone in the Proteins & Microbes and associated labs: Dr J Hobbs, Dr J McKenzie, Dr E Summers, Dr J Burrows, Associate Professor I McDonald, Dr M Till, Dr M Cumming, Dr S Hanning, Dr M Liddament, Dr J Steemson, Line, Ashley, Ali, Tifany, Jo D, Emily, Abby, Erica, Chelsea, Joel and Vikas, I would not have got through this without you. Jo M, thanks for showing me the ropes and giving me confidence in myself, Judith, thank you for being our lab Mum and keeping us going with your yummy baking, and Jo H, thank you for taking the time to critically read over this thesis.

Family and friends thank you for your love and support. Thanks for always being happy to listen to me go on 'protein this....protein that...' Mum, Dad and Jack, thanks for believing in me and for your ongoing encouragement. Dad thanks for your amazing Sunday roasts which not only got us together but set me up well for the week ahead. Mum thanks for our regular coffee dates and Jack thanks for our weekly games of squash. Jen and Laura, thanks for our catch ups that kept me in the real world.

Finally Shannon, thank you for your ongoing understanding, patience, love and support. Thank you for always being there for me, I would have not made it through this roller coaster ride without you.

I would like to thank the University of Waikato and the New Zealand Federation of Graduate Women for my scholarships, Universities New Zealand and the Maurice and Phyllis Paykel Trust for travel funding.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

SI (Systeme Internationale d'Unités) abbreviations for units and standard notations for chemical elements and formulae are used throughout this thesis. Other abbreviations are listed below.

3D	three dimensional
5'RACE	5' rapid amplication of cDNA ends
aa	amino acid
AAA	ATPases associated with diverse cellular activities
AAA+	ATPases associated with diverse cellular activities (variation)
ABC	ATPase Binding Cassette
ACN	Acetonitrile
ADC	albumin dextrose catalase
AMP-PNP	Adenylyl-imidodiphosphate
ASCE	Additional strand catalytic E
ATP	adenosine triphosphate
bp	base pair(s)
BLAST	basic local alignment search tool
BS-YlaK	*Bacillus subtilis*-YlaK
CFU	colony forming unit
COG	cluster of orthologous genes
CMR	Comprehensive microbial resource
C-terminal	carboxy terminus
Da	Daltons
DEPC	diethylpyrocarbonate
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
dNTP	deoxynucleotide triphosphate
ds	double-stranded
DTT	dithiothreitol
EDTA	ethylene diamine tetraacetic acid (disodium salt)
EV	empty vector
F	forward

FPLC	fast performance liquid chromatography
g	times the force of gravity
GITC	guanidium isothiocyanate
GSH	glutathione
GSSG	glutathione disulfide
HI	Walker A
H2	Walker B
HCCA	α-cyano-4-hyoxycinnamic acid
HEPES	N-2-hydroxyethylpiperazine-N‟-2-ethanesulphonic acid
His-tag	poly-histidine tag
HIV	human immunodeficiency virus
HMM	hidden markov model
IMAC	immobilised metal affinity chromatography
IPA	isopropanol
IPTG	isopropylthio-β-D-galactosidase
kb	kilobase
kDa	kilo dalton
KG	kinase-GTPase
KH	K-homology domain
kV	kilo volt
LB	luria bertani
LBT	luria bertani Tween-80
MALDI	matrix assisted laser desorption ionisation
mAU	milli-absorbance units
MCS	multiple cloning site
MDR	multidrug resistant
MES	2-(N-morpholino)-ethanesulfonic acid
MME	mono methyl ether
MPD	2-methyl 2,4-pentanediol
mRNA	messenger RNA
MS	mass spectrometry
MSMEG	*Mycobacterium smegmatis*
$MSMEG_{alt}$	*Mycobacterium smegmatis* from alternative start site
MTB	*Mycobacterium tuberculosis*

MTB_{alt}	*Mycobacterium tuberculosis* from alternative start site
MW	molecular weight
MWCO	molecular weight cut off
Native-PAGE	non-denaturing PAGE
NCBI	National Center for Biotechnology Information
NDP	Nucleotide diphosphate
NMD	nonsense mediated decay
NMR	nuclear magnetic resonance
N-terminal	amino terminus
Nt	nucleotide
NTP	nucleotide triphosphate
NYN	Nedd4-BP1, YacP nuclease
OD	optical density
ORF	open reading frame
PAGE	polyacrylamide gel electrophoresis
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PDB	protein data bank
PEG	polyethylene glycol
pI	isoelectric point
Pi	phosphate
PIN	PilT N-terminal domain
P-loop	Phosphate loop
Pst	Phosphate specific transporter
R	reverse
RE	restriction enzyme
rpm	revolutions per minute
RNA	ribonucleic acid
RNAP	RNA polymerase
RNase	ribonuclease
rRNA	ribosomal RNA
RT	room temperature
RT-PCR	Reverse transcriptase PCR
SI	Sensor I

SII	Sensor II
SDS	sodium dodecyl sulphate
Ser (S)	Serine
SF	superfamily
SPITC	4-sulfophenyl isothiocynate
SRH	Second region of homolgy
sRNA	small RNA
Super-DHB	2-hyroxy-5-methoxybenzoic acid
ss	single stranded
TA	toxin antitoxin
TAE	tris-acetate-EDTA
TB	tuberculosis
TBIS	*Thermobispora bispora*
Tc	Tetracycline
TE	tris EDTA buffer
TEM	Transmission electron microscopy
TEMED	N, N, N, N,-tetramethylethylenediamine
TFA	trifluoroacetic acid
TOF	time of flight
Tm	annealing temperature
$tRNA^{fMET}$	formylmethionine tRNA
TSP	transcriptional start point
UTR	untranslated region
UV	ultra violet
Vap	virulence associated protein
v/v	volume per volume
WHO	world health organisation
WT	wild type
w/v	weight per volume
XDR	extensively drug resistant

Chapter One: Introduction

1.1 Infectious diseases and the problem of persistence

Devastating life threatening illnesses are caused by a number of pathogenic organisms. One of the most common and serious bacterial infections is caused by *Mycobacterium tuberculosis.* The disease tuberculosis (TB) is transmitted in the air when infected individuals with active infection expel TB bacteria. Typically, it affects the lungs causing pulmonary disease but it can also affect other sites of the body and, if left untreated, the rate of mortality is high (WHO 2012). Treatment involves a 6 month course of combinations of four first-line antibiotics: isoniazid, rifampicin, ethambutol and pyrazinamide. For multi-drug resistant TB (MDR-TB) (resistant to isoniazid and rifampicin) the course is longer and requires more costly and toxic treatments/regimes (WHO 2012). Globally, the burden of this disease is high. In 2011, the disease caused 1.4 million deaths and 5.8 million new incidences of TB were reported to national TB control programmes; 13 % of these were individuals also infected with HIV, and 3.7 % of these incidences were MDR-TB (WHO 2012).

The prevalence of this disease is in part due to this organism's ability to enter a non-replicative, persistent state that is less susceptible to treatment with antibiotics (Connolly et al. 2007), and the more recent escalation of MDR and extensively drug resistant (XDR) strains of *M. tuberculosis*, along with the vulnerability of HIV-infected patients (WHO 2007).

M. tuberculosis has to survive in a number of stressful environments such as under hypoxia and nutrient stress, in order to persist within its host. In the genome of *M. tuberculosis* there is a high occurrence of proteins belonging to the PIN-domain family (Arcus et al. 2011). Members of this family are proposed to be stress response proteins and have been implicated in a number of stress related states such as the thermal stress reponse of *Sulfolobus solfataricus* (Cooper et al. 2009). They are also involved with pathogen trafficking (Mattison et al. 2006), and with metabolic regulation in *M. smegmatis* by virtue of their biochemical activity (McKenzie et al. 2012b).

The repertoire of PIN-domain proteins found within *M. tuberculosis* has led many researchers to speculate on the involvement of this protein family in the stress response and persistence of *M. tuberculosis* (Gerdes 2000; Hayes 2003; Arcus et al. 2005; Buts et al. 2005; Gerdes et al. 2005).

1.2 PhoH2 from *M. tuberculosis*

In the genome of *M. tuberculosis* there is a single copy of a *phoH2* gene (*Rv1095*). Annotations from the Comprehensive Microbial Resource (CMR) and the National Center for Biotechnology Information (NCBI) for this protein include: "unknown function but similar to proteins described as PhoH-like"; "phosphate starvation inducible protein" and "putative PhoH-like protein". Figure 1.1 shows the relative organisation of the two domains that make up *phoH2*.

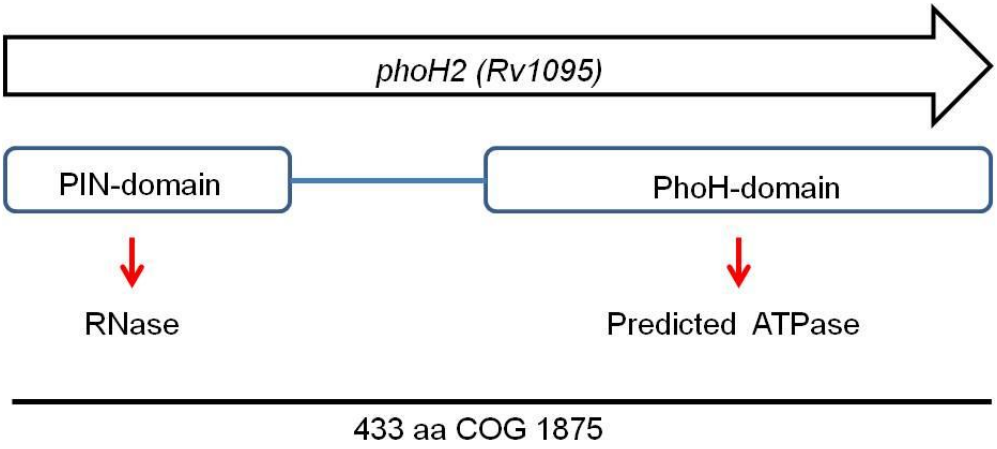

Figure 1.1 PhoH2 protein domain annotation
Domains identified in PhoH2 and their known and predicted biochemical functions. (http://www.ncbi.nlm.nih.gov/)

1.2.1 *phoH2* gene architecture

PFam lists 967 sequences with PIN-PhoH (PIN_4-PhoH) architecture (Punta et al. 2012). These sequences are found in many diverse bacterial phyla including Chlamydiae, beta, delta, and gamma Proteobacteria, Cyanobacteria, Lentispharae, Elusimicrobia, Verruccmicrobia, Thermotogae, Spirochaeta, Planctomycetes and Deferribacteres, and in the archaeal phylum Euyarchaeota. No sequences fall within the eukaryotes. The domain arrangement makes up COG 1875: NYN (Nedd4-BPI, YacP nuclease) ribonuclease and ATPase of PhoH family domain

(http://www.ncbi.nlm.nih.gov/). PhoH-like proteins are predicted to be ATPases that are induced under phosphate starvation (Marchler-Bauer et al. 2011). They are categorised in PFam as clan CL0023, which is part of the P-loop NTPase superfamily consisting of 193 families (Punta et al. 2012).

1.3 PIN-domain proteins

The PIN-domain family of proteins are found in a diverse range of organisms, and were first observed and annotated based on sequence similarity to the N-terminal domain of the type IV pilus protein, PilT (PilT N-terminus), from *Myxococcus xanthus* (Wall & Kaiser 1999). PIN-domain proteins are small proteins of ~130 amino acids that share three strictly conserved acidic residues at positions 4, 40 and 93, and a less well conserved residue at position 112 (Figure 1.2). Small polar residues (Asp N, Ser S, Thr T) following the first conserved aspartic acid at positions *i* + 1 or *i* + 2, also play a structural role within this family (positions 5 and 6 in Figure 1.2). Structurally, PIN-domain proteins are described as a 3-layer α/β/α sandwich with a central 5-stranded parallel β sheet (Andreeva et al. 2008). This conformation causes the conserved acidic residues to group together to form a putative active site (Arcus et al. 2011), which binds Mg^{2+} or Mn^{2+} and facilitates cleavage of RNA (Arcus et al. 2004; Arcus et al. 2005).

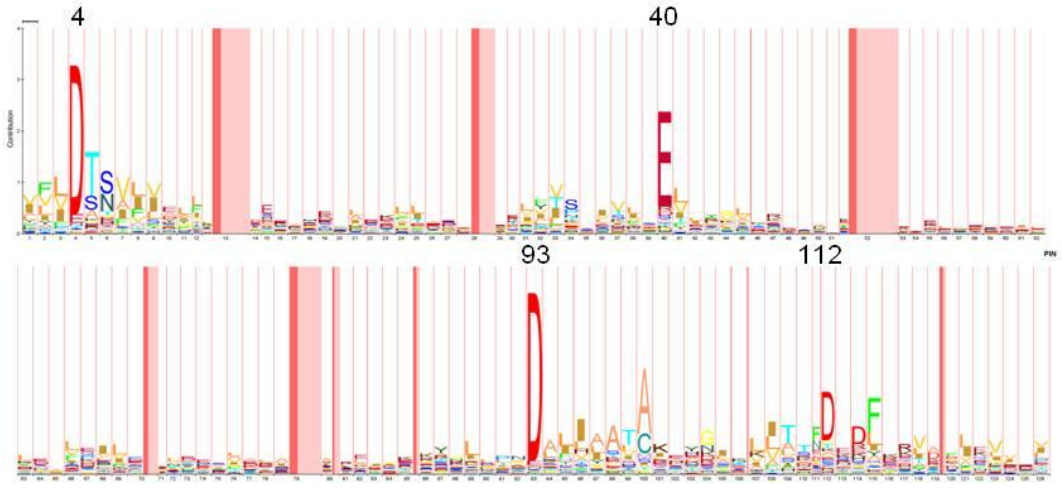

Figure 1.2 Hidden Markov Model that defines PIN-domain proteins
Based on a multiple sequence alignment the height of each letter defines the significance of the amino acid at that position within the PIN-domain family. The width of each letter defines the significance of this position in the identification of the PIN-domain family. The dark and light pink regions are where insertions have occurred. The 4 highly conserved amino acids (D, E, D, D) at positions 4, 40, 93 and 112 characterise the PIN-domain protein family. Generated using PFam (Finn et al. 2010).

Structural studies indicate that PIN-domain proteins have homology to 5'- 3' nucleases, T4 RNase H, *Taq* polymerase and flap endonucleases (Arcus et al. 2004), suggesting that PIN-domains are likely to exhibit nuclease activity. The majority (~95 %) of PIN-domain proteins are single domain proteins that are found in the genome with a transcription factor upstream that creates an operon with the PIN-domain gene. In ~5% of cases, however they are fused with TRAM domains, KH domains and AAA+ ATPase domains (Arcus et al. 2011). TRAM and KH domains are annotated as RNA-binding domains (Musco et al. 1997; Anantharaman et al. 2001) and the AAA+ ATPase domains belong to the AAA+ superfamily of ring-shaped P-loop NTPases (Iyer et al. 2004), which oligomerise into hexameric rings and are involved in remodelling or translocating macromolecules in an energy dependent manner.

Eukaryotic PIN-domain proteins are reported to be involved with nonsense mediated decay (NMD) of RNA (Takeshita et al. 2007; Huntzinger et al. 2008) and processing of rRNA pre-18S fragments (Lamanna & Karbstein 2009). The human PIN-domains SMG5 and SMG6 are involved in pathways that degrade mRNAs that have premature stop codons. Both have similar overall folds, however SMG6 contains all four conserved acidic residues required for activity, whereas SMG5 is missing three of the four residues (Glavan et al. 2006). SMG6 displays endoribonuclease activity on single-stranded RNA (ssRNA) in the presence of Mn^{2+} and, to a lesser extent, in the presence of Mg^{2+} but not on ssDNA or double-stranded RNA (dsRNA), and SMG5 does not show catalytic activity (Glavan et al. 2006). The yeast Nob1 protein also contains a PIN-domain that is required for processing of pre-18S rRNA fragments. The PIN-domain binds to cleavage site D of the ss pre-18S rRNA fragment and is responsible for cleavage at the ss 3' end (Lamanna & Karbstein 2009). Human, Est1A protein also contains a PIN-domain but is involved in the regulation of telomere elongation and NMD of RNA (Takeshita et al. 2006).

M. tuberculosis has 48 PIN-domain proteins in its genome (Anantharaman & Aravind 2003; Arcus et al. 2005; Arcus et al. 2011). These PIN-domains are

exactly conserved in *Mycobacterium bovis* (Pandey & Gerdes 2005), yet in other mycobacterial genomes such as that of *Mycobacterium smegmatis* and *Mycobacterium leprae*, far fewer or even no homologues of PIN-domains are found (Arcus et al. 2011).

Most structures of prokaryotic PIN-domains are dimers (Figure 1.3b). This arrangement positions the active sites into a groove along the long axis of the structure (Figure 1.3b). The PIN-domain PAE_{2754} from *Pyrobaculum aerophilum* forms a tetramer of two dimers and so organises the active sites into a tunnel conformation, that is suitable for the binding of ssRNA to the positively charged lysine residues that protrude into the tunnel (Figure 1.3c) (Arcus et al. 2004; Bunker et al. 2008).

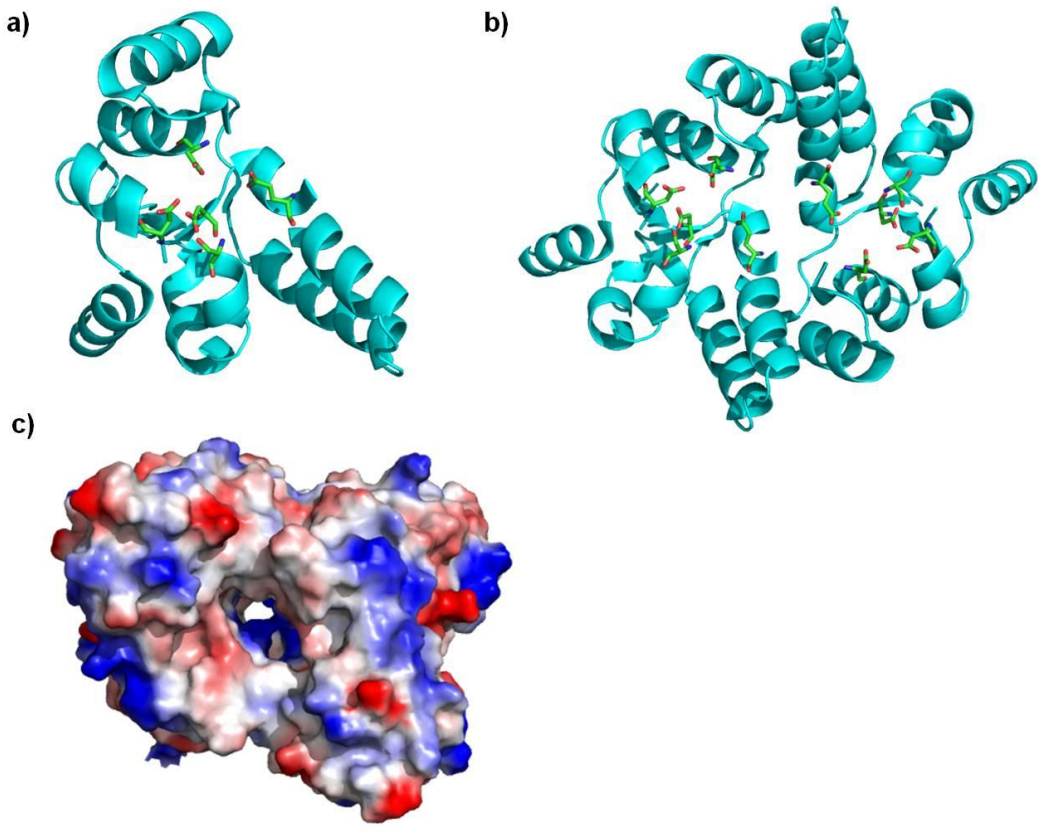

Figure 1.3 PIN-domain protein structures
a) Ribbon diagram of PAE_{2754} monomer. The four conserved acidic residues and polar residue are shown as sticks. b) Ribbon diagram of PAE_{2754} dimer. The four conserved acidic residues and polar residue are shown as sticks which form a groove. c) Electrostatic surface diagram of PAE_{2754} tetramer. Negative (red) and Positive (blue) charges show where ssRNA may bind to positively charged residues at the entrance of the tunnel. Images were made in PyMOL using PDB IV8P.

The structure of the PIN-domain of MJ_{1533} from *Methanocaldococcus jannaschii* (PDB 3I8O) adopts a different monomeric and higher order conformation to previously crystallised PIN-domain proteins (Figure 1.4). This PIN-domain is found at the N-terminus of MJ_{1533} which is annotated as a PIN_VirB11L-ATPase (http://www.ncbi.nlm.nih.gov/). VirB11 proteins are part of the type IV secretory pathway involved in T-pilus biogenesis and virulence (Marchler-Bauer et al. 2011) and these ATPases form dynamic hexameric structures (Savvides et al. 2003). The structure of this PIN-domain protein forms an eight stranded β barrel with four-fold symmetry which creates a central channel (Figure 1.4c and d).

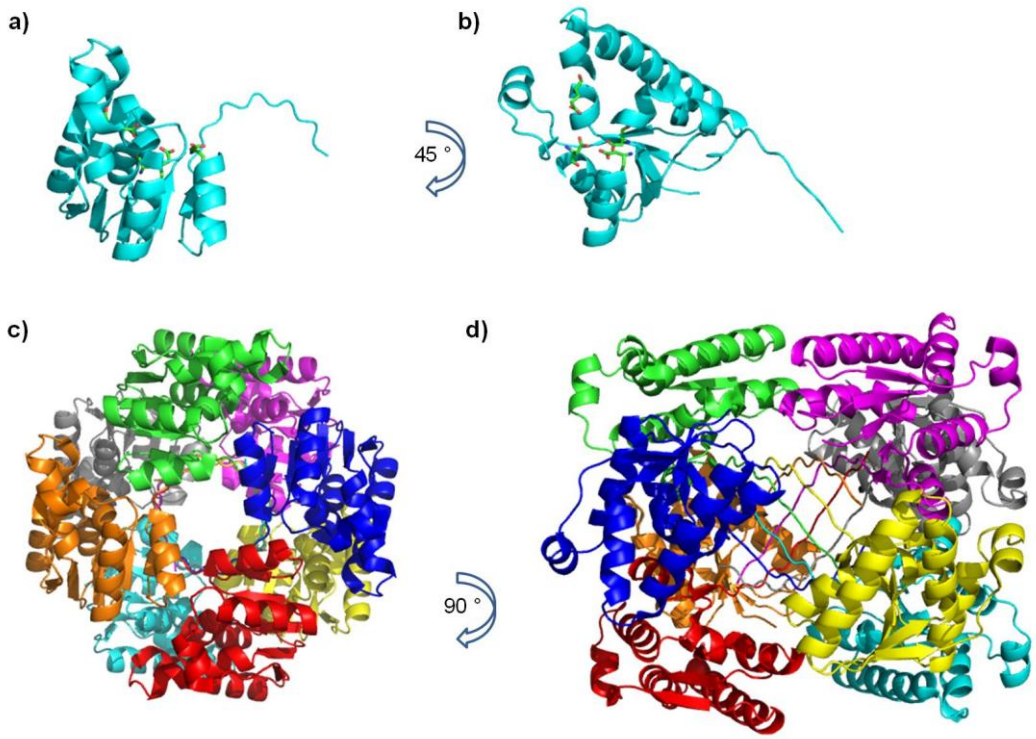

Figure 1.4 Structure of PIN-domain protein of MJ_{1533}
a) Ribbon diagram of MJ_{1533} PIN-domain monomer showing the long tail. The four conserved acidic residues are shown as sticks. b) The same structure rotated 45 ° to show the position of the conserved residues. c) Symmetry mates of monomer conformation showing the possible interaction of multiple monomers and the central channel that is formed. d) The same structure rotated 90 ° to show the tunnel formed by the tails. Images were made in PyMOL using PDB 3I8O.

It is unknown how the PIN-domain in this higher order oligomeric state would interact with its ATPase domain, if the ATPase domain is hexameric. A hexameric protein would not adopt four-fold symmetry. This PIN-domain structure, however demonstrates that PIN-domain proteins can adopt a range of oligomeric states.

Despite being found in many important pathogens such as *M. tuberculosis* (Arcus et al. 2005; Ramage et al. 2009), the functional role of PIN-domain proteins in this bacterium is not understood.

1.4 P-loop NTPases, helicases and PhoH-like ATPases

1.4.1 P-loop NTPases

The P-loop NTPases are a superfamily of proteins involved in a diverse range of cellular functions. They are classified based on their structure into two main classes: the kinase-GTPase (KG) class and the additional strand catalytic E (ASCE) class. Proteins belonging to these classes include the Ras-like GTPases and the circularly permutated YlqF-like GTPase proteins, and the ATPase Binding Cassette (ABC), DExD/H-like helicases, 4Fe-4S iron sulphur cluster binding proteins of the NifH family, RecA-like F1-ATPases, and AAA+ ATPases, respectively (Marchler-Bauer et al. 2011).

Proteins that belong to the P-loop NTPase superfamily are characterised by two conserved nucleotide phosphate-binding motifs: the Walker A (GxxxxGK[ST], where x is any residue) and Walker B motifs (hhhh[DE], where h is a hydrophobic residue) (Figure 1.5). The Walker A motif adopts a flexible loop conformation that enables correct positioning of the triphosphate moiety of the bound nucleotide triphosphate (NTP). The conserved aspartic acid (D) in the Walker B motif is involved with coordination of a Mg^{2+} cation and securing the two phosphate binding motifs through a bond to the threonine or serine (T/S) of Walker A (Walker et al. 1982; Leipe et al. 2002; Marchler-Bauer et al. 2011).

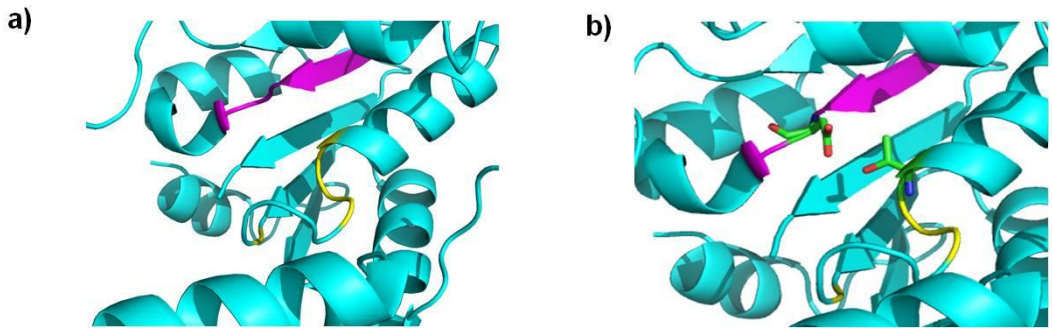

Figure 1.5 Walker A and B motifs of P-loop NTPase
a) Respective positions of the Walker A (yellow) and Walker B (magenta) motifs and b) positions of the conserved threonine (T) of Walker A and aspartic acid (D) of the Walker B motifs involved with positioning of the two nucleotide phosphate binding motifs (Walker et al. 1982). Images were made in PyMOL using PDB 3B85.

The P-loop superfamily contains a number of families which contain helicase proteins and other motor proteins.

1.4.2 Helicase superfamilies

Helicases are enzymes that couple the hydrolysis of ATP to the unwinding of double stranded nucleic acid into its respective single strands (Singleton et al. 2007).

Amino acid sequence comparisons were used in the early classification of helicases which revealed three distinct and diverse superfamilies (SF1, SF2 and SF3) and two smaller families (SF4 and SF5) (Gorbalenya & Koonin 1993). All helicases carry the Walker A and Walker B motifs. SF1 and SF2 are the largest of the superfamilies, and contain similar sets of seven conserved motifs scattered across the sequence. The third superfamily (SF3) is made up of viral helicases that share four conserved motifs. The DnaB-like family (SF4) contains helicases that carry the Walker A and B motifs plus three distinct conserved motifs. The Rho-like family (SF5), despite its similarity to SF4 helicases, is categorised as a separate family on the basis of sequence (Singleton et al. 2007). Singleton (2007) has proposed the addition of the AAA+ class as SF6 to encompass the nucleic acid motors that are categorised in this family. The signature sequences of amino acids used for the early identification of helicase proteins found by Gorbalenya & Koonin (1993) are listed in Table 1.1.

Table 1.1 Signature amino acid sequences used to identify helicase proteins

Motif	Amino acid pattern	Helicase Group
I	$hx_4GxGK[TS]x_2[hA]$	SF1
II/V	$[hH][hA]D[DE]x_n[TSN]x_4[QK]Gx_7[hA]$	
VI	$x_n[VT]x[UTC][TS]R$	
I, II, VI (relaxed pattern)	$GxGK[ST]x_n$	SF2
	$[hA][hAC]DEx[DHQ]x_n$	
	$[QH]x_3Rx_2R$	
I, II, VI (strict pattern)	$GxGK[ST]x_n$	SF2
	$[hA][hAC]DEx[DH]x_n$	
	$[QH]Gx_2Rx_2R$	
I/II	$UUx_3UGK[ST]x_nUUhxD[YH]Ux_nUx_2U$	DnaB family
III	$x_3U[KR][GA]hA$	

h - bulky hydrophobic residue (I, L, V, M, F, Y, W), U - bulky aliphatic amino acid residue (I, L, V, M), x - any residue, alternate residues in square brackets.

Helicases are either toroidal, hexameric structures, or non-ring forming (Singleton et al. 2007). SF1 and SF2 helicases are non-ring forming and the remaining SF3-5 families plus the AAA+ class, contain those helicases that are toroidal in nature. Translocation of nucleic acid can occur with 3' to 5' (type A) or 5' to 3' (type B) polarity. SF1, SF2 and the AAA+ class contain examples of both types A and B. Thus far, SF3 are type A and all members of SF4 and SF5 are type B.

1.4.2.1 SF1 and SF2

The SF1 and SF2 helicases contain a signature core of tandem RecA-like folds (α/β/α) either within one polypeptide chain or between subunits (Singleton et al. 2007). The highest level of sequence conservation between SF1 and SF2 occurs in the residues that coordinate binding and hydrolysis of the triphosphate (motifs I, II and VI) (Table 1.1) identified by Gorbalenya & Koonin (1993). These motifs are positioned in the cleft between the two conserved RecA-like domains. Figure 1.6 illustrates the respective sequence and locations of the seven conserved motifs identified for the SF1 helicase PcrA (Caruthers & McKay 2002).

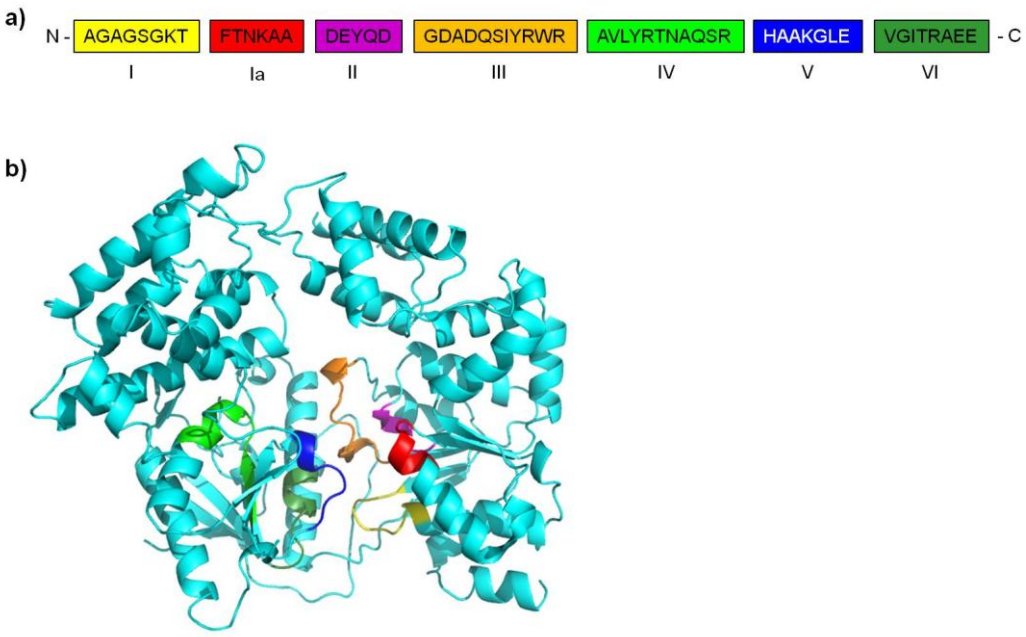

Figure 1.6 Structure and conserved motifs of the representative SF1/SF2 helicase PcrA
a) Illustration of the conserved motifs (Gorbalenya & Koonin 1993) with sequences from *Bacillus stearothermophilus* PcrA (SF1). Motifs I and II are the Walker A and B motifs respectively. b) Ribbon diagram of PcrA. The conserved motifs are coloured as in a). Images were made in PyMOL using PDB 1PJR.

Residues in motif Ia, III, IV, and V are involved in nucleic acid binding (Caruthers & McKay 2002). The signature motifs of SF1 and SF2 helicases have increased to include motifs such as TxGx (nucleic acid binding) that precedes the Walker B motif, and the Q motif (coordination of adenine base) that precedes the Walker A motif, which can be specific to each superfamily or subfamily (Singleton et al. 2007). Despite some complexity around the motifs, across all families the motifs contain universal features: a) residues involved in the binding and hydrolysis of the NTP that are equivalent to the Walker A and B motifs and b) an 'arginine finger' (motif VI) that plays a role in energy coupling.

The remaining helicase families (SF3-5 and AAA+) are toroidal composed of six individual RecA-like folds.

1.4.2.2 SF3

The SF3 helicases were initially identified in small DNA and RNA viruses (Gorbalenya & Koonin 1993). Helicases from this family form single or

double hexamers and unwind nucleic acid with 3' to 5' directionality. SF3 helicases share four conserved motifs: Walker A, Walker B, B' (oligonucleotide interactions) and C (analogous to sensor I AAA+). The conserved arginine finger is located after motif C (Hickman & Dyda 2005; Singleton et al. 2007). Figure 1.7a shows the five conserved motifs identified for the SF3 helicase SV40 (Hickman & Dyda 2005). The monomeric and hexameric structures (Figure 1.7b and c) show the position of the motifs and the nature of the toroidal structure.

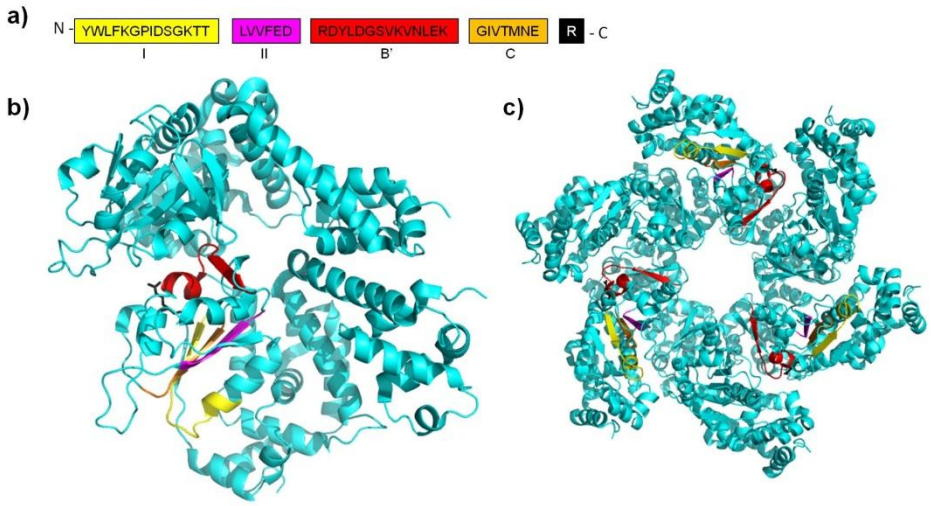

Figure 1.7 SF3 helicase SV40 Large T antigen helicase domain
a) Illustration of the conserved motifs with sequences from Simian virus 40 (SF3). Motifs I and II are the Walker A and B motifs respectively. b) Ribbon diagram of SV40. The conserved motifs are coloured as in a). The conserved arginine residue is shown as a stick (black). c) Hexamer of SV40. Images were made in PyMOL using PDB IN25.

SF3 helicases have an *ori* DNA-binding domain preceding the helicase domain, and the helicase domain is a modified AAA+ core that contains the nucleotide binding motifs in the central β strands and a set of inserted loops that are involved with DNA binding or protein-protein interactions (Singleton et al. 2007).

1.4.2.3 SF4

Helicases in this family were initially identified in bacteria and bacteriophages, and are replicative helicases. Gorbalenya & Koonin (1993) classed these helicases as DnaB-like and they are functionally and physically associated with DNA primases (Gorbalenya & Koonin 1993). These helicases

contain the Walker A (H1) and B motifs (H2) plus three distinct motifs (H1a, H3 and H4). These latter motifs have no obvious counterparts in any of the other helicase families (Singleton et al. 2007). Figure 1.8 shows the monomeric and hexameric structure of the SF4 helicase T7 gp4D. The coloured regions in the central pore shown in Figure 1.8c are thought to be involved with DNA binding (excluding H1 and H2, coloured yellow and magenta respectively) (Donmez & Patel 2006).

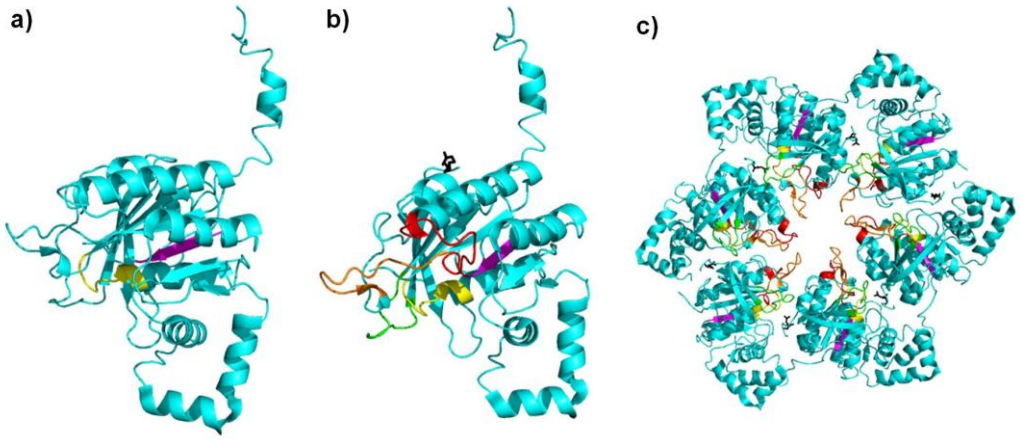

Figure 1.8 SF4 helicase T7 gp4D
Ribbon diagrams of a) T7 gp4D monomer showing the positions of the H1 and H2 motifs (Walker A and B motifs, coloured yellow and magenta respectively). b) T7 gp4D monomer showing the additional motifs thought to be involved with DNA binding coloured in red, orange and green (residues 424-429, 464-475, 503-513 respectively (Donmez & Patel 2006)). c) T7 gp4D hexamer. The motifs thought to be involved with DNA binding are positioned in the channel of the central pore. Images were made in PyMOL using PDB 1E0K.

1.4.2.4 SF5 Rho

The Rho-like helicases are categorised as a separate family based on differences in their amino acid sequences, despite their close relatedness to SF4 (Singleton et al. 2007). Rho carries out the termination of bacterial transcription through binding to specific sequences on the nascent RNA and then unwinding the DNA/RNA hybrid (Singleton et al. 2007). Figure 1.9 depicts the hexameric structure of Rho.

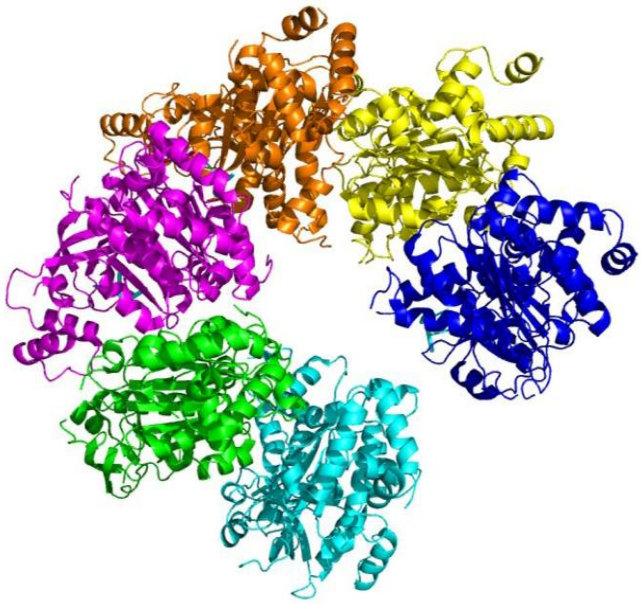

Figure 1.9 SF5 helicase Rho
Hexameric structure of Rho SF5 helicase. The colours show the six individual RecA-like folds that coordinate into the hexameric ring. Image was made in PyMOL using PDB 1PV4.

Rho adopts an open notched hexameric ring which is thought to represent its natural functional state for the loading of RNA into the hole of the hexamer (Yu et al. 2000).

1.4.2.5 AAA and AAA+ (SF6)

A subset of NTPases are distinguished as the additional strand catalytic E (ASCE) based on the insertion of a β-strand between the Walker A and B motifs, as well as conservation of a second conserved acidic residue in the Walker B motif (DE) (Erzberger & Berger 2006). Many are described as being RecA-like as the first ASCE protein to be characterised was RecA. The differences between ASCE protein families arise from small insertions and rearrangements of secondary structural elements that are specific to distinct groups and that help determine the relative organisation of subunits within higher order assemblies.

For AAA+ proteins, the primary distinguishing feature is the absence of β strand additions to the ASCE core and a small helical bundle fused to the C-terminus of the central α/β/α fold (Erzberger & Berger 2006). In AAA+

proteins, the arginine finger occupies a defined position generally at the end of α5 and is known as the second region of homology (SRH). Two other nucleotide interaction motifs, sensor I, a polar residue (N, D, T, or H) at the top of β4, and sensor II, an arginine residue at the base of α7 in the helical lid, compose the conserved motifs (Erzberger & Berger 2006) (Figure 1.10).

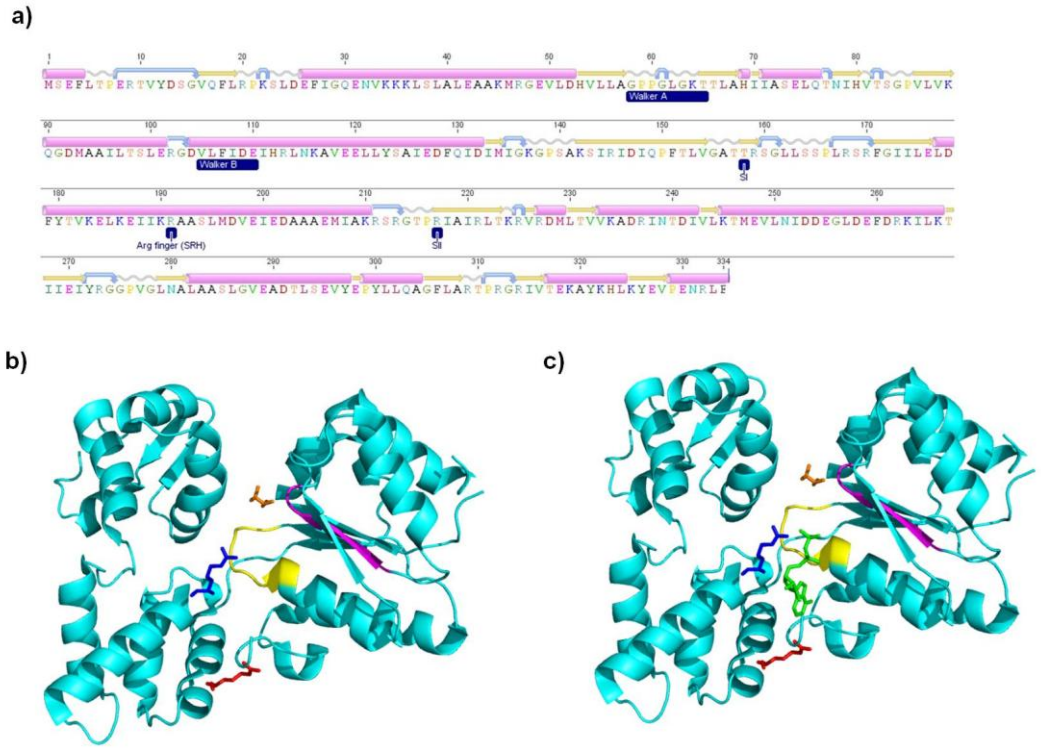

Figure 1.10 SF6/AAA+ helicase RuvB
a) Annotated RuvB amino acid sequence showing the locations of the conserved motifs. Secondary structure predictions: α helix (pink), β strand (yellow), coil (grey), turn (blue). b) Ribbon diagram of RuvB monomer showing the positions of the Walker A and B motifs (coloured yellow and magenta respectively), sensor I (orange stick) and II (blue stick), and arginine finger SRH (red stick). c) RuvB monomer with ADP (green), the other motifs coloured as in b). Annotated sequence was made using GeneiousPro (V 5.1.7). Images were made in PyMOL using PDB 1IN4.

For AAA proteins, the point of difference lies with the SRH which is positioned between β4 and β5 (Lupas & Martin 2002). Both AAA+ and AAA proteins have β strand topology 51432 (Figure 1.11).

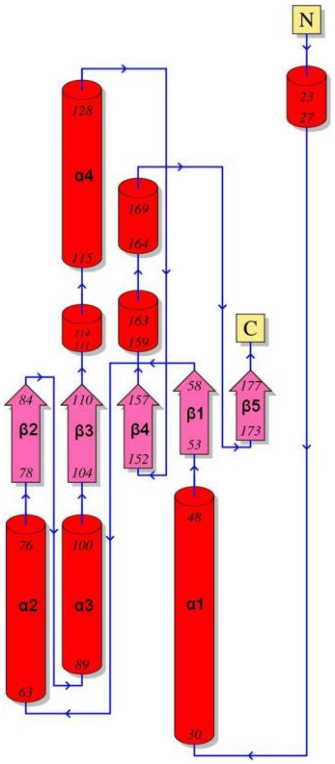

Figure 1.11 AAA+ helicase RuvB N-terminal domain topology
Topology diagram of AAA+ helicase RuvB N-terminal domain showing β strand topology 51432.
Image was generated using PDBsum (http://www.ebi.ac.uk/pdbsum/).

1.4.2.6 Mechanisms of NTP hydrolysis

The ASCE assembly of hexameric helicases generally places the ATP binding motifs of one subunit near the γ phosphate sensor elements of an adjacent subunit, leading to a radial arrangement of active sites that cooperate to hydrolyse ATP (Thomsen & Berger 2009). The cycles of nucleotide binding and hydrolysis are coupled to conformational rearrangements of substrate-binding elements in all ring-translocases and three major types of mechanisms have been proposed for the progression of nucleotide turnover within the system: 1) Sequential 2) Concerted and 3) Stochastic (Lyubimov et al. 2011) (Figure 1.12).

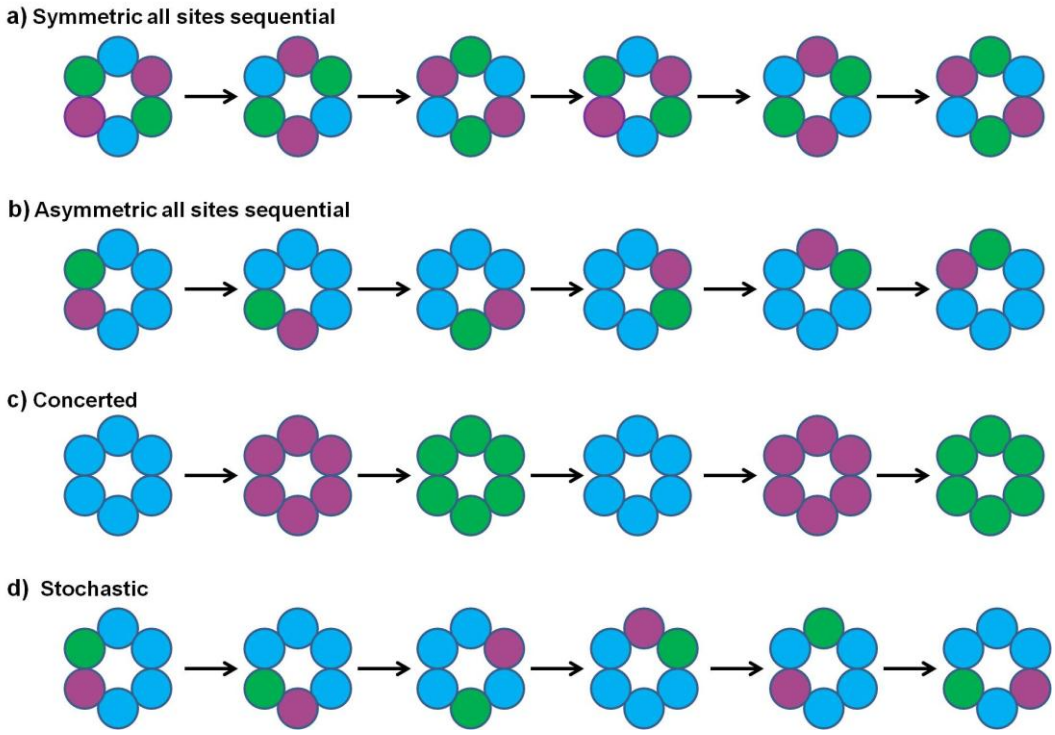

Figure 1.12 Models of NTP turnover by hexameric helicases
a) All sites sequential (symmetric) b) All sites sequential (asymmetric) c) Concerted and d) Stochastic. Each subunit of the helicase is represented by a coloured circle. NTP bound (Blue), NDP bound (Purple) and Empty (Green). Models were adapted from Singleton (2007), Rabhi (2010) and Lyubimov (2011).

In the contemporary sequential model, all subunits in the hexamer are catalytic and cycle through NTP bound, nucleotide diphosphate (NDP) bound and empty states (Figure 1.12a and b). NTP hydrolysis is coordinated between adjacent subunits either NTP bound, NDP bound and empty states at any given time. This model may be symmetric or asymmetric. In the presence of DNA or RNA the nucleotide-binding pattern and symmetry of the system may be affected, and all subunits would interact with the substrate at some point during the translocation cycle (Singleton et al. 2007).

The concerted model proposes that all six subunits are in the same state and that all interchange between NTP bound, NDP bound and empty simultaneously (Figure 1.12c). This model is suggested to represent SF3 SV40 and may be applicable to AAA+ translocases (Singleton et al. 2007).

For the stochastic model all subunits are active, however, there is no stringent sequential order of their firing. In this case any single subunit may

independently hydrolyse ATP and translocate the substrate (Figure 1.12d). A consequence of this type of activity is that stalling or inactivation of any single subunit will not completely inhibit the enzyme. This model is reported to work well for ClpX but not for any helicases (Singleton et al. 2007).

A more detailed mechanism has been elucidated for E1 and Rho proteins whereby the NTPase state influences the polarity of unwinding (Thomsen & Berger 2009; Rabhi et al. 2010; Lyubimov et al. 2011). The organisation of the ATPase state with each subunit dictates the relative position of hairpins or loops that line the central channel. These form a right handed 'spiral staircase' conformation, which curls around the 5' or 3' face of the phosphodiester backbone and acts to tug and chaperone the substrate through the central pore. This occurs in a single direction owing to the sequential firing around the helicase ring and arrangement of substrate-binding loops. The direction of movement is influenced by reversing the substrate binding orientation, reversing the chirality of the pore loop staircase and by reversing the NTPase cycle progression (Lyubimov et al. 2011). Both these proteins share a common substrate binding orientation, however based on the differences in NTPase state upon substrate binding and the consequent firing order, the directionality is reversed between these two proteins.

For Rho, the six catalytic ATPase sites are split into four distinct states representing an empty or nucleotide exchange state (E), a tightly NTP-bound state (T), a weakly NTP-bound state (D) and a hydrolysis competent state (T*). The relative order of flanking active sites in E1 is flipped with respect to Rho (Figure 1.13).

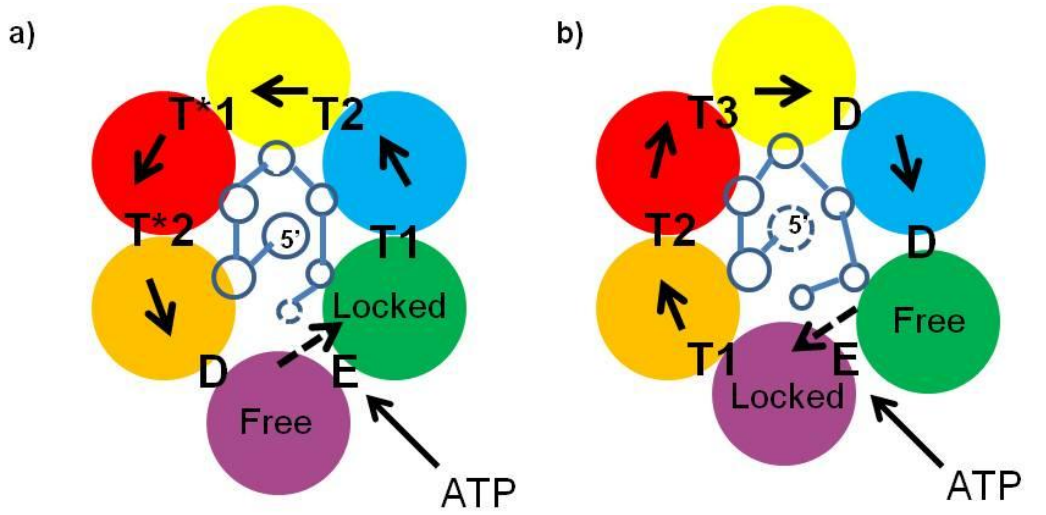

Figure 1.13 Schematic of Rho and E1
Schematic of sequential ATP hydrolysis cycle for a) Rho and b) E1. Each subunit of the hexamer is coloured differently. Phosphates from the nucleic acid are shown as open circles with the incoming phosphate shown with a dashed open circle. The interface between subunits represents the bipartite active site which is either empty or exchange (E), NTP tightly bound (T), NTP weakly bound (D) or in a hydrolysis competent state (T*). Subunits that lock (touch) represent the insertion of the arginine finger in the ATP bound state. The solid arrows dictate the direction of ATPase cycle. The dotted arrow represents the movement of the mobile subunit upon NTP binding, in the direction of a partner subunit that is locked in place by ATP-dependent contacts in the ring. Models were adapted from (Thomsen & Berger 2009).

The exact mechanism of how hexameric motor proteins coordinate ATP turnover between 6 catalytic sites to processive translocation of a polymeric substrate is unknown.

1.4.3 PhoH-like ATPases

The PhoH family of ATPases are thought to have evolved in bacteria (Anantharaman et al. 2002). As a result of a gene duplication event within the bacterial lineage (Figure 1.14), there are two orthologous groups of PhoH-like ATPases: PhoH and YlaK (Anantharaman et al. 2002).

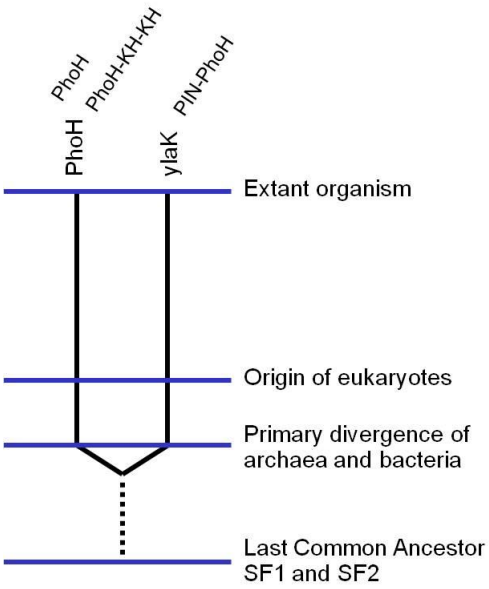

Figure 1.14 Predicted evolution of PhoH-like ATPases
Scheme adapted from Anantharaman (2002) "an evolutionary scheme for RNA helicases and related ATPases". K-homology domain (KH).

An *in silico* study has shown that orthologues of *Bacillus subtilis* YlaK (BS-YlaK) proteins are present in most aerobic species and in actinobacteria, where they were shown to be linked to regulons involved with fatty-acid beta-oxidation (Kazakov et al. 2003). Positional analysis showed that BS-YlaK orthologues are part of non-conserved loci, however they are found co-localised with genes encoding oxidative stress proteins. It was later identified that BS-YlaK orthologs were only observed in organisms with aerobic metabolism (Kazakov et al. 2003).

Using computer methods, Koonin and Rudd (1996) tested the hypothesis that proteins exist that are homologous to the N-terminus and C-terminus of SF1 helicases. They suggested that the evolution of PhoH-like ATPases has come about through the loss of the C-terminal α/β domain of SF1 and SF2 helicases that was apparent in the last common ancestor of these two superfamilies. A conserved sequence was identified that is shared specifically by PhoH from *Escherichia coli* and its homologue in *M. leprae*, $hx_2[GA]x_2GxGK[TS]x_n$ $hx_2DExQX_nhx_2GDx_2Q$ (x any amino acid and h any hydrophobic amino acid (A, V, L, I, P, M, F, W)) which is similar to motifs I, II and III of SF1 helicases (Koonin & Rudd 1996) (Figure 1.15).

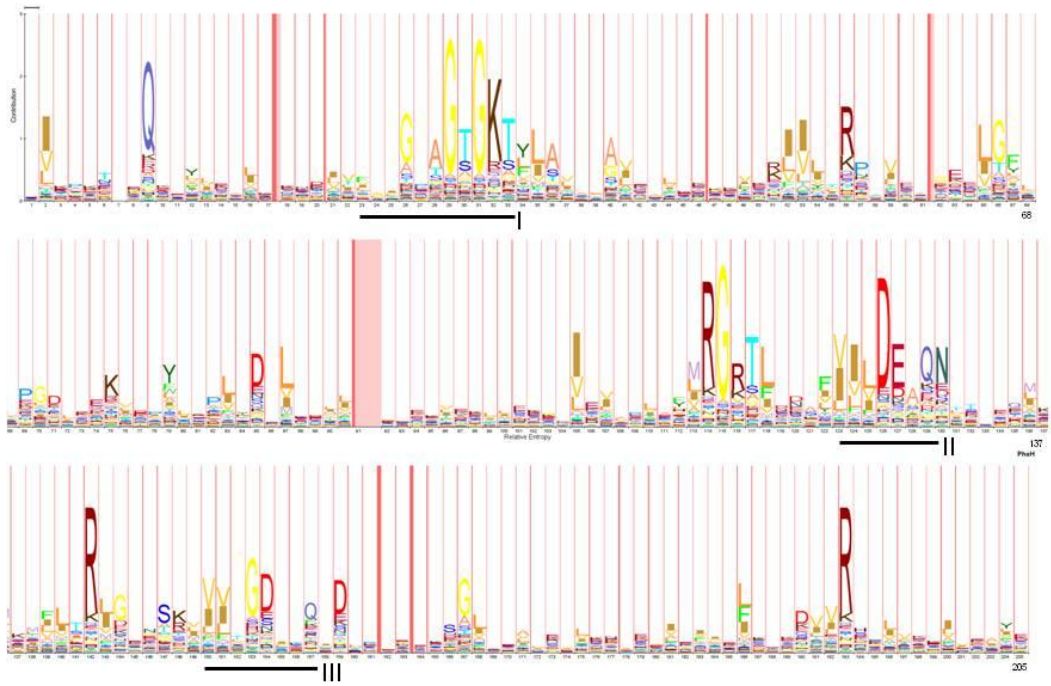

Figure 1.15 Hidden Markov Model showing conserved motifs of PhoH proteins that are similar to SF1 helicases
Based on a multiple sequence alignment the height of each letter defines the significance of the amino acid at that position within the PhoH family. The width of each letter defines the significance of this position in the identification of the PhoH family. The dark and light pink regions are where insertions have occurred. The black underlined regions mark the conserved tripartite sequence specific to PhoH from *E. coli* and *M. leprae*. The first and second (I and II) contain the Walker A and B motifs respectively (Koonin & Rudd 1996). Generated using PFam (Finn et al. 2010).

Motif I Walker A, Motif II Walker B, and motif III in SF1 are involved with the coordination of the NTP and nucleic acid binding (Fairman-Williams et al. 2010). This analysis suggests that PhoH proteins are missing the recognised helicase domain motifs V and VI of SF1 helicases, and led Koonin and Rudd (1996) to propose that a stand-alone PhoH protein is comparable to the N-terminal helicase domain of SF1 helicases.

Upon closer examination of the PhoH protein HMM and comparisons with the sequence of the one PhoH structure (PDB 3B85), *cg2513* from *Corynebacterium glutamicum*, it is apparent that there are a number of additional conserved residues, many of which are similar to conserved motifs of other helicase families (Figure 1.16 and Table 1.2).

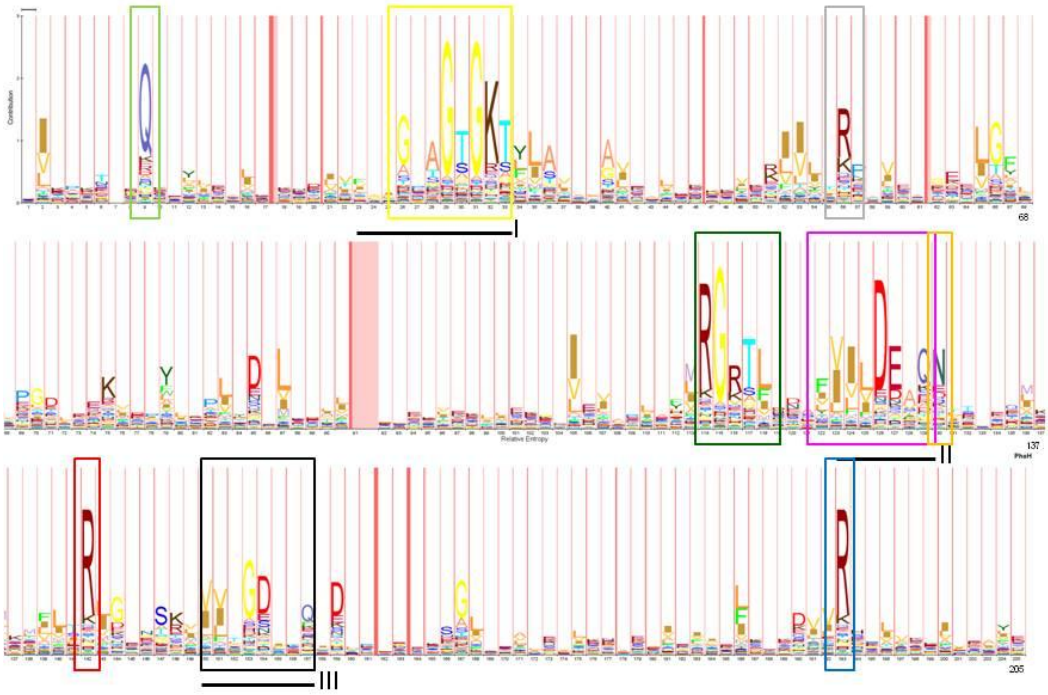

Figure 1.16 Hidden Markov Model showing the conserved motifs of PhoH proteins
Based on a sequence alignment the height of each letter defines the significance of the amino acid at that position within the PhoH family. The width of each letter defines the significance of this position in the identification of the PhoH family. The dark and light pink regions are where insertions have occurred. The black underlined regions mark the conserved tripartite sequence specific to PhoH from *E. coli* and *M. leprae* (Koonin & Rudd 1996). The boxed regions show sites of high amino acid conservation that are similar to that of other helicase families: Q motif (green), Walker A (yellow), R residue (light grey), RGRTL (dark green), Walker B (magenta), N residue (orange), R residue (red), motif III (black) and R residue (blue). Generated using PFam (Finn et al. 2010).

Two of the nine motifs shown in Figure 1.16 appear to be unique to PhoH proteins, the arginine residue (Figure 1.16 light grey box) and the RGRTL sequence (Figure 1.16 dark green box). The remaining conserved regions can be matched to characteristic motifs of other helicase families (Table 1.2).

Table 1.2 Comparison of conserved amino acids in PhoH with other conserved helicase SF motifs

	Motif	Position in HMM	Position in 3B85	Identified in SF	Function
Green	Q motif	9	124	Some SF1/2 SF2 DEAD box	Adenine coordination
Yellow	Walker A	26-33 GxxGxGKT	141-148 GPAGSGKT	All	NTP coordination and hydrolysis
Light grey	R residue	56	171 SO_4 binding	Unique	
Dark green	RGRTL	114-118 RGRTL	229-233 SO_4 binding	Unique	
Magenta	Walker B	123-129 hhhhDExQ	237-244 FVILDEAQ	All	NTP coordination and hydrolysis
Orange	N residue	130	245 Top of β4	AAA+ SI Polar residue top of β4	Water coordination with WB for nucleophilic attack on γ phosphate
Red	R residue	142	257 End of α5	AAA SRH Arg finger end of α5 (between β4 and β5)	Energy coupling
Black	Motif III	150-157 VVxGDxxQ	265-272	SF1 Motif III	Coordination of NTP and nucleic acid binding
Blue	R residue	193	308 Base of α7	AAA+ SII Base of α7	Interaction with γ phosphate

Q motif (Singleton et al. 2007), WA and WB (Walker et al. 1982; Leipe et al. 2002; Marchler-Bauer et al. 2011), SI AAA+ (Erzberger & Berger 2006), SRH AAA (Lupas & Martin 2002), Motif III SF1 (Caruthers & McKay 2002), SII AAA+ (Erzberger & Berger 2006).

A topology diagram of the PhoH protein sequence from PDB 3B85 and the 3D structure of PhoH protein (*cg2513*) from *C. glutamicum* show the positions of the conserved residues and motifs identified in the HMM, and by SO_4 ion binding in the structure of PDB 3B85 (Figure 1.17). Both show that PhoH adopts the traditional P-loop NTPase α/β protein structure of α-β units with centrally, most parallel β-sheets surrounded by α-helices (Leipe et al. 2002), and has SF1 helicase family β strand topology, 615423 (Figure 1.17a) (Fairman-Williams et al. 2010). It is apparent that PhoH has the SRH (arginine finger) positioned similar to that of AAA proteins between β4 and

β5 (amino acid 257 3B85) (Lupas & Martin 2002) (Figure 1.17). Figure 1.17c and d show the relative position of the conserved residues and motifs.

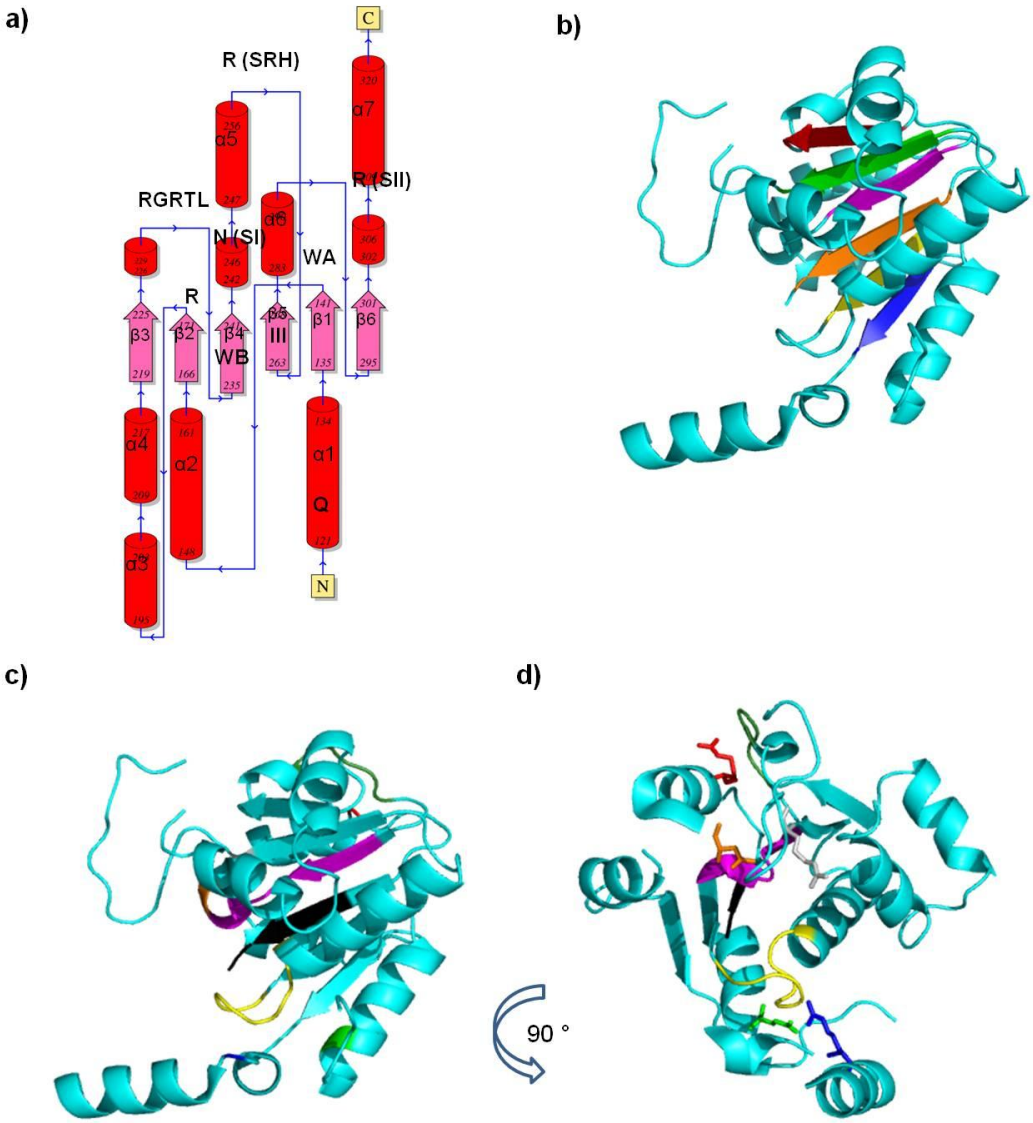

Figure 1.17 PhoH protein topology and PhoH-like ATPase monomer protein structure

a) Topology diagram of PhoH protein 3B85 showing SF1/2 β strand topology (615423) and the positions of the conserved residues and motifs identified in the HMM: Q motif (Q), Walker A (WA), arginine residue (R), RGRTL motif, Walker B (WB), Sensor I (N) (SI), arginine second region of homology (R) (SRH), Motif III (III), Sensor II (R) (SII). b) Monomeric PhoH protein from *C. glutamicum* (*cg2513*) showing the characteristic SF1/2 family domain β strand topology, 615423 (yellow β1, green β2, red β3, magenta β4, orange β5 and blue β6) also shared by PhoH proteins (Fairman-Williams et al. 2010). c) Monomeric PhoH2 protein showing the positions of conserved residues/motifs Q (green), Walker A (yellow), R residue (light grey), RGRTL (dark green), Walker B (magenta), N residue (SI) (orange), R residue (SRH) (red), III (black) and R residue (SII) (blue). d) Monomeric PhoH2 protein turned 90 ° to show the positions of the conserved residues/motifs (coloured as in b). Images were generated using PDBsum (www.ebi.ac.uk/pdbsum/) and were made in PyMOL using PDB 3B85.

Based on the structure of PDB 3B85, it is likely that PhoH-like ATPases adopt a hexameric macromolecular conformation (Figure 1.18). Furthermore, this structure suggests that multiple hexamers may work together to perform mechanical work (Figure 1.18).

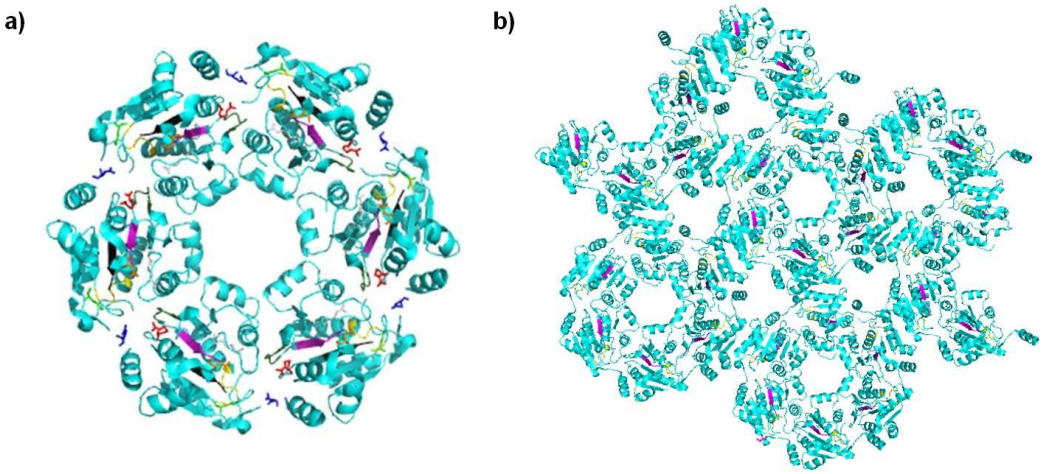

Figure 1.18 PhoH-like ATPase hexamer and symmetry mate protein structures
a) Hexamer of PhoH protein from *C. glutamicum* (*cg2513*) showing the positions of the conserved residues/motifs Q (green), Walker A (yellow), R residue (light grey), RGRTL (dark green), Walker B (magenta), N residue (SI) (orange), R residue (SRH) (red), III (black) and R residue (SII) (blue). b) Symmetry mates of hexamer conformation showing the possible interaction of multiple hexamers, the Walker A and B motifs are coloured yellow and magenta respectively. Images were made in PyMOL using PDB 3B85.

Overall, PhoH proteins share a number of conserved motifs and residues with the other helicase families, as well as two motifs that are unique to PhoH: an arginine residue (position 56 in HMM and position 171 in 3B85) and conserved sequence RGRTL located between β3 and β4. Like SF1/2 helicases, PhoH has core β strand topology as well as a motif similar to motif III of SF1. They also have the defining features of AAA+ proteins (SI, SII) and the SRH in a position similar to AAA proteins. Furthermore, PhoH has a Q motif characteristic of DEAD box RNA helicases and many members of SF1/2. Based on these observations, it is likely that PhoH proteins will possess helicase activity.

1.4.3.1 phoH *from* E. coli

The gene for *phoH* in *E. coli* is identical to *psiH*, which has been identified as a phosphate starvation-inducible gene (Metcalf et al. 1990). In *E. coli*,

promoters of Pho regulon genes share a consensus sequence known as the *pho* box: CTGTCATA (A/T) A (A/T)CTGT(C/A) A (C/T) (Kim et al. 1993). In *E. coli*, this sequence is located ~150 bp upstream of the translational start site. The *phoH* gene in *E. coli* has two reported transcriptional start sites: the first site includes a *pho* box, and PhoB, the transcriptional activator of the *pho* regulon in *E. coli*, was shown to bind to the *pho* box of the *phoH* gene under phosphate starvation conditions (Kim et al. 1993). Transcription from the second site was constitutive and was independent of *phoB*. Further, Kim et al. (1993) have also demonstrated experimentally ATP binding by PhoH from *E. coli*, suggesting PhoH has ATPase characteristics. Whether this activity was dependent on a secondary nucleic acid substrate remains to be determined.

It has been speculated that PhoH-like ATPases are likely to play a role in RNA metabolism due to the addition of RNA binding PIN and KH domains to some PhoH family members. They may also demonstrate helicase function or be involved with ATP-dependent activities with uncharacterised ribonucleoprotein (RNP) complexes in bacteria (Anantharaman et al. 2002). It is also likely that PhoH proteins are involved in the cellular response to phosphate starvation.

1.5 *phoH2* from *C. glutamicum* and the phosphate starvation response in mycobacteria

In *C. glutamicum* the *NCgl0065* (*phoH2*) gene was identified as one of a subset of genes that shows elevated expression after a shift from phosphate-sufficient to phosphate-limiting conditions (Ishige et al. 2003). In a further study, none of these prominent phosphate starvation genes previously identified including *cg0085 phoH1* (homologous to *NCgl0065*) were induced by a shift from phosphate-sufficient to phosphate-limiting conditions in a Δ*phoRS* knockout strain. This indicates that two-component regulatory system PhoRS controls transcriptional activation of these genes and that it is involved with the adaptation of *C. glutamicum* to phosphate starvation (Kocan et al. 2006). These findings raise the possibility that *phoH2*

genes in other organisms are under the control of PhoRS or other phosphate response systems.

1.5.1 PhoPR in *M. tuberculosis*

The PhoPR two-component system has been described as essential for virulence and pathogenesis of *M. tuberculosis* (Ryndak et al. 2008). In *M. tuberculosis*, the PhoPR two-component system has a variety of functions including involvement with lipid metabolism and secretion, and the initial and enduring hypoxic response (Gonzalo-Asensio et al. 2008; Ryndak et al. 2008). Unlike PhoRS in *C. glutamicaum*, PhoPR has not been reported to be involved with the phosphate starvation response. When global gene expression was compared between wild type and a PhoP mutant strain grown in 7H9 medium, the *phoH2* gene (*Rv1095*) did not show any differential expression, suggesting that it is not under the direct control of the PhoPR system in *M. tuberculosis* (Walters et al. 2006). A further study to confirm the regulation of DosR/DevR (dormancy survival regulon) by PhoP compared expression between a PhoP mutant strain and wild type, DosR expression was reduced in the mutant strain which was reversed when complemented with PhoP, to wild type expression levels (Gonzalo-Asensio et al. 2008). PhoP is suggested to be a link between early and enduring hypoxic responses of *M. tuberculosis*.

Given the observations with PhoP, this raises questions as to the regulation of the *phoH2* gene and protein in *M. tuberculosis*.

1.5.2 Phosphate starvation response and SenX3-RegX3 in mycobacteria

In *M. tuberculosis*, the two-component regulatory system SenX3-RegX3 controls expression of the phosphate specific transport operon *pstS3-pstC2-pstA1* which is induced under phosphate starvation (Rifat et al. 2009). It has been identified that SenX3 contains a PAS-like domain. These domains are found in a number of prokaryotes and eukaryotes, and function as sensors of oxygen and redox state (Rickman et al. 2004). It was later found that SenX3-RegX3, in addition to being responsible for expression of

phosphate-dependent genes, has a role in modulating *cydB* and *gltA1* genes involved with aerobic respiration (Parish et al. 2003b; Roberts et al. 2011). In relation to *phoH2*, Roberts et al. (2011) have shown that there is no difference in the levels of expression of *phoH2 (Rv1095)* with a RegX3 mutant compared with wild type expression under static and hypoxic growth conditions, where other genes showed differential expression. This suggests that *phoH2* is not under the direct control of the SenX3-RegX3 two-component system in its response to hypoxia in *M. tuberculosis*. Similarly, in the study of Rifat et al. (2009) which looked at phosphate dependent gene expression under phosphate limiting conditions using microarrays, *phoH2* was not listed as having differentially regulated gene expression at 24 and 72 hours after phosphate starvation. The complete list of genes, up or down regulated under these growth conditions was not provided therefore these results do not rule out a role of *phoH2* in the response to phosphate starvation in *M. tuberculosis*.

In *M. smegmatis*, two systems are responsible for assimilation of phosphate under nutrient stress: the high affinity ABC-transport system *pst* (phosphate specific transporter) and *phn* system, both of which have a similar affinity for phosphate (Gebhard et al. 2006). Within the *phn* operon, an additional gene coding for a GntR-like transcriptional regulator, *phnF*, has been identified. This gene was found to act as a repressor of the *phn* operon but did not affect the expression of the *pst* system (Gebhard & Cook 2008). Additional expression analyses also showed that the *phnF* and *phn* operon are regulated by the SenX3-RegX3 regulon. In *M. smegmatis*, this two-component regulatory system is responsible for expression of phosphate-dependent genes (Glover et al. 2007). Berney and Cook (2010) also showed with microarray data that the SenX3-RegX3 two-component regulatory system is upregulated under hypoxic growth conditions.

In summary, the precise role and regulation of PhoH2 proteins is unclear. It is plausible that they are involved in the response to phosphate starvation and may be linked with growth under hypoxic conditions.

1.6 Toxin-Antitoxin systems

Toxin-antitoxin (TA) loci that contain PIN-domain proteins are found in a diverse range of prokaryotes (Gerdes & Wagner 2007). TA systems were first identified in *E. coli* where they were shown to be involved in plasmid maintenance post-segregation (Gerdes et al. 1986). TA systems have also been found on bacterial chromosomes, suggesting additional biological roles (Buts et al. 2005). These systems are reported to be involved with antibiotic resistance and tolerance, virulence, pathogenicity islands, bacterial persistence and pathogen trafficking (Gerdes 2000; Hayes & Sauer 2003; Hayes 2003; Mattison et al. 2006)

TA systems are classified based on the nature of their antitoxin. In type I TA systems, the antitoxin is a small RNA (sRNA) that base pairs with the mRNA of the toxin gene. This base pairing prevents synthesis of the toxin protein (Fozo et al. 2010) (Figure 1.19a). In type II TA systems, the antitoxin is a protein which binds to the toxin protein forming a benign complex and in doing so offsets the toxic effects of the toxin protein (Fozo et al. 2010) (Figure 1.19b). More recently, a third RNA-protein TA system has been reported, ToxIN, in which a 36 nucleotide motif found in the RNA antitoxin encoded by the *tox*I gene inhibits the toxic effect of the protein encoded by *tox*N post-protein expression (Fineran et al. 2009) (Figure 1.19c).

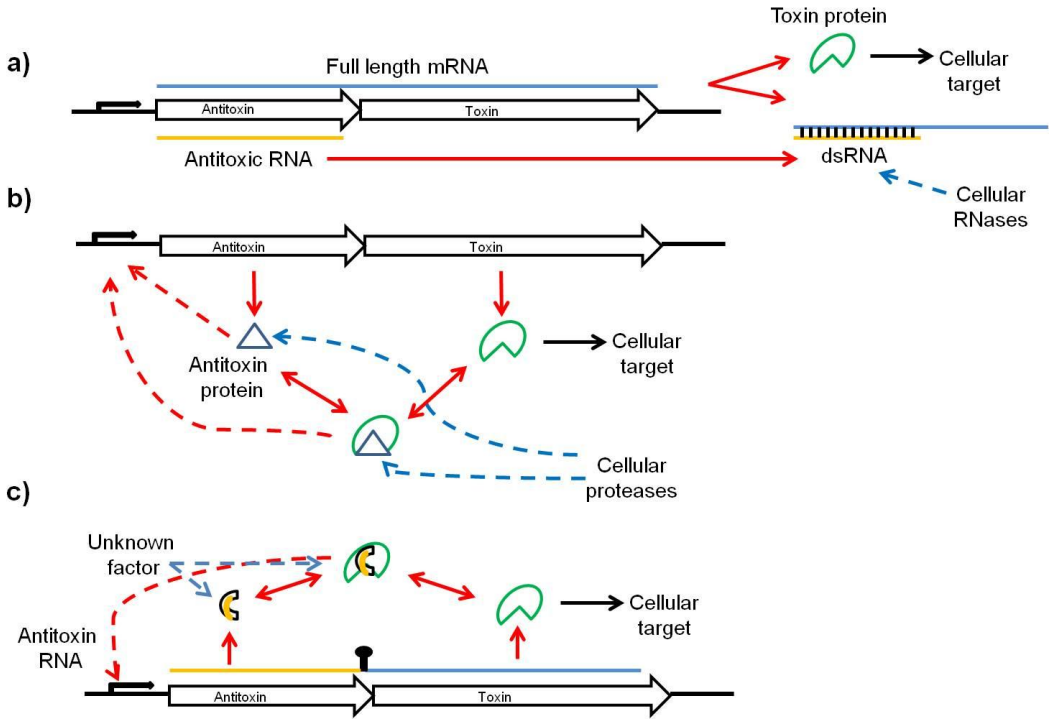

Figure 1.19 TA systems
TA systems are organised in an operon where the antitoxin gene precedes the toxin gene. a) In type I TA systems the sRNA antitoxin base pairs with the mRNA of the toxin gene suppressing translation and causes the resulting duplex mRNA to be susceptible to degradation by cellular RNases. b) In type II TA systems the antitoxin protein binds to the toxin protein forming a benign protein complex which binds back to the promoter region in so regulating transcription of the operon. The antitoxin protein is susceptible to cellular proteases which cause release of the toxin. c) In type III TA systems the antitoxic RNA binds to the toxin protein forming a benign RNA-protein complex. A transcriptional terminator located between the two genes regulates the expression levels of *tox*I and *tox*N. To regulate transcription of the operon the complex binds to the promoter region. Unknown cellular factors promote the activation of the toxin through degradation of the antitoxin. Adapted from (Blower et al. 2011).

Generally, type I toxin proteins are smaller and more hydrophobic than the toxin proteins of type II TA systems (Buts et al. 2005). Each toxin protein of the various TA systems has different biochemical activities; the MazF and VapC (PIN-domain) toxins, for example, are sequence specific RNases that have been shown to target mRNA and to inhibit translation (Inouye 2006; Robson et al. 2009; McKenzie et al. 2012b).

1.6.1 VapBC TA systems

The most widespread TA system in bacteria and archaea is the VapBC TA family. In many unrelated organisms there has been an expansion in the number of *vapBC* operons. For example, in the archaeon *Pyrococcus kodakaraensis*, there are 29 annotated *vapBC* operons and 42 are annotated

in *Sinorhizobium meliloti* (Sevin & Barloy-Hubler 2007; Makarova et al. 2009). Within mycobacteria, the expansion of the VapBC family has occurred most considerably in species that are pathogenic to mammals (McKenzie et al. 2012b). Of the 48 PIN-domain proteins in *M. tuberculosis*, 47 are annotated as putative VapBC TA systems (Arcus et al. 2005; Ramage et al. 2009; Arcus et al. 2011).

Several VapBC TA systems have been biochemically characterised including *FitAB* from *Neisseria gonorrhoeae*, *NtrPR* from *S. meliloti* and *VapBCs* from mycobacteria, *Haemophilus influenzae, Leptospira interrogans,* enterobacteria, *S. solfataricus* and two VapC proteins from *P. aerophilum* (Puskas et al. 2004; Zhang et al. 2004; Bodogai et al. 2006; Mattison et al. 2006; Daines et al. 2007; Cooper et al. 2009; Robson et al. 2009; Winther & Gerdes 2009; McKenzie et al. 2012a; McKenzie et al. 2012b). Each VapC toxin belongs to the PIN-domain family and exerts its effect via RNase activity. NtrR is predicted to have RNase activity due to the conservation of the four acidic residues characteristic of PIN-domains for enzymatic activity (Arcus et al. 2011).

VapC proteins have been implicated in the control of bacterial growth within intracellular environments (Hopper et al. 2000), growth inhibition of *E. coli* (Daines et al. 2007) and disruption of translation in *M. smegmatis* (Robson et al. 2009). Biochemically, VapC proteins target RNA, for example, VapC-1 from *H. influenzae* cleaves total RNA (Daines et al. 2007), and VapCs from *Shigella flexneri* 2a and *Salmonella enterica* cleave formylmethionine tRNA ($tRNA^{fMET}$) (Winther & Gerdes 2011). VapC from *M. smegmatis* shows sequence specific endoribonuclease activity, cleaving the ssRNA sites AUAU and AUAA both *in vitro* and *in vivo* (McKenzie et al. 2012b). The VapC proteins $VapC_{Rv0065}$ and $VapC_{Rv0617}$ from *M. tuberculosis*, like VapC from *M. smegmatis* also cleave RNA at specific sites (Ahidjo et al. 2011). VapC protein from *vapBC* operon *Rv0596c-Rv0595c* specifically targets ssRNA at ACGC and AC(A/U)GC (McKenzie et al. 2012a; Sharp et al. 2012), and the VapC proteins $VapC_{Rv0065}$ and $VapC_{Rv0617}$ target (G/C)G(G/C)(G/C/A) sequences (McKenzie et al. 2012a). Data suggest that within *M. smegmatis*, VapC plays a role in the rate of

glycerol utilisation, which is coupled to growth in order to match anabolic and catabolic demands of the cell. This allows for post-transcription tuning of catabolic metabolism (McKenzie et al. 2012b).

It is evident that VapC PIN-domain proteins are involved in the regulation of cellular metabolism by virtue of their ribonuclease activity. Thus far, all characterised prokaryotic PIN-domain proteins have been part of TA systems, whereas the other classes of PIN-domain proteins remain unexamined.

1.7 Aims and objectives

At the beginning of this project, a bioinformatics survey (unpublished) indicated an association between VapBC TA systems in bacterial genomes and AAA+ ATPase proteins. A BLAST search for AAA+ ATPases or similar in *M. tuberculosis* at this time revealed PhoH2 (*Rv1095*). What was striking about PhoH2 was that it was an example of a PIN-domain protein fusion with a PhoH-domain, and this PIN-domain protein accounted for the one remaining PIN-domain in the *M. tuberculosis* genome not associated with a toxin-antitoxin operon. The biological function of PhoH2 from *M. tuberculosis* and *M. smegmatis* and of PhoH2 proteins in general at the start of this project was unknown.

The overall aim of this research was to investigate the biological function of PhoH2 proteins by examining PhoH2 from *M. tuberculosis*. The characterisation of PhoH2 was carried out based on three main objectives as outlined below.

Objective 1 – Investigate the transcriptional context of *phoH2* genes from *M. tuberculosis* and *M. smegmatis* and the biological effect of PhoH2 protein expression.

Objective 2 – Clone, express, purify, crystallise and solve the structure of PhoH2 proteins from *M. tuberculosis* and *M. smegmatis*.

Objective 3 – Determine the biochemical activity of PhoH2 proteins from *M. tuberculosis* and *M. smegmatis*.

Chapter Two: Materials and Methods

Details of all media and bacterial strains used in this study can be found in Appendix A.

2.1 DNA and RNA manipulations

M. bovis BCG Pasteur 1173P2 was kindly provided by Dr Ray Cursons, and *M. smegmatis* mc²155 from Dr Joanna McKenzie for culture.

2.1.1 *M. bovis* BCG and *M. smegmatis* Genomic DNA Extraction

M. bovis BCG strain Pasteur 1173P2 and *M. smegmatis* mc²155 were cultured in LB Tween-80 broth (LBT) for a minimum 18-21 days or 2-3 days respectively. Cells were added with 1/10 the volume of 5 M guanidium isothiocynate (GITC) pH 7 (usually 1 ml cells and 100 µl 5 M GITC) to a 2 ml screw cap tube containing approximately 0.3 g of 0.1 mm and 2.5 mm zirconia beads. Cells were ruptured with three rounds of bead beating using a FastPrep cell disrupter (FP120 Thermo Savant) for 30 seconds (setting 6), left to cool for 1 minute and the process repeated twice more. Fifty µl of 2 M Na acetate pH 4, and an equal volume of phenol/chloroform (made fresh) was added and the samples were shaken well. For deproteinisation, samples were rotated for 10 minutes at RT. The samples were then centrifuged at 13 000 rpm at RT for 1 minute to separate the phases. The top layer was transferred to a new tube, and an equal volume of isopropanol (IPA) was added. This was incubated for 10-20 minutes at RT. After incubation samples were centrifuged at 13 000 rpm at RT for 20 minutes, the supernatant was removed and discarded, and the pellet washed with 1 ml 70 % ethanol. All traces of ethanol were removed and the pellet was air dried at RT. Once dried the pellet was resuspended in 50 µl TE buffer (10 mM Tris HCl pH 8, 1 mM EDTA).

2.1.2 *E. coli* Vector DNA Extraction

E. coli vector DNA was extracted from 5 ml overnight cultures using the QIAprep Spin Miniprep Kit (Qiagen, Netherlands) according to manufacturer's instructions. Vector DNA was eluted in 50 µl of elution buffer.

2.1.3 Agarose Gel Electrophoresis

DNA fragments were separated using agarose gel electrophoresis. The percent of agarose used in the gel depended on the size of DNA; as a general rule samples <200 bp were run on a 2 % gel, 200-400 bp 1.5 %, 400-1000 bp 1 % and >1000 bp 0.8 % agarose in 1x TAE buffer (40 mM Tris-acetate, 20 mM EDTA). Samples were mixed with 5 x DNA loading dye (25 % glycerol, 0.2 % bromophenol blue) prior to loading onto the gel. Agarose gels were either stained with 0.5 µg/ml ethidium bromide or 1 x SYBR Safe™ DNA gel stain (Invitrogen, USA). Gels were visualised by UV light (ethidium bromide) or on a blue light box (Invitrogen, USA) and images captured. Band sizes were determined by comparison with the 1kb-Plus DNA ladder (Invitrogen, USA).

2.1.4 DNA Quantification

DNA was quantified using a Nanodrop ND-1000 Spectrophotometer (Nanodrop Technologies, USA). This measures absorbance of DNA at 260 nm, quantifying the amount of DNA.

2.1.5 Polymerase Chain Reaction (PCR)

2.1.5.1 Primers

All primers were a minimum of 16 bp in length. Primers used for ligase-based cloning were designed with 5' and 3' flanking restriction enzyme sites that included an extra 2-4 bases to allow for efficient cleavage (the number of bases included was dependent on the restriction enzymes used). Primers for Gateway cloning were designed as recommended in the manufacturer's instructions. All primers were analysed using either Vector Nti (Invitrogen, USA) or Geneious Pro (Version 5.1.7) (BioMatters Ltd, NZ) to check for secondary structure such as primer dimers or hairpins. Primers were supplied by Invitrogen (USA), Sigma (USA) or IDT (USA) and were dissolved in 1 x TE buffer to a 100 µM concentration.

2.1.5.2 PCR for RT-PCR and amplification of vector inserts

The *Taq* or Platinum *Pfx*® DNA polymerase system (Invitrogen, USA) was used for RT-PCR and for the amplification of gene inserts. For each set of primers a gradient PCR was carried out to determine the optimal Tm for the

PCR reaction. Five °C above and below the calculated Tm of the primers was selected as the upper and lower limits of the gradient. A Tm that resulted in a product of the expected size was selected for use in PCR for amplification. The reactions were carried out in 15 or 25 µl volumes. The following concentrations and reaction conditions were used:

Platinum *Pfx*® PCR	***Taq* PCR**
1 x *Pfx* amplification buffer	1 x PCR buffer
0.3 mM deoxynucleotide mix (dATP, dCTP, dTTP & dGTP)	0.2 mM deoxynucleotide mix (dATP, dCTP, dTTP & dGTP)
1 mM $MgSO_4$	1.5 mM $MgCl_2$
1-2 U Platinum *Pfx*® DNA polymerase	2.5 U *Taq* DNA Polymerase
0.2 µM each primer	0.2 µM each primer
20-120 ng template DNA	20-120 ng template DNA
Cycling conditions (min:sec) x29 cycles	
95 °C 2:00 (initial denaturation)	95 °C 2:00 (initial denaturation)
95 °C 0:15	95 °C 0:15
Tm* 0:30	Tm* 0:30
68 °C 0:45**	72 °C 0:45**
68 °C 7:00 (final extension)	72 °C 5:00 (final extension)

*Initial denaturation was carried out for 2 minutes

*Tm determined by gradient PCR

**Elongation time dependent of the insert size, as a rule 1 kb = 1 minute

PCR reactions were performed and visualised as in section 2.1.3.

2.1.5.3 Colony PCR

The *Taq* DNA polymerase system (Invitrogen, USA) was used to test positive transformants for the correct insertion of vector insert DNA. Positive transformant colonies were resuspended in a dilute antibiotic solution and used as DNA template for PCR using T7 or M13 F and R primers (Table 2.1) and the method described in section 2.1.5.3. For the T7 primers, a Tm of 52.5 °C was used and for the M13 primers a Tm of 55 °C.

Table 2.1 Primers used for colony PCR

Primer		Primer Sequence
F	T7	TAATACGACTCACTATAGGG
	M13	TGTAAAACGACGGCCAGT
R	T7	TAGTTATTGCTCAGCGGTGG
	M13	CAGGAAACAGCTATGACC

2.1.5.4 PCR for difficult templates

For the amplification of difficult DNA templates the KAPA HiFi™ HotStart DNA polymerase was used using kit recommended reaction conditions and cycling parameters.

KAPA HiFi™ HotStart PCR
1 x KAPA HiFi GC buffer (contained 2 mM Mg^{2+})
0.3 mM deoxynucleotide mix
(dATP, dCTP, dTTP & dGTP)
5 % DMSO
0.5 U KAPA HiFi HotStart DNA polymerase
0.3 μM each primer
20-120 ng template DNA
Cycling conditions (min:sec) x29 cycles
95 °C 2:00 (initial denaturation)
98 °C 0:20
65 °C 0:15
72 °C 0:45
72 °C 5:00 (final extension)

PCR reactions were performed and visualised as in section 2.1.3.

2.1.6 Purification of PCR Products and other reactions from Solution

PCR or enzymatic reactions were pooled and purified using a High Pure PCR Product Purification Kit (Roche Applied Science, Switzerland) or a Qiaquick PCR Product Purification Kit (Qiagen, Netherlands) according to manufacturer's instructions.

2.1.7 Purification of PCR Products from agarose gels

PCR reactions of the desired size were excised and purified using a High Pure PCR Product Purification Kit (Roche Applied Science, Switzerland) or a Qiagen Gel Purification Kit (Qiagen, Netherlands) according to manufacturer's instructions.

2.1.8 Restriction Enzyme Digest

Restriction enzymes were purchased from Invitrogen (USA) and Roche Applied Science (Switzerland). Digestion of DNA with restriction enzymes was carried out according to manufacturer's instructions. Buffers were chosen based on manufacturer's recommendations to maximise the activity of the restriction enzyme and reactions were incubated for 1 to 4 hours at the appropriate temperature. Double digestions (two restriction enzymes in one reaction) were used if the reaction buffer for each enzyme was the same. If not, one reaction with one restriction enzyme was performed then the DNA purified according to method in section 2.1.6 and the second reaction with the other restriction enzyme performed. Digested DNA was purified from solution according to method in section 2.1.6.

2.1.9 DNA Ligation

DNA ligations were performed according to manufacturer's recommendations using 5 U of T4 DNA Ligase (Invitrogen, USA) in 20 µl reactions and incubated for 6 hours at room temperature or at 18 °C overnight.

2.1.10 DNA Transformation

2.1.10.1 Preparation of Electrocompetent E. coli

A glycerol stock of the desired strain of *E. coli* was streaked onto an LB agar plate and incubated at 37 °C overnight. A single colony was then used to start a 5 ml LB culture which was grown at 37 °C in a shaking incubator at 200 rpm overnight. The overnight culture was used to seed a 500 ml LB culture, which was grown at 37 °C in a shaking incubator at 200 rpm until an OD_{600} of 0.5-0.7 was reached. The cells were then chilled on ice for at least 20 minutes before being transferred to a pre-chilled sterile centrifuge bottle and centrifuged at 4000 g at 4 °C for 15 minutes. The supernatant was discarded and the cell pellet resuspended in 500 ml ice cold sterile 10 % glycerol. The cells were centrifuged again at 4000 g at 4 °C for 15 minutes and the supernatant discarded. The cell pellet was resuspended in 250 ml ice cold sterile 10 % glycerol and centrifuged at 4000 g at 4 °C for 15 minutes. The supernatant was discarded and the cell pellet resuspended in 20 ml ice cold

sterile 10 % glycerol and transferred to a pre-chilled 50 ml falcon tube and centrifuged at 4000 g at 4 °C for 15 minutes. The supernatant was discarded and the pellet resuspended in 1 to 2 ml ice cold sterile 10 % glycerol. Fifty µl aliquots of resuspended cells were then flash frozen in liquid nitrogen and stored at -80 °C.

2.1.10.2 Electroporation of E. coli

One - two µl of vector DNA or ligation reaction was added to 50 µl freshly thawed (on ice) electrocompetent *E. coli* cells of the desired strain. The mixture was placed in a 2 mm electroporation cuvette (BioRad Laboratories, USA), tapped to settle the mixture between the electrodes and electroporated with a Bio-Rad Gene Pulser™ (Bio-Rad Laboratories, USA) at 2.5 kV, 25 µF capacitance and 200 Ω resistance. SOC medium (1 ml) was added immediately to the electroporated cells and incubated at 37 °C shaking at 300 rpm for 30 minutes. An aliquot of the cells was plated on LB agar plates containing the appropriate antibiotic for selection and incubated at 37 °C overnight. Colonies from these agar plates were then used to start 5 ml LB broth cultures containing the appropriate antibiotic. These cultures were incubated at 37 °C shaking at 200 rpm overnight and used to isolate vector DNA according to the method in section 2.1.2.

2.1.10.3 Preparation and transformation of chemically competent E. coli *(Gateway®)*

One colony of DH5α cells cultured overnight was grown in 5 ml of LB medium at 37 °C at 200 rpm for 3.5 hours. After incubation 1.5 ml of cells were harvested by centrifuging at 13 000 rpm for 1 minute. The supernatant was removed and the cells were resuspended in 100 µl of 0.1 M $CaCl_2$, 1% PEG 8000 solution and incubated on ice for 5 minutes. Ten µl of vector-insert was added and the suspension incubated further on ice for 20 minutes with periodic gentle mixing. The suspension was then heat shocked at 42 °C for 45 seconds, 900 µl of LB medium was added and the heat shocked suspension was incubated at 37 °C for 1 hour. The cells were harvested by centrifuging at 13 000 rpm for 1 minute. The supernatant was removed and the cells were resuspended in 100 µl of LB medium and spread onto an LB agar plate

supplemented with 50 µg/ml kanamycin ($pDONR_{221}$) or 100 µg/ml ampicillin ($pDEST_{17}$) for incubation at 37 °C overnight.

2.1.10.4 Preparation of Electrocompetent M. smegmatis

A glycerol stock of the desired *M. smegmatis* strain was streaked onto an LBT agar plate and incubated at 37 °C overnight. A single colony from this plate was used to start a 5 ml LBT culture which was grown overnight at 37 °C in a shaking incubator at 200 rpm until an OD_{600} of around 0.7 was reached. One ml of the overnight culture was used to seed 100 ml of 7H9/ADC + 0.05 % (v/v) Tween-80 medium. The 100 ml culture was grown at 37 °C in a shaking incubator at 200 rpm until an OD_{600} of 0.5-0.7 was reached. The cells were then chilled on ice for 1.5 hours before being transferred to a pre-chilled sterile centrifuge bottle and centrifuged at 4000 g at 4 °C for 20 minutes. The supernatant was discarded immediately as mycobacterial cell pellets are delicate and likely to break up. The cell pellet was resuspended in 100 ml ice cold sterile 10 % glycerol. The cells were centrifuged again at 4000 g at 4 °C for 20 minutes and most of the supernatant was discarded leaving a small volume of supernatant to resuspend the cells in. The cell pellet was resuspended in the remaining supernatant, transferred to a sterile 50 ml falcon tube and centrifuged at 4000 g at 4 °C for 20 minutes. The supernatant was discarded and the cell pellet resuspended in 0.2 ml ice cold sterile 10 % glycerol. Forty µl aliquots of resuspended cells were then flash frozen in liquid nitrogen and stored at -80 °C.

2.1.10.5 Electroporation of M. smegmatis

Electroporation of *M. smegmatis* required adjustment of transformation parameters. One µl of plasmid DNA or ligation reaction was added to 40 µl freshly thawed (on ice) electrocompetent *M. smegmatis* cells of the desired strain. The mixture was placed in a 2 mm electroporation cuvette (Bio-Rad Laboratories, USA), tapped to settle the mixture between the electrodes and electroporated with a Bio-Rad Gene Pulser™ (Bio-Rad Laboratories, USA) at 2.5 kV, 25 µF capacitance and 1000 Ω resistance. One ml of 7H9/ADC + 0.05 % (v/v) Tween-80 medium was added immediately to the electroporated cells and incubated at 37 °C shaking at 200 rpm for 3 hours.

An aliquot of the cells was plated on 7H10/ADC + 0.05 % (v/v) Tween-80 agar plates containing the appropriate antibiotic for selection and incubated at 37 °C for 3 days.

2.1.10.6 Glycerol Stocks

Glycerol stocks for the long term storage of transformed bacteria were made by the addition of 0.5 ml of overnight culture (LB (*E. coli*) or LBT (*M. smegmatis*) + appropriate antibiotic) to 0.5 ml of sterile 50 % glycerol. Glycerol stocks were stored at -80 °C.

2.1.11 Isolation of Total RNA from *M. bovis* BCG and *M. smegmatis*

Method developed by Ali Ruthe

Proteins & Microbes Laboratory, University of Waikato

For the preparation of total RNA, *M. bovis* BCG strain Pasteur 1173P2 and *M. smegmatis* mc^{2}155 were cultured in LBT broth for a minimum of 21 days or 2-3 days at 37 °C respectively. Cells were immediately added to four times the volume of 5 M GITC pH 7 (usually 2 ml cells with 8 ml 5 M GITC) and pelleted by centrifugation. The supernatant was removed and cells resuspended in 0.5 ml of fresh 5 M GITC and transferred to a 2 ml screw cap tube containing approximately 0.3 g of 0.1 mm and 2.5 mm zirconia beads. The cells were ruptured with four rounds of bead beating in a Fast Prep cell disrupter (FP120 Thermo Savant) for 20 seconds (setting 6.5), left to cool for one minute and the process repeated for increasing periods (i.e. 20, 25, 30 and 35 seconds). Samples were centrifuged briefly at 4 °C to remove foam. *One tenth of the volume (50 µl) of 2 M Na acetate pH 4 was added and mixed gently. An equal volume (0.5 ml) of water-saturated phenol was added, vortexed then rotated for 10 minutes at RT. One hundred µl of 1-bromo, 3-chloro propanate was added, shaken for one minute and incubated on ice for 5 minutes. Samples were centrifuged to separate the phases and the top layer removed and the method repeated from the * above with volumes decreased accordingly. The top layer was removed and was added to an equal volume of isopropanol and chilled on ice for 30 minutes. The precipitated RNA was then centrifuged at 13 000 g at 4 °C for 30 minutes, 1 ml 70 % ethanol added, centrifuged at 13 000 g at 4 °C for 15 minutes,

supernatant removed, 1 ml 100 % ethanol added, centrifuged at 13 000 g at 4 °C for 15 minutes, supernatant removed. The pellet was then centrifuged again briefly and any residual ethanol removed. The pellet was resuspended in 25 µl of 10 mM Tris-HCl pH 7, 0.5 mM $MnCl_2$. RNA concentrations were determined using a Nanodrop ND-1000 spectrophotometer (Nanodrop Technologies, USA).

2.2 Protein Purification and Manipulation

2.2.1 SDS-Polyacrylamide Gel Electrophoresis (SDS-PAGE) Protein Analysis

SDS-PAGE gels were cast in a Hoeffer gel casting system. SDS-PAGE gels were made up of a resolving gel (10, 12, 15 or 18 % acrylamide depending on protein size), that was overlaid with a stacking gel (5 % acrylamide). All SDS-PAGE gels were made up with 30 % acrylamide with an acrylamide:bisacrylamide ratio of 37:5:1 (Bio-Rad Laboratories, USA) and included 0.1 % (w/v) SDS, and were polymerised by the addition of 0.05 % (w/v) ammonium persulfate (APS) and 0.05 % (v/v) TEMED.

Protein samples were mixed in a 3:1 ratio with 4 x SDS loading buffer (250 mM Tris HCl pH 6.8, 20 % glycerol, 4 % SDS, 10 % β-mercaptoethanol, 0.025 % (w/v) bromophenol blue) and heated to 95 °C for 5 minutes before loading onto gel. Gels were run in 1 x SDS-PAGE running buffer (25 mM Tris, 250 mM glycine, 0.1 % (w/v) SDS) at 70 V until the dye front entered the resolving gel and then at 150 V until the dye front reached the end of the gel.

2.2.2 Native Polyacrylamide Gel Electrophoresis (Native-PAGE) Protein Analysis

Native-PAGE gels were cast and run as for SDS-PAGE gels as in section 2.2.1, except without SDS in the gel, running buffer (25 mM Tris, 250 mM glycine) or loading dye (300 mM Tris HCl pH 6.8, 50 % glycerol, 0.05 % (w/v) bromophenol blue).

2.2.3 Coomassie Blue Staining for Protein Gel Electrophoresis

Gels were stained using the quick coomassie blue staining method (Wong et al. 2000) unless otherwise stated. Gels were placed in a microwaveable box with 50 ml Fairbanks staining solution A (0.05 % (w/v) coomassie blue R-250, 25 % (v/v) isopropanol, 10 % (v/v) acetic acid), microwaved for 30 seconds on full power, then cooled to RT with gentle shaking. The stain was then removed and Fairbanks staining solution B (0.005% (w/v) coomassie blue R-250, 10% (v/v) isopropanol, 10% (v/v) acetic acid) was added and microwaved as above. This was then repeated for Fairbanks staining solutions C & D (C – 0.002 % (w/v) coomassie blue, 10 % (v/v) acetic acid and D - 10 % (v/v) acetic acid), solution D being the destaining solution. Protein sizes were estimated by comparison with the protein sizes of the Precision Plus Protein™ Unstained Standard (Bio-Rad Laboratories, USA).

2.2.4 Measurement of Protein Concentration

Protein concentration was measured using the Biorad Protein Assay (Biorad Technologies, USA), and/or Nanodrop ND-1000 spectrophotometer (Nanodrop Technologies, USA). The Biorad Protein Assay is based on the Bradford Assay and was performed according to manufacturer's instructions. Comparison of samples with a standard curve provides a relative measurement of protein concentration. The Nanodrop ND-1000 measures absorbance at 280 nm and the accompanying software calculates protein concentration using the Beer-Lambert equation: $A=\varepsilon.c.l$

A = absorbance at 280 nm

ε = theoretical molar extinction coefficient ($M^{-1}cm^{-1}$)

c = concentration (M)

l = pathlength (cm)

The theoretical molar extinction coefficients were calculated using the online tool ProtParam (http://web.expasy.org/protparam), by providing the amino acid sequence.

2.2.5 Concentration of Protein Samples

Protein samples were concentrated using 20 ml, 2 ml or 500 µl Vivaspin concentrators (Sartorius AG, Germany) with molecular weight cut offs of 10 kDa. Concentrators were pre-equilibrated with 0.05 % (v/v) Tween-80 and purification buffer. Protein samples were added to the upper reservoir of a pre-equilibrated concentrator and spun at 3 000 g (20 ml), 3 600 g (2 ml) or 9 000 g (500 µl) at 4 °C until the desired volume or concentration was achieved.

2.2.6 Dialysis of Protein Samples

Protein samples were dialysed using Spectra Por® dialysis tubing of 6-8 kDa molecular weight cut off (Spectrum Laboratories, USA). The dialysis tubing was pre-wet with target buffer, the protein sample was then added to the tubing and the ends of the tubing sealed with dialysis clips. The protein sample was then left to dialyse against 1 L of target buffer with gentle stirring for 2-3 hours at 4 °C. The target buffer was then replaced with fresh buffer and the protein dialysed with gentle stirring overnight at 4 °C.

2.2.7 Purification of His-tagged Proteins using IMAC

The supernatant from lysed large scale expression cultures was filtered through 1.2, 0.45 and 0.2 µm Minisart filters (Sartorius AG, Germany). If the supernatant was very viscous as was generally the case with *M. smegmatis* expression cultures, it was filtered only through 1.2 and 0.45 µm Minisart filters. The filtered supernatant was then loaded onto an IMAC column, which was pre-equilibrated with the appropriate lysis buffer at a flow rate of 1-2 ml/min. The column was then attached to either a ÄKTA Prime™, ÄKTA Basic™ or ÄKTA Purifier™ FPLC system. The column was washed with 15-25 ml of lysis buffer, at a flow rate of 1 ml/min to remove unbound proteins. Bound proteins were removed by running a gradient of 0-50 % elution buffer (lysis buffer + 1 M imidazole) over 50 ml at a flow rate of 1 ml/min. Fractions containing the desired protein were then analysed by SDS-PAGE as in section 2.2.1.

2.2.8 Size Exclusion Chromatography

Size exclusion chromatography was used as required to further purify proteins. Size exclusion chromatography was performed using either a HiLoad™ 16/60 Superdex™ or a 10/300 Superdex™ column (analytical size exclusion column) (GE Healthcare, UK). Protein solutions were concentrated to between 1 and 5 ml and filtered with a Nanosep® MF 0.2 µm filter (Pall, USA) before loading onto a pre-equilibrated column.

2.2.9 Calibration of the Size Exclusion Column S200 10/300

High and low molecular weight gel filtration kits (GE Healthcare, UK) were used to calibrate the S200 10/300 analytical size exclusion column. Briefly, all proteins were reconstituted to the recommended concentration with 50 mM phosphate buffer pH 7.4, 200 mM NaCl with the exception of carbonic anhydrase, which was resuspended in water as per kit instructions. Two protein mixes were made up: Protein Mix A contained ferritin, conalbumin, carbonic anhydrase and ribonuclease with final concentrations of 0.3 mg/ml for ferritin and 3 mg/ml for the remaining proteins; and Protein Mix B contained aldolase, ovalbumin, ribonuclease A and aprotinin with final concentrations of 4 mg/ml for aldolase and ovalbumin, and 3 mg/ml for ribonuclease A and aprotinin. Each protein mix was fractionated and the elution volume for each protein determined. The void volume of the column was determined by the elution volume of Blue Dextran 2000 (1 mg/ml).

A calibration curve was then determined for the column by calculation of the K_{av} (gel phase distribution coefficient) values for the calibration kit proteins using the equation:

$K_{av} = (V_e - V_o) / (V_c - V_o)$

(V_o = column void volume, V_e = elution volume, V_c = geometric column volume).

The calibration curve was made by plotting K_{av} against the log molecular weight and an equation determined to calculate the molecular weight of sample proteins.

2.3 Methods relating to Chapter 3

2.3.1 RT-PCR

For the preparation of total RNA from *M. bovis* BCG strain Pasteur 1173P2 and *M. smegmatis* mc^2155, RNA was extracted using the phenol-GITC method for the isolation of RNA as described in section 2.1.11.

To prepare cDNA, 1 µg of RNA was treated with 1 U of RQ1 RNase-Free DNase (Promega, USA) in the presence of 10 mM Tris-HCl pH 7, 0.5 mM $MnCl_2$ according to manufacturer's instructions. Briefly reactions were incubated at 37 °C with shaking at 600 rpm for 30 minutes. The DNase reaction was stopped with of 1.1 µl EGTA stop solution and samples heated to 65 °C for 10 minutes. Synthesis of cDNA was carried out using the Superscript III First-Strand Synthesis system (Invitrogen, USA) according to kit instructions with and without Superscript III to exclude DNA contamination (+/-RT). Random hexanucleotides (2 µl of 50 ng/µl) were added and incubated at 65 °C for 5 minutes then placed on ice immediately for 1 minute. Buffer 5x (4 µl), 0.1 M DTT (1 µl), 10 mM dNTP (1 µl) (-RT) and SSIII (0.75 µl of 200 U/µl) (+RT) were added and mixed gently. Reactions were incubated at 25 °C for 10 minutes then at 50 °C for 1 hour and 85 °C for 10 minutes. cDNA was stored at -20 °C.

M. bovis BCG and *M. smegmatis* cDNA (+ and – RT) were used as templates for PCR reactions with 6 primer combinations that spanned the upstream sequence of *Rv1095* and *MSMEG_5247*, including observed possible alternative translational start sites upstream of *phoH2*. Figure 2.1 illustrates the location of each RT primer set.

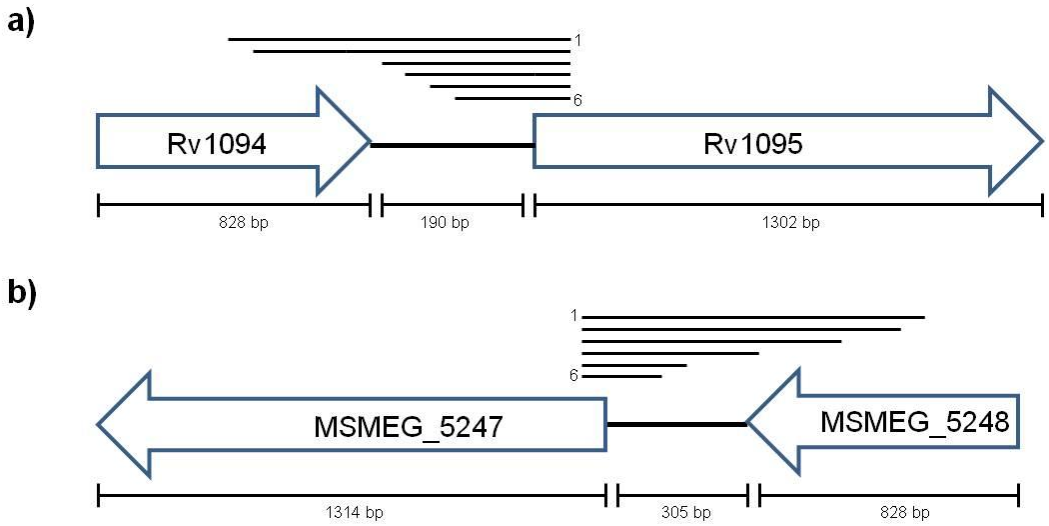

Figure 2.1 Genomic position of primers used in RT-PCR
Genomic location of each RT primer set for a) *M. tuberculosis* (*M. bovis* BCG) and b) *M. smegmatis*. RT F 1-6 primers were each used with RT R positioned within the *phoH2* gene (*Rv1095*/*MSMEG_5247*).

Primers used for RT-PCR and the corresponding expected PCR product sizes are listed in Table 2.2.

Table 2.2 Primers used for RT-PCR

Organism	Primer		Primer sequence	Expected band size
M. bovis BCG	F	RT 1	GCTACCGAGCCGAGAAGTACAC	724 bp
		RT 2	TTCTACGAGCGCTGCG	674 bp
		RT 3	CGGAGTAGGTCGCACTCGAT	275 bp
		RT 4	CACTCGATGGCTAGCGACATGCTC	263 bp
		RT 5	GCTAGCGACATGCTCTGCTGCCA	254 bp
		RT 6	TCATGACAAGGGCAGGACCGGC	204 bp
	R	RT	CACCACCAACGGAACCACCACAT	
M. smegmatis	F	RT 1	GCTACCGCGCCGACCGCTTCA	880 bp
		RT 2	TTCTTCGAGCGCTCGC	830 bp
		RT 3	CTTCTCCAACCTCGTCAGCTA	714 bp
		RT 4	AGTGCGGGTTCGTGGCGC	471 bp
		RT 5	GTCGAAGTCGATGGCTAGCGAC	315 bp
		RT 6	GGGGCCGCATGAAGAGCG	205 bp
	R	RT	GCCGCGCGAACCAGCCGA	

In addition to + and –RT templates, genomic DNA was tested to confirm the size of the PCR products. Initially gradient *Taq* polymerase PCR cycling parameters were screened from 52-62 °C as in section 2.1.5.2 and 58 °C was found to be a suitable annealing temperature for all primer sets. PCR products were analysed by agarose gel electrophoresis as described in

section 2.1.3. The experiment was repeated at least 3 times to confirm the presence or absence of +RT products.

2.3.2 Conditional expression

To create tetracycline inducible expression constructs of *phoH2* and *phoH2*$_{alt}$ (*phoH2* cloned from alternative start site 1 (Figure 3.6)) from *M. tuberculosis* and *M. smegmatis*, the genes for *Rv1095*, *Rv1095*$_{alt}$ and *MSMEG_5247* were amplified from genomic DNA using primers listed in Table 2.3. The PCR products were digested and ligated between the *Bam*HI/*Spe*I sites of the tetracycline-inducible vector pMIND and were transformed into *E. coli* TOP10 cells for selection on low salt LB medium supplemented with 50 µg/ml hygromycin B. Sequencing confirmed the correct integration of the insert. Primers pTB F, pTB$_{alt}$ F, pMS F also introduced a synthetic RBS sequence upstream of the *phoH2* gene to ensure translation.

Table 2.3 Primers used for cloning into pMIND

Primer		Primer Sequence RE + **RBS**	RE
F	pMS	AAATTTGGATCC**GGAGG**AATAATGGTGACTGAGCAAGCTGTCCG	*Bam*HI
	pTB	AAATTTGGATCC**GGAGG**AATAATGGTGACCGATACCCGCACGTA	*Bam*HI
	pTB$_{alt}$	AAATTTGGATCC**GGAGG**AATAATGGCTAGCGACATGCTCTG	*Bam*HI
R	pMS	AAATTTACTAGTTCAGGGCAGGGCGCCGGGGC	*Spe*I
	pTB	AAATTTACTAGTTCAGCGCGGCCCGGTGATCT	*Spe*I

RE – restriction site **RBS** – ribosome binding site

Sequenced constructs were transformed into wild type *M. smegmatis* mc^{2}155 as described in section 2.1.10.5, as well as the intact empty pMIND vector (EV), to generate *M. smegmatis* mc^{2}155 harbouring pMIND$_{EV}$, pMIND$_{MSMEG_5247}$, pMIND$_{Rv1095}$, and pMIND$_{Rv1095alt}$.

For overexpression analysis, all *M. smegmatis* strains carrying vectors pMIND$_{EV}$, pMIND$_{MSMEG_5247}$, pMIND$_{Rv1095}$, and pMIND$_{Rv1095alt}$ were grown to an OD$_{600}$ between 0.2-0.4 in LBT supplemented with 50 µg/ml hygromycin B. Each LBT starter culture was used to seed a second starter culture in Hartman's-de Bont minimal medium (HdB) supplemented with 0.2 % glycerol, 50 mM MES, 0.05 % (v/v) Tween-80 and 50 µg/ml hygromycin B.

HdB cultures were grown to an OD_{600} between 0.1-0.2 and diluted to an OD_{600} of 0.0025 in 100 ml HdB medium in a 250 ml flask. The growth of all strains was monitored until an OD_{600} of between 0.1-0.15 was reached, and then protein expression was induced with 20 ng/ml tetracycline (Tc), which had been previously reported to be sufficient for induction from the tetracycline-inducible promoter with limited toxicity to the cells (Robson 2010). Samples of culture were taken periodically to monitor the growth (OD_{600}) of the culture and cell viability (CFUs). For cell viability, samples of culture were serially diluted 10-fold from 10^{-1} to 10^{-6} in 1x phosphate buffered saline (PBS) (0.02 M phosphate (0.0038 M NaH_2PO_4), 0.15 M NaCl, pH 7.4) and each dilution was spotted onto plates supplemented with hygromycin B and tetracycline (+Tc), and hygromycin B only (-Tc).

2.4 Methods relating to Chapter 4

2.4.1 Cloning into pYUB1049

The annotated ORF containing the *Rv1095* sequence was amplified from *M. bovis* BCG strain Pasteur 1173P2 genomic DNA (100 % identity to *Rv1095* from *M. tuberculosis*) using two different primer sets $Rv1095_C$ F and $Rv1095_C$ R; and $Rv1095_N$ F and $Rv1095_N$ R (Table 2.4) using the method described in section 2.1.5.2. The amplified PCR products were purified as in section 2.1.6 and digested with *Nco*I/*Bam*HI and *Nde*I/*Bam*HI respectively, and then ligated into the pYUB1049 shuttle vector as in section 2.1.9, enabling expression of a C-terminal and N-terminal his-tag attached to PhoH2 (*Rv1095*).

Table 2.4 Primers used for cloning *Rv1095* and *Rv1095alt* with a C- or N-terminal his-tag into pYUB1049

Primer		Primer Sequence RE	RE
F	$Rv1095_C$	TAACTCCATGGGTGACCGATACCCGCAC	*Nco*I
	$Rv1095_N$	TCTCGTACATATGGTGACCGATACCCG	*Nde*I
	$Rv1095_{altC}$	TAACTCCATGGCTAGCGACATGCTCTGC	*Nco*I
	$Rv1095_{altN}$	TCTCGTACATATGGCTAGCGACATGCTCTG	*Nde*I
R	$Rv1095_C$	TACAGGATCCGCGCGCGGCCCGGTGATC	*Bam*HI
	$Rv1095_N$	TACAGGATCCTCAGCGCGCGGCCCGGTGATC	*Bam*HI

RE – restriction enzyme

Ligations were transformed into *E. coli* TOP10 cells as in section 2.1.10 and plated on low salt LB agar medium supplemented with 0.05 % (v/v)

Tween-80 and 50 µg/ml hygromycin B to select for positive transformants. To test for gene insertion, colony PCR was carried out using T7 F and R primers that flanked the insert site as in section 2.1.5.3. Positive clones were cultured overnight, their vector DNA was extracted as in section 2.1.2. and sequenced with T7 F and R primers prior to transformation into *M. smegmatis* $mc^2$4517 cells (section 2.1.10.5). Positive transformants of *M. smegmatis* $mc^2$4517 were selected by plating onto 7H10 agar medium supplemented with ADC, 0.05 % (v/v) Tween-80 and 50 µg/ml kanamycin and hygromycin B. Selected colonies were cultured in 5 ml of PA-0.5G medium supplemented with 0.05 % (v/v) Tween-80, 50 µg/ml kanamycin and hygromycin B with shaking at 200 rpm at 37 °C for 48 hours. These cultures were used to seed 5 ml ZYP-5052 autoinduction medium supplemented with 0.05 % (v/v) Tween-80 and 50 µg/ml kanamycin and hygromycin B, at a 1:100 dilution. Expression cultures were incubated at 37 °C with shaking at 200 rpm for 96 hours.

The same approaches were taken to clone and express *Rv1095$_{altC}$* and *Rv1095$_{altN}$* using Rv1095$_{altC}$ F and Rv1095$_{C}$ R, and Rv1095$_{altN}$ F and Rv1095$_{N}$ R (Table 2.4).

2.4.2 Cloning into Gateway pDONR$_{221}$, pDEST$_{SMG}$, and pDEST$_{17}$

For cloning into Gateway vectors, the ORFs encoding *Rv1095*, *Rv1095$_{alt}$*, the *PIN-domain* of *Rv1095* and the *PhoH-domain* of *Rv1095* sequences were amplified from *M. bovis* BCG strain Pasteur 1173P2 genomic DNA using flanking gene specific primers listed in (Table 2.5) and RT 5 F (Table 2.2), using the method described in section 2.1.5.2. The amplified PCR products were purified as in section 2.1.6 and used as template for a second round of PCR using gene specific recombination primers to gain *attB*-flanked PCR products.

Table 2.5 Primers used for Gateway flanking and recombination reactions

Flanking Primer		Primer Sequence
F	Flank 1	CGGAGTAGGTCGCACTCGAT
	PhoH FL_{GW}	GGGTCAATGCGCATAAACGTGTT
R	Rv1095 FL_{GW}	GCGCAGTGCGATTCTGTGTC
	PIN FL_{GW}	TCCTGCGCGTGGTACTCGTCG
Recombination Primer		**Primer Sequence**
F	$Rv1095_{GWR}$	GGGGACAAGTTTGTACAAAAAAGCA GGCTTCGTGACCGATACCCGCACGTA
	$Rv1095_{altGWR}$	GGGGACAAGTTTGTACAAAAAAGCA GGCTTCATGGCTAGCGACATGCTCTG
	$PhoH_{GWR}$	GGGGACAAGTTTGTACAAAAAAGCA GGCTTCGAGGCGTTCGGTCTGCGTG
R	$Rv1095_{GWR}$	GGGGACCACTTTGTACAAGAAAGCTG GGTCTCAGCGCGGCCCGGTGATCTCC
	PIN_{GWR}	GGGGACCACTTTGTACAAGAAAGC TGGGTCCTAAACGCAGCGGAATG

The recombination PCR products were purified as in section 2.1.6 and inserted into $pDONR_{221}$ via the Gateway BP reaction. Briefly, 100 femtomoles of *att*B-PCR product was incubated with 2 µl of $pDONR_{221}$ (150 ng/µl) and 4 µl of BP clonase™ in a 20 µl reaction made up with TE buffer pH 8 for 1 hour at 25 °C. The reactions were stopped by incubating at 37 °C for 10 minutes with the addition of 2 µl of Proteinase K. These reactions were transformed into chemically competent DH5α cells (section 2.1.10.3) and plated onto LB agar medium containing 50 µg/ml kanamycin. Colony PCR was carried out using M13 F and R primers that flanked the insertion site to check for the insert as in section 2.1.5.3. Successful clones were cultured overnight and vectors extracted as in section 2.1.2.

For expression of constructs in *M. smegmatis*, these were inserted into $pDEST_{SMG}$ via the Gateway LR reaction. Briefly, 100 ng of $pDONR_{Rv1095}$, $pDONR_{Rv1095alt}$, $pDONR_{PIN}$ and $pDONR_{PhoH}$ were incubated with 2 µl (300 ng) of destination vector ($pDEST_{SMG}$) and 4 µl of LR clonase™ in a 20 µl reaction made up with TE buffer pH 8 for 1 hour at 25 °C. The reactions were stopped as before with Proteinase K. To verify that the recombination with $pDEST_{SMG}$ had been successful, the expression constructs were transformed into *E. coli* TOP10 cells as in section 2.1.10.2 and plated on low salt LB agar medium supplemented with 50 µg/ml hygromycin B. Colonies were tested for the presence of insert via colony PCR using T7 F and R primers as in section

2.1.5.3. Successful clones were cultured overnight and vector DNA extracted as in section 2.1.2. These constructs were sequenced with T7 F and R primers prior to transformation into *M. smegmatis* $mc^2$4517 cells as described in section 2.4.1.

For expression of $pDONR_{221}$ constructs in *E. coli*, a second Gateway LR reaction was performed with $pDEST_{17}$. The same LR reaction was performed as before, and to verify that recombination with $pDEST_{17}$ had been successful the expression constructs were transformed into electrocompetent, or chemically competent DH5α cells and plated onto LB agar medium containing 100 µg/ml ampicillin. Colonies were tested for the presence of the insert via colony PCR and successful clones were cultured overnight and vector DNA extracted as in section 2.1.2. These were sequenced prior to transformation into *E. coli* BL21 (DE3) cells (section 2.1.10.2). Positive *E. coli* transformants were selected by plating onto LB agar medium containing 100 µg/ml ampicillin. Colonies were cultured in 5 ml of LB medium supplemented with 100 µg/ml ampicillin with shaking at 200 rpm at 37 °C for 24 hours, these cultures were used to seed 5 ml LB broth at a 1:100 dilution. Overexpression cultures were incubated at 37 °C with shaking at 200 rpm until an OD_{600} 0.4-0.6 was reached. They were induced with 1 mM IPTG (final concentration) and were incubated at 37 °C with shaking at 200 rpm overnight.

All of these constructs enabled expression of an N-terminal his-tag attached to each of the proteins.

2.4.3 Cloning into pYUB28b

The ORFs containing *Rv1095*, the *PIN-domain* of *Rv1095*, *PhoH-domain* of *Rv1095*, and PIN_{alt}*-domain* of *Rv1095* were amplified from *M. bovis* BCG strain Pasteur 1173P2 genomic DNA or previously cloned constructs using the primers listed in Table 2.6 as in section 2.1.5.2.

Table 2.6 Primers used for cloning *Rv1095*, *PIN*, *PhoH*, and PIN_{alt} domains with a C-terminal his-tag into pYUB28b

Primer		Primer Sequence RE	RE
F	$Rv1095/PIN_{28b}$	TAACTCCATGGTGACCGATACCCGC	*Nco*I
	$PhoH_{28b}$	TAACTCCATGGAGGCGTTCGGTCTGCG	*Nco*I
	PIN_{alt28b}	TAACTCCATGGCTAGCGACATGCTCT	*Nco*I
R	$Rv1095/PhoH_{28b}$	TAACTAAGCTTGCGCGGCCCGGTGATC	*Hind*III
	PIN_{28b}	TAACTAAGCTTAACGCGCAGCGGAAT	*Hind*III

RE – restriction enzyme

In addition, the ORFs containing *MSMEG_5247*, $MSMEG\_5247_{alt}$ (cloned from alternative start site 1 (Figure 3.6)) the *PIN-domain* of *MSMEG_5247*, $PhoH_{+1}$ and $PhoH_{+2}$*-domains* of *MSMEG_5247* were amplified from *M. smegmatis* mc²155 genomic DNA using the primers listed in Table 2.7 as in section 2.1.5.2.

Table 2.7 Primers used for cloning *MSMEG_5247*, $MSMEG\_5247_{alt}$, *PIN*, $PhoH_{+1}$ and $PhoH_{+2}$ domains with a C-terminal his-tag into pYUB28b

Primer		Primer Sequence RE	RE
F	$MS5247_{28b}$ F	TAACTCCATGGTGACTGAGCAAGCTGTC	*Nco*I
	$MS5247_{alt28b}$ F	TAACTCCATGGCTAGCGACCTGCT	*Nco*I
	$PhoH_{MS128b}$ F	TAACTCCATGGAAGTGTTCGGCCTC	*Nco*I
	$PhoH_{MS228b}$ F	TAACTCCATGGATCGCGAAGTGTTC	*Nco*I
R	$MS5247_{28b}$ R	TAACTAAGCTTGGGCAGGGCGCCGGG	*Hind*III
	PIN_{MS28b} R	TAACTAAGCTTGACGCGCAGCGGGAT	*Hind*III

RE – restriction enzyme

The amplified products were purified as in section 2.1.2 and digested with *Nco*I/*Hind*III, then ligated into the pYUB28b shuttle vector enabling expression of a C-terminal his-tag attached to each of the proteins. The successful clones were transformed into *E. coli* TOP10 cells as in section 2.4.1 and screened by colony PCR as in section 2.1.5.3 using T7 F and R primers. Positive clones were sequenced with T7 F and R primers prior to transformation into *M. smegmatis* mc²4517 cells (section 2.1.10.5).

2.4.4 Cloning into pET28b-*Pst*I

The amino acid sequence of *TBIS_3092* was submitted to GeneArt® (Life Technologies) for gene synthesis with codon optimisation for protein expression in *E. coli*. The resulting lyophilised plasmid DNA was resuspended in MQ as per manufacturers' instructions. The gene was amplified from the

GeneArt® plasmid using primers TBIS F and R (Table 2.8) as described in section 2.1.5.2.

Table 2.8 Primers used for cloning GeneArt synthesised *TBIS_3092* with a C-terminal his-tag into pET28b-*Pst*I

Primer		Primer Sequence RE	RE
F	TBIS	AACTGCAGATGGTTGCAGTTAGCAGCA	*Pst*I
R	TBIS	CCCAAGCTTGATGGTAATATCTTGCAGC	*Hind*III

RE – restriction enzyme

The amplified products were purified as in section 2.1.2 and digested with *Pst*I/*Hind*III, then ligated (section 2.1.9) into the pET28b-*Pst*I expression vector enabling expression of a C-terminal his-tag attached to the *TBIS_3092*. Ligations were transformed into DH5α cells as in section 2.1.10.2 and plated on LB agar medium supplemented with 50 µg/ml kanamycin to select for positive transformants. These were screened by colony PCR using the above gene specific primers and sequenced prior to transformation into *E. coli* BL21 (DE3) cells (section 2.1.10.2). Positive clones were selected by plating onto LB agar medium supplemented with 50 µg/ml kanamycin. These colonies were cultured in 5 ml LB broth supplemented with 50 µg/ml kanamycin and were incubated with shaking at 200 rpm at 37 °C for 24 hours. These cultures were used to seed 5 ml LB broth overexpression cultures and were grown and induced as in section 2.4.2 for protein expression in *E. coli*.

2.4.4.1 Small scale expression testing and purification from **M. smegmatis *and* E. coli**

After successful protein expression strains were obtained in either *M. smegmatis* or *E. coli*, Ni^{2+}-sepharose beads were used to test for the expression and his-tag binding affinity of the proteins. A sample of 2.5 ml of each 5 ml overexpression culture was spun down in an Eppendorf tube at 13 000 rpm. The cells were resuspended in 200 µl of 50 mM Na-phosphate buffer pH 7.4, 200 mM NaCl, 20 mM imidazole (lysis buffer A), and sonicated on ice using an Ultra sonic processor XL 2020 sonicator (Misonix incorporation, USA) (micro tip, setting 3, 4 x 15 seconds). The sonicated cell lysate was then centrifuged at 13 000 rpm at 4 °C for 20 minutes in order to separate the insoluble and soluble fractions. The Ni^{2+}-sepharose beads were

washed with 1 ml of lysis buffer A, left to settle, and were centrifuged at 3000 rpm at room temperature (RT) for 30 seconds. The wash buffer was then removed and replaced with the supernatant from the centrifuged cell lysate. Samples were incubated with shaking (1000 rpm) at RT for 15 minutes. After incubation, Ni^{2+}-sepharose beads were left to settle and were then centrifuged at 3000 rpm at RT for 30 seconds. The supernatant was removed and kept, and the beads were washed a total of three times with 1 ml lysis buffer A as before. Typically, the expression culture, the initial supernatant, the supernatant from the Ni^{2+}-sepharose beads after incubation, wash one and the Ni^{2+}-sepharose beads were analysed on a 15 % SDS-PAGE gel.

2.4.5 Large scale protein expression and purification from *M. smegmatis*

Autoinduction medium (ZYP-5052) supplemented with 0.05 % (v/v) Tween-80 and 50 µg/ml kanamycin and hygromycin B was used for protein expression. As described in section 2.4.1 a single transformed colony was selected and used to inoculate a PA-0.5G seeder culture. This culture was grown for 48 hours at 37 °C with shaking at 200 rpm and was used at a 1:100 dilution to inoculate a ZYP-5052 expression culture which was grown for 96 hours at 37 °C with shaking at 200 rpm.

The cells from large scale expression cultures were harvested by centrifuging at 4600 rpm at 4 °C for 20 minutes. Cell pellets were resuspended in 50 mM Na-phosphate buffer pH 7.4, 200 mM NaCl, and centrifuged again. For purification, cells were resuspended in an appropriate lysis buffer (generally 50 mM Na-phosphate buffer pH 7.4, 200 mM NaCl, 20 mM imidazole (lysis buffer A)) and sonicated and harvested as in section 2.4.4.1 using the large tip, with setting 5-7, for 30 second bursts for a total of 4-5 minutes. The soluble fraction containing his-tagged $PhoH2_{alt\ MTB}$, $PhoH2_{alt\ MSMEG}$, or $PhoH2_{MSMEG}$ was loaded onto a HisTrap HP column (GE Healthcare Life Sciences) and purified as in section 2.2.7. The fractions containing purified protein were concentrated as in section 2.2.5 and if required, purified further by size exclusion chromatography using either a HiLoad™ 16/60 Superdex™ or a 10/300 Superdex™ column (analytical size exclusion column) (GE

Healthcare, UK) as in section 2.2.8., typically equilibrated with the same lysis buffer minus imidazole (50 mM Na-phosphate buffer pH 7.4, 200 mM NaCl).

2.4.6 Large scale protein expression and purification from *E. coli*

LB broth supplemented with 50 µg/ml kanamycin was used for protein expression. As described in section 2.4.4 a single transformed colony was selected and used to inoculate a 5 ml LB seeder culture. This culture was grown for 24 hours at 37 °C with shaking at 200 rpm and was used at a 1:100 dilution to inoculate 1 L of LB medium supplemented with 50 µg/ml kanamycin. Overexpression cultures were incubated at 37 °C with shaking at 200 rpm until an OD_{600} 0.4-0.6 was reached. Cultures were induced with a final concentration of 1 mM IPTG and were incubated at 37 °C with shaking at 200 rpm overnight.

The cells from large scale expression cultures were harvested and purified as in section 2.4.5 using 50 mM Tris or Na-Hepes buffer pH 7.4, 200 mM NaCl, buffers.

2.4.6.1 Lysis buffer screens

Lysis buffer screens were performed to test for protein solubility. Overexpression cultures were tested similarly as described in section 2.4.4.1. Briefly, aliquots of overexpression culture were resuspended in 200 µl of buffer listed in Table 2.9. Cells were sonicated then centrifuged, and samples of the supernatant and cell pellet were analysed on a 15 % SDS-PAGE gel.

Table 2.9 Lysis buffers used in lysis buffer screen

Buffer	pH	NaCl
50 mM Na-Acetate	5.0	150 mM
50 mM Na-Acetate	5.0	500 mM
50 mM Na-Hepes	7.0	150 mM
50 mM Na-Hepes	7.0	500 mM
50 mM Na-Phosphate	7.4	200 mM
50 mM Tris	8.0	150 mM
50 mM Tris	8.0	500 mM
50 mM Na-bis-tris propane	9.0	150 mM
50 mM Na-bis-tris propane	9.0	500 mM

2.4.6.2 Growth temperature screen for protein expression

Growth temperature screens were trialled in an attempt to remedy poor protein expression with cultures grown at 37 °C. For protein expression in *M. smegmatis*, the temperature for incubation of autoinduction cultures was reduced to 30 °C. For protein expression in *E. coli* cultures were incubated at 28 °C, 22 °C or 18 °C post-induction with 1 mM final IPTG.

2.4.7 Denaturing purification of PIN and PhoH-domains from *E. coli*

Large scale protein expression of the PIN and PhoH-domain from *Rv1095* that had been cloned into $pDEST_{17}$ was carried out in LB medium. One litre of LB broth supplemented with 100 μg/ml ampicillin cultures were seeded with 1:100 dilution of seeder culture, induced with IPTG as before at an OD_{600} 0.4-0.6 and incubated at 37 °C overnight. Cells from overexpression cultures were harvested as above for the purification from *M. smegmatis* and *E. coli*. Cell pellets were resuspended in 50 mM Tris buffer pH 7.5, 200 mM NaCl, 8 M urea and sonicated on ice as in section 2.4.5 for a total of 2-3 minutes. The cell lysate was centrifuged at 13 000 rpm at 4 °C for 20 minutes, and the insoluble fraction was resuspended in 50 mM Tris buffer pH 7.5, 200 mM NaCl, 8 M urea and incubated on ice with gentle shaking for 1 hour. The resuspended pellet was then centrifuged at 10 000 rpm for 10 minutes, the supernatant removed and loaded onto a HisTrap HP column (GE Healthcare Life Sciences) and the protein purified as in section 2.2.7.

2.4.8 Protein refolding

An adapted version of the Hampton refolding screen (Hampton Research, USA) of 16 solutions was used to test for protein refolding (Appendix A). The solutions screened the effect of: divalent cations (Mg^{2+}/Ca^{2+} and EDTA), buffer pH and composition (MES/Tris-HCl), salt concentration (NaCl/KCl), PEG; detergent (Triton X), polar (arginine) or non polar (sucrose) additives, chaotropes (guanidine-HCl), and oxidising/reducing agents (GSH/GSSG or DTT) on refolding.

2.4.8.1 Small scale refolding screen

The PIN and PhoH-domain proteins were purified as in section 2.2.7 under denaturing conditions (section 2.4.7). These proteins were dialysed in 20 µl dialysis buttons sealed with 6-8 MWCO dialysis membrane against 2 ml of the refolding screen (Appendix A) solutions at 4 °C for 1-2 days as in section 2.2.6. The dialysed protein was examined for precipitate and analysed by Native-PAGE (section 2.2.2). Positive protein refolding was determined by the absence of precipitate after dialysis and by being able to successfully run into native polyacrylamide gel.

2.4.8.2 Large scale refolding

Protein refolding conditions that showed the most promise in the small scale refolding screen were scaled up in 10 fold steps from 20 µl to 20 ml. Denatured protein was dialysed in 1000x the protein volume of appropriate refolding buffer at 4 °C for 1-2 days. The dialysed protein was then dialysed at 4 °C overnight into storage buffer (refolding buffer base plus 200 mM NaCl) and analysed by size exclusion chromatography as in section 2.2.8.

2.4.9 Solubilisation trials of PIN and PhoH-domains from *E. coli*

2.4.9.1 Small scale solubilisation trials

Small scale expression cultures were set up as in section 2.4.2. For solubilisation testing, 3 mL of culture was spun down and resuspended initially in 250 µl 50 mM Na-phosphate buffer pH 7.4, 200 mM NaCl, 1.5 % (w/v) N-lauroylsarcosine or 50 mM Tris buffer pH 7.5, 200 mM NaCl, 20 mM imidazole, 1.5 % (w/v) N-lauroylsarcosine. Buffer without N-lauroylsarcosine or triton X-100 was used as a control. The suspension was sonicated and centrifuged at 13 000 rpm at 4 °C for 20 minutes. The supernatant was removed from the cell pellet and was mixed with 1 %, 2 % or 3 % triton X-100. Small scale purification trials were performed as in section 2.4.4.1. For elution of the protein from the Ni resin, 100 µl 50 mM Tris buffer pH 7.5, 200 mM NaCl, 500 mM imidazole was added to the resin and pipetted to mix for 2 minutes. The resin was left to settle and the supernatant was removed and kept. Both the supernatant and the beads were analysed by SDS-PAGE and Native-PAGE.

2.4.9.2 Large scale solubilisation trials

For large scale solubilisation trials, 1 L expression cultures were set up as in section 2.4.6 supplemented with 100 µg/ml ampicillin. The cells were harvested as described in section 2.4.5 and resuspended in 20 mL 50 mM Tris buffer pH 7.5, 200 mM NaCl, 20 mM imidazole. Once resuspended, 1.5 % final (w/v) N-lauroylsarcosine was added and mixed by inversion. The suspension was sonicated and harvested as in section 2.4.6. The supernatant was removed and mixed with 1 % final concentration of triton X-100. The suspension was analysed by IMAC purification as in section 2.2.7.

2.4.10 Initial crystallisation screens

Initial crystallisation trials for $PhoH2_{alt\ MTB}$ protein purified in 50 mM Na-phosphate buffer pH 7.4, 200 mM NaCl were set up using a medium-throughput robotic system (Moreland et al. 2005) at the School of Biological Sciences at the University of Auckland. This trial consisted of five in-house screens with a total of 480 conditions from various screens (Moreland et al. 2005) in five 96 well Intelli-plates (Hampton Research, USA). A Multiprobe® II HTEX dispensing robot (Perkin-Elmer, USA) was used to dispense 75 µl of each condition into the reservoir wells of the Intelli-plate. A Honeybee Sitting Drop crystallisation robot (Cartesian™ Dispensing Systems, USA) was used to dispense protein and reservoir solutions (100 nl each) into the crystallisation well and the plate was then sealed with Clearseal™ film (Hampton Research, USA).

Crystallisation trials were carried out with $PhoH2_{MSMEG}$ purified in 50 mM Na-phosphate buffer pH 7.4, 200 mM NaCl and 50 mM Na-Hepes buffer pH 7.4, 200 mM NaCl using the following Hampton Research crystallisation screens (Hampton Research, USA): Crystal (HR2-130), PEG Rx (HR2-086), Salt Rx (HR2-136) and Index (HR2-134). One hundred microlitres of each screen condition were pipetted into four 96-well Intelli-plates (Hampton Research, USA). A Mosquito® crystallisation robot (TTP LabTech Ltd, UK) was used to dispense protein and reservoir solutions (100 nl each) into the crystallisation well, and the plates were sealed with Clearseal™ film (Hampton Research, USA).

Screens were also performed for $PhoH2_{TBIS}$ purified in 50 mM Tris buffer pH 8, 200 mM NaCl and 50 mM Na-Hepes buffer pH 7.5, 200 mM NaCl in the presence of 1 mM adenylyl-imidodiphospahte (AMP-PNP), and after a limited proteolytic digest of $PhoH2_{TBIS}$ with trypsin at a 1:100 ratio for 30 minutes at RT using the procedure described for $PhoH2_{MSMEG}$.

2.4.11 Optimisation of crystallisation conditions

When promising crystallisation conditions were identified from the robot screen, fine screens were performed to optimise the condition for protein crystal growth.

2.4.11.1 Fine screening - Hanging drop vapour diffusion

For hanging drops, the tops of the wells of a Crystalquick 24-well plate (Greiner Bio-one, Germany) were greased with glisseal grease and 500 µl of each fine screen condition was added to the wells. One microlitre of protein was mixed 1:1 with the reservoir solution and pipetted onto a 22 mm cover slip (siliconised glass, square) (Hampton Research, USA). The cover slip was then inverted and placed over the reservoir solution and pressed down gently to seal.

2.4.11.2 Additive screening

For additive screening a selection of 14 additives were incorporated into the precipitant solution. Drops were set up as in section 2.4.11.1.

Table 2.10 Additives used in additive screen

Additive	Final concentration
$Co(II)Cl_2.6H_2O$	0.01 M
$MgCl_2.6H_2O$	0.01 M
Ni (II) $Cl_2.6H_2O$	0.01 M
$CsCl_2$	0.1 M
Glycine	0.1 M
EDTA	0.01 M
Glucose	3 %
Sucrose	3 %
Glycerol	3 %
PEG 400	5 %
Ethanol	3 %
Methanol	3 %
Acetonitrile	4 %
Butanol	0.7 %

2.4.11.3 Seeding

For macro and streak seeding, hanging crystallisation drops were set up as in section 2.4.11.1 and left to equilibrate for 2-3 hours before seeding. For macro seeding, an already formed crystal from a source drop was removed to the new pre-equilibrated drop. For streak seeding, a dry cat's whisker cleaned with 70 % ethanol was streaked through a source drop containing crystals and then brushed through the new pre-equlibrated drop. For batch seeding, a source drop containing crystals was added to 18 µl of crystallisation condition and vortexed. This seed stock was diluted 1:10, 1:100 and 1:1000 in crystallisation condition, mixed 1:1 with protein and placed over the reservoir crystallisation solution.

2.4.11.4 Testing of Crystals by X-Ray Diffraction

Prior to testing for diffraction, crystals were looped out of their drop and moved into a cryo protectant solution which was made up of the crystallisation solution plus 5-20 % glycerol, before being snap frozen in liquid nitrogen. X-ray diffraction testing was performed at the Home-Source at the School of Biological Sciences, University of Auckland using CuKα radiation (λ=1.5418 Å). X-rays were generated by a MicroMax™ -007HF generator (Rigaku, Japan) operated at 50 kV and 100 mA with a rotating copper anode and Osmic VariMaz optics. The cryo-loop was mounted onto the goniometer using a MAR345dtb goniometer setup with an Easymount™ automatic sample mounting system (MAR Research, Germany). Crystals were kept at 110-113 K with a cold stream of nitrogen gas produced by a Cobra cryosystem (Oxford Cryosystems, UK).

Crystals were tested and X-ray diffraction data collected at the Australian synchrotron on beamline MX1, equipped with a ADSC Quantum 210r detector (Area Detector Systems Corporation, USA) with λ=0.9537 Å.

2.4.11.5 Data processing

Data were analysed using iMOSFLM for indexing, cell refinement and integration. The resulting mtz file was scaled using Scala or Aimless, and the

Matthews coefficient and self rotation function analysed using Matthews and MOLREP in the CCP4i suite.

2.4.12 Transmission electron microscopy (TEM) of PhoH2$_{TBIS}$

Size exclusion purified PhoH2$_{TBIS}$ protein was prepared at 0.1, 0.03 mg/ml for analysis by TEM in 50 mM Tris buffer pH 8, 200 mM NaCl. Protein samples were negatively stained onto grids with 2 % uranyl acetate. Electron microscope grids were examined using BIRU Technai G^2 electron microscope and low dose images were taken.

2.4.13 Cloning of alt peptide into pYUB28b

The alt peptide sequence from *M. tuberculosis* was amplified from pYUB1049 habouring the *Rv1095$_{alt}$* gene sequence using primers ALT F and R (Table 2.11) and was cloned into pYUB28b as in section 2.4.3 and transformed into *M. smegmatis* 4517 cells as in section 2.1.10.5.

Table 2.11 pYUB28b *M. tuberculosis* alt cloning primers

Primer		Primer Sequence RE	RE
F	ALT	TAACTCCATGGCTAGCGACATGCTCT	*Nco*I
R	ALT$_{28b}$	TAACTAAGCTTCGGAGCGCTCCTCGAGCT	*Hind*III

RE – restriction enzyme

Small scale expression tests were carried out as in section 2.4.4.1 to observe protein expression and his-tag binding affinity.

2.4.14 MALDI-TOF Mass Spectrometry (University of Waikato)

2.4.14.1 Preparation of tryptic digest and whole protein samples

Tryptic digest samples for analysis by MALDI-TOF MS were prepared as follows. Purified protein was separated on a 12 % SDS-PAGE gel and the protein band of the correct size was excised. The gel piece containing protein was destained overnight in 1:1 acetonitrile (ACN):25 mM NH_4HCO_3 and then washed with fresh 1:1 ACN:25 mM NH_4HCO_3. The supernatant was discarded and the gel piece was dehydrated with ACN for 15 minutes. The supernatant was removed again and the sample was dried under vacuum. One µl of 1 µg/µl trypsin was resuspended in 199 µl of 25 mM NH_4HCO_3 and was used to rehydrate the gel piece at 4 °C for 1 hour. Excess trypsin solution was

removed and the gel piece was incubated at 37 °C for 4-5 hours. A solution of 20 % ACN:0.1 % trifluoroacetic acid (TFA) (15 µl) was added to the sample and left to incubate overnight at RT. Samples were sonicated in a waterbath sonicator (Elma, Germany) at RT for 10 minutes prior to mixing 1:1 (0.5 µl:0.5 µl) with matrix and spotted onto an AnchorChip™ MALDI-TOF target plate (Bruker Daltonics, USA) and left to air dry. Whole protein samples for analysis by MALDI-TOF MS were prepared by dialysis into MQ overnight and then mixed 1:1 with matrix.

2.4.14.2 Preparation of matrix

A 9:1 mixture of 2, 5-dihydrozybenzoic acid and 2-hydroxy-5-methoxybenzoic acid (Super-DHB) was used as matrix for whole protein samples and α-Cyano-4-hydroxycinnamic acid (HCCA) as matrix for peptides. Solid matrix (10 mg) was added to 30 µl of 2:1 0.1 % TFA or ACN: 0.1 % TFA respectively and mixed well before sonicating in a waterbath sonicator (Elma, Germany) at RT for 10 minutes. The matrix solution was centrifuged at 13 000 rpm for 10 minutes and briefly centrifuged again before mixing with calibrant (peptide calibration standard or the whole protein calibration standard II (Bruker Daltonics, USA) or sample, and spotting onto the anchor chip.

2.4.14.3 MALDI-TOF Set Up

An Autoflex™ II MALDI-TOF mass spectrometer (Bruker Daltonics, USA) was used to analyse samples. Samples were analysed in linear mode, with the mass range selector set to “medium range”, pulsed ion extraction of 450 ns, gain to 2500 V, acceleration voltage to 20 kV and a range of 5-60 kDa collected depending on application. Laser power was typically around 60-80 %. Spectra for the peptide calibration standard or the whole protein calibration standard II (Bruker Daltonics, USA) were collected first and the spectrometer calibrated with an automatic polynomial correction. All spectra were searched against the SWISS-PROT amino acid sequence database using the Mascot search engine (http://www.matrixscience.com). The search was set up for full tryptic peptides with a maximum of four missed cleavage sites.

2.4.15 N-terminal sequencing

For the method performed by CPR, University of Otago, freshly IMAC purified $PhoH2_{alt\ MTB}$ and $PhoH2_{alt\ MSMEG}$ were couriered immediately after purification for N-terminal sequencing analysis.

The samples were prepared for analysis on a PVDF membrane. For sequencing the N-terminus half, the samples on the membrane were subjected to guanidination and sulfonation with 4-sulfophenyl isothiocyanate (SPITC) according to the method developed by Chen et al. (2004). All samples (modified and unmodified) were digested on-membrane with trypsin (Shevchenko et al. 1996) and the eluted peptides collected and dried. Samples were resuspended in 30 % ACN:0.1 % TFA and were mixed with α-cyano-4-hydroxycinnamic acid (HCCA) matrix before spotting onto an Opti-TOF 384 well plate (Applied Biosystems, MA) for analysis using an 4800 MALDI tandem time-of-flight mass spectrometer (MALDI TOF/TOF, Applied Biosystems, MA).

For analysis by LTQ-Orbitrap the same sample preparation was undertaken, however the trypsin digested samples were fractionated by nano-flow reverse phase liquid chromatography and injected directly via an emitter tip into the ion source of a LTQ-Orbitrap hybrid mass spectrometer for analysis.

All MS spectra with MALDI TOF/TOF were gained in positive-ion mode with 800 laser pulses per sample spot. The 15-20 strongest precursor ions of each sample spot were selected for MS/MS collision-induced dissociation (CID) analysis. CID spectra were gained with 5000 laser pulses per selected precursor using the 2 kV mode and air as the collision gas at a pressure of $1x10^{-6}$ torr.

All MS/MS spectra (from either mass spectrometer) of modified peptides that had reasonable peak intensities were manually interpreted. Unambiguous sequence tags were used for homology searches using the BLAST program at NCBI (http://www.ncbi.nlm.nih.gov/BLAST/). Further, all MS/MS spectra were searched against the SWISS-PROT amino acid sequence database using

the Mascot search engine (http://www.matrixscience.com). The search was set up for full tryptic peptides with a maximum of four missed cleavage sites. Carboxyamidomethyl cysteine, oxidized methionine, pyroglutamate (E, Q), guanidination (K) and SPITC-sulfonation (N-terminus) were included as variable modifications. The precursor mass tolerance threshold was 75 ppm and the maximum fragment mass error was 0.4 Da. The peptide score distribution ions score was -10log(*P*), where *P* is the probability that the observed match is a random event. Individual ions scores > 32 indicate identity or extensive homology where $p < 0.05$.

Samples were also screened for their intact protein mass by MALDI TOF/TOF. Proteins in solution were diluted 1/27, mixed with matrix and spotted onto the anchor chip.

2.5 Methods relating to Chapter 5

2.5.1 ATPase Methods

KH_2PO_4 phosphate standards were prepared by diluting a 10 mM KH_2PO_4 stock to a 100 µM (phosphate) working stock. The working stock was then serially diluted 1:1 in assay buffer (20 mM Na-Hepes buffer pH 7.4, 20 mM NaCl) from 50 µM to 1.56 µM.

Pilot trials with $PhoH2_{alt\ MTB}$ showed that 1.25 µg of protein in a 100 µl reaction (0.25 µM) was suitable for sufficient colour development using substrate concentrations from 0-6 µM ATP.

All reactions contained a final concentration of 1 mM $MgCl_2$ and 0-6 µM ATP, and were made up to a total volume of 100 µl with assay buffer. Control EDTA reactions contained a final concentration of 2 mM EDTA. Each reaction was pre-equilibrated at 37 °C prior to the addition of protein. Reactions were incubated at 37 °C for 10 minutes before being stopped with 30 µl stop solution (1.5 % ammonium molybdate, 0.18 % (v/v) Tween-80, 0.1 % malachite green and 14 % sulphuric acid). Colour development was measured at 620 nm after incubation at RT for 10 minutes.

For assays performed in the presence of RNA, reactions included 0.0025 μM total *M. smegmatis* RNA or 5'3 RNA (Table 2.12). For time course assays, 1 μM ATP was used for all reactions and reactions were incubated for 0-8 minutes before being stopped with stop solution.

All assays were performed in triplicate and standardised against a blank reaction containing no protein. The molar concentration of phosphate in the reactions was calculated by comparing with the phosphate (KH_2PO_4) standard (0-100 μM).

2.5.2 Unwinding and ribonuclease Methods

Self-annealing RNA and DNA oligonucleotides were ordered from Sigma and IDT (Custom Science). The RNA-oligonucleotides were made up to a working concentration (2 μM) in diethylpyrocarbonate (DEPC) or nuclease free water. The oligonucleotides were radio-labelled with adenosine 5'-triphosphate γ-^{33}P (SciMed) using T4 polynucleotide kinase according to kit instructions and then diluted 1:1 in hybridisation buffer (20 mM Tris-HCl pH 7.5, 500 mM NaCl, 1 mM EDTA). These were incubated at 95 °C for 10 minutes then left to hybridise overnight to RT. DNA-oligonucleotides were made up to a working concentration (12.5 μM) and were radio-labelled as above then diluted 4:1 in annealing buffer (50 mM Tris-HCl pH 8, 150 mM NaCl, 1 mM EDTA). Samples were incubated at 95 °C for 5 minutes and left to anneal for 45 minutes to RT.

Purified PhoH2 (1 μM, 10 pmol) was incubated with 10 fmol (0.01 μM) of RNA/DNA oligonucleotide (Table 2.12) in a reaction containing 1 mM ATP made up to 10 μl with assay buffer (50 mM Na-phosphate pH 7.4, 20 mM NaCl, 10 mM $MgCl_2$). Control reactions to test for ATP-dependent activity (-ATP), cofactor-dependent activity (+EDTA), and the requirement for ATP hydrolysis (AMP-PNP used in place of ATP), and test reactions were incubated at 37 °C for 30 minutes and then stopped with an equal volume (10 μl) of stop solution (20 % glycerol (w/v), 0.2 % SDS, 4 mM EDTA, 0.1 % (w/v) bromophenol blue) or with 2x formamide stop solution (80 % formamide (v/v), 5 mM EDTA, 0.1 % (w/v) bromophenol blue, 0.1 % (w/v) xylene cyanol FF). The control for positive unwinding (oligonucleotide with

assay buffer) was boiled at 95 °C for 10 minutes prior to being stopped with stop solution. All reactions were analysed directly by electrophoresis on either a native 1xTBE-10 % polyacrylamide sequencing gel or a 1xTBE-10 % minigel. Gels were run in 1x TBE (0.089 M Tris HCl, 0.089 M boric acid, 20 mM EDTA). The results were visualised by autoradiography. The gels were dried and reactions were exposed to film (Amersham Hyperfilm™ MP) for 2-24 hours. Films were developed for at least 2 minutes in Kodak Developer and replenisher (Sigma Aldrich), and fixed with Kodak Fixer and replenisher (Sigma Aldrich) for at least 30 seconds in the dark.

Table 2.12 RNA and DNA oligonucleotides

Oligonucleotide		Sequence 5'-3 (**duplex**)	Predicted ΔG (kcal/mol^{-1})
RNA	5'	GAAUGUACAUCA**GAGUGCGCACUC**	-4.33
	3'	**GAGUGCGCACUC**GAAUGUACAUCA	-4.74
	ss	GAAUGUACAUCAGAAUGUACAUCA	-7.82
	Blunt	**GAAUGUACAUCAUGAUGUACAUUC**	-11.06
	3'AU	**GAGUGCGCACUC**GAAUGUACAU	-4.34
	3'UA	**GAGUGCGCACUC**GAAUGUA	-4.07
	5'AC	ACAUGUACAUCA**GAGUGCGCACUC**	-4.12
	3'AC	**GAGUGCGCACUC**ACUACAUGUACA	-4.31
	5'AC blunt	**ACAUGUACAUCAUGAUGUACAUGU**	-11.74
	5'ACAA	ACAACAACAUCA**GAGUGCGCACUC**	-2.60
	5'1	UCAUGUACAUCA**GAGUGCGCACUC**	-4.12
	5'2	AGAUGUACAUCA**GAGUGCGCACUC**	-4.46
	5'3	ACUUGUACAUCA**GAGUGCGCACUC**	-4.08
	5'4	ACAAGUACAUCA**GAGUGCGCACUC**	-3.86
	5'5	ACAUCUACAUCA**GAGUGCGCACUC**	-2.63
DNA	5'	GAATGTACATCA**GAGTGCGCACTC**	
	3'	**GAGTGCGCACTC**GAATGTACATCA	
	Blunt	**GAATGTACATCATGATGTACATTC**	
	ss	GAATGTACATCAGAATGTACATCA	
	5'AC	ACATGTACATCA**GAGTGCGCACTC**	

5' or 3' indicates position of the ss tail with respect to the duplex.
The predicted ΔG value of each RNA oligonucleotide was calculated using http://www.basic.northwestern.edu/biotools/oligocalc.html.

2.5.3 UTR assays

Upstream substrates MTB_{alt} and $MSMEG_{alt}$ were amplified from *M. bovis* BCG strain Pasteur 1173P2 and *M. smegmatis* mc^{2}155 genomic DNA using MTB_{alt} F with MTB_{alt} R, and $MSMEG_{alt}$ F with $MSMEG_{alt}$ R. The PCR products were purified as in section 2.1.6 and were used as template for further PCR amplification with MTB_{alt}T7 with MTB_{alt} R, and $MSMEG_{alt}$T7 F with $MSMEG_{alt}$

R (Table 2.13) to generate $MTB_{alt}T7$ and $MSMEG_{alt}T7$. These products were used as template for transcription of RNA using the MEGAscript transcription reaction kit (Ambion) according to the kit instructions. The resulting RNA was dephosphorylated using shrimp alkaline phosphatase (SAP) (Roche) according to the kit instructions and was radio labelled with γ-^{33}P as in section 2.5.2.

Table 2.13 Primers for MTB and MSMEG alt substrates

Primer		Primer Sequence T7 promoter sequence
F	MTB_{alt}	ATGGCTAGCGACATGCTCTGCTGCC
	$MSMEG_{alt}$	ATGGCTAGCGACCTGCTCTGCTGTC
	$MTB_{alt}T7$	TAATACGACTCACTATAGGGATGGCTAGCGACATGCTCTG
	$MSMEG_{alt}T7$	TAATACGACTCACTATAGGGATGGCTAGCGACCTGCTCTG
R	MTB_{alt} R	CGGAGCGCTCCTCGAGCTAACG
	MS_{alt} SMEG R	GTCACGTGGCGCTCCTAGGGGA

For activity assays, 0.01 µM of RNA was incubated with 1 µM protein in a reaction containing 1 mM ATP made up to 10 µl with assay buffer (50 mM Na-phosphate pH 7.4, 20 mM NaCl, 10 mM $MgCl_2$) for 15 minutes at 37 °C. All reactions were incubated at 95 °C for 5 minutes then incubated on ice immediately for at least 1 minute prior to being analysed directly by electrophoresis on a native 1xTBE-10% polyacrylamide gel and visualised by autoradiography.

Chapter Three: Bioinformatics and transcriptional context of *phoH2*

3.1 Bioinformatics

3.1.1 *phoH2* genes in *M. tuberculosis* and *M. smegmatis*

The genomes of *M. tuberculosis* and *M. smegmatis* each contain a single copy of a *phoH2* gene (*Rv1095* and *MSMEG_5247*). The proteins encoded by these genes are 90.2 % identical, where most differences are conservative, single amino acid changes. Both amino acid sequences are of similar lengths (433 aa for PhoH2 from *M. tuberculosis* and 437 aa for PhoH2 from *M. smegmatis*) and share the conserved acidic residues characteristic of PIN-domain proteins at positions ~10-12, ~47, ~102 and ~127, and each of the conserved motifs identified for PhoH proteins in section 1.4.3 (Figure 3.1).

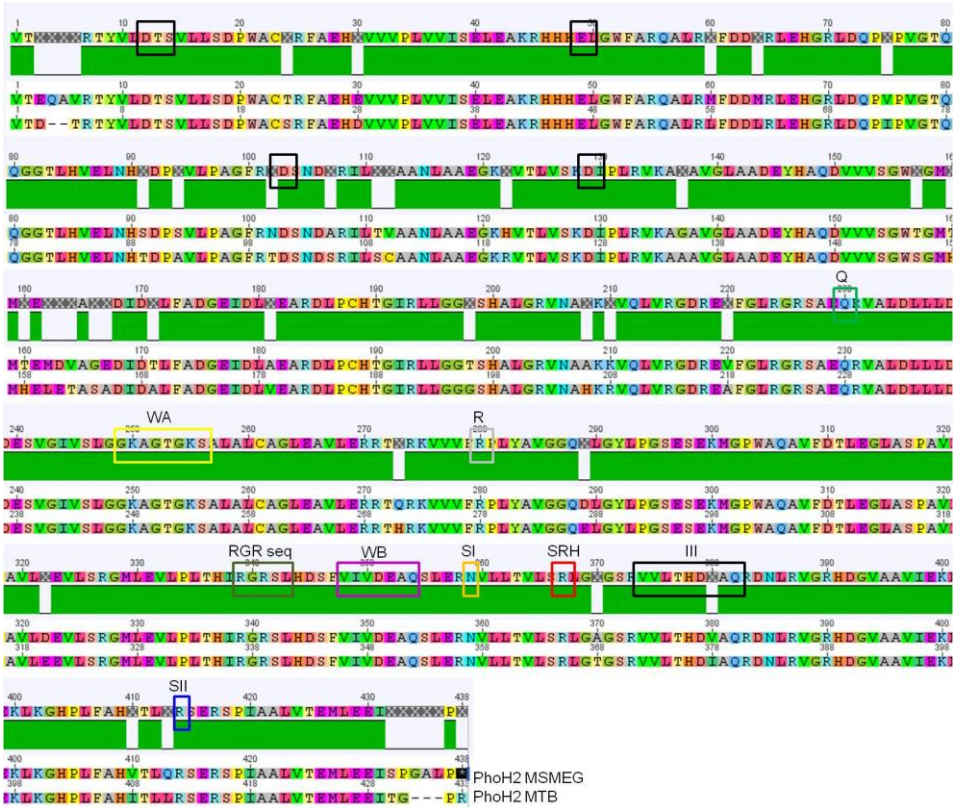

Figure 3.1 Alignment between PhoH2 protein sequences from *M. tuberculosis* (*Rv1095*) and *M. smegmatis* (*MSMEG_5247*)
The four N-terminal black boxed regions indicate the positions of the four conserved acidic residues within the PIN-domain of PhoH2. The latter boxes show the conserved residues and motifs of PhoH proteins: Q (green), Walker A (WA, yellow), R residue (R, light grey) RGRSL (RGR seq, dark green), Walker B (WB, magenta), N residue (SI, orange), R residue (SRH, red), III (black), R residue (SII, blue). Alignment was generated in GeneiousPro (V 5.1.7) using ClustalW.

3.1.2 Genomic context

The genomic context between *M. tuberculosis* and *M. smegmatis* is very similar (Figure 3.2).

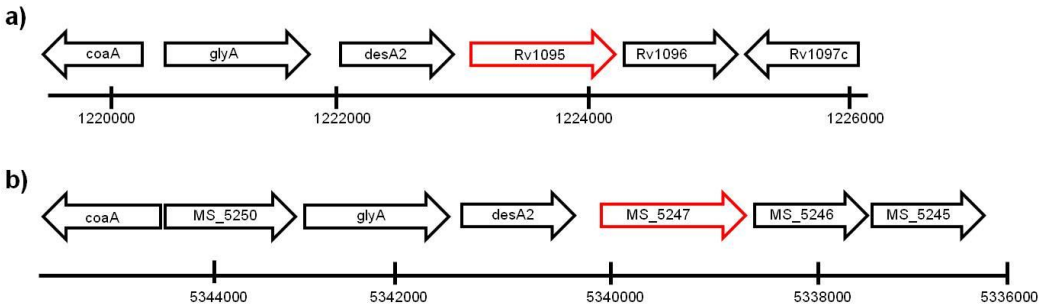

Figure 3.2 Genomic context of *Rv1095* and *MSMEG_5247*
Genomic context of *phoH2* genes from a) *M. tuberculosis* (*Rv1095*) and b) *M. smegmatis* (*MSMEG_5247*). Red arrow indicates position of *phoH2* gene. Adapted from Xbase (http://www.xbase.ac.uk/) and BLAST (NCBI).

For both genomic contexts, the genes *coaA* and *glyA* encode for a pantothenate kinase and serine hydroxymethyltransferase, respectively. The remaining genes surrounding *Rv1095* encode a putative acyl-desaturase (desA2), a glycosyl hydrolase (*Rv1096*) and putative membrane glycine and proline rich proteins (*Rv1097c)*. Most of these proteins are involved with fatty acid biosynthesis and metabolism (Nagai & Bloch 1968; Jaworski & Stumpf 1974; Stover & Schirch 1990; Appaji et al. 2003). The genes *coaA* and *glyA* are found in a similar position in *M. smegmatis*. There is, however an additional gene inserted between them (*MSMEG_5250*). This gene is listed as CalR9, which when analysed by PSI-BLAST is annotated as a hypothetical protein of unknown function. The gene *MSMEG_5246* is listed as a hypothetical protein and does not align with *Rv1096*. PSI-BLAST analysis annotates *MSMEG_5246* as a nitroreductase. *MSMEG_5245* is a universal stress protein (USP-like) and *MSMEG_5244* (not shown) encodes the transcriptional regulator DevR. DevR in *M. smegmatis* is involved with the adaptation to oxygen starvation and heat stress during stationary phase (O'Toole et al. 2002). *MSMEG_5241* (not shown) harbours DevS, the histidine kinase associated with DevR. DevRS (*Rv3132*, *Rv3133c*) in *M. tuberculosis* is found at a remote region of the genome from *phoH2* gene (O'Toole et al. 2002).

In *M. smegmatis*, microarray data revealed an upregulation of several genes including the gene cluster *MSMEG_5241* through *MSMEG_5246* when cultured under slow growth conditions and 0.6 % air (Berney & Cook 2010). Upon closer examination of these microarray data, *MSMEG_5247* was also found to be upregulated >2 fold under these growth conditions, suggesting the possible involvement of *phoH2* in the response to oxidative stress.

The STRING database (http://string-db.org) provides information on protein-protein interactions, each of which is provided with a confidence score and accessory information (Szklarczyk et al. 2011). The STRING database in the case of *Rv1095* and *MSMEG_5247* lists 10 predicted functional partners for both *phoH2* genes (Table 3.1). The first six functional partners of *phoH2* in Table 3.1 are shared by both *M. tuberculosis* and *M. smegmatis.* The latter genes are unique to each organism.

Table 3.1 Predicted functional partners of *phoH2*

Partner	Type of partner	Annotation
Rv3419c gcp *MSMEG_1580 gcp*	Neighbourhood	Putative DNA binding/iron metalloprotein/ AP endonuclease
Rv1094 desA2 *MSMEG_5248*	Neighbourhood Coexpression	Acyl-[acyl-carrier protein] desaturase
Rv1863c *MSMEG_3612*	Neighbourhood	Integral membrane protein
Rv2883c pyrH *MSMEG_2540 pyrH*	Neighbourhood	Uridylate kinase – pyrimidine metabolism
Rv0807 *MSMEG_5817*	Cooccurrence	Hypothetical
Rv1010 rsmA *MSMEG_5438 ksgA*	Neighbourhood	Dimethyladenosine transferase
Rv0430	Cooccurrence	Hypothetical
Rv1096	Neighbourhood	Glycosyl hydrolase
Rv0476	Cooccurrence	Transmembrane protein
Rv3416 whiB3	Cooccurrence	Transcriptional regulatory protein
MSMEG_5246	Neighbourhood	Hypothetical protein
MSMEG_5245	Neighbourhood	Universal stress protein
MSMEG_5244	Neighbourhood	Two-component transcriptional regulator DevR
MSMEG_4173	Neighbourhood	MaoC family protein

AP – Apurinic/apyrimidinic, Rv – *M. tuberculosis*, MSMEG – *M. smegmatis*

WhiB3 in *M. tuberculosis* is regulated by DosR/S/T (DevRS), and controls the differential production of virulence associated lipids in reponse to fluctuations in the intracellular redox environment (Singh et al. 2009). This modulates the response of macrophages to infection and maintains intracellular redox homeostasis (Singh et al. 2009). In *M. smegmatis* WhiB3 shows a slight increase in expression under slow growth conditions and 0.6 % air (Berney & Cook 2010), however not to the same degree as the cluster around *MSMEG_5247*. Predicted functional partner *MSMEG_1580* (putative DNA binding/metalloprotein/AP endonuclease) is upregulated under these growth conditions (Berney & Cook 2010) further suggesting the plausibility of *phoH2* involvement in the oxidative stress response.

3.1.3 Predicted transcriptional regulators and regulatory sites

The Tuberculosis Database (TBDB) has predicted four transcriptional regulator binding sites responsible for the expression of *Rv1095* (Figure 3.3).

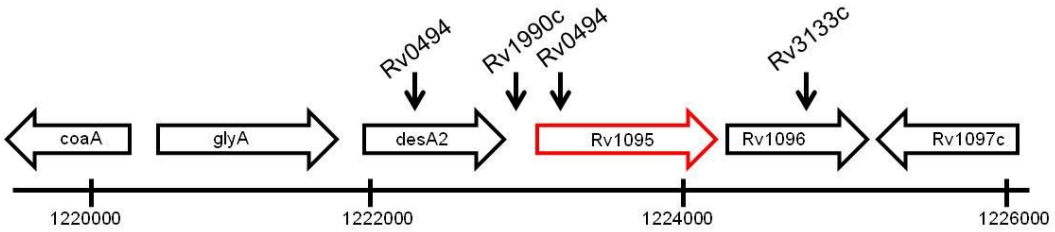

Figure 3.3 Genomic positions of predicted transcription factor binding sites
Arrows indicate the position of the four predicted transcription factor binding sites. The red arrow indicates the position of the *PhoH2* gene. Adapted from TBDB (http://www.tbdb.org/).

Rv0494 is a transcriptional regulator that belongs to the *GntR* family and is predicted to target *Rv1093-Rv1095* at -755 and +204 relative to the start of *Rv1095*. *Rv1990c* is a transcriptional regulator of unknown function and is predicted to target *Rv1094* and *Rv1095* at position -95 and *Rv3133c* encodes DevR which is predicted to target *Rv1095-Rv1097c* at position +1935. In *M. smegmatis* DevR is found two genes away from *phoH2* (*MSMEG_5244*). Likely homologues of *Rv0494* (*MSMEG_0179*) and *Rv3133c* (*MSMEG_5244*) in *M. smegmatis* show an elevated level of expression when cultured under slow growth conditions and 0.6 % air (Berney & Cook 2010). No homologue was found to *Rv1990c* in *M. smegmatis*. In a SenX3-RegX3 mutant strain of

M. tuberculosis, global expression data have shown that *Rv1990c* expression was increased compared with levels in the wild type, suggesting control of this transcriptional regulator by Sen3-RegX3 (Parish et al. 2003a).

Based on these observations and those made in sections 1.5.1, 1.5.2 and 3.1.2, it is plausible that *phoH2* may be involved in the response of the cell to hypoxic growth conditions and may play a role in virulence associated lipid biosynthesis. This may be achieved in *M. tuberculosis* through indirect regulation by: 1) PhoP, through its control of the Dos/Dev regulon (section 1.5.1) 2) WhiB3, that is under control of Dos/Dev (3.1.2) 3) SenX3-RegX3, through its control of *Rv1990c* (section 3.1.3) or 4) *Rv0494*, which is implicated in the hypoxic response (section 3.1.3). In *M. smegmatis*, through: 1) Dos/Dev and 2) SenX3-RegX3, both which showed elevated expression under hypoxic growth conditions along with *phoH2*.

In relation to these potential regulators, Kazakov et al. (2003), using an iterative signal PSI-SITE search predicted potential regulatory sites upstream of orthologous *BS-YlaK* genes. In the genomes of *M. tuberculosis* and *M. bovis BCG* the conserved 18 bp candidate site was predicted to be a pseudopalindrome (AGGACCGGCCCCGGTCCT) and located at position -87 (Figure 3.4a). The same sequence is found upstream of *MSMEG_5247* in *M. smegmatis* at position -86 (Figure 3.4b).

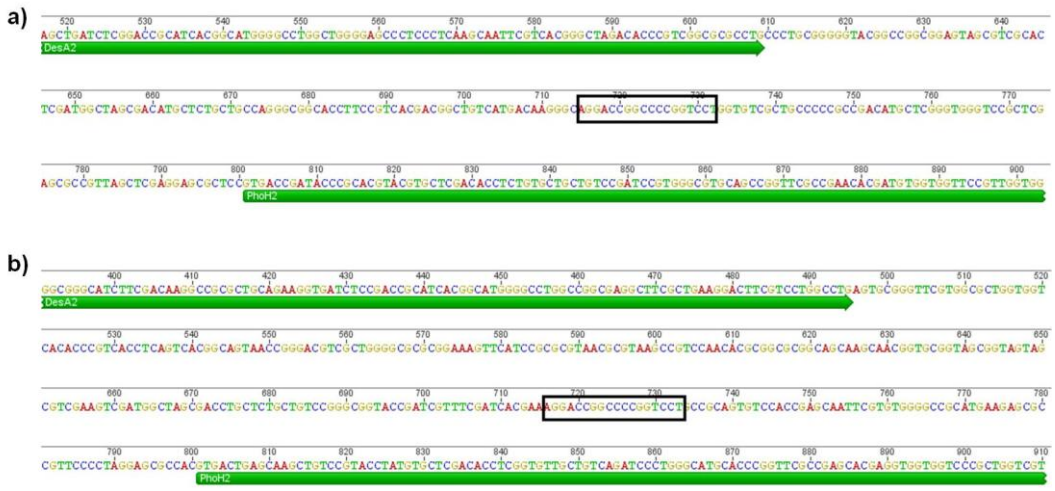

Figure 3.4 Relative position of candidate regulatory site upstream of *Rv1095* and *MSMEG_5247*
Positions of candidate regulatory sites upstream of a) *M. tuberculosis*/*M. bovis BCG* and b) *M. smegmatis* (Kazakov et al. 2003). Annotated sequence was made using GeneiousPro (V 5.1.7).

Overall, there are a number of candidate transcriptional regulator sites and potential regulatory sequences within the 5'UTR for the regulation of *phoH2*. It is striking that there exists strict conservation at -87 and -86 in *M. tuberculosis* and *M. smegmatis*.

3.1.4 Candidate alternative start sites upstream of *phoH2*

As PIN-domains, when they are found as single domain proteins are coexpressed with an upstream transcription factor (Arcus et al. 2011) and given the position of the PIN-domain at the N-terminus of PhoH2 it is possible that with *phoH2* there may exist an analogous transcriptional arrangement to that of *vapBC*. In the case of *vapBC*, there is a coding region upstream of *vapC* (PIN-domain), *vapB* (antitoxin) that is out of frame of and overlaps the translational start of the VapC protein. Figure 3.5 illustrates the possible arrangement of a coding region upstream of *phoH2*.

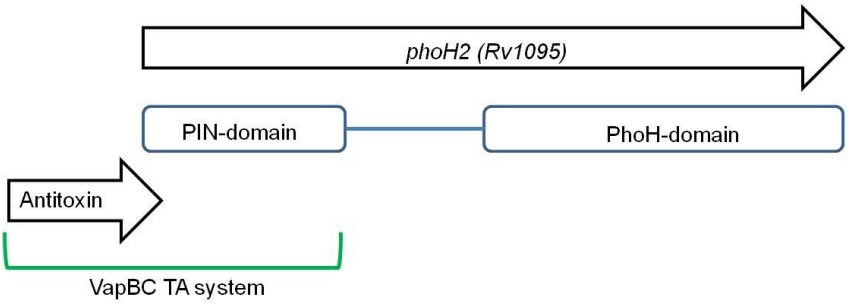

Figure 3.5 PhoH2 protein schematic with possible antitoxin upstream
PhoH2 protein schematic illustrating the possible position of an antitoxin protein upstream, analogous to *vapB* of *vapBC*.

Upon closer examination of the sequences upstream of the *phoH2* gene in *M. tuberculosis* and *M. smegmatis*, a number of possible alternative start sites for short potentially coding ORFs were identified that are out of frame of *Rv1095* and *MSMEG_5247* (Figure 3.6), analogous to VapB.

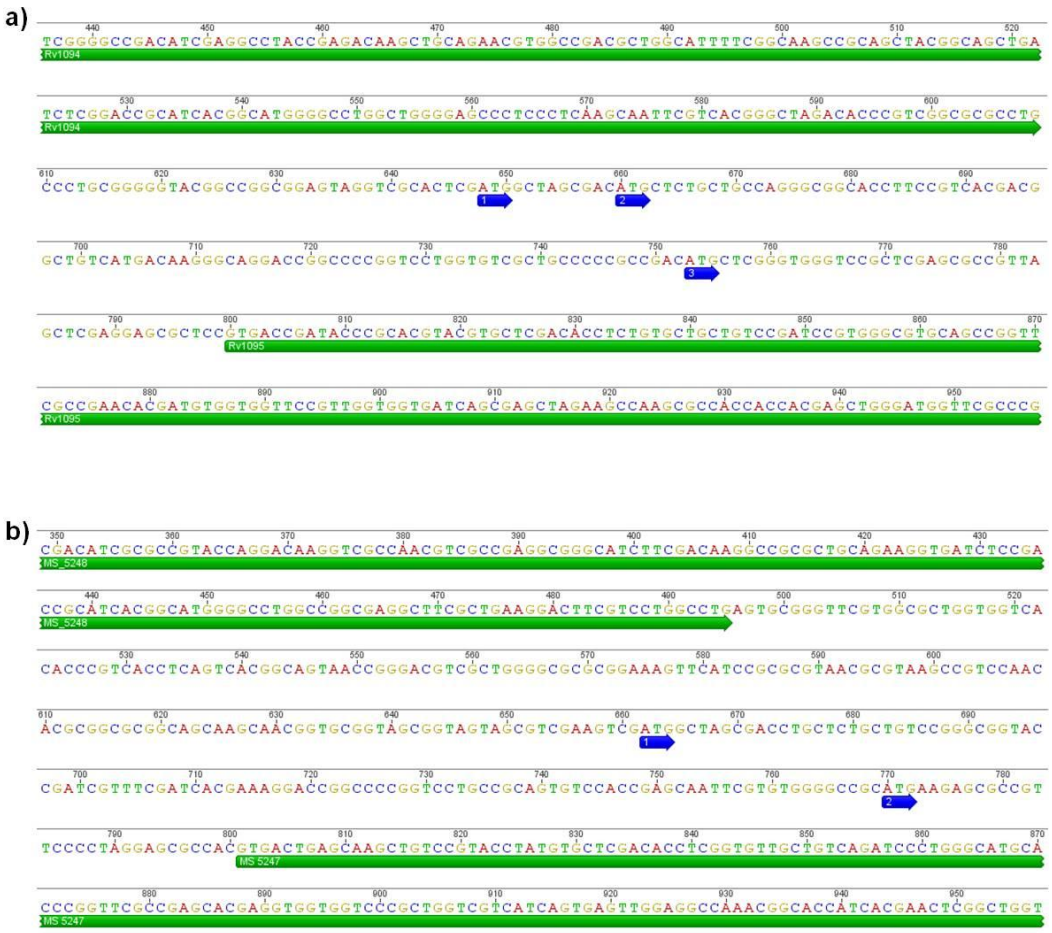

Figure 3.6 Position of possible alternative start sites upstream of *Rv1095* and *MSMEG_5247*
Locations of possible alternative start sites for short potentially coding ORFs upstream of a) *Rv1095* and b) *MSMEG_5247*.

This raised the possibility of a coding region that is not currently annotated in either genome, and the possibility of a gene that may code for a protein of similar nature to an antitoxin protein. A BLAST search of the non-redundant protein sequence database revealed that the sequence taken from the farthest alternative start site (1) in *M. tuberculosis* is highly conserved in closely related mycobacterial species (Figure 3.7).

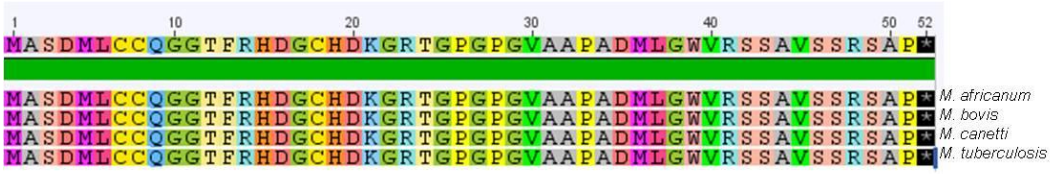

Figure 3.7 Amino acid sequence alignment of mycobacterial sequences from alternative start site (1)
Amino acid alignment from the alternative start site to the stop site 1-2 amino acids in from the annotated PhoH2 start site from *Mycobacterium africanum*, *M. bovis*, *Mycobacterium canetti* and *M. tuberculosis*. Alignment was generated in GeneiousPro (V 5.1.7) using ClustalW.

Across other mycobacterial species, sequence homology from this alternative start site in *M. tuberculosis* is observed for *Mycobacterium marinum* and *Mycobacterium ulcerans* 80 %, *M. leprae* 72 %, and *Mycobacterium avium* 77 % sequence identity. This further suggests the likelihood of a protein coding transcript. This sequence conservation was limited to the mycobacterium family. In some genomes this ORF is annotated as protein coding, such as for *M. marinum* (MMAR_4372) (Figure 3.8).

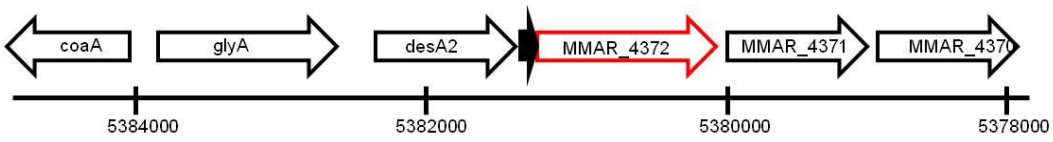

Figure 3.8 Genomic context of *MMAR_4342*
Solid black arrow indicates the position of annotated *MMAR_4373* ORF and the red arrow indicates the position of the *phoH2* gene. Adapted from Xbase (http://www.xbase.ac.uk/).

The sequence directly upstream of *MSMEG_5247* from *M. smegmatis* is 69 % similar to the equivalent sequence in *M. tuberculosis* (Figure 3.9). This may be a reflection of the genetic differences between *M. tuberculosis* and

M. smegmatis which is likely due to pathogenic differences between these two organisms.

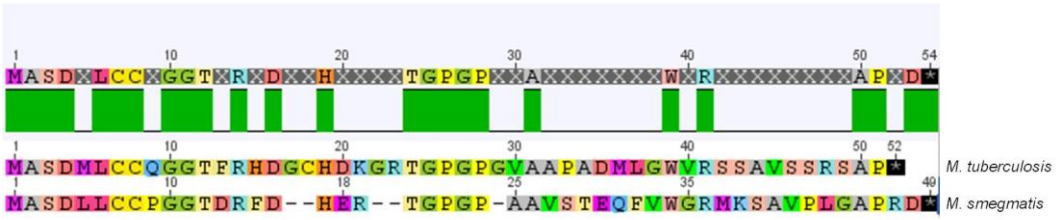

Figure 3.9 Amino acid sequence alignment of *M. tuberculosis* and *M. smegmatis* upstream sequences
Amino acid alignment from the alternative start site to the stop site 1-2 amino acids in from the annotated PhoH2 start site from *M. tuberculosis* and *M. smegmatis*.

These observations prompted the analysis of transcription of *phoH2* genes from *M. tuberculosis* and *M. smegmatis*.

3.2 Transcriptional context of *phoH2*

3.2.1 Introduction

PIN-domain proteins are commonly expressed as part of an operon with an antitoxin gene located upstream of the gene for the PIN-domain. In these systems, expression and translation of the antitoxin protein is required to offset the potentially toxic activity of the PIN-domain protein (Arcus et al. 2011). In the case of PIN-domain proteins expressed as part of a domain fusion, it is unknown whether there is a requirement for a preceding antitoxin-like element for successful protein expression, translation or regulation, and whether the domain fusion is enough to compensate for the potentially toxic activity of the PIN-domain.

The remainder of this chapter will focus on the investigation of transcription of *Rv1095* and *MSMEG_5247*, and the effect of conditional PhoH2 protein expression on *M. smegmatis* mc^{2}155 cell growth and viability.

RT-PCR and 5'RACE are two methods that can be used to study mRNA transcripts. RT-PCR is a method used to amplify cDNA copies of RNA (Sambrook & Russell 2001). The use of random hexanucleotides in the amplification of cDNA leads to the production of fragments of the entire

population of RNA and ensures a uniform representation of all the RNA sequences. These products can then be analysed using specific primers in order to map mRNA transcripts.

Once the approximate mRNA transcript length has been obtained, to determine the exact transcriptional start site of an mRNA transcript, 5'RACE can be used. By tailing cDNA that has been generated using a gene specific primer with dATP, and subsequent PCR amplification and cloning with a dT oligo plus a second gene specific primer, the 5' terminal sequence can be identified.

To study the effect of protein expression on wild type growth and viability conditional expression experiments can be performed. Conditional expression vectors allow for lower levels of protein expression under the control of an inducible promoter.

The conditional overexpression of PIN-domain proteins, such as VapC from *M. smegmatis* has been shown to result in reduced cell growth and viability (Robson et al. 2009). To determine whether PhoH2 from *M. tuberculosis* and *M. smegmatis* had a similar effect when overexpressed in wild type *M. smegmatis* mc^2155, inducible expression constructs were made that contained the *phoH2* gene from *M. smegmatis* and *M. tuberculosis* under the control of a tetracycline-inducible promoter using the pMIND vector. In order to investigate any effects of cloning from the alternative start site, for the short ORF upstream of *phoH2*, the $Rv1095_{alt}$ ($phoH2_{alt}$) sequence from *M. tuberculosis* was also cloned into pMIND.

3.2.2 Results

3.2.2.1 RT-PCR

RT-PCR was used map the mRNA transcripts associated with *Rv1095* and *MSMEG_5247* from *M. tuberculosis* and *M. smegmatis*. Using 6 primer combinations that spanned the upstream region of each *phoH2* gene (Figure 2.1) both *Rv1095* and *MSMEG_5247* were shown to be transcribed as part of a longer mRNA transcript (Figure 3.10 and Figure 3.11). With *Rv1095*, each

primer combination that covered the possible alternative start sites (Figure 3.6) and extended into the preceding gene, with the exception of primer combination (1) gave a positive result for an mRNA transcript (Figure 3.10a and c) suggesting a leader length of approximately 569 bases (Figure 3.10b). With *MSMEG_5247*, each primer combination with the exception of (1) and (2) gave a positive +RT result (Figure 3.11a and c). These results suggested a leader length of approximately 548 bases (Figure 3.11b) for *M. smegmatis*.

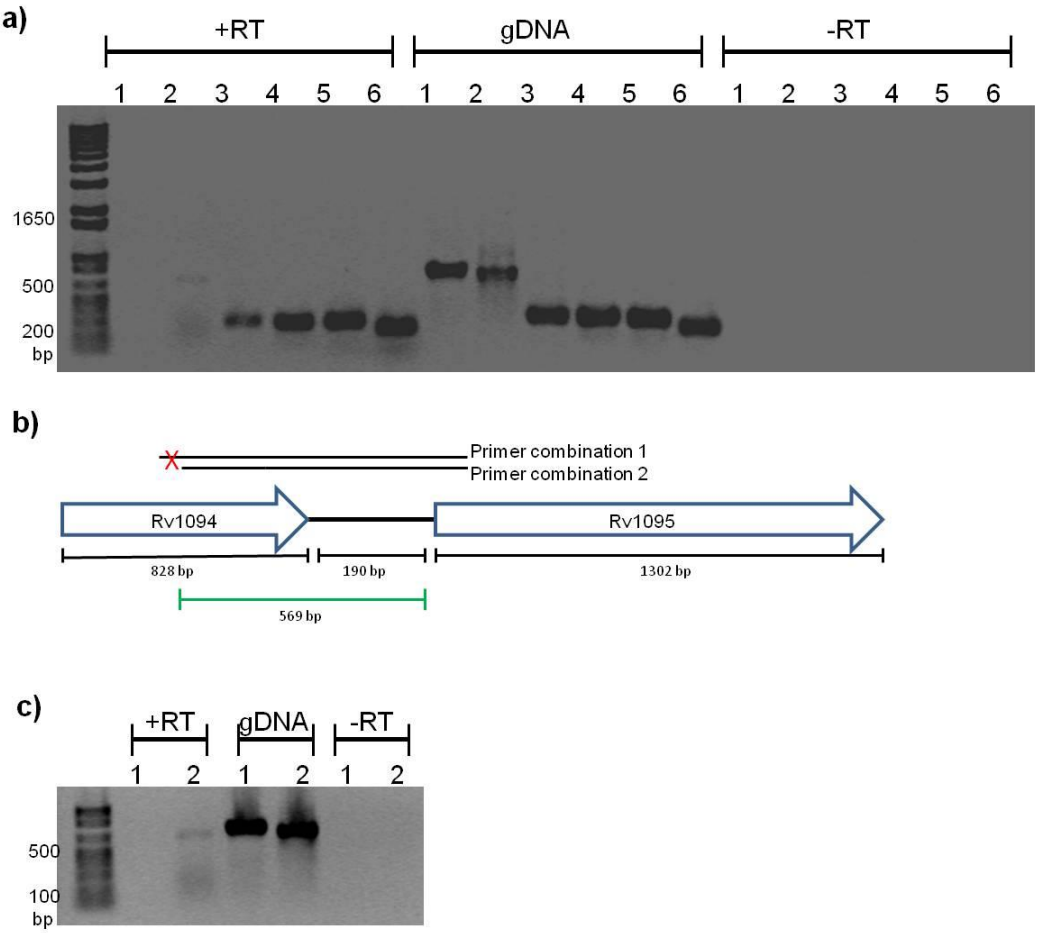

Figure 3.10 Length of *Rv1095* leader
a) RT-PCR results with each of the 6 RT primer combinations (1-6). (+RT) cDNA, (gDNA) genomic DNA and (–RT) absence of RT during cDNA synthesis, used as template for PCR reactions (gDNA to confirm PCR product sizes and –RT to exclude DNA contamination during cDNA synthesis). b) Schematic of primer positions that yielded a positive (2) and negative (1) result for an mRNA transcript. c) Repeat test of primer sets 1 and 2 confirming absence of +RT product with primer combination 1.

The +RT PCR product with primer combination 2 may represent partial primer binding due to limited overlap between the primer and the cDNA template. This may be caused by the lack of cDNA template available for the 5' end of the primer to bind to as the 5' end of the mRNA transcript may fall within the primer sequence, this ultimately leading to a weak PCR product.

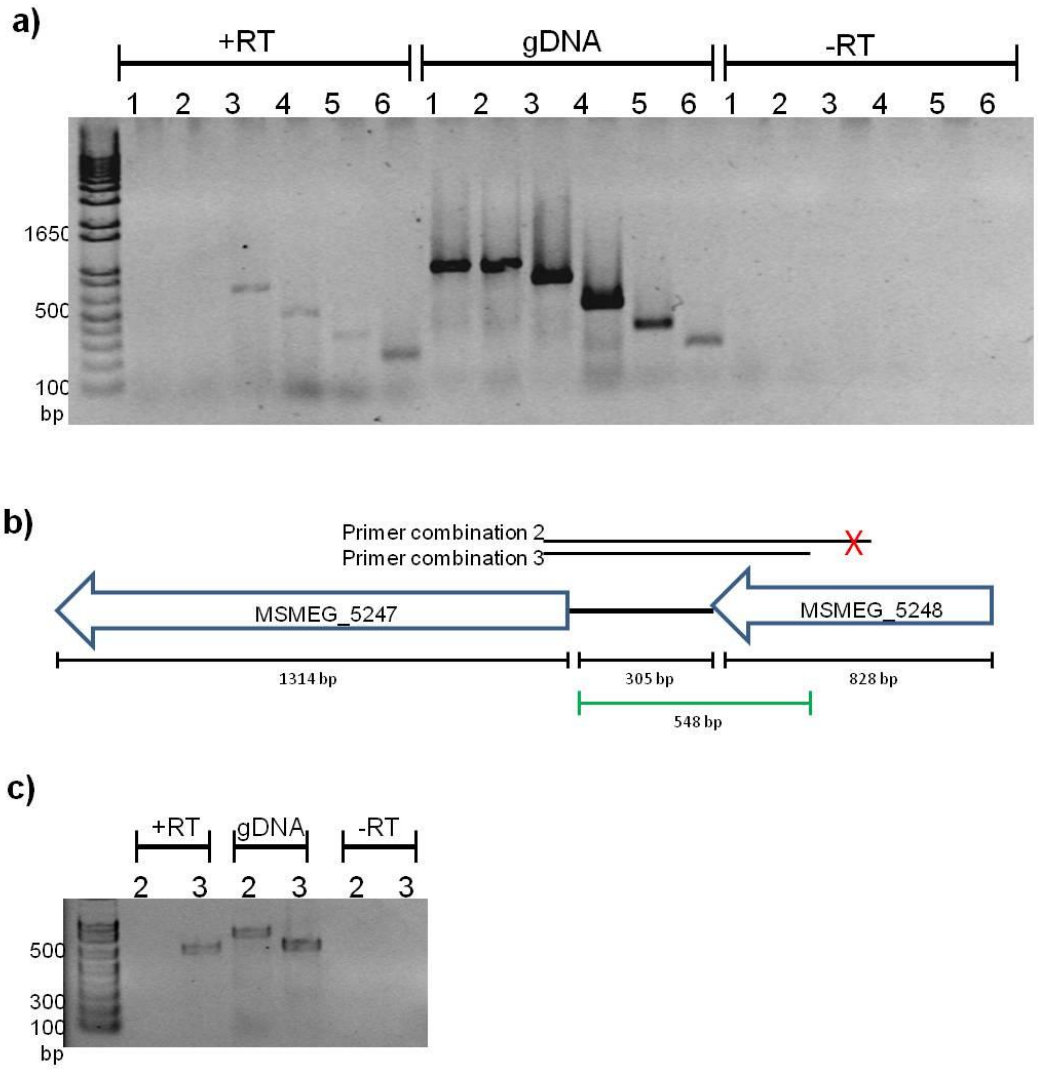

Figure 3.11 Length of *MSMEG_5247* leader
a) RT-PCR results with each of the 6 RT primer combinations (1-6). (+RT) cDNA, (gDNA) genomic DNA and (–RT) absence of RT during cDNA synthesis, used as template for PCR reactions (gDNA to confirm PCR product sizes and –RT to exclude DNA contamination during cDNA synthesis). b) Schematic of primer positions that yield a positive (3) and negative (2) result for an mRNA transcript. c) Repeat test of primer sets 2 and 3 confirming absence of +RT product with primer combination 2.

Given the positive result for a long mRNA transcript associated with both *phoH2* genes, a closer examination of the mRNA transcripts (+RT products) was undertaken to identify any distinguishing transcriptional elements.

In prokaryotes it is common for a hexameric sequence to be located around the -10 region upstream of the transcriptional start point (TSP). This promoter element is known as a Pribnow box. In mycobacteria it has been identified that these promoter sequences are dissimilar from those found in *E. coli* (Bashyam et al. 1996). It was found by Bashyam (1996) that

mycobacterial Pribnow boxes contain conserved nucleotides at specific positions, for *M. tuberculosis* the sequence T 80 %, A 90 %, Y 60 %, G 40 %, A 60 % and T 100 %, and for *M. smegmatis* T 100 %, A 93 %, T 50 %, A 57 %, A 43 % and T 71 % was derived by aligning mycobacterial promoter sequences on the basis of their TSP, where Y is a pyrimidine. It is evident that most conservation lies with positions 1, 2, and 6, the other positions are thought to be more variable due to the high GC content of mycobacterial DNA (Bashyam et al. 1996). It is further reported that no single -35 sequence in mycobacteria is identical suggesting that mycobacteria can tolerate a greater variety of sequences at this position compared with *E. coli* (Bashyam et al. 1996). The -35 sequences in mycobacteria were investigated by (Agarwal & Tyagi 2006), here it was found that a specific -35 sequence increased RNAP activity (TTGCGA) and that 18 nucleotides between the -10 and -35 sites was more optimal for promoter activity than 17 nucleotides which is favoured by *E. coli*.

Figure 3.12 outlines the hypothetical -10 and -35 sites of both *phoH2* and *alt* genes from *M. tuberculosis* and *M. smegmatis*. The -10 and -35 sites upstream of the *phoH2* TSP are dissimilar between *M. tuberculosis* and *M. smegmatis*. However, there is conservation of a GGAG Shine Delgarno sequence at the -10 position. In contrast, the -10 sites upstream of the *alt* gene TSP are identical and located in the exact same position in these two organisms, 8 nucleotides from the TSP. Further, these -10 sites fulfil the consensus 1st, 2nd and 6th positions (T, A and T) identified by Bashyam et al. (1996). The -35 sites do not match the sequence found to increase promoter activity, and these sites are dissimilar between these two organisms.

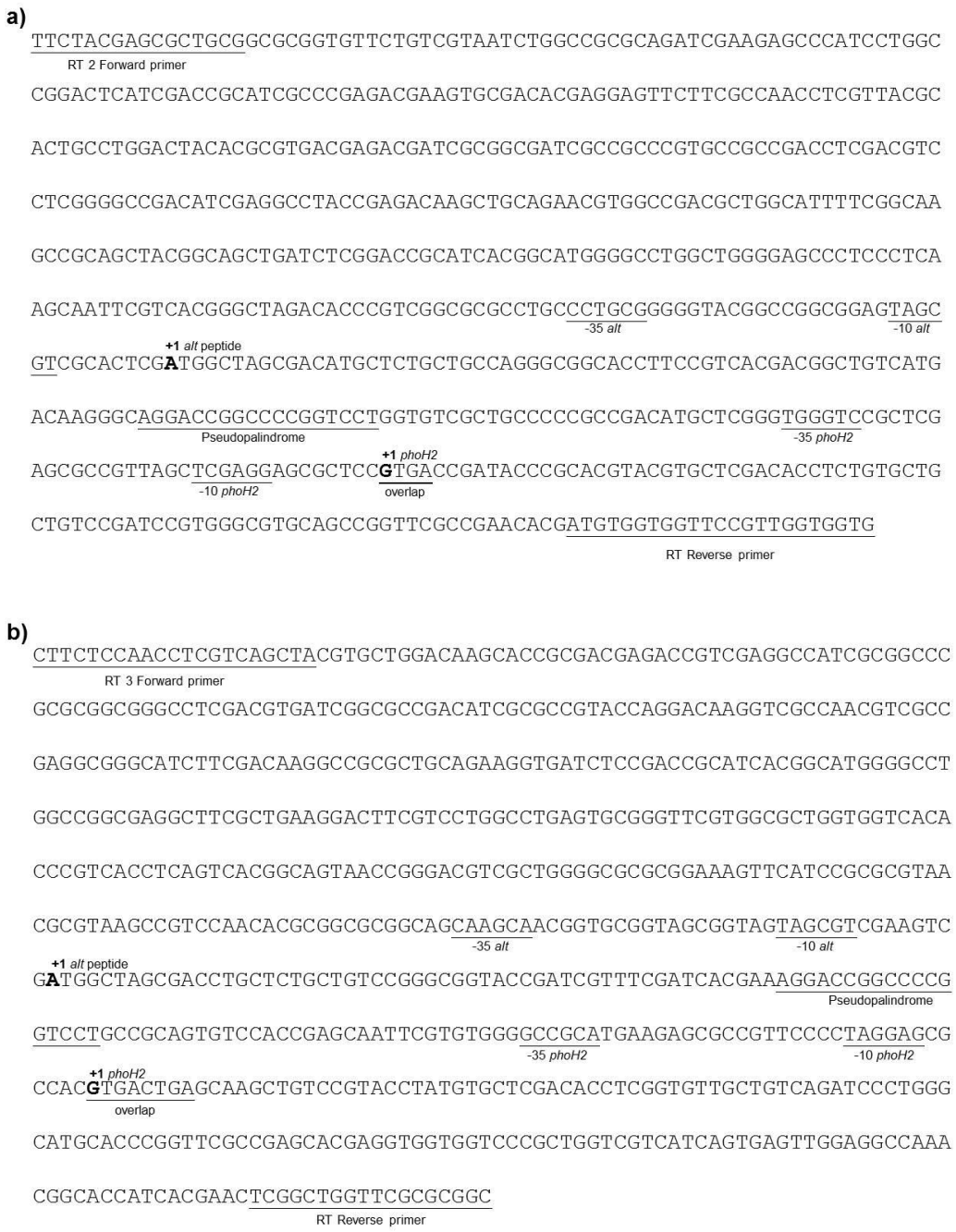

Figure 3.12 Putative promoter elements of *phoH2* from *M. tuberculosis* and *M. smegmatis*

mRNA transcripts found to be associated with *phoH2* from a) *M. tuberculosis* and b) *M. smegmatis* (+RT product) were analysed for distinguishing putative promoter elements. The underlined nucleotides at the beginning and end of each sequence represent the sites of the RT primers that yielded the +RT product with RT-PCR. The further underlined nucleotides represent putative promoter elements (-10 and -35) and the site of the pseudopalindromic sequence identified by Kazakov (2003). The TSP is indicated (+1, bold) for both the *phoH2* and *alt* genes.

These mRNA transcripts were also analysed for consensus sigma factor sequences (Rodrigue et al. 2006). None were identified in either sequence.

3.2.2.2 5'RACE

Many attempts were made to use 5'RACE as an approach to determine the actual transcriptional start site of *phoH2* from both *M. tuberculosis* and *M. smegmatis*. Efforts with Roche 5'RACE kit as per kit instructions and with the steps optimised as required were exhaustive but unsuccessful likely due to the extensive secondary structure within the GC rich upstream region. RNA secondary structure has been shown to cause strong hindrance of reverse transcription and result in the build up of spurious products or inhibition of reverse transcriptase activity (Tuerk et al. 1992; Olsen et al. 1994).

3.2.2.3 Conditional protein expression

To determine the biological effect of PhoH2 protein expression in *M. smegmatis*, conditional expression constructs that contained *phoH2* genes from *M. tuberculosis* and *M. smegmatis* ($pMIND_{Rv1095}$, $pMIND_{Rv1095alt}$ and $pMIND_{MSMEG\_}5247$) were cloned into tetracycline inducible pMIND conditional expression vector and transformed into *M. smegmatis* mc^2155. This enabled the effect of PhoH2 and $PhoH2_{alt}$ protein expression on the growth and viability of *M. smegmatis* mc^2155 to be examined by monitoring OD_{600} and CFUs respectively.

The conditional overexpression of $pMIND_{MSMEG\_5247}$ (PhoH2 from *M. smegmatis*) had no effect on the growth or viability of *M. smegmatis* (Figure 3.13). The expression of $pMIND_{Rv1095}$ (PhoH2 from *M. tuberculosis*) however caused less growth and reduced cell viability, which was reversed by the presence of the alt peptide ($pMIND_{Rv1095alt}$/$PhoH2_{alt}$). This suggests a negative effect of protein expression by $pMIND_{Rv1095}$ (PhoH2) that is alleviated by the alt peptide.

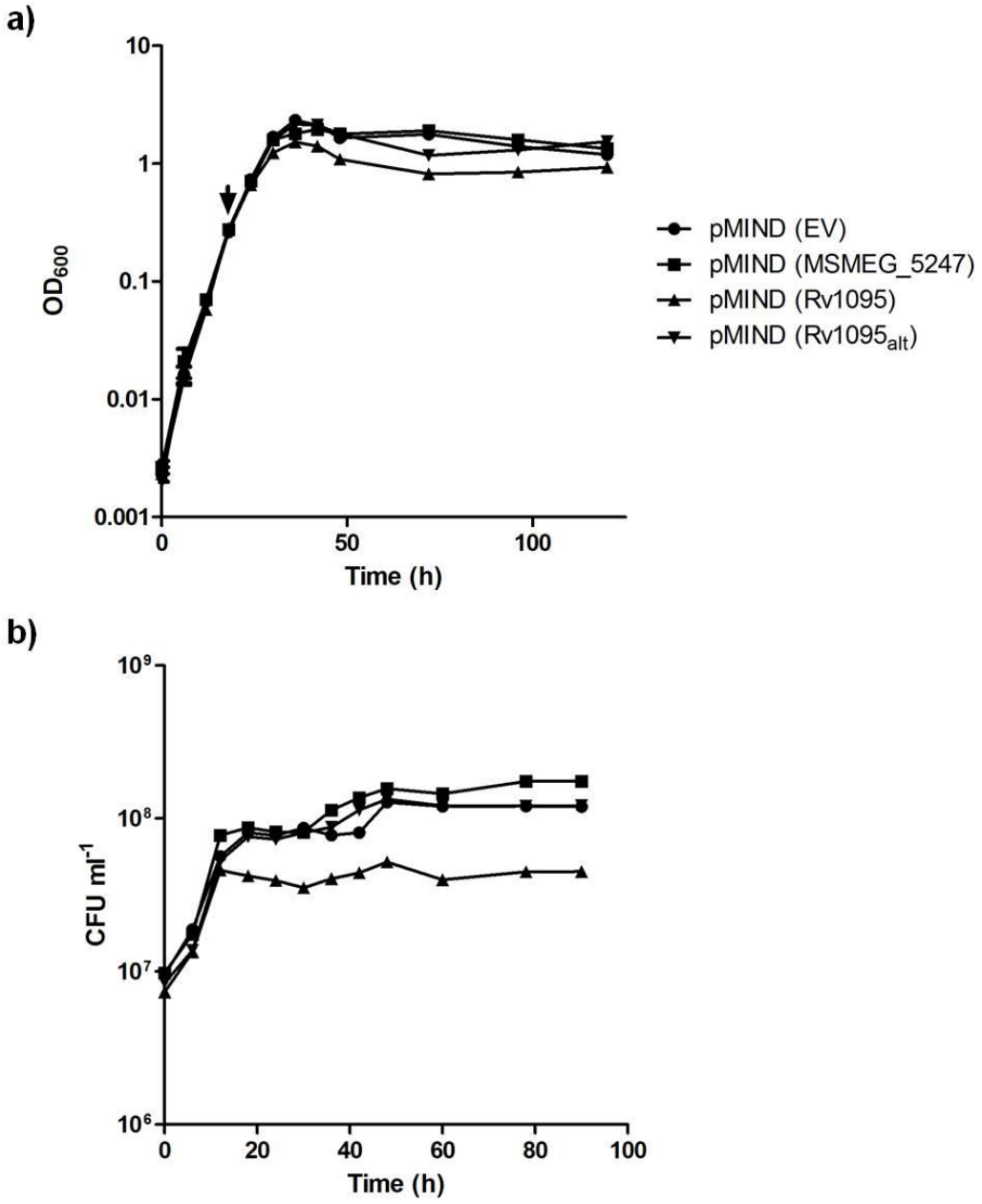

Figure 3.13 The effect of conditional expression from $pMIND_{EV}$, $pMIND_{MSMEG\_5247}$, $pMIND_{Rv1095}$, and $pMIND_{Rv1095alt}$ on cell growth and viability of *M. smegmatis* mc^2155

Cellular growth and the effect of protein expression were examined for *M. smegmatis* mc^2155 strains harbouring pMIND constructs $pMIND_{EV}$ (empty vector), $pMIND_{MSMEG\_5247}$, $pMIND_{Rv1095}$, and $pMIND_{Rv1095alt}$ by monitoring optical density a) OD_{600} and cell viability b) CFU ml^{-1} in the presence of Tc. During early exponential phase (OD_{600} of ~0.1-0.2) protein expression was induced in all strains by the addition of tetracycline (indicated by black arrow). Graphs are shown from data from one experiment of three cultures for each strain. For OD_{600} the results are shown as the mean and $^{+/-}$SD of the three cultures. The viability was determined by CFUs from point of induction of expression by the addition of tetracycline for 96 hours, the results are shown as the mean and $^{+/-}$SD of three technical replicates taken at each time point from each of the three cultures.

Culture samples that had been induced with tetracycline were also spotted onto LBT plates supplemented with hygromycin B only (-Tc) to look for

differences between colonies grown on plates containing tetracycline and those without tetracycline. If there is an effect of protein expression on growth and colony formation it should be apparent on those plates containing tetracycline as PhoH2 protein expression should be maintained. If PhoH2 protein expression is bactericidal, growth on the plates that do not contain tetracycline should be similar to the growth on plates that contain tetracycline. If PhoH2 protein expression is bacteriostatic (reversible) then culture samples grown on plates that do not contain tetracycline should adopt 'normal' colony size. Figure 3.14 shows that immediately after induction of the cultures with tetracycline no differences are observed between those colonies grown in the presence (+Tc) or absence (-Tc) of tetracycline (0 hours panel). For *M. smegmatis* mc^2155 strain expressing $pMIND_{Rv1095}$, culture samples plated out after 72 hours of growth in the presence of tetracycline, onto agar plates supplemented with tetracycline (+Tc) showed smaller pinpoint-sized colonies compared with the normal colony growth observed for the same samples grown on plates in the absence (-Tc) of tetracycline (Figure 3.14c (72 hours panel)). Slight differences were observed for this strain at earlier time points, however not to the degree seen at 72 hours.

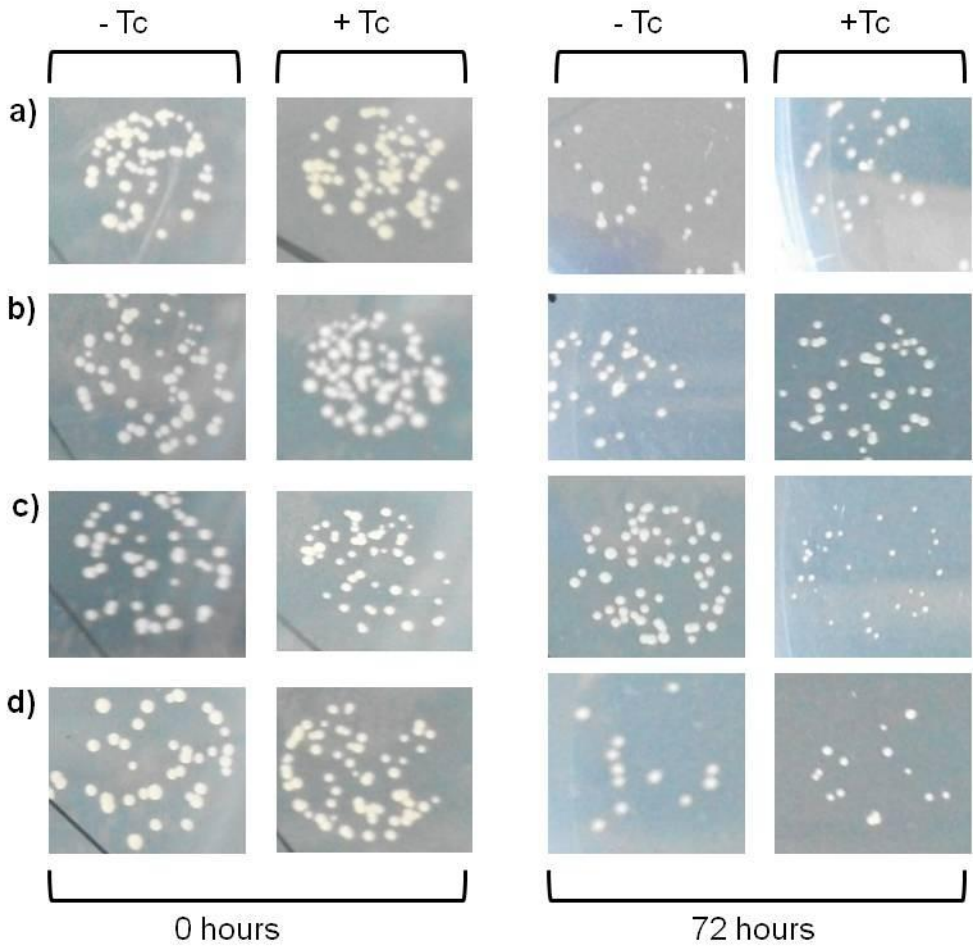

Figure 3.14 Colony morphology of *M. smegmatis* mc²155 strains harbouring pMIND$_{EV}$, pMIND$_{MSMEG_5247}$, pMIND$_{Rv1095}$, and pMIND$_{Rv1095alt}$
Colony growth on agar plates of *M. smegmatis* mc²155 harbouring pMIND constructs a) pMIND$_{EV}$ (empty vector) b) pMIND$_{MSMEG_5247}$ c) pMIND$_{Rv1095}$ and d) pMIND$_{Rv1095alt}$ in the presence (+Tc) and absence (-Tc) of tetracycline at 0 and 72 hours post-culture induction with tetracycline. Spots are representative of the 10^{-4} dilution of each culture sample.

The difference in colony morphology observed for *M. smegmatis* mc²155 pMIND$_{Rv1095}$ supports the negative effect of pMIND$_{Rv1095}$ protein expression on the viability (CFUs) of *M. smegmatis*.

3.2.3 Discussion

The genes and genomic contexts of *phoH2* in *M. tuberculosis* and *M. smegmatis* are similar. In addition, the predicted functional partners and transcriptional regulators implicate *phoH2* in the cells response to phosphate and oxidative stress as well as with virulence associated lipid biogenesis in response to redox environments. Both share a pseudopalindrome sequence

at nearly identical positions upstream and alternative start sites possibly a protein coding sequence out of frame of the *phoH2* gene.

The results of the RT experiments that aimed to map the mRNA transcript associated with *phoH2* from *M. tuberculosis* and *M. smegmatis* show that the mRNA transcript length of *phoH2* is similar between these two organisms. These results suggest the possibility of an upstream gene due to positive RT (+RT) products obtained for each primer set that spanned the possible alternative start sites. Further, the putative -10 sites upstream of both possible alt genes in *M. tuberculosis* and *M. smegmatis* are the conserved and share the consensus nucleotides identified for mycobacterial -10 sites, this was not the case for the -10 site upstream of the *phoH2* gene. The -35 sites were dissimilar in both positions which is not surprising given previous observations that no two mycobacterial -35 regions are the same (Bashyam et al. 1996).

It is possible that these results may also suggest regulation by transcription factors that are predicted to control the expression of *Rv1094* and *MSMEG_5248*, such as the predicted transcriptional regulator *Rv0494* (section 3.1.3). This regulator is thought to target *Rv1094* and *Rv1095* at positions -755 from the annotated start site of *Rv1095* and at +204 within the *Rv1095* gene. The predicted binding site at -755 is 185 nucleotides upstream of the 5' end of the +RT PCR product.

For genes that do not have a leader sequence, the AUG site is the only site recognised by the ribosome *in vivo* for translation (Kozak 2005). Long 5' leaders can adopt a number of possible roles including presentation of structural elements that confer structural stability to the overall transcript, and sequence and structural components that guide intracellular mRNA localisation. Irnov (2010) states that the average length of long 5' leaders which contain *cis*-acting elements is around 360 nucleotides in *B. subtilis.* The reported average shorter leader length of 35 nucleotides does not possess regulatory elements. It is further suggested that where 5' leader regions occur they contain specialised functions within them (Irnov et al. 2010). In

addition to *cis*-acting elements, small ORF's located near the 5' of the mRNA can down regulate expression of the mRNA transcript. These can be overcome by changes in mRNA structure (Kozak 2005). Certain RNase enzymes have been shown to interact with their 5' leader and in so doing, autoregulating their own expression (Schuck et al. 2009). It is possible that PhoH2 acts on its 5' leader region and that short alternative 5' ORF's are involved with protein expression. These hypotheses are further investigated in Chapter 4 and 5.

Conditional expression experiments which examined the biological effect of PhoH2 protein expression in *M. smegmatis* showed a negative effect of *Rv1095* protein expression that was alleviated by the presence of the alt peptide and reversed in the absence of tetracycline (-Tc) indicating the effect was bacteriostatic in nature. With the *M. smegmatis* mc^2155 strain expressing the $Rv1095_{alt}$ construct from *M. tuberculosis* (*phoH2* cloned from alternative start site 1), no effect on growth or viability was observed. This may suggest that this construct includes transcriptional and/or translational regulatory factors within the sequence upstream of the annotated start site of the *phoH2* gene that are responsible for 'normal' protein expression and translation. Conversely, there may be an inhibitory role of a possible protein or transcript encoded in this region that is offsetting the activity of the PhoH2 protein analogous to the toxin-antitoxins. It is apparent that the fusion between the PIN and PhoH-domain does not play a compensatory role with regard to its toxic activity.

For *M. smegmatis* mc^2155 strain containing $pMIND_{MSMEG\_5247}$ growth and viability were similar to wild type *M. smegmatis* containing $pMIND_{EV}$ (empty vector). Protein expression did not have a negative effect as seen for *Rv1095*, despite the similarity between the two proteins from these two mycobacterial organisms. It is likely that *MSMEG_5247* may have its toxicity masked by compensation by endogenous levels of *MSMEG_5247* antitoxin (equivalent upstream sequence to that of $phoH2_{alt}$), due to expression in *M. smegmatis* mc^2155. Alternatively, these two proteins may differ in their modes of regulation.

3.2.4 Conclusions

This investigation of the transcriptional context of *phoH2* from *M. tuberculosis* and *M. smegmatis* has revealed both similarities and differences between these two organisms. My results have shown that both genes are transcribed as part of a long, extensively structured mRNA transcript which is likely to contain regulatory elements or short 5'ORFs involved in successful expression, regulation and translation of the PhoH2 protein. In addition *Rv1095* protein expression had a negative effect on the growth of *M. smegmatis* mc^2155 which is bacteriostatic in nature as in the absence of inducer (tetracycline) 'normal' colony growth was resumed. This negative effect was alleviated by the presence of the alt peptide as it was not observed with the expression of $Rv1095_{alt}$ ($phoH2_{alt}$) or *MSMEG_5247* from *M. smegmatis*, which raises the possible role of elements present in the short sequence directly upstream of *Rv1095* that may be involved with protein expression, translation or mRNA transcript regulation.

Chapter Four: Protein cloning, expression, purification & crystallisation

4.1 Introduction

4.1.1 Protein expression in *M. smegmatis*

Expression of mycobacterial proteins in *E. coli* can be challenging due to the formation of insoluble protein often found in inclusion bodies. Previous work in our lab showed that often mycobacterial proteins express insolubly in *E. coli*. One approach to overcome the expression of insoluble protein is to express the protein in its host or in a host that is more closely related to the organism from which the protein is derived. *M. smegmatis* is a non-pathogenic mycobacterial species which is relatively fast growing compared with *M. tuberculosis*, with a doubling time of ~3 hours (Snapper et al. 1990). The *M. smegmatis* overexpression strain mc²4517, in combination with the traditional mycobacterial expression vector pYUB1049, can yield 7 mg of soluble protein per litre of culture (Bashiri et al. 2007). This strain of *M. smegmatis* can be grown in autoinduction medium using methods established for *E. coli*, which negates the need to monitor optical density for induction and also reduces the cost of culture by eliminating the need for albumin dextrose catalase supplement (ADC) enrichment traditionally used for mycobacterial culture growth (Bashiri et al. 2007).

4.1.2 Protein expression using Gateway®

A cloning system which can help to overcome the obstacles associated with recombinant *E. coli* protein expression is Gateway. The expression vector $pDEST_{SMG}$ is adapted from the mycobacterial shuttle vector pYUB1049 (Goldstone et al. 2008) and allows for ease of cloning associated with the Gateway system. It also makes use of protein expression in the *M. smegmatis* mc²4517 expression strain. In this system BP clonase is used to catalyse the recombination of the gene insert amplified with flanking *att*B sites and the donor vector which inserts the gene into the *att*P recombination sites. The gene can then be transferred to the destination vector using LR clonase which recombines and inserts the gene into the *att*R recombination sites.

With the cloning of genes initially into pDONR$_{221}$ and subsequently into a destination vector, proteins can be tested for expression in an alternative host which can be useful for further applications.

4.1.3 Protein refolding

If the issue of insoluble protein expression cannot be solved through the use of different expression constructs one approach to recover insoluble protein is through denaturation and renaturation. Chaotropic agents such as urea or guanidine HCl, thiocyanate salts or detergents like SDS can be used to denature proteins (De Bernardez 1998). Generally, 6-8 M urea or 6-7 M guanidine HCl is necessary to unfold and solubilise proteins (Tsumoto et al. 2003). A number of methods are available to promote protein refolding including one-step dialysis, step-wise dialysis, gel filtration, dilution, mixing and solid phase refolding (De Bernardez 1998 ; Tsumoto et al. 2003). With one-step dialysis, denatured protein samples are dialysed against a refolding buffer. Over the course of dialysis the concentration of denaturant decreases, promoting the folding of the protein into intermediate and native structures (Tsumoto et al. 2003). Renaturation buffers need to promote the correct formation of disulphide bonds. Reagents such as glutathione (GSH), glutathione disulphide (GSSG) and dithiothreitol (DTT) are commonly used as oxidising and reducing agents to encourage disulphide bonds to break and reform.

4.1.4 Protein solubilisation with detergents

As well as being used to fully denature proteins prior to refolding, detergents can also be used to disrupt protein complexes for solubilisation from inclusion bodies, and can leave the disrupted protein in a more ordered state (Tsumoto et al. 2003). It has been reported that detergent N-lauroylsarcosine in combination with triton X-100 can solubilise *E. coli* expressed GST-fusion proteins, vascular endothelial growth factor and GST-pyruvate kinase muscle 2, without the need for further renaturation (Zou et al. 2009; Park et al. 2011). It is possible that other protein constructs may be able to be solubilised in this manner.

4.1.5 Protein crystallisation and Electron Microscopy

Once soluble protein has been obtained the determination of the three-dimensional (3D) structure of a protein can provide insights into the molecular organisation and conformation. Nuclear magnetic resonance spectroscopy (NMR) and X-ray crystallography are two approaches used to gain high resolution protein structures. NMR is limited to smaller protein complexes in solution (<30 kDa), whereas X-ray crystallography is not limited by protein size but does rely on the growth of single protein crystals suitable for diffraction. Electron microscopy (EM) is a tool that can be used to, at low resolution, gain visualisation of the conformation and oligomeric state of proteins. EM generally relies on a uniform protein sample in solution.

4.1.6 Mass spectrometry for protein analysis

Mass spectrometry (MS) is a comprehensive and versatile tool for the study of proteins. The development of soft ionisation techniques, such as electrospray ionisation (ESI) and matrix-assisted laser desorption/ionisation (MALDI) enabling the ionisation of both proteins and peptides, has revolutionised the use of MS in proteomics (Han et al. 2008). The MALDI-TOF MS at the University of Waikato is generally used for protein identification after an in-gel proteolytic digestion. The MALDI TOF/TOF MS is used for this application as well and further allows for identification of sequence tags by collision induced dissociation (CID) MS/MS (CPR, University of Otago). LTQ-Orbitrap MS makes use of either MALDI or ESI as an ionisation source, has a higher mass accuracy and is generally used in top down proteomics and for protein identification from complex peptide mixtures (CPR, University of Otago) (Han et al. 2008).

This chapter will focus on the cloning, expression, purification and investigation of PhoH2 proteins from *M. tuberculosis*, *M. smegmatis* and thermophilic homologue from *Thermobispora bispora* for crystallisation, characterisation and functional studies. A thermophilic PhoH2 protein was chosen as due to its thermophilic origins, it should have increased stability and may be able to be expressed and purified in higher yields than mycobacterial proteins.

4.2 Results

For functional and crystallographic studies, soluble PhoH2 protein needed to be obtained. To gain soluble protein a number of PhoH2 protein constructs were made to try to gain PhoH2 as a whole, and as two separate domains (PIN-domain and PhoH-domain) from both *M. tuberculosis* and *M. smegmatis*.

4.2.1 Cloning into pYUB1049, pDONR221, $pDEST_{SMG/17}$, pYUB28b, and pET28b-*Pst*I

Each *phoH2* gene or sequence variant was successfully cloned into its respective vector for protein expression as listed in Table 4. 1.

Table 4. 1 Genes cloned for protein expression and predicted protein MW

Vector	His-tag position	Gene	Expected MW (Da)
pYUB1049	C	*Rv1095*	49274.0
		$Rv1095_{alt}$	54445.9
	N	*Rv1095*	49176.0
		$Rv1095_{alt}$	54216.6
$pDONR_{221}$ $pDEST_{SMG}$	N	*Rv1095*	49432.3
		$Rv1095_{alt}$	54604.2
		PIN-domain	17203.4
		PhoH-domain	25957.8
$pDONR_{221}$ $pDEST_{17}$	N	*Rv1095*	49574.5
		PIN-domain	17345.6
		PhoH-domain	26100.0
pYUB28b	C	*Rv1095*	48532.4
		PIN-domain	16303.5
		PhoH-domain	25057.9
		$PIN\text{-}domain_{alt}$	21344.1
		MSMEG_5247	48657.5
		$MSMEG\_5247_{alt}$	53728.2
		PIN-domain	16554.8
		$PhoH\text{-}domain_{+1}$	25116.9
		$PhoH\text{-}domain_{+2}$	25388.2
pET28b-*Pst*I	C	*TBIS_3092*	49285.5

MW was calculated using ProtParam (http://web.expasy.org/protparam/).

For mycobacterial protein expression using pYUB1049 or pYUB28b and restriction based cloning, the ligation reactions were first transformed into electrocompetent *E. coli* TOP10 cells as this strain allowed for selection of transformants with hygromycin B and ease of vector purification for sequencing prior to transformation into *M. smegmatis* $mc^2$4517 cells. Purification of vector DNA is challenging with mycobacterial strains due to

the nature of their cell wall which is rich in mycolic acids and is generally thicker than the cell wall of other bacteria (Bhamidi 2009).

Successful insertion of genes cloned with each cloning system was confirmed by sequencing vector DNA from positive *E. coli* transformants with T7 or M13 forward and reverse primers that flanked the multiple cloning site. This provided sequences in both the forward and reverse direction.

4.2.2 Small scale expression tests

Small scale expression tests were used to evaluate protein expression and his-tag binding affinity. Protein expression of PhoH2 and sequence variants in *M. smegmatis* had mixed success summarised in Table 4.2.

Table 4.2 Summary of small scale expression tests and his-tag binding

Vector	His-tag	Gene	Protein	None	Insol	Sol
pYUB1049	C	*Rv1095*	$PhoH2_{MTB}$	**X**		
		$Rv1095_{alt}$	$PhoH2_{alt\ MTB}$			X
	N	*Rv1095*	$PhoH2_{MTB}$	**X**		
		$Rv1095_{alt}$	$PhoH2_{alt\ MTB}$	**X**		
$pDONR_{221}$ $pDEST_{SMG}$	N	*Rv1095*	$PhoH2_{MTB}$	**X**		
		$Rv1095_{alt}$	$PhoH2_{alt\ MTB}$	**X**		
		PIN-domain	PIN_{MTB}	**X**		
		PhoH-domain	$PhoH_{MTB}$		X	
$pDONR_{221}$ $pDEST_{17}$	N	*Rv1095*	$PhoH2_{MTB}$	**X**		
		PIN-domain	PIN_{MTB}		X	
		PhoH-domain	$PhoH_{MTB}$		X	
pYUB28b	C	*Rv1095*	$PhoH2_{MTB}$	**X**		
		PIN-domain	PIN_{MTB}		X	
		PhoH-domain	$PhoH_{MTB}$	**X**		
		$PIN\text{-}domain_{alt}$	$PIN_{alt\ MTB}$	**X**		
		MSMEG_5247	$PhoH2_{MSMEG}$			X
		$MSMEG\_5247_{alt}$	$PhoH2_{alt\ MSMEG}$			X
		PIN-domain	PIN_{MSMEG}		X	
		$PhoH\text{-}domain_{+1}$	$PhoH_{+1\ MSMEG}$		X	
		$PhoH\text{-}domain_{+2}$	$PhoH_{+2\ MSMEG}$		X	
pET28b-*Pst*I	C	*TBIS_3092*	$PhoH2_{TBIS}$			X

Abbreviations: No protein expression (None), Insoluble protein expression (Insol) Soluble protein expression (Sol).

Figure 4.1 shows that the cloning of *Rv1095* ($PhoH2_{MTB}$) from its annotated start site into pYUB1049 for expression in *M. smegmatis* with either a C- or N-terminal his-tag resulted in no clear protein expression neither soluble or insoluble (Figure 4.1a and b). This absence of expression prompted full

evaluation of the sequence upstream of *phoH2* for possible alternative start sites for a short ORF given the length of the leader sequence identified by RT-PCR (Figure 3.10) and the analysis for putative transcriptional elements in the +RT products of both *M. tuberculosis* and *M. smegmatis*. It was hypothesised that the leader sequence may encode regulatory elements such as a short 5'ORF which may be involved in protein expression, and that if the ORF was acting as an antitoxin protein of PhoH2, soluble protein expression should be obtained as the possible negative effects of protein expression on growth of the culture should be overcome. The sequence from alternative start site 1 (Figure 3.6), $Rv1095_{alt}$, was cloned into pYUB1049 for expression with a C- or N-terminal his-tag. Clear soluble protein expression was gained with cloning of $Rv1095_{alt}$ ($PhoH2_{alt\ MTB}$) with a C-terminal his-tag (Figure 4.1c) but not with an N-terminal his-tag (Figure 4.1d).

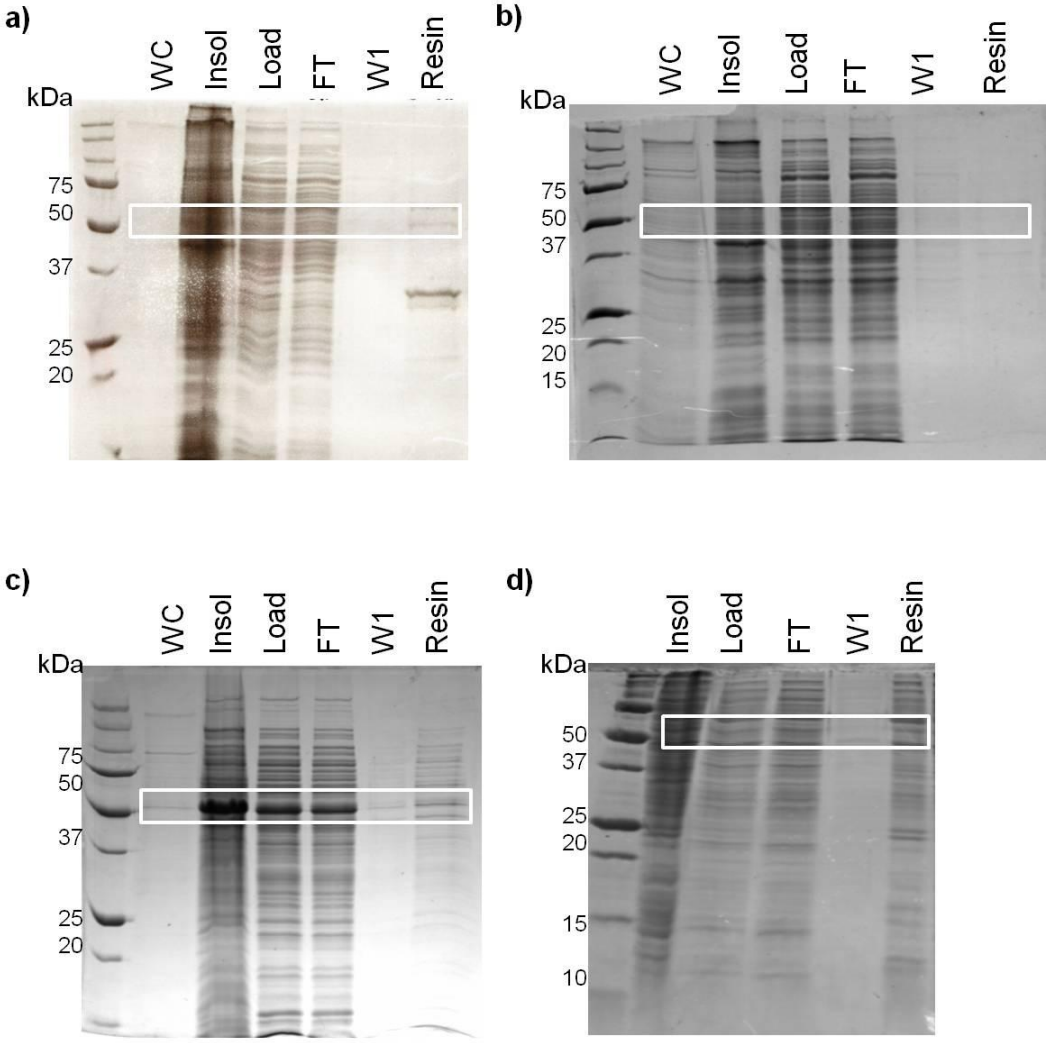

Figure 4.1 Small scale expression screen of PhoH2$_{MTB}$ and PhoH2$_{alt\ MTB}$ (pYUB1049 constructs) in *M. smegmatis*
Expression and his-tag binding of a)PhoH2$_{C\ MTB}$ b) PhoH2$_{N\ MTB}$ c)PhoH2$_{altC\ MTB}$ and d) PhoH2$_{altN\ MTB}$. White boxes indicate expected band size. Labels: whole cell fraction (WC), insoluble fraction (Insol), soluble fraction loaded onto Ni resin (Load), flow through (FT), wash one (W1), Ni resin with protein bound after wash steps (Resin).

PhoH2$_{alt\ MTB}$ was subjected to a lysis buffer screen (section 2.4.6.1) that tested for protein solubility in a range of lysis buffers with different pH and NaCl concentrations, and PhoH2$_{alt\ MTB}$ was found to be soluble in each of the buffers screened (not shown).

To confirm the requirement of cloning from the alternative start site and determine whether cloning with another system would yield further soluble protein, *Rv1095,* the *PIN* and *PhoH*-domain from *Rv1095* were cloned into Gateway pDONR$_{221}$ then into pDEST$_{SMG}$ and pDEST$_{17}$ for expression testing in

M. smegmatis and *E. coli* respectively. The $Rv1095_{alt}$ sequence was also cloned into $pDEST_{SMG}$ for expression trials in *M. smegmatis*. No clear protein expression was observed with the $pDEST_{SMG}$ constructs except with the PhoH-domain which was insoluble (Figure 4.2c). In this system, inclusion of the upstream segment did not give soluble expression. The PIN and PhoH-domain expressed insolubly in *E. coli* (Figure 4.2a and b).

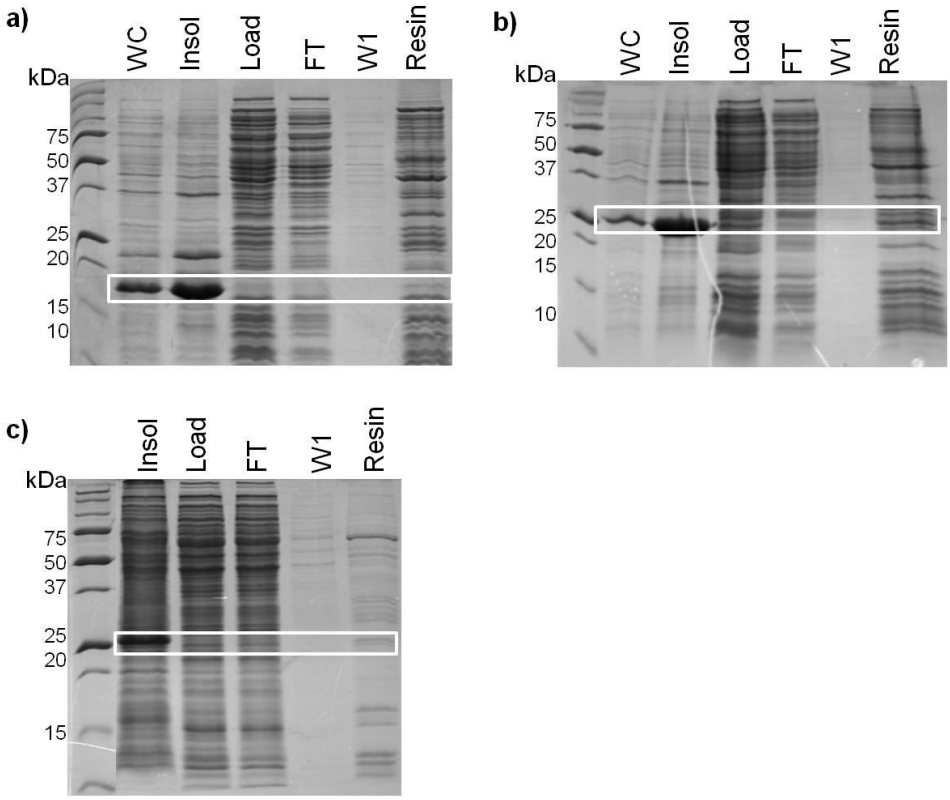

Figure 4.2 Small scale expression screen of PIN_{MTB} and $PhoH_{MTB}$ domains ($pDEST_{17}$ constructs) in *E. coli* and $PhoH_{MTB}$ domain ($pDEST_{SMG}$) in *M. smegmatis*

Expression and his-tag binding of a) $PIN_{N\ MTB}$ b)$PhoH_{N\ MTB}$ in *E. coli* and c)$PhoH_{N\ MTB}$ in *M. smegmatis*. White boxes indicate expected band size. Labels: whole cell fraction (WC), insoluble fraction (Insol), soluble fraction loaded onto Ni resin (Load), flow through (FT), wash one (W1), Ni resin with protein bound after wash steps (Resin).

The insoluble PIN and PhoH-domains (Figure 4.2) were subjected to a lysis buffer screen (section 2.4.6.1) and reduced incubation temperatures for protein expression (section 2.4.6.2) as an attempt to improve solubility. These did not remedy insoluble protein expression.

After receiving the new modified pYUB1049 vector, pYUB28b, which had the sparse MCS replaced with the MCS from pET28b, *Rv1095*, *PIN*, PIN_{alt}

(PIN-domain cloned from the alternative start site upstream of *Rv1095*) and *PhoH-domain* from *M. tuberculosis* were cloned into pYUB28b and tested for protein expression. No expression was observed with PhoH2 (*Rv1095*) and a similar observation was made for $PIN_{alt\ MTB}$ and the PhoH-domain. The PIN-domain expressed insolubly (Figure 4.3).

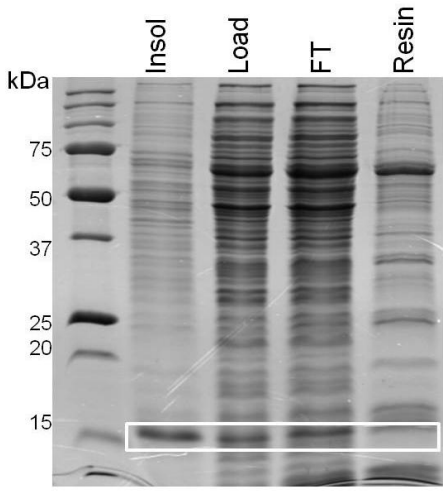

Figure 4.3 Small scale expression screen of PIN_{MTB} (pYUB28b construct) in *M. smegmatis*
Expression and his-tag binding of $PIN_{C\ MTB}$. White box indicates expected band size. Labels: insoluble fraction (Insol), soluble fraction loaded onto Ni resin (Load), flow through (FT), wash one (W1), Ni resin with protein bound after wash steps (Resin).

Given the difficulties with expressing PhoH2 from *M. tuberculosis* it was hypothesised that the gene for PhoH2 from *M. smegmatis* (*MSMEG_5247*) may be more suited for protein expression, given that the protein is to be expressed in the organism from which it is derived. Genes for *MSMEG_5247* and *MSMEG_5247$_{alt}$* from *M. smegmatis*, and the *PIN* and *PhoH-domain* ($PhoH_{+1/+2}$) from *MSMEG_5247* were cloned into pYUB28b. $PhoH_{+1/+2}$ were extended sequences of the PhoH-domain that contained 1 or 2 additional amino acids at the N-terminus obtained from the linker region between the PIN and PhoH-domain. These were cloned from *M. smegmatis* to try to improve protein solubility. Figure 4.4 shows that both $PhoH2_{MSMEG}$ (*MSMEG_5247)* and $PhoH2_{alt\ MSMEG}$ (*MSMEG_5247$_{alt}$)* from *M. smegmatis* expressed solubly. The extension of the PhoH-domain sequence did not remedy insoluble protein expression (Figure 4.4d and e). The PIN-domain also expressed insolubly (Figure 4.4c).

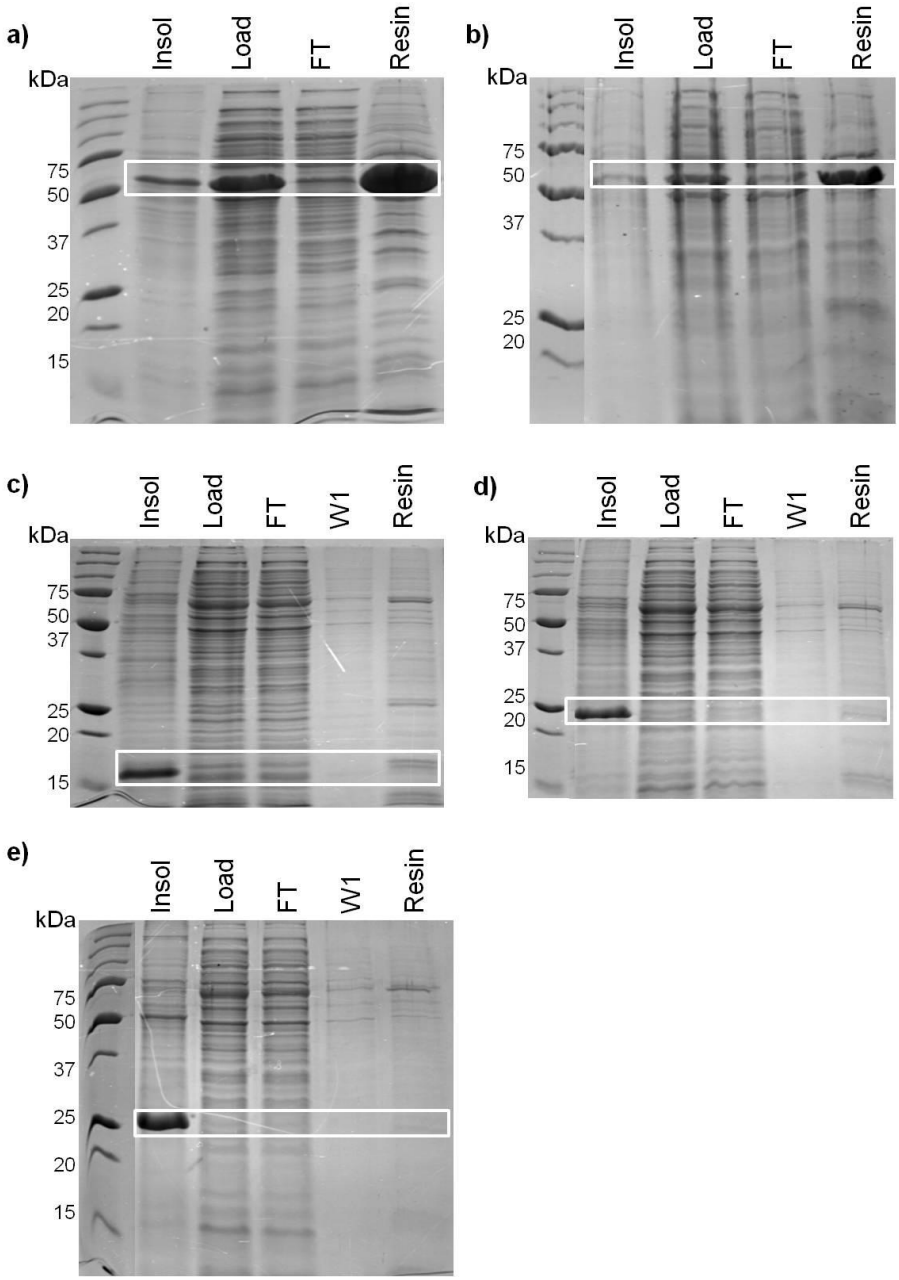

Figure 4.4 Small scale expression screen of PhoH2$_{MSMEG}$, PhoH2$_{alt\ MSMEG}$, PIN$_{MSMEG}$, PhoH$_{+1\ MSMEG}$ and PhoH$_{+2\ MSMEG}$ (pYUB28b constructs) in *M. smegmatis*
Expression and his-tag binding of a) PhoH2$_{C\ MSMEG}$ b) PhoH2$_{altC\ MSMEG}$ c) PIN$_{C\ MSMEG}$ d) PhoH$_{+1C\ MSMEG}$ e) PhoH$_{+2C\ MSMEG}$. White boxes indicate expected band size. Labels: insoluble fraction (Insol), soluble fraction loaded onto Ni resin (Load), flow through (FT), wash one (W1), Ni resin with protein bound after wash steps (Resin).

Despite gaining soluble expression of PhoH2$_{MSMEG}$ and PhoH2$_{alt\ MTB/MSMEG}$, these proteins proved difficult to work with (inconsistent expression and low protein yield). Therefore to help overcome challenges met working with PhoH2 and PhoH2$_{alt}$ from *M. tuberculosis* and *M. smegmatis*, the nucleotide sequence that coded for the amino acid sequence of a thermophilic homologue of PhoH2 in the actinobacterium *T. bispora* (*TBIS_3092*) (75.5 %

identical to PhoH2 from *M. tuberculosis*) was obtained from GeneArt® optimised for protein expression in *E. coli*. This gene was cloned into pET28b-*Pst*I and produced soluble protein expression in *E. coli* (Figure 4.5).

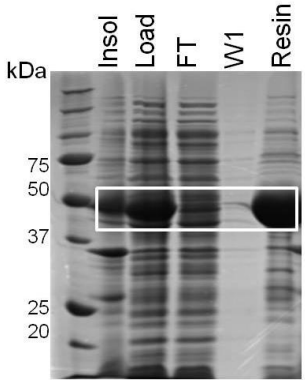

Figure 4.5 Small scale expression screen PhoH2$_{TBIS}$ (pET28b-*Pst*I construct)
Expression and his-tag binding of PhoH2$_{TBIS}$. White box indicates expected band size. Labels: insoluble fraction (Insol), soluble fraction loaded onto Ni resin (Load), flow through (FT), wash one (W1), Ni resin with protein bound after wash steps (Resin).

4.2.3 Large scale expression and protein purification from *M. smegmatis* and *E. coli*

Following evidence of soluble protein expression of PhoH2$_{alt\ MTB}$, PhoH2$_{alt\ MSMEG}$, PhoH2$_{MSMEG}$ and PhoH2$_{TBIS}$ in small scale screens (Figure 4.1, Figure 4.4 and Figure 4.5), protein expression cultures were scaled up as in section 2.4.5 for *M. smegmatis* and 2.4.6 for *E. coli*, and each protein was purified by IMAC. The chromatograms and corresponding SDS-PAGE gels in Figure 4.6 depict the purification, column load and flow through fractions. With the purification of proteins expressed in *M. smegmatis*, non-specific proteins eluted early off the column at a low concentration of imidazole, whereas PhoH2$_{alt\ MTB}$, PhoH2$_{alt\ MSMEG}$ and PhoH2$_{MSMEG}$ eluted later in the gradient with a greater concentration of imidazole (~200-250 mM). Purification of PhoH2$_{TBIS}$ expressed in *E. coli* eluted early (~120-180 mM imidazole) and as a single peak.

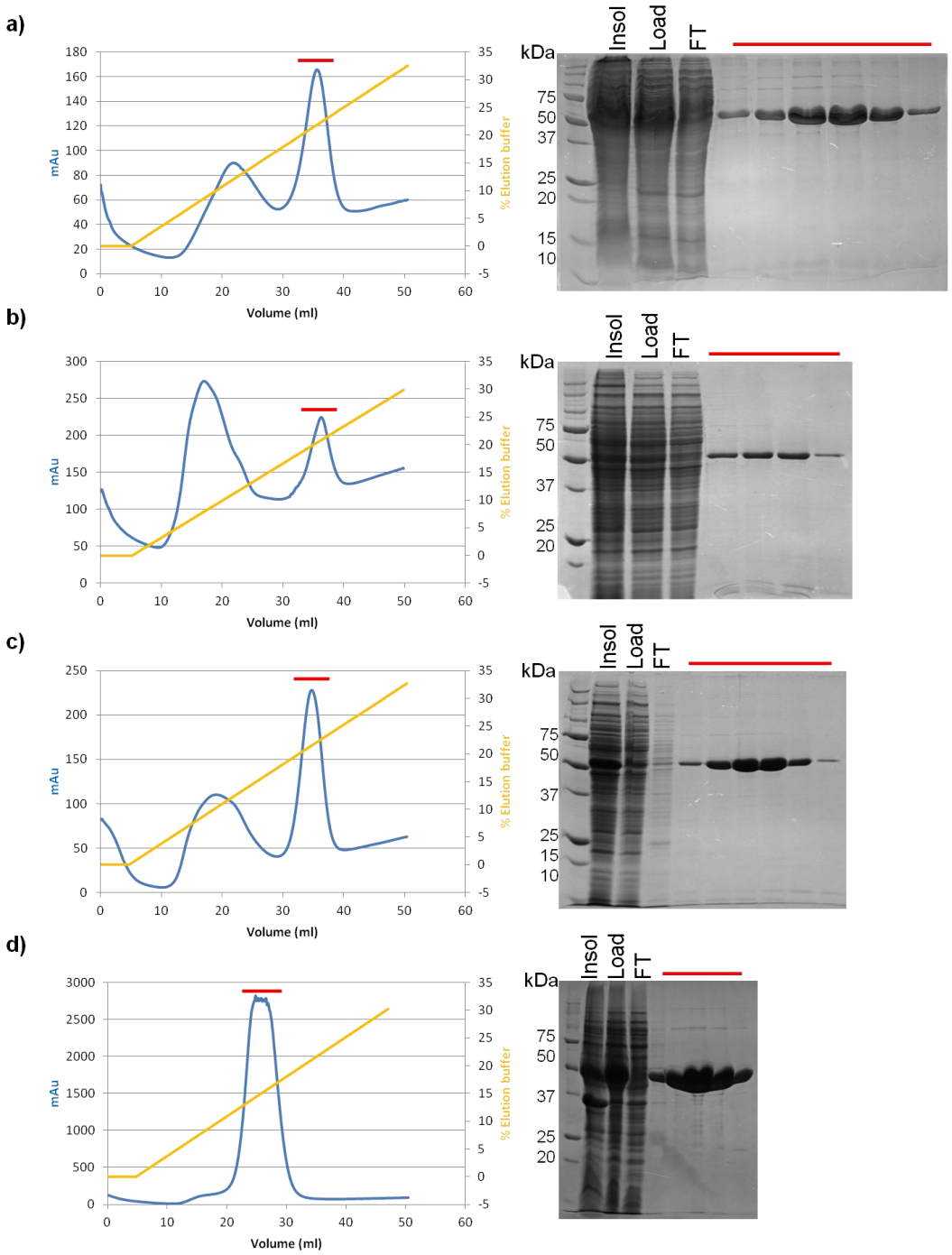

Figure 4.6 IMAC purification and corresponding 12 % SDS-PAGE gels
IMAC purified a) PhoH2$_{alt\ MTB}$ b) PhoH2$_{alt\ MSMEG}$ c) PhoH2$_{MSMEG}$ d) PhoH2$_{TBIS}$
The chromatogram depicts the UV absorbance (blue) and elution profile of each protein as the concentration of elution buffer increases (orange). The corresponding SDS-PAGE gels depict the insoluble fraction (Insol), soluble fraction loaded onto the column (Load), and flow through (FT). Fractions and their elution position are shown by the red bar.

Each protein was further purified by size exclusion chromatography. Each protein eluted as a single peak on an analytical size exclusion column (Figure 4.7).

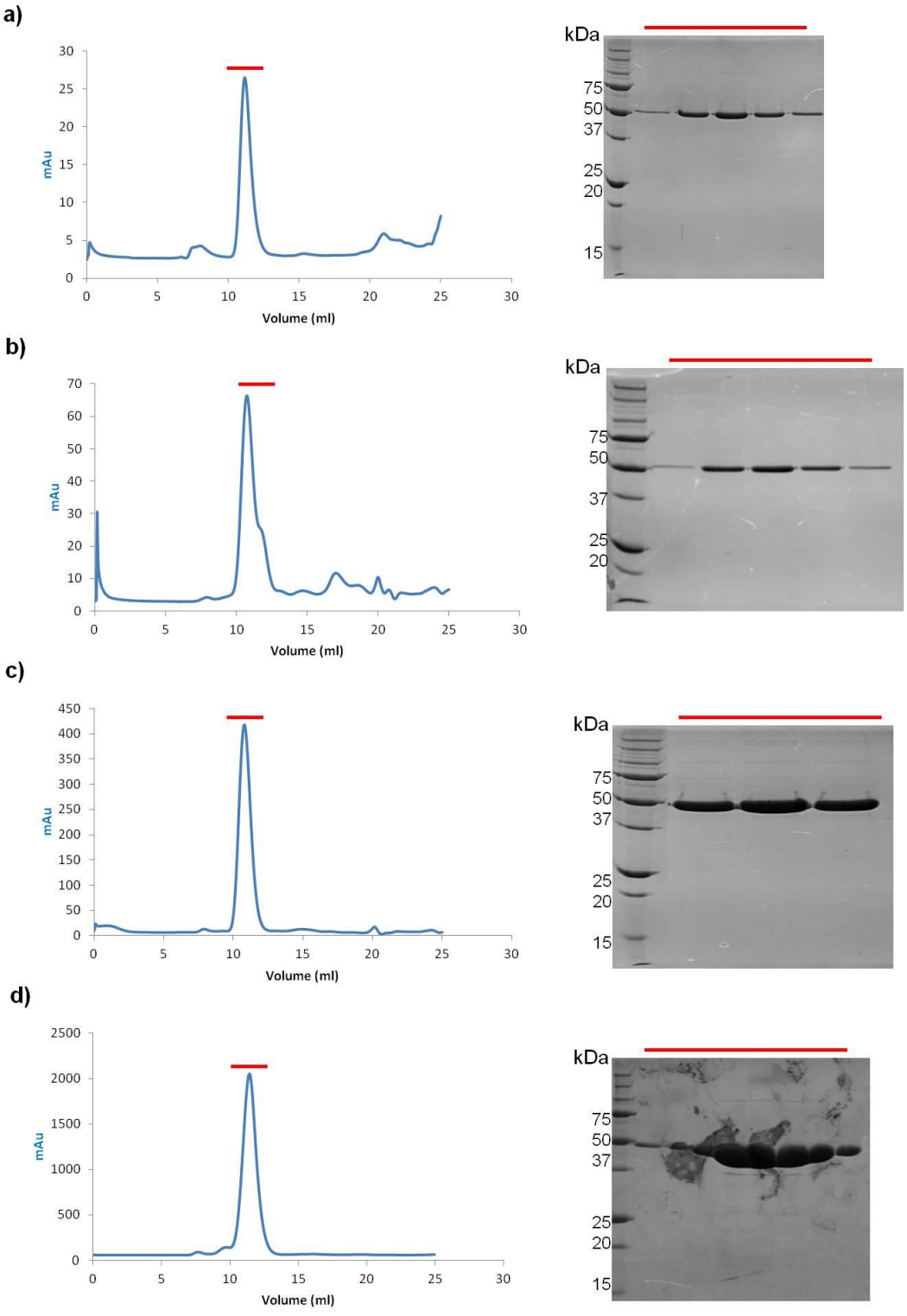

Figure 4.7 Size exclusion chromatography and corresponding 12 % SDS-PAGE gels

Size exclusion purified a) PhoH2$_{alt\ MTB}$ b) PhoH2$_{alt\ MSMEG}$ c) PhoH2$_{MSMEG}$ d) PhoH2$_{TBIS}$. The chromatogram depicts the UV absorbance profile. The SDS-PAGE gel depicts the fractions, and their elution position is shown by the red bar.

An equation to determine the molecular weight (MW) of a protein using the Superdex™ S200 10/300 size exclusion column was derived from the calibration of the column, using high and low molecular weight Gel Filtration Calibration kits (GE Healthcare, UK). The kit proteins were passed through the column and the elution profiles were used to calculate K_{av} values for each. These values were plotted against the log molecular weight to generate an equation that can be used to calculate the MW of sample proteins.

For sample proteins the K_{av} (gel phase distribution coefficient) for each protein was determined using the following equation:

$K_{av} = (Ve - Vo) / (Vc – Vo)$

Where

Ve = elution volume

Vo = column void volume = 8.11 ml

Vc = geometric column volume = 24 ml

The K_{av} value for the protein was then substituted into the equation derived from the curve below (Figure 4.8) to determine the MW of the protein.

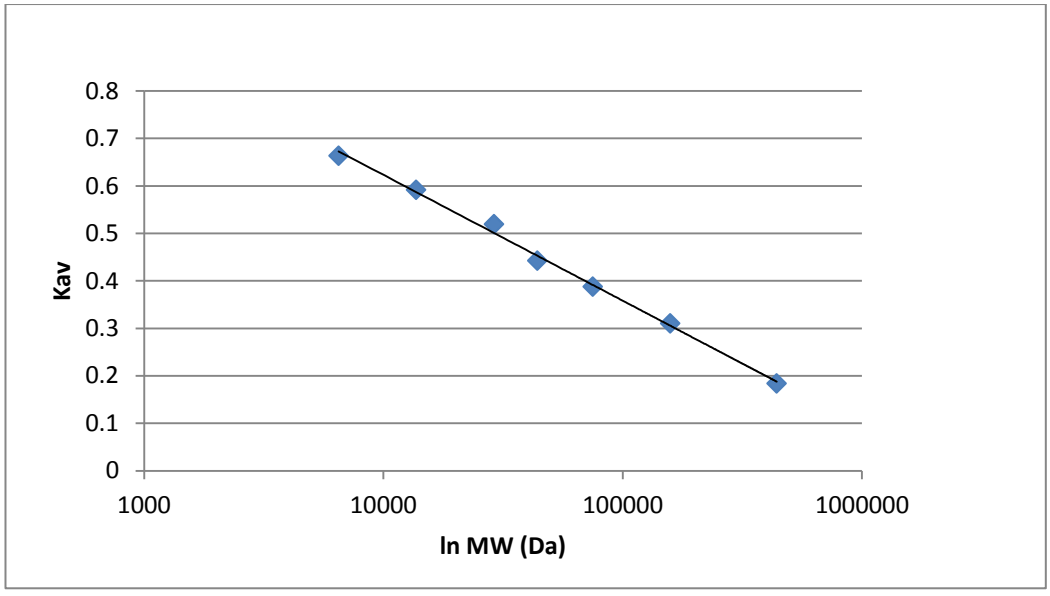

Figure 4.8 Calibration Curve for S200 10/300 size exclusion column
K_{av} values of kit proteins plotted against the log molecular weight.

The equation from the calibration curve (y = -0.115x + 1.6833) was rearranged to calculate the molecular weight of each PhoH2 protein as below:

K_{av} = -0.115 logMW + 1.6833

logMW= (K_{av}-1.6833)/-0.115

MW = $e^{[(Kav - 1.6833) / -0.115)]}$

Table 4.3 denotes the elution volume of each PhoH2 protein and its corresponding MW calculated using this equation.

Table 4.3 Calculated MW of PhoH2 proteins

Protein	Elution volume (ml)	K_{av}	MW (kDa)	Predicted MW (kDa)	MW/Pred
$PhoH2_{alt\ MTB}$	11.23	0.1965	411.7	54.5	7.5
$PhoH2_{alt\ MSMEG}$	10.80	0.1655	539.3	53.7	10.0
$PhoH2_{MSMEG}$	10.74	0.1693	521.9	48.7	10.7
$PhoH2_{TBIS}$	11.46	0.2108	363.7	49.3	7.4

Predicted MWs were calculated using ProtParam (http://web.expasy.org/cgi-bin/protparam/protparam).

These results suggest a large protein complex that may be hexameric, heptameric or octomeric. The sizes obtained for $PhoH2_{alt\ MTB}$ and $PhoH2_{TBIS}$ are similar. The predicted MW of PhoH protein (*cg2513* PDB 3B85) using ProtParam is 37.6 kDa, which when in a hexameric conformation would have a MW of ~225.8 kDa (37.6 x 6). With the addition of the PIN-domain and linker region (~23.5 kDa) to the PhoH protein a combined MW of ~366.8 kDa can be expected (23.5 x 6 = 141 kDa + 225.8 kDa = 366.8 kDa) which is very similar to the MW of $PhoH2_{TBIS}$ and $PhoH2_{alt\ MTB}$ suggesting these PhoH2 proteins may be hexameric. The sizes obtained for $PhoH2_{alt\ MSMEG}$ and $PhoH2_{MSMEG}$ may reflect a larger "heptameric" or "octameric" oligomer.

Samples of $PhoH2_{alt\ MTB}$, $PhoH2_{alt\ MSMEG}$, $PhoH2_{MSMEG}$ and $PhoH2_{TBIS}$ were analysed by Native-PAGE as in section 2.2.2.

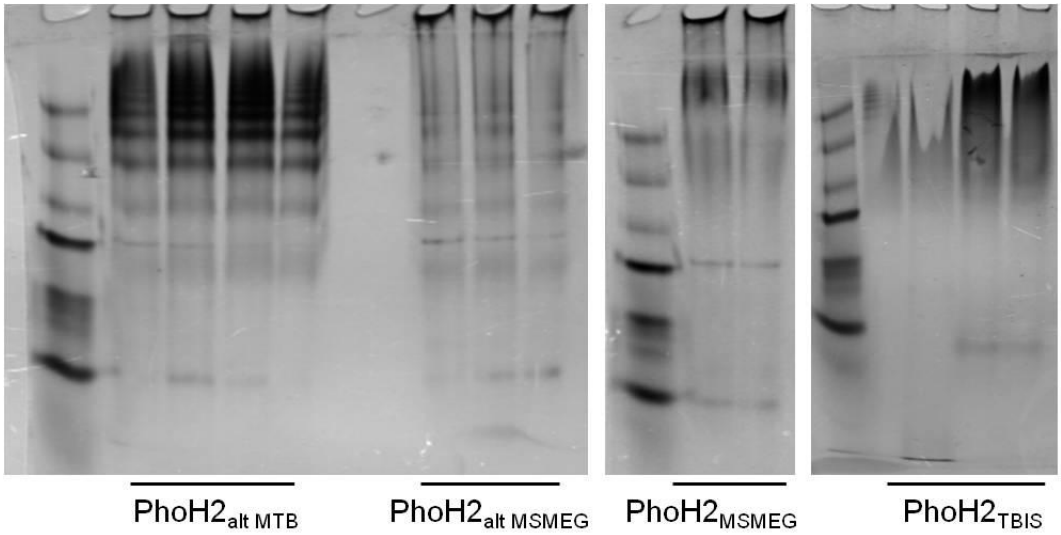

Figure 4.9 Native-PAGE gels of PhoH2$_{alt\ MTB}$, PhoH2$_{alt\ MSMEG}$, PhoH2$_{MSMEG}$ and PhoH2$_{TBIS}$
IMAC purified PhoH2$_{alt\ MTB}$, PhoH2$_{alt\ MSMEG}$, PhoH2$_{MSMEG}$ and PhoH2$_{TBIS}$ protein samples were loaded and run on a 10 % Native-PAGE gel as described in section 2.2.2.

PhoH2$_{alt\ MTB}$ and PhoH2$_{alt\ MSMEG}$ have distinct banding patterns which are not present with PhoH2$_{MSMEG}$ or PhoH2$_{TBIS}$. This banding may suggest an effect on native protein conformation with the cloning and expression of PhoH2 from the alternative start site compared with the cloning and expression from their annotated start site.

4.2.4 Effect of PhoH2$_{MTB}$ expression on growth of *M. smegmatis* mc^{2}4517

In order to investigate whether the lack of protein expression of PhoH2$_{MTB}$ from *M. tuberculosis* in *M. smegmatis* was due to a potential toxic effect, growth of the *M. smegmatis* mc^{2}4517 strains harbouring the pYUB28b (empty vector), pYUBRv1095 (PhoH2$_{MTB}$), pYUBRv1095$_{alt}$ (PhoH2$_{alt\ MTB}$), pYUBMS5247 (PhoH2$_{MSMEG}$) and pYUBMS5247$_{alt}$ (PhoH2$_{alt\ MSMEG}$) constructs were compared during overexpression in ZYP-5052 autoinduction medium. Cultures were set up in triplicate and OD600 readings were taken periodically over a 96 hour period (Figure 4.10).

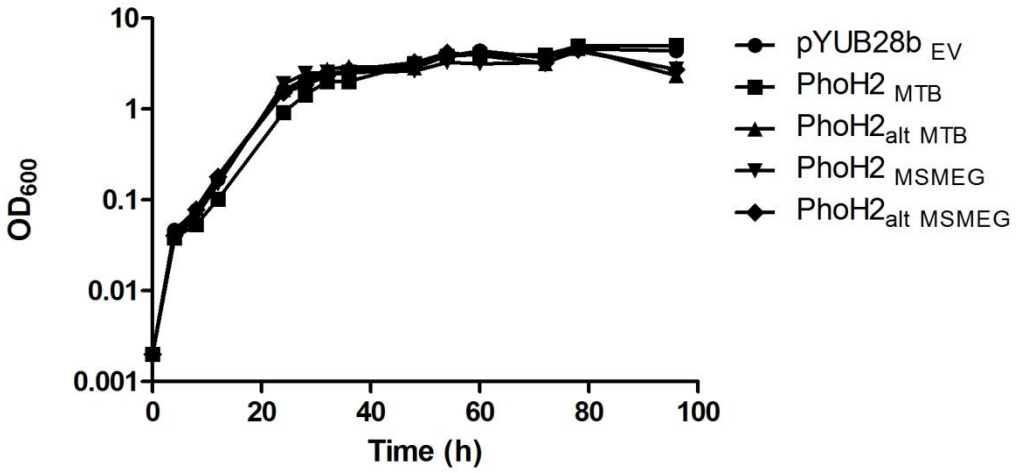

Figure 4.10 Growth of *M. smegmatis* mc²4517 expression strains over a 96 hour period
Results are the mean of three technical replicates taken at each time point from each culture, the error bars represent +/-SD.

There was no difference in growth between the strains harbouring $PhoH2_{MTB/MSMEG}$ and $PhoH2_{alt\ MTB/MSMEG}$ from *M. tuberculosis* or *M. smegmatis*, and the control (empty vector). The strain harbouring the pYUBRv1095 ($PhoH2_{MTB}$) construct that does not express PhoH2 shows no difference in growth compared to the other strains above. A similar observation was made for VapC from *M. smegmatis* which did not express and did not appear to have an effect on growth of an overexpression culture. In this case, a 2 bp insertion in the *vapC* gene was identified that led to a frame shift and insertion of a premature stop codon downstream (McKenzie 2011). The difference observed between these results and the results of conditional expression (Figure 3.13) may be due to the low level of protein expression with pMIND compared with the high levels of protein expression associated with overexpression strains. A negative effect on cell viability with conditional expression was also observed for VapC from *M. smegmatis* (Robson 2010).

$PhoH2_{alt\ MTB}$, $PhoH2_{alt\ MSMEG}$ and $PhoH2_{MSMEG}$ were expressed at high levels with no obvious effect on the growth of the cells (Figure 4.10).

4.2.5 Small and large scale solubilisation trials

The combination of N-lauroylsarcosine and triton X-100 was used to try to solubilise the PIN and PhoH-domains of $PhoH2_{MTB}$ that had been cloned into Gateway vector $pDEST_{17}$ and expressed insolubly in *E. coli* (Figure 4.11). During screening, 1.5 % (w/v) N-lauroylsarcosine was added during cell lysis and 1-3 % triton X-100 was mixed with the cell lysate prior to screening for solubilisation. Figure 4.11a and b show that the PhoH-domain was solubilised under these conditions compared with the buffer only control. Figure 4.11c shows that the PIN-domain was also solubilised by the addition of N-lauroylsarcosine and triton X-100.

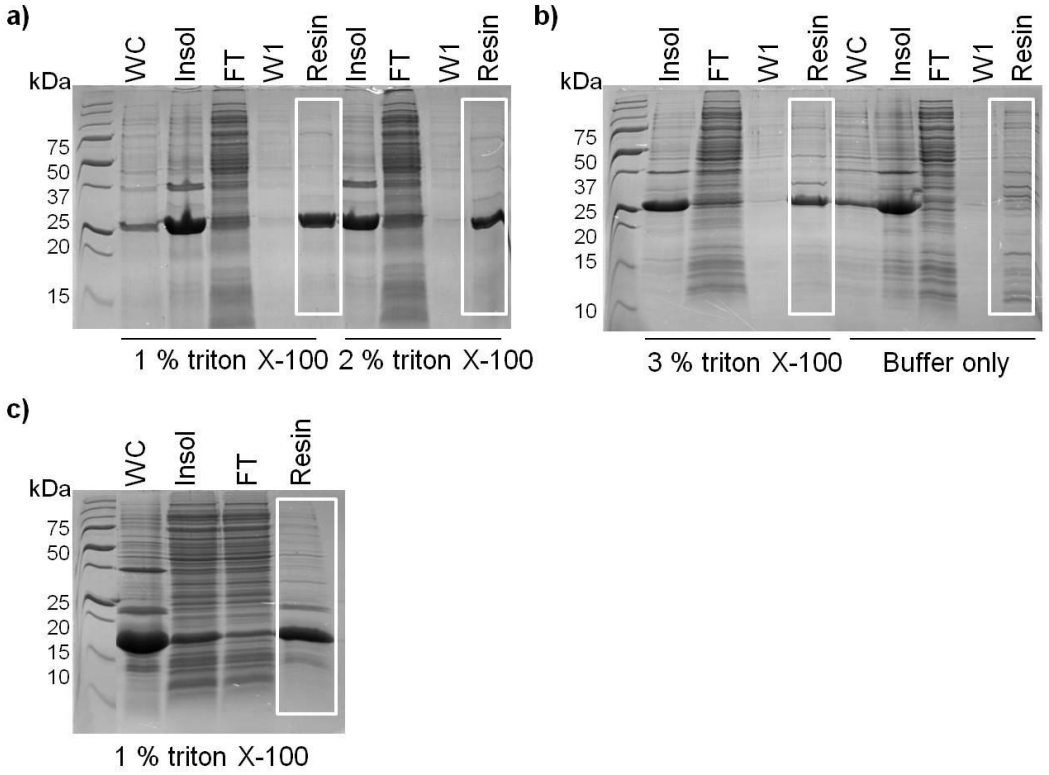

Figure 4.11 Small scale solubilisation trial with the PhoH and PIN-domain
a) PhoH-domain purified with N-lauroylsarcosine and 1 % and 2 % triton X-100. b) PhoH-domain purified with N-lauroylsarcosine and 3 % triton X-100, and buffer only control. c) PIN-domain purified with 1.5 % N-lauroylsarcosine + 1 % triton X-100. White boxes indicate where solubilisation has occurred. Labels: whole cell fraction (WC), insoluble fraction (Insol), flow through (FT), wash one (W1), Ni resin with protein bound after wash steps (Resin).

After evidence of protein solubilisation on a small scale, larger cultures were tested for solubilisation using the same concentration of N-lauroylsarcosine and triton X-100 that led to soluble protein in the small scale tests (Figure

4.11) using the method described in section 2.4.9.2. These proteins were analysed for protein solubilisation by IMAC purification. No protein was observed on the chromatogram for either the PIN or PhoH-domain after treatment with both detergents (not shown). Therefore, small scale protein elution tests were performed to test for elution of the protein from the Ni resin as described in section 2.4.9.1. Figure 4.12 shows that in this trial most of the protein remained insoluble with little to no protein binding to the Ni resin, and neither protein eluted after treatment with imidazole. It is unknown why with the scaling up and repeats of this experiment that protein solubilisation was lost.

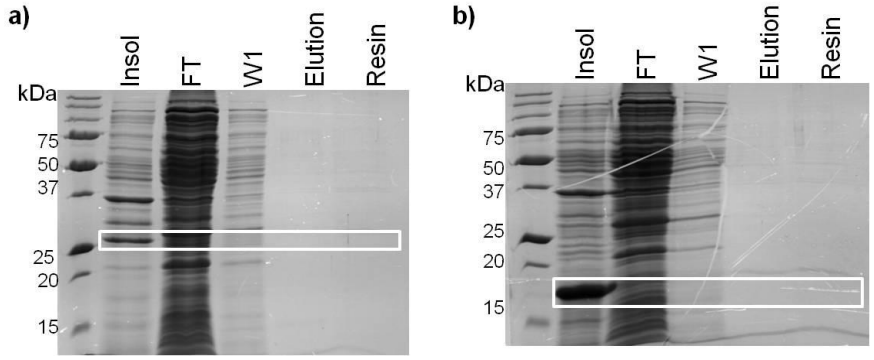

Figure 4.12 Small scale solubilisation and purification of the PhoH and PIN-domain
a) PhoH-domain and b) PIN-domain purified with 1.5 % N-lauroylsarcosine and 1 % triton X-100. White boxes indicate the expected band size. Labels: insoluble fraction (Insol), flow through from Ni resin (FT), wash one (W1), elution treatment from Ni resin (Elution), Ni resin after elution treatment (Resin).

4.2.6 Denaturing purification of PIN and PhoH-domains

Expression of the PIN and PhoH-domains from *M. tuberculosis* in *E. coli* resulted in insoluble protein. As no soluble expression had been gained for these two domains from either *M. tuberculosis* or *M. smegmatis* with further cloning using pYUB28b, a denaturing purification of the PIN and PhoH-domain ($pDEST_{17}$ constructs) was performed as in section 2.4.7 by adding 8 M urea to lysis and elution buffers. The PIN and PhoH-domain were purified by IMAC as in section 2.2.7 under denaturing purification conditions. The chromatogram and corresponding SDS-PAGE gel (Figure 4.13) show that the PIN-domain eluted late in the elution gradient at ~ 40-50 % elution buffer corresponding to 400-500 mM imidazole, whereas the PhoH-domain eluted earlier, around ~ 20-35 % elution buffer (~200 mM imidazole).

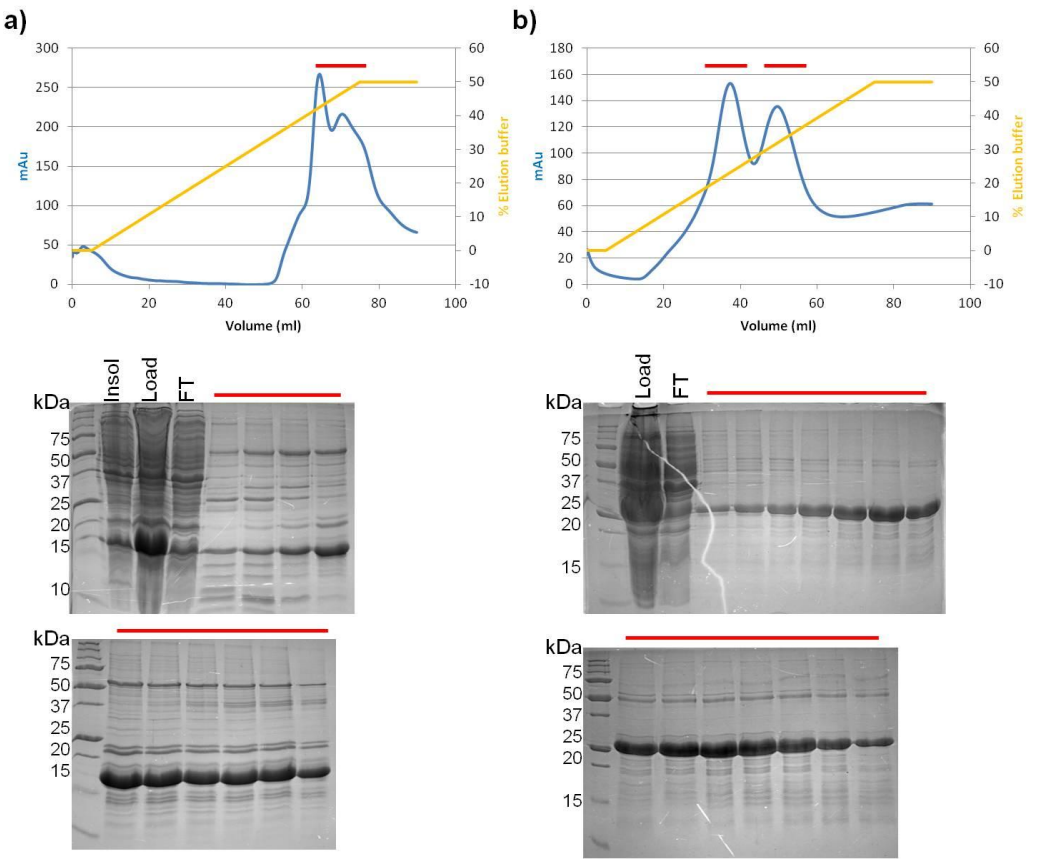

Figure 4.13 Denaturing IMAC purification of the PIN and PhoH-domains expressed in *E. coli* and corresponding 15 % SDS-PAGE gels
The chromatogram depicts the UV absorbance (blue) and elution profile (orange) of the a) PIN and b) PhoH-domains under denaturing conditions. The SDS-PAGE gel depicts the insoluble fraction (Insol), the soluble fraction (Load) and the flow through (FT). Fractions and their elution positions are shown by the red bar.

4.2.7 Small and large scale refolding

Once purified under denaturing conditions, the PIN and PhoH-domains were tested for refolding in a protein refolding screen (section 2.4.8.1). Small scale refolding screens of the PIN-domain yielded promising results for refolding solution sixteen (Figure 4.14b) (50 mM Tris pH 8.2, 250 mM NaCl, 2 mM $MgCl_2$, 2 mM $CaCl_2$, 0.05 % PEG 3350 (w/v), 0.3 mM dodecyl maltoside, 500 mM guanidine HCl, 200 mM sucrose, 200 mM L-arginine, 1 mM GSH, 0.1 mM GSSG).

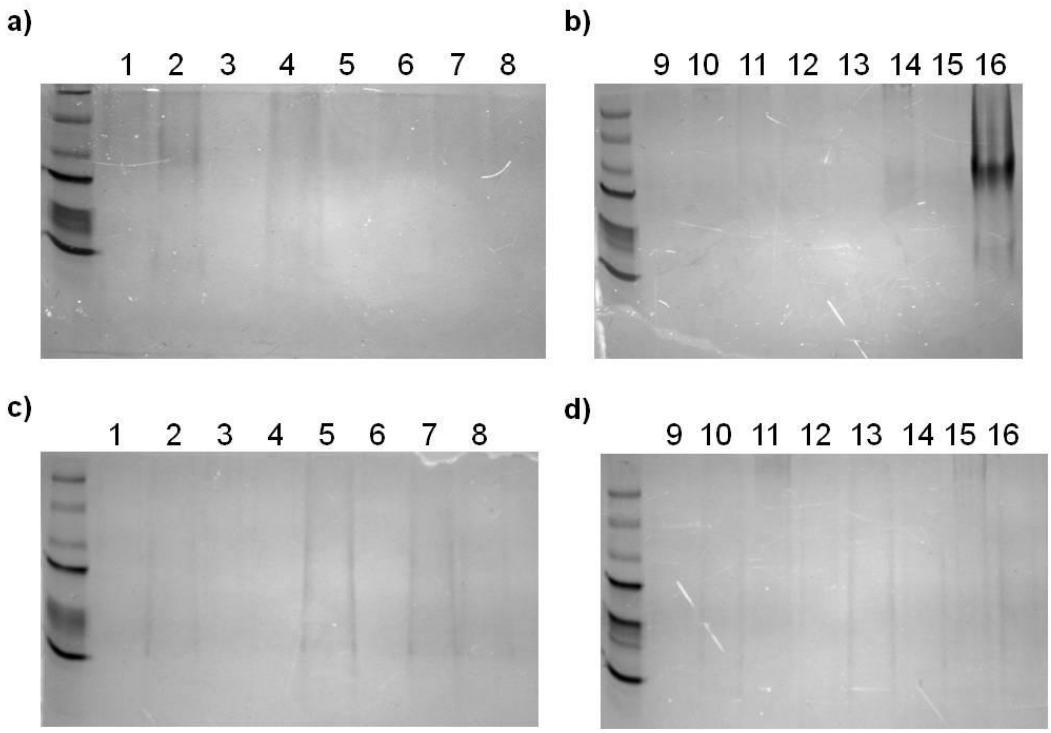

Figure 4.14 Small scale refolding screen analysis by 10 % Native-PAGE
PIN-domain refolding with refolding conditions a) 1-8 and b) 9-16, and PhoH-domain refolding with refolding conditions c) 1-8 and d) 9-16.

The successful refolding treatment of the PIN-domain was scaled up to a 200 µl dialysis button and further in 10 fold increments from 2 ml to 20 ml. The refolded protein was then dialysed into a storage buffer for analysis with size exclusion chromatography. No refolded PIN-domain protein was observed on a native-PAGE gel with scaling up from 20 µl, and no protein was eluted following size exclusion chromatography.

No further cloning and expression trials were continued at this time.

4.2.8 Initial crystallisation screens

An initial crystallisation screen was carried out at the University of Auckland for $PhoH2_{alt\ MTB}$ that had been purified by IMAC and size exclusion in 50 mM Na-phosphate buffer pH 7.4, 200 mM NaCl and concentrated to 1.96 mg/ml. This trial yielded no distinctive crystals, however, there were a few conditions where flecks or edges were apparent and these conditions were followed up with fine screens. Crystals did not successfully grow in these fine screens and in some of the negative drops (reservoir crystallisation solution mixed with buffer only), small flecks of a crystalline nature appeared,

suggesting the formation of salt crystals. A second screen was performed with $PhoH2_{alt\ MTB}$ purified in 50 mM Na-Hepes buffer pH 7.4, 200 mM NaCl ~ 100 mM imidazole with the aim of reducing the likelihood of salt crystals forming. One condition resulted in needle formation and a fine screen was set up to optimise this growth condition. However, these crystals were deemed likely imidazole salt crystals due to the formation of the same crystals in the negative drops (buffer only).

Due to the relatively low protein concentration obtained for crystallisation of $PhoH2_{alt\ MTB}$, $PhoH2_{MSMEG}$ was trialled in crystallisation screens at the University of Waikato, once purified in 50 mM phosphate buffer pH 7.4, 200 mM NaCl and secondly purified 50 mM Na-Hepes pH 7.4, 200 mM NaCl. No real crystal growth was observed for $PhoH2_{MSMEG}$ purified in phosphate buffer, and fine screens around conditions in which flecks were observed did not gain any crystals. Two conditions were further screened for $PhoH2_{MSMEG}$ purified in 50 mM Na-Hepes buffer pH 7.4, 200 mM NaCl and one condition (0.1 M $CaCl_2$ 0.2 M Bis-Tris pH 6.5 45 % MPD) led to crystal growth in only the protein drop and not in the negative drop (buffer only). This crystal was tested for diffraction at the Home-Source at the University of Auckland but was found to be salt.

It was evident with the low protein concentrations obtained for mycobacterial PhoH2 proteins despite efforts to scale up to obtain greater protein yield, and the lack of promising conditions in robot screens, that a related thermophilic homologue of PhoH2 may be a better option for crystallisation trials.

$PhoH2_{TBIS}$ was purified by IMAC and size exclusion chromatography in 50 mM Tris buffer pH 8, 200 mM NaCl and concentrated to 9.9 mg/ml and trialled in crystallisation screens. Fine screens were set up around promising conditions, however better crystal growth than observed previously was not obtained. The initial screen was re-checked after 6 weeks and 'true' crystals (non-fleck or edge only) had grown in the crystallisation condition (0.1 M imidazole pH 7, 25 % PEG 550 mono methyl ether (MME)). A fine screen was

performed around this condition and one drop (0.15 M imidazole pH 7, 18.75 % PEG 550 MME), gave small crystals in the protein drop and none in the negative drop (buffer only). This condition was re-screened with protein concentrated to 5 mg/ml and yielded crystals suitable for X-ray diffraction testing (Figure 4.15 c and d). An additive screen was also laid down to try to improve crystal growth, however no crystals grew in the presence of additives.

Crystals grown from repeats of this fine screen condition (0.15 M imidazole pH 7, 18.75 % PEG 550 MME) were tested for X-ray diffraction at the Australian Synchrotron which confirmed that they were protein crystals (Figure 4.15 e).

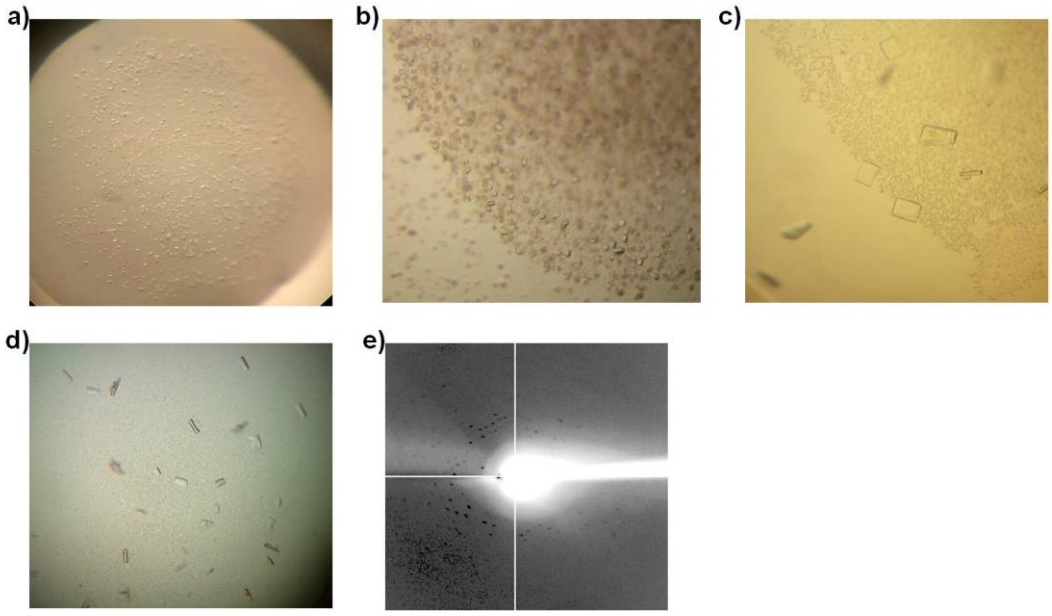

Figure 4.15 PhoH2$_{TBIS}$ protein crystals
a) Robot screen drop 0.1 M imidazole pH 7, 25 % PEG 550 MME. b) Crystals grown in fine screen in 0.15 M imidazole pH 7, 18.75 % PEG 550 MME. c) and d) repeat condition of b) with 5 mg/ml protein. e) X-ray diffraction obtained from Australian synchrotron showing protein diffraction spots to ~ 8 Å.

Further crystallisation screens were trialled for PhoH2$_{TBIS}$; PhoH2$_{TBIS}$ + 1 mM final AMP-PNP (concentrated to 16.94 mg/ml prior to addition of AMP-PNP), and PhoH2$_{TBIS}$ was subjected to a limited proteolytic digest with trypsin (1:100) prior to laying down screens. No significant crystal growth was observed in these trials.

4.2.9 Protein crystal growth optimisation

For optimisation of PhoH2$_{TBIS}$ protein crystals, seeding methods were adopted with the aim of improving crystal quality. Macro, streak and batch seeding were trialled for PhoH2$_{TBIS}$ concentrated to 15 and 29 mg/ml as described in section 2.4.11.3. Batch seeding led to the most favourable crystal growth (Figure 4.16).

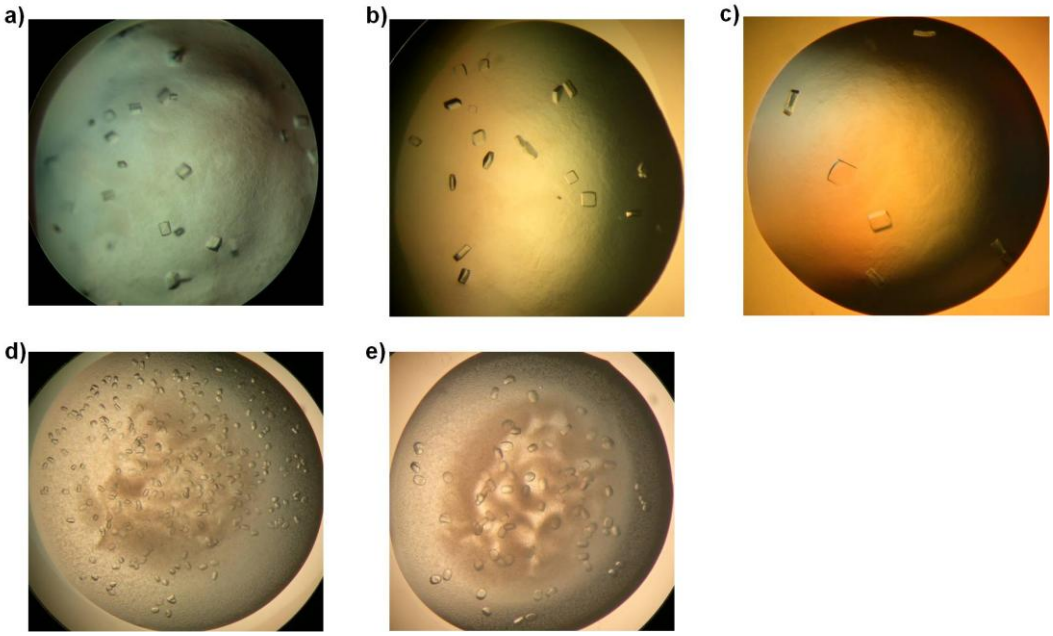

Figure 4.16 PhoH2$_{TBIS}$ protein crystals optimised using seeding
Examples of the types of crystals obtained with seeding a) Streak seeding (15 mg/ml) b) Batch seeding 1:100 (15 mg/ml) c) Batch seeding 1:1000 (15 mg/ml) d) Batch seeding 1:100 (29 mg/ml) and e) Batch seeding 1:1000 (29 mg/ml).

A range of crystals from batch seeding drops 1:1000 (Figure 4.16c and e) were tested for X-ray diffraction at the Australian Synchrotron. Data to 7.5 Å (Figure 4.17) were collected for one crystal taken from batch seeding drop 1:1000 (15 mg/ml) (Figure 4.16c). Frames from 1 – 360 ° were taken with 0 % attenuation at 1 ° oscillations and 2 second exposures.

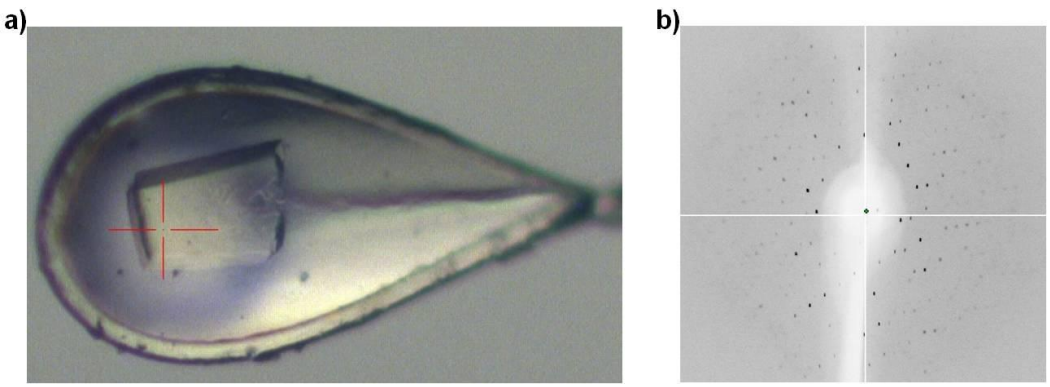

Figure 4.17 PhoH2$_{TBIS}$ Crystal and X-ray diffraction to 7.5 Å
a) PhoH2$_{TBIS}$ batch seeding crystal used to gain improved X-ray diffraction b) X-ray diffraction of one frame obtained during data collection at the Australian Synchrotron (plate resolution 4.51 Å).

4.2.10 Preliminary data processing

Data analyses in iMOSFLM and CCP4i provided the following data.

Table 4.4 Crystallographic data

	Data set 1	Data set 2
Space group	C2	I222
Unit cell dimensions		
a	186.45	181.91
b	297.22	205.21
c	174.25	208.67
α	90.0	90.0
β	121.84	90.0
γ	90.0	90.0
Matthews coefficient	2.33	3.23/2.15
Number of molecules/asymmetric unit	18	6/9
Solvent content %	47.17	61.9/42.86
Resolution range (overall) Å	93.82-7.5	48.78-8.34
Number of observations	341055	27070
Number of unique reflections	45970	3829
Completeness	99.7	98.2
Multiplicity (overall)	7.4	7.1

The maximum resolution obtained for the first data set that indexed in space group C2 was ~ 7.5 Å. The second data set to 8.3 Å were collected from a further batch seeding drop (1:100).

These data were analysed in MOLREP using the self rotation function which provides information as to the contents and organisation of the asymmetric unit. This analysis is suggestive of hexameric symmetry (Figure 4.18).

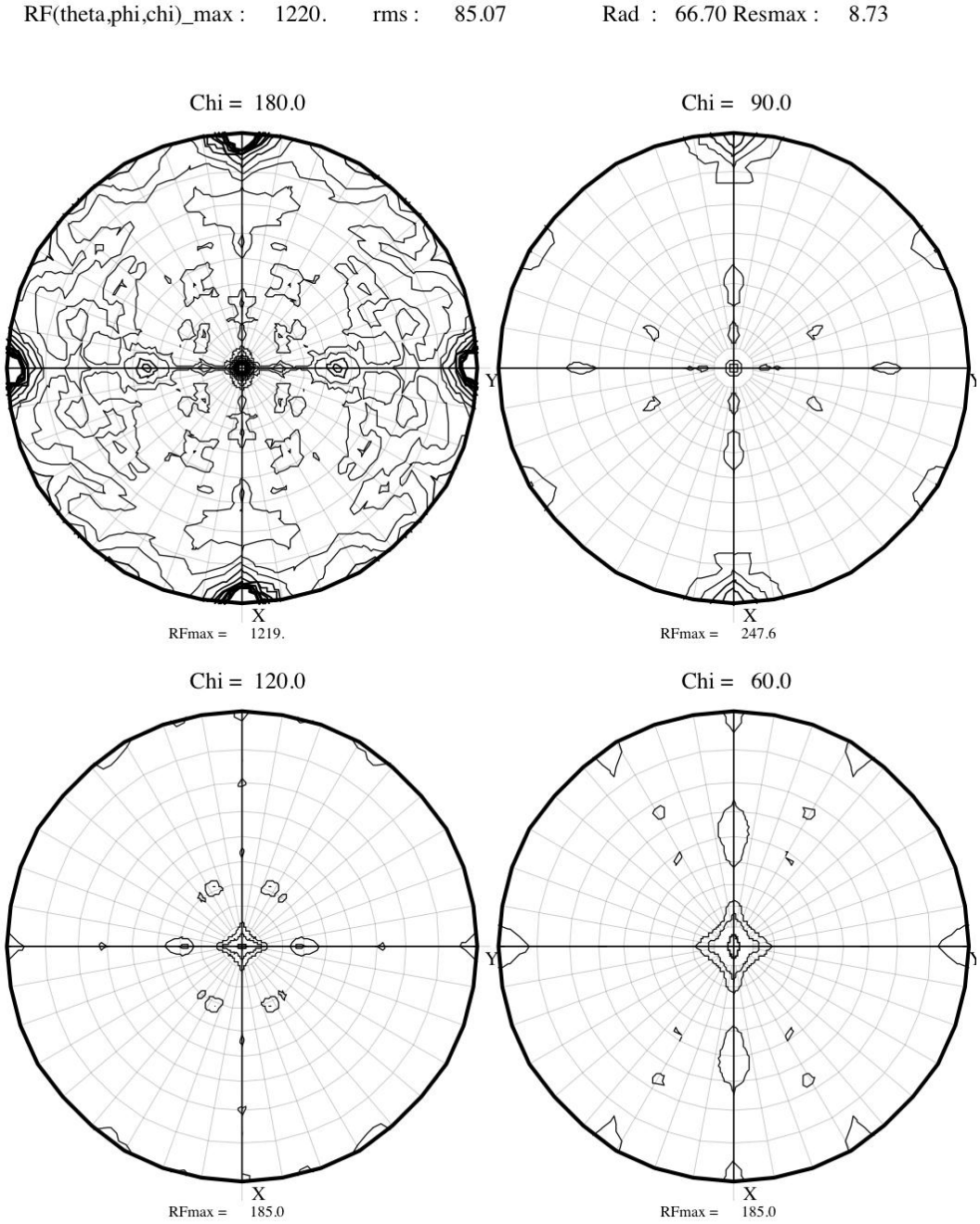

Figure 4.18 Self rotation function map of PhoH2$_{TBIS}$
Self rotation function maps of PhoH2 data set 2 in space group I222 at different rotations (Chi). Peaks mark the positions where rotational symmetry axes enter and exit the sphere. The orientations of the axes in the sphere indicate the orientation of the n-fold symmetry axes in the crystal. Six axes of symmetry are observed together with the symmetry gained from the I222 crystallographic space group for PhoH2. Maps were generated using MOLREP Self Rotation Function using data set 2.

The number of molecules in the asymmetric unit (Matthews probability) is likely to be 6 or 9 depending on the solvent content of the crystals. This is consistent with the probable hexameric symmetry revealed with the rotational symmetry (Figure 4.18) and the MW obtained for PhoH2$_{TBIS}$ (Table

4.3) which is very similar to the MW of PhoH protein (3B85) which shows a hexamer. PhoH2 probably adopts a hexameric oligomer.

4.2.11 Electron microscopy investigation of PhoH2$_{TBIS}$

EM was used to further investigate the oligomerisation state of PhoH2 proteins. PhoH2$_{TBIS}$ was purified by IMAC and size exclusion chromatography and diluted to 0.1 mg/ml and 0.03 mg/ml for analysis at the University of Auckland. The first image taken at the higher protein concentration revealed both circular assemblies and aggregates in solution (Figure 4.19).

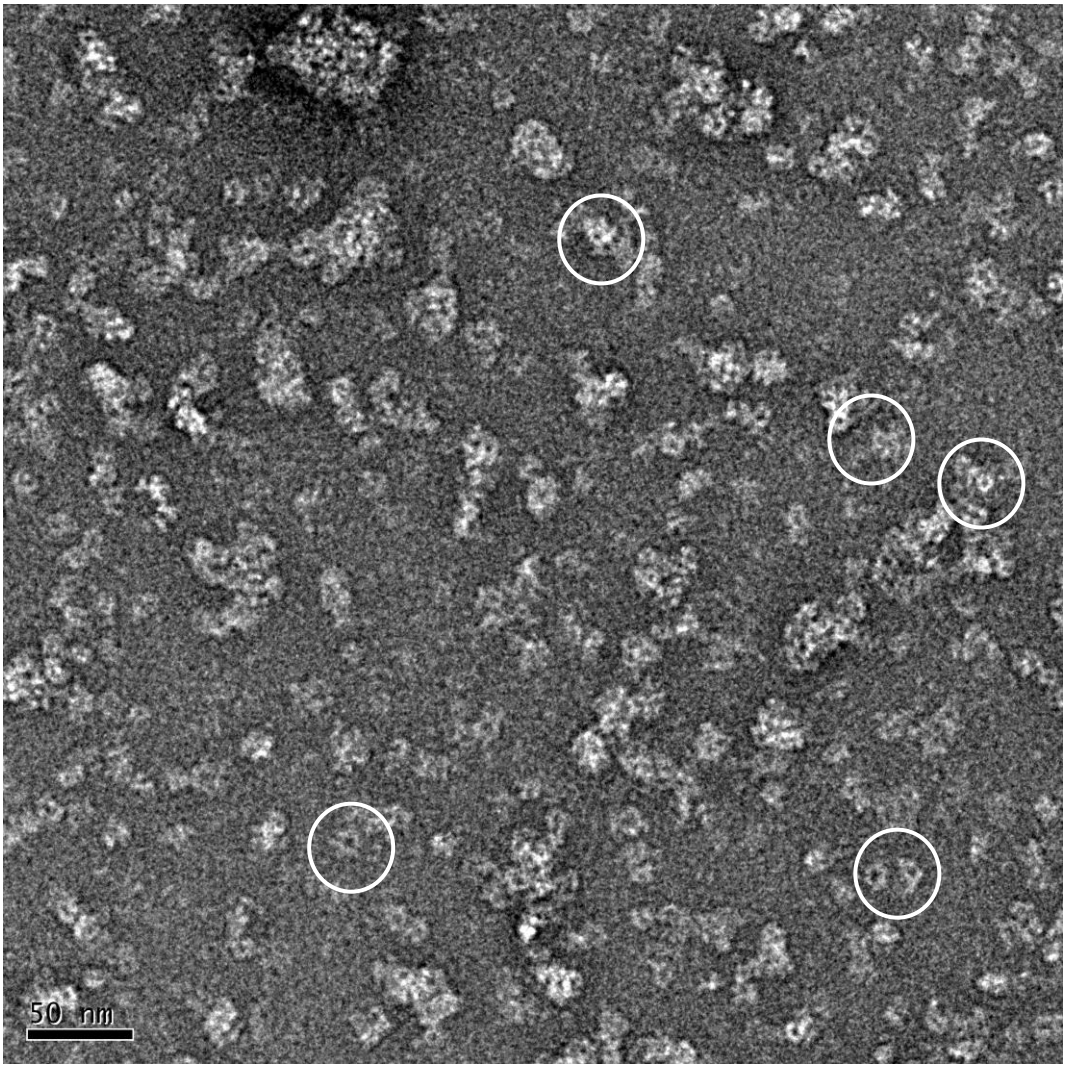

Figure 4.19 PhoH2$_{TBIS}$ visualised by TEM (0.1 mg/ml)
White circles identify some of the circular assemblies observed.

The second image taken of PhoH2$_{TBIS}$, at the lower concentration allowed for better visualisation of specific assemblies. This image revealed a variety of oligomeric states (Figure 4.20).

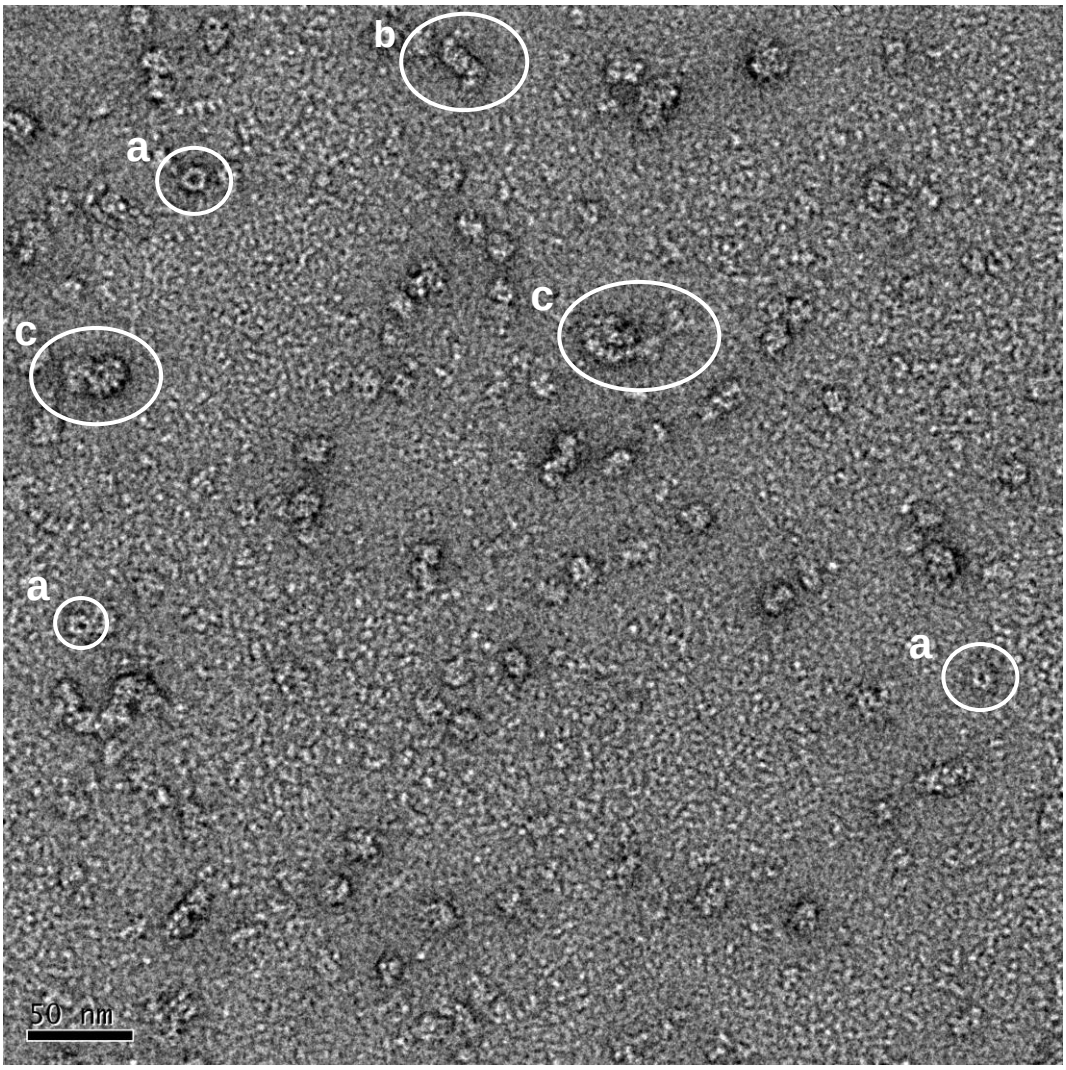

Figure 4.20 PhoH2$_{TBIS}$ visualised by TEM (0.03 mg/ml)
The white circles identify a) some of the circular ring-like assemblies, b) a helical assembly and c) spiral assemblies of PhoH2$_{TBIS}$ protein in solution.

Taken together, PhoH2$_{TBIS}$ adopts both circular ring-like oligomers composed of 4, 5 or 6 subunits, and larger helical and spiral assemblies.

4.2.12 Translational investigation of PhoH2$_{alt}$ proteins

To gain soluble protein expression of PhoH2 from *M. tuberculosis*, cloning from an alternative start site located upstream and out of frame with the annotated translational start site was required. It was unclear throughout much of this research whether this short sequence was being translated. Therefore whether translation was occurring from the alternative or annotated translational start site or both sites, with PhoH2$_{alt\ MTB}$ and PhoH2$_{alt\ MSMEG}$, which were both cloned from the alternative start site was unknown. When each of the PhoH2 proteins is run on a SDS-PAGE gel they

are fairly similar in size (Figure 4.6 and Figure 4.7). It is apparent, when these proteins are analysed by Native-PAGE, that there is an effect on native protein conformation with the cloning and expression of PhoH2 from the alternative start site, compared with the cloning and expression from their annotated start site (Figure 4.9). These results may reflect differences in protein translation or effects of regulatory factors within the additional transcript.

To investigate the translational context of $PhoH2_{alt\ MTB}$ and $PhoH2_{alt\ MSMEG}$, protein expression and mass spectrometry (MS) based analyses were carried out at the University of Waikato and at the Centre for Protein Research (CPR) at the University of Otago, who specialise in mass spectrometry based protein identification techniques.

4.2.12.1 Cloning and expression testing of the alt sequence in pYUB28b/M. smegmatis

The alt peptide sequence from *M. tuberculosis* (alt_{MTB}) was successfully amplified and cloned into pYUB28b, and transformed into *M. smegmatis* $mc^2$4517 cells. Small scale expression tests revealed a lack of peptide protein expression as shown by the absence of enrichment at the expected molecular weight (6.7 kDa) (Figure 4.21). It was possible that the peptide was too small to detect.

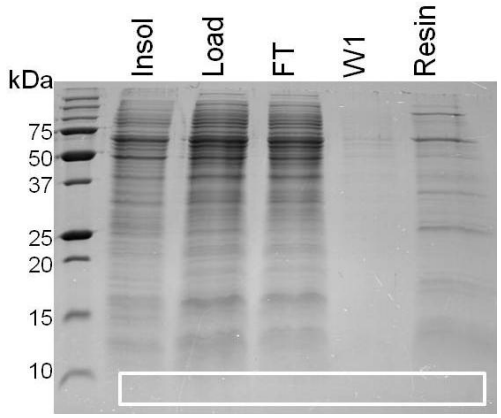

Figure 4.21 Small scale expression screen of alt peptide from *M. tuberculosis*
Expression and his-tag binding of alt_{MTB}. White box indicates expected band size. Labels: Insoluble fraction (Insol), Soluble fraction loaded onto Ni resin (Load), flow through (FT), wash one (W1), Ni resin with protein bound after wash steps (Resin).

4.2.12.2 Mass spectrometry – University of Waikato

MALDI-TOF MS was used to analyse trypsin digested and whole protein samples of PhoH2$_{alt\ MTB}$ to confirm the cloning, expression, and translation of the full length construct. The resulting peptides were mapped to the PhoH2$_{alt\ MTB}$ amino acid sequence. Figure 4.22 shows the moderate sequence coverage of the matched tryptic peptides.

Figure 4.22 Sequence coverage of PhoH2$_{alt\ MTB}$
Pink bars indicate sequence coverage of PhoH2$_{alt\ MTB}$ (37 %). Arrow indicates the position of the annotated translational start site.

PhoH2 from *M. tuberculosis* was identified with 37 % sequence coverage. No peptides were found that matched the sequence between the alternative and annotated start site.

PhoH2$_{MTB}$ was also analysed as a whole, Figure 4.23 shows that a mass of 49207.2 Da was gained for PhoH2$_{MTB}$ by this method at the University of Waikato.

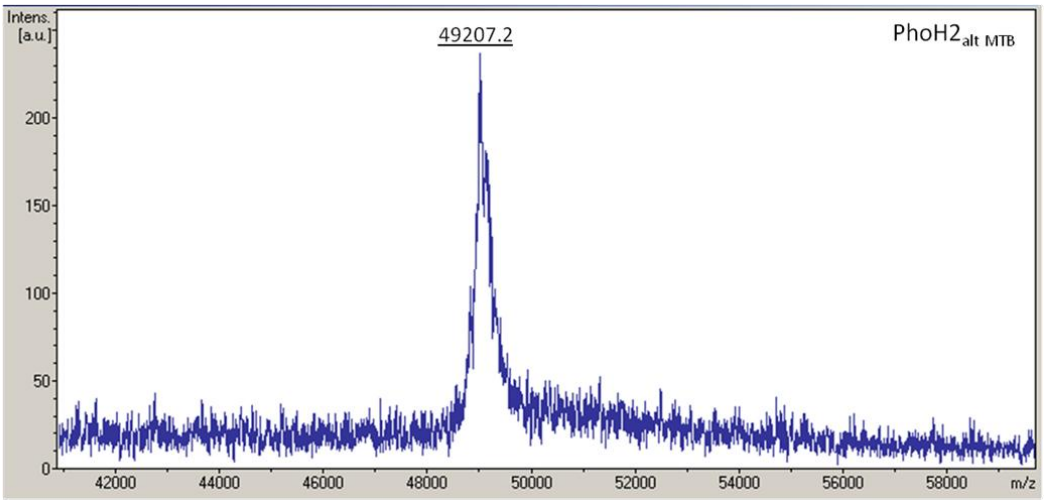

Figure 4.23 Whole protein MALDI-TOF MS of PhoH2$_{alt\ MTB}$
PhoH2$_{alt\ MTB}$ as a whole was analysed by MALDI-TOF MS. A mass of 49207.2 Da was gained.

This mass may correspond to PhoH2 translation from its annotated start site or loss of the peptide during MALDI-TOF testing. The expected MW of PhoH2$_{alt\ MTB}$ was 54445.9 Da and the expected MW of the alt peptide is 5189.8 Da, (54445.9 Da– 5189.8 Da = 49256.1 Da).

These analyses were also attempted for PhoH2$_{alt\ MSMEG}$ and PhoH2$_{MSMEG}$ but either no signal was obtained or the intensity was too weak to distinguish from background signal.

4.2.12.3 N-terminal sequencing – University of Otago

Due to the inconclusive results gained with MS at Waikato, PhoH2$_{alt\ MTB}$ and PhoH2$_{alt\ MSMEG}$ were tested at the Centre of Protein Research (CPR) at the University of Otago, who specialise in N-terminal sequencing of tryptic protein peptides.

For this purpose, freshly purified protein samples of PhoH2$_{alt\ MTB}$ and PhoH2$_{alt\ MSMEG}$ were couriered for analysis. For N-terminal sequencing, protein samples were chemically modified by guanidination and sulfonation prior to digestion with trypsin. Both modifications aimed to increase the efficiency and selective detection of peptide masses. Samples were analysed by MALDI TOF/TOF and by LTQ-Orbitrap MS.

The aim of these analyses was to identify the N-terminus and gain sequence coverage upstream of the annotated translational start site for both PhoH2$_{alt\ MTB}$ and PhoH2$_{alt\ MSMEG}$ if indeed, an antitoxin-like peptide was being translated. Figure 4.24 shows the combined sequence coverage of PhoH2$_{alt\ MTB}$ gained with MALDI TOF/TOF and LTQ-Orbitrap analysis of peptides. LTQ-Orbitrap MS revealed a peptide mass that mapped within the upstream sequence ahead of the annotated translational start site (ions score 45, individual ions scores >32 indicate identity or extensive homology $p < 0.05$) suggesting that translation can take place from the alternative start site which is positioned out of frame of the annotated translational start site.

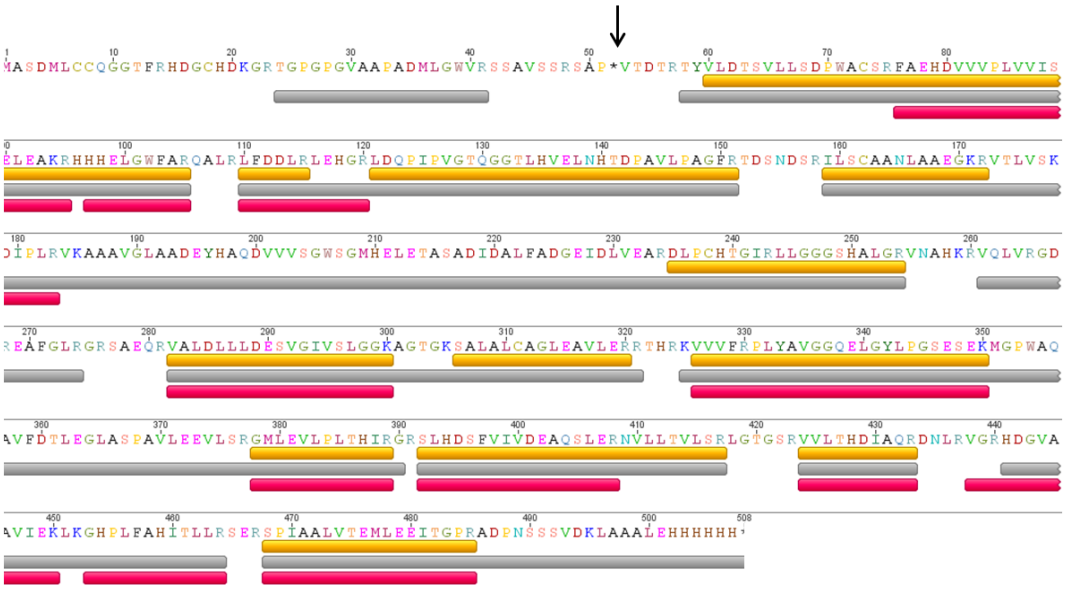

Figure 4.24 Sequence coverage of PhoH2$_{alt\ MTB}$
Sequence coverage of PhoH2$_{alt\ MTB}$. Yellow bars indicate MALDI peptide analysis (53 % coverage), grey bars Orbitrap peptide analysis (83.8 % coverage), pink bars MALDI peptide analysis (University of Waikato). Arrow indicates the position of the annotated translational start site.

Two other peptides towards the N-terminus were also detected by LTQ-Orbitrap MS (ion scores of 101 and 65 respectively) (Figure 4.25).

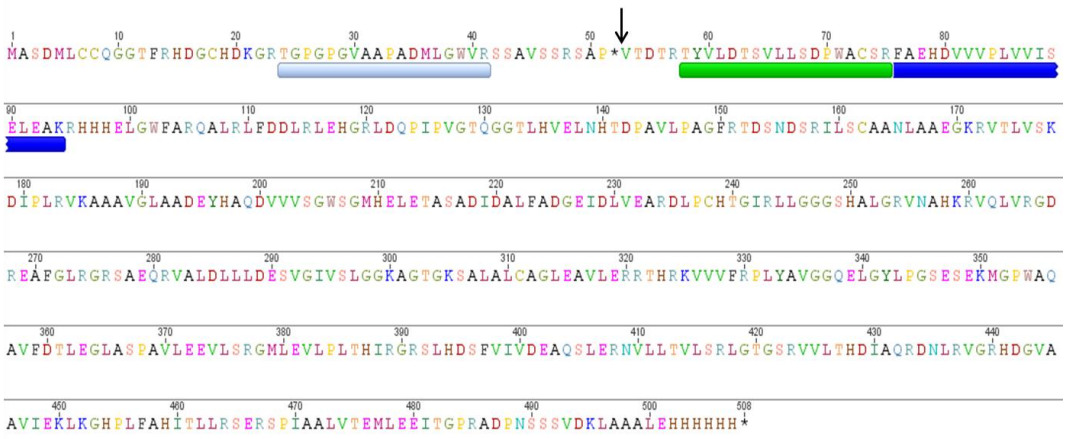

Figure 4.25 Position of the three most N-terminal matched peptides
PhoH2$_{alt\ MTB}$ sequence with the positions of the three most N-terminal peptides. Grey bar most N-terminal peptide (ion score 45), green bar second most N-terminal peptide (ion score 101) and blue bar third most N-terminal peptide (ion score 65) detected significantly. Arrow indicates the position of the annotated translational start site.

Neither of the two most N-terminal peptides mapped to PhoH2$_{alt\ MTB}$ were identified with their N-terminal modification (SPITC) perhaps due to full tryptic cleavage of the modification resulting in small fragments that were unable to be detected. The third most N-terminal peptide in contrast was identified with the modification.

MALDI TOF/TOF and LTQ-Orbitrap MS analyses were also performed with PhoH2$_{alt\ MSMEG}$.

Figure 4.26 shows the combined sequence coverage of PhoH2$_{alt\ MSMEG}$ which confirms that translation of PhoH2$_{alt\ MSMEG}$ also does take place from the alternative start site.

Figure 4.26 Sequence coverage of PhoH2$_{alt\ MSMEG}$
Sequence coverage of PhoH2$_{alt\ MSMEG}$. Yellow bars indicate MALDI peptide analysis (59 % coverage) and grey bars indicate LTQ-Orbitrap analysis (86.75 %). The two most N-terminal peptides had ions scores of 42 and 45 respectively. Arrow indicates the position of the annotated translational start site.

Intact protein measurements were also made using MALDI TOF/TOF. Figure 4.27 shows a number of intact masses for PhoH2$_{alt\ MTB}$ despite the presence of only one protein band on a SDS-PAGE gel. The largest reported mass was (49162.99 Da) which is comparable to the whole protein mass gained previously 49207. 2 Da (Figure 4.23). PhoH2$_{alt\ MSMEG}$ too gave a range of intact mass measurements (Figure 4.28), the largest being 54817.07 Da which is similar to the theoretical MW of the full length construct (full translation alternative start site) of 53728.2 Da.

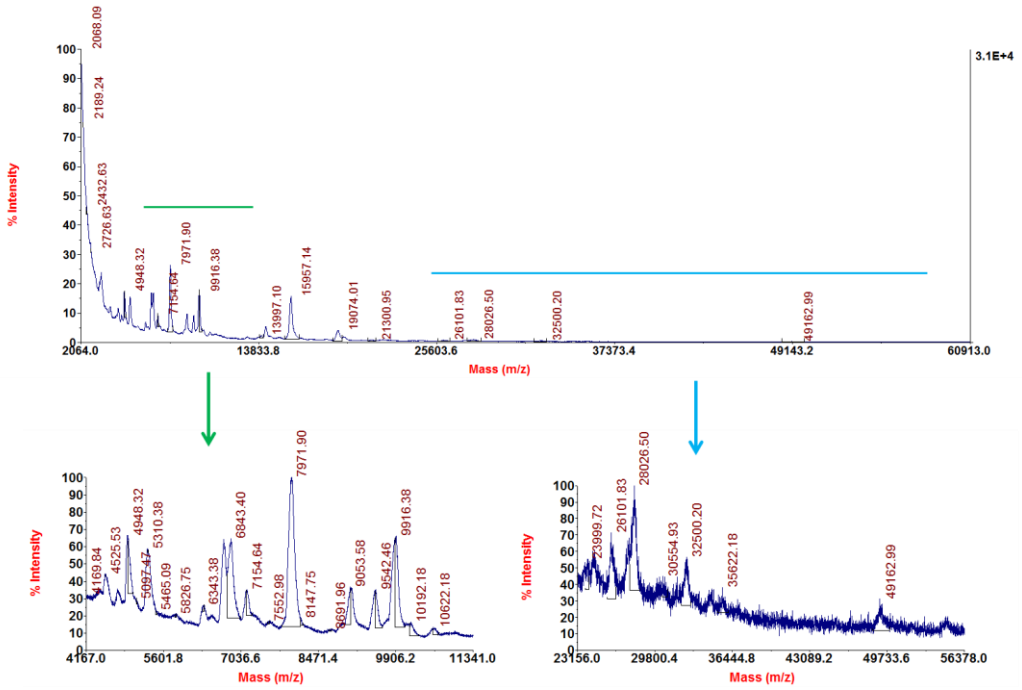

Figure 4.27 Intact mass measurements of PhoH2$_{\text{alt MTB}}$
Whole protein mass spectra of PhoH2$_{\text{alt MTB}}$ (Spectra obtained from CPR, University of Otago). A range of masses were observed, the largest mass being 49162.99 Da. Lines above the masses indicate area shown in the enlargement.

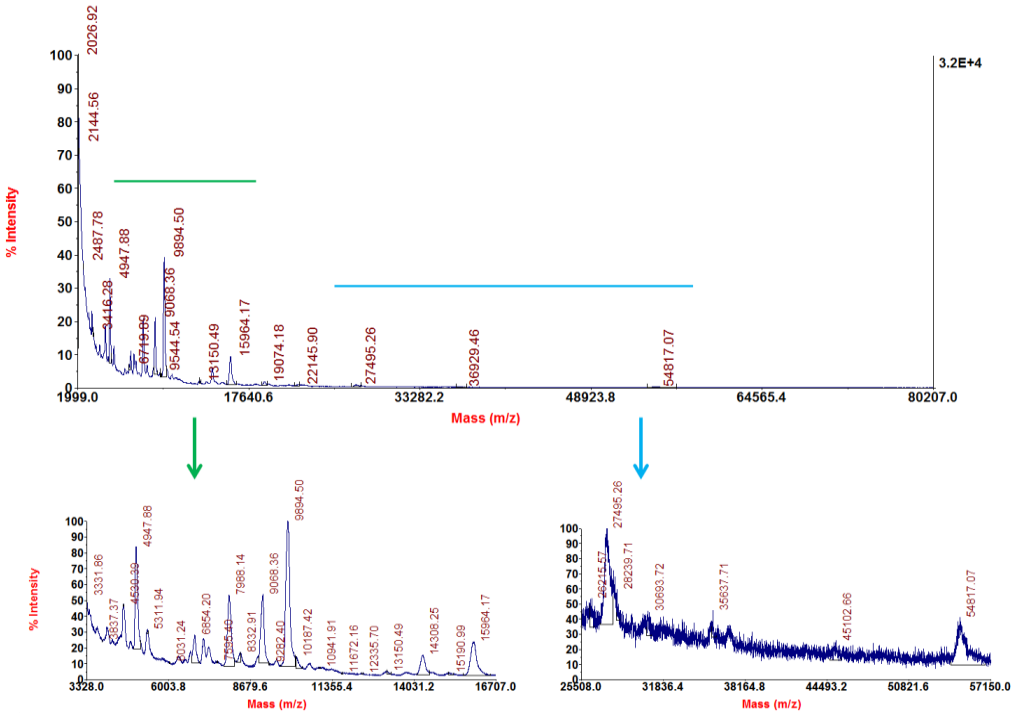

Figure 4.28 Intact mass measurements of PhoH2$_{\text{alt MSMEG}}$
Whole protein mass spectra of PhoH2$_{\text{alt MSMEG}}$ (Spectra obtained from CPR, University of Otago). A range of masses were observed, the largest mass being 54817.07 Da. Lines above the masses indicate area shown in the enlargement.

Taken together the MALDI TOF/TOF peptide analysis and LTQ-Orbitrap results reflect a pool of the same protein but of different lengths, and suggest that translation can occur from the alternative start site.

A summary of all analyses and their main results are listed in Table 4.5.

Table 4.5 Summary of all analyses

Protein	Analysis	Main results
$PhoH2_{alt\ MTB}$	MALDI TOF MS – tryptic digest	37 % sequence coverage
	MALDI TOF MS – whole protein	Intact mass of 49207.2 Da
	MALDI TOF/TOF MS – SPITC	53 % sequence coverage
	LTQ-Orbitrap MS	83.8 % sequence coverage
	MALDI TOF/TOF MS – whole protein	Intact mass of 49162.99 Da
$PhoH2_{alt\ MSMEG}$	MALDI TOF/TOF MS – SPITC	59 % sequence coverage
	MALDI TOF/TOF MS – whole protein	Intact mass of 54817.07 Da
	LTQ-Orbitrap MS	86.75 % sequence coverage

4.3 Discussion

Structure-function studies of PhoH2, and the domains that make up PhoH2, from *M. tuberculosis* and *M. smegmatis* would enable a greater understanding of the role of these proteins in mycobacteria. Obtaining soluble protein expression is the first step in achieving this.

4.3.1 Cloning and expression using pYUB1049

The traditional mycobacterial shuttle vector used for cloning mycobacterial proteins into *M. smegmatis* via *E. coli* is pYUB1049. This vector was the first cloning vector used for protein expression of PhoH2 from *M. tuberculosis* with a C- or N-terminal his-tag in *M. smegmatis* $mc^2$4517 cells. Soluble expression of PhoH2 was gained only with the cloning of the $Rv1095_{alt}$ gene sequence with a C-terminal his-tag, which included a short segment of sequence upstream of *Rv1095* deemed a possible alternative translational start site by RT-PCR and promoter analysis. $PhoH2_{alt\ MTB}$ was tested in a lysis buffer which indicated that the protein was soluble in each of the buffers screened (data not shown), however, 50 mM Na-phosphate buffer pH 7.4, 200 mM NaCl was later found to be the only buffer suitable for concentrating protein, other buffers prevented the successful concentrating of $PhoH2_{alt\ MTB}$.

It was assumed that there was no expression of $PhoH2_{MTB}$ from its annotated start site due to its potential toxicity to the cells or vector, as when expressed with a possible regulatory or inhibitory sequence, expression was observed. This is analogous to PIN-domain (VapC) from *M. smegmatis* which does not express in the absence of its antitoxin (VapB) (McKenzie 2011).

4.3.2 Cloning and expression using $pDONR_{221}$ and $pDEST_{SMG/17}$

The Gateway cloning system for expression of proteins was the second cloning system used for protein expression of $PhoH2_{MTB}$, and the PIN and PhoH-domains of PhoH2 from *M. tuberculosis*, and to test the requirement for cloning from the alternative start site for $PhoH2_{MTB}$ by cloning $PhoH2_{alt\ MTB}$. Two destination vectors were used; $pDEST_{SMG}$ designed for the overexpression of mycobacterial proteins in *M. smegmatis* $mc^2$4517, and $pDEST_{17}$ for the overexpression of proteins in *E. coli*.

No protein expression was observed for $PhoH2_{MTB}$, $PhoH2_{alt\ MTB}$, or the PIN-domain, and insoluble protein expression was observed with overexpression of the PhoH-domain in *M. smegmatis*. Those protein constructs that did not show any signs of expression were re-transformed into *M. smegmatis* $mc^2$4517 cells, and the PhoH-domain which had insoluble protein expression was subject to a lysis buffer screen (section 2.4.6.1) and reduced incubation temperature for overexpression culture (section 2.4.6.2). $PhoH2_{MTB}$, $PhoH2_{alt\ MTB}$ and the PIN-domain proteins did not express despite the second round of transformation, and the PhoH-domain remained insoluble.

$PhoH2_{MTB}$, the PIN-domain and PhoH-domain from *M. tuberculosis* were also cloned into $pDEST_{17}$ for overexpression in *E. coli*. $PhoH2_{MTB}$ did not express, however, there was evidence of enrichment around the expected band size present on the Ni resin. This band was observed with each of the $pDEST_{17}$ *E. coli* small scale expression trials. $PhoH2_{MTB}$, and both the PIN and PhoH-domain proteins that expressed insolubly, were subjected to reduced incubation temperatures for protein expression, and a lysis buffer screen.

These approaches did not remedy the lack of protein expression or insoluble protein expression.

It is possible given the lack of protein expression of $PhoH2_{MTB}$ and $PhoH2_{alt\ MTB}$ when cloned with a N-terminal his-tag in pYUB1049 that the position of the his-tag at the N-terminus of the protein was responsible for difficulties with protein expression. The length of the linker sequence between the N-terminal his-tag and the start of the protein sequence may have also contributed to the poor expression and insolubility of these proteins (Woestenenk et al. 2004; Kurz et al. 2006).

4.3.3 Cloning and expression using pYUB28b

One of the difficulties with using the mycobacterial shuttle vector pYUB1049 is the restriction based removal of a ~1000 bp segment from the multiple cloning site which is replaced with the gene of interest. This requires gel purification of the digested vector away from the ~1000 bp piece, which is generally inefficient, to ensure re-ligation of the vector with the ~1000 bp segment does not occur. In order to recover enough vector for ligations, large quantities of vector needed to be digested and gel purified. The mycobacterial shuttle vector pYUB28b overcomes the need to digest away the large ~1000 bp segment. Vector pYUB28b is a modified pYUB1049 vector which has had the original fragmented multiple cloning site replaced with the multiple cloning site from pET28b (Bashiri et al. 2010). This replacement has eliminated the need to gel purify digested vector away from the large ~1000 bp segment prior to ligation and allows ease of C-terminal his-tag expression, compared with pYUB1049 which required inclusion of additional bases in the forward primer to ensure the gene would be in frame for subsequent protein expression.

The pYUB28b vector was the third cloning vector used to try to express PhoH2 from *M. tuberculosis* with a C-terminal his-tag in *M. smegmatis* mc^24517 cells. This vector was also used to express PhoH2 and $PhoH2_{alt}$ from *M. smegmatis* which were soluble. The solubility observed with PhoH2 from

M. smegmatis is likely due to endogenous levels of the alt peptide present, due to expression of this protein in the organism from which it is derived.

This vector additionally was used to try to solubly express the PIN, PhoH, and PIN_{alt} domains of PhoH2 from *M. tuberculosis*, and the PIN, $PhoH_{+1}$ and $PhoH_{+2}$ domains that were cloned with additional amino acids at the N-terminus from *M. smegmatis*. This aimed to increase the likelihood of solubilisation given the insolubility observed earlier with cloning of the PhoH-domain from *M. tuberculosis.* This approach was unsuccessful and no further proteins were gained using this system.

4.3.4 Large scale protein expression and purification

Large scale expression of $PhoH2_{alt}$ from *M. tuberculosis*, and PhoH2 and $PhoH2_{alt}$ from *M. smegmatis* expressed in *M. smegmatis* $mc^2$4517 cells using autoinduction medium had mixed success with regard to consistent successful expression and subsequent purification. In general, when expression and IMAC purification were successful the protein yield was relatively poor (0.1-0.5 mg/ml), and to concentrate the proteins, the proteins required purification in Na-phosphate buffer and the concentrator had to be blocked with 0.05 % (v/v) Tween-80. The highest concentration obtained for these proteins following concentration was 1-2 mg/ml. For further purification, the smaller Superdex™ S200 10/300 size exclusion column was used in place of the Superdex™ S200 16/60 size exclusion column to try to recover as much protein as possible.

In light of the low yields for PhoH2 from mycobacteria, a thermophilic homologue was used from *T. bispora*. This organism is classified in the actinobacteria, and is an obligate aerobe with an optimal temperature range of 50-65 °C (Liolios et al. 2010). It was hypothesised that expression and purification of a thermophilic PhoH2 protein would have increased stablity due to the thermophilic origins of this protein, and that expression would gain high protein yields owing to the optimisation of the amino acid sequence for expression in *E. coli*. This idea has subsequently been extended to the recently characterised *Mycobacterium thermoresistibile* a thermophile from

mycobacteria (Edwards et al. 2012). With large scale expression of $PhoH2_{TBIS}$ in *E. coli* using IPTG induction, expression and subsequent purification were successful. This protein was able to be purified and concentrated in a range of buffers and the highest protein concentration obtained was greater than 25 mg/ml. This protein was stored at RT as precipitation was observed with storage at 4 °C.

4.3.5 Small and large scale solubilisation trials and refolding

It has been reported that N-lauroylsarcosine in combination with triton X-100 can solubilise previously insoluble proteins without the need for further refolding or renaturation efforts (Zou et al. 2009; Park et al. 2011).
The use of these detergents in the solubilisation and purification of the PIN and PhoH-domains of PhoH2 from *M. tuberculosis* that were cloned into $pDEST_{17}$ and expressed in *E. coli* initially looked promising, however, further experimentation and up scaling was unsuccessful and soluble protein was not obtained. The results suggested the possibility that with the initial promising treatment of the PIN and PhoH-domains with detergents had caused soluble aggregation of the proteins and therefore unsuccessful solubilisation.

The refolding of the PIN and PhoH-domain of PhoH2 from *M. tuberculosis* was unsuccessful. In this case, the correct conditions for correct refolding were not met upon scaling up.

4.3.6 Protein crystallisation and Electron microscopy

Initial crystallisation trials with mycobacterial PhoH2 proteins were challenging due to the low yield and buffer requirements of these proteins. Crystallisation screens using the thermophilic homologue $PhoH2_{TBIS}$, however, were more favourable due to the ease of purification in a range of buffers and with concentrating leading to a far greater protein concentration than with mycobacterial PhoH2 proteins. After fine screening around one promising condition obtained from the crystallisation screen, protein crystals were identified and optimisation attempts were successful to a point for data collection. A maximum resolution of ~7.5 Å was obtained. Preliminary data

analyses performed on the two sets of diffraction data led to the identification of possible space groups and with the self rotation function analysis, probable hexameric non-crystallographic symmetry, together with the calculated molecular weights gained from size exclusion chromatography. Electron microscopy revealed a variety of oligomeric assemblies. The symmetry mate structure of PhoH protein (3B85) shows an array of hexamers (Figure 1.18) suggesting the possibility of multiple oligomeric states. These oligomeric states may be similar to the assemblies revealed by EM of $PhoH2_{TBIS}$. The heterogeneity of $PhoH2_{TBIS}$ purified by size exclusion chromatography determined by EM, can further account for the resolution limitations met with protein crystallisation.

4.3.7 Investigation of $PhoH2_{alt}$ proteins

Initial analysis using mass spectrometry of $PhoH2_{alt\ MTB}$ at the early stages of this research identified that PhoH2 from *M. tuberculosis* had been successfully cloned and expressed in *M. smegmatis* $mc^2$4517 cells. It was unclear, however whether translation was occurring from the alternative start site. With VapC PIN-domain proteins that are cloned and expressed as a VapBC complex in which the gene for the VapB protein is out of frame of the VapC protein, separation of the antitoxin and toxin is observed when purified VapBC protein is run on a SDS-PAGE gel (McKenzie 2011), this is not observed in the case of $PhoH2_{alt}$ proteins.

When further PhoH2 proteins were cloned, expressed and purified from alternative and annotated start sites (*M. smegmatis* and *T. bispora*) it was evident that each ran fairly similarly and as a single band on a SDS-PAGE gel (Figure 4.6 and Figure 4.7). It was necessary therefore to investigate $PhoH2_{alt\ MTB}$ and $PhoH2_{alt\ MSMEG}$. $PhoH2_{alt\ MTB}$ and $PhoH2_{alt\ MSMEG}$ were therefore tested for their N-terminus at the CPR, University of Otago.

In this approach to identify the N-terminus, both proteins were subjected to a two-step chemical modification prior to proteolytic digestion and analysis by MALDI TOF/TOF. The sulfonation step aimed to introduce a negatively charged sulfonic acid group on the N-terminus of the protein (Chen et al.

2004), which can be identified during peptide analysis. In this case the N-terminal modification was not identified, however peptides were identified both tryptic and non-tryptic, indicating the possibility of an N-terminal peptide. As a result the tryptic peptides gained by MALDI TOF/TOF MS each protein was analysed by LTQ-Orbitrap with the aim to improve sequence coverage. Additional peptide masses were identified which mapped onto the $PhoH2_{alt}$ sequence. For $PhoH2_{alt\ MTB}$, one peptide was identified within the sequence between the alternative and annotated start site with an ion score of 45. Additional peptides towards the N-terminus were identified significantly (ions score 101 and 65) one with and one without the N-terminal modification (Figure 4.25 (blue bar)). It is possible that the other N-terminal peptides (grey and green bars) found without their respective N-terminal tag had lost their modification due to full tryptic cleavage given the positions of lysine (K) and arginine (R) residues near the N-terminus. Two peptides were matched between the alternative and annotated start site for $PhoH2_{alt\ MSMEG}$, neither carried the SPITC N-terminal modification.

Intact mass measurements gained by MALDI TOF/TOF reported a mass for $PhoH2_{alt\ MTB}$ similar to that gained previously (Figure 4.23 and Figure 4.27), and for $PhoH2_{alt\ MSMEG}$ a mass close to the predicted MW (53728.2 Da) for the full length construct (54817.07 Da) (Figure 4.28). These results together confirm that protein expression of $PhoH2_{alt\ MTB}$ and $PhoH2_{alt\ MSMEG}$ can occur from the alternative start site.

There are two possible hypotheses that lead to the quaternary structure of $PhoH2_{alt}$ proteins. The first is that translation of the alt peptide and PhoH2 protein is independent leading to the production of two polypeptides that interact with one another post-translation. The second is that translation occurs from the alternative site and when the ribosome reaches the end of the alt transcript it falls off releasing the alt peptide, then binds to the nearby Shine-Delgarno sequence and resumes translation of the PhoH2 transcript. In both cases it is likely that the two proteins are transcriptionally coupled due to the overlap of the stop and start codons. Upstream of the alt peptide sequence in both *M. tuberculosis* and *M. smegmatis* there is conservation of

the -10 promoter sequences (Figure 3.12). After translation, the two proteins associate to form a stable complex that survives purification. The presence of a peptide identified by MS that matched between the alternative and annotated start sites is suggestive of an interaction between the alt peptide and the PhoH2 protein.

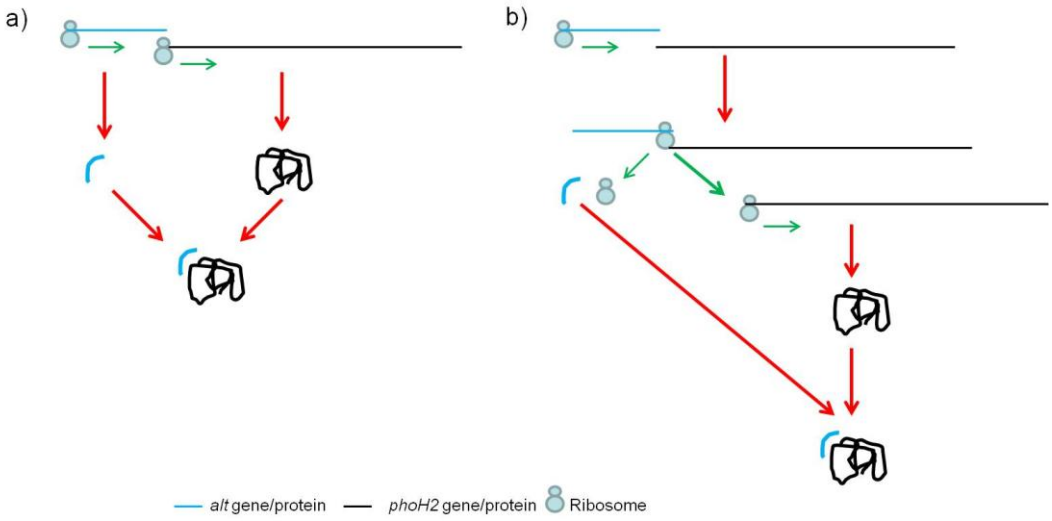

Figure 4.29 Schematic of PhoH2$_{alt}$ protein translation
Schematic depicting the two possible hypotheses for PhoH2$_{alt}$ protein translation. a) The ribosome begins translation from the alternative start site and the annotated start site leading to the production of two polypeptides that associate with one another post-translation. b) The ribosome begins translation from the alternative start site and when it reaches the end of the alt peptide the ribosome either releases the peptide and falls off, or it releases the peptide and binds to the nearby Shine-Delgarno sequence and starts translation from the annotated PhoH2 start site. These hypotheses lead to the production of two polypeptides that associate with one another post-translation.

The differences observed between PhoH2 and PhoH2$_{alt}$ proteins when analysed by Native-PAGE can be accounted for by the translation from the alternative start site. The apparent absence of a negative functional effect with full translation of PhoH2$_{alt}$ suggests that alt peptide may be susceptible to degradation post-translation more rapidly than the PhoH2 protein. The mixed population of purified PhoH2$_{alt}$ protein which was identified as the same protein but of different lengths identified during mass spectrometry analyses also suggest a labile nature of the translated alt peptide.

The extensive attempts made to obtain soluble expression of PhoH2 from mycobacteria have provided some information regarding the nature of these

proteins. A number of cloning, expression and purification approaches were taken to try to gain a variety of soluble protein constructs for function and structure studies. The cloning of the *Rv1095*$_{alt}$ from *M. tuberculosis*, and either *MSMEG_5247*$_{alt}$ or *MSMEG_5247* from *M. smegmatis* in mycobacterial expression cells, led to soluble protein.

This result enabled characterisation of the biochemical activity of PhoH2 from mycobacteria, which is the focus of Chapter 5.

Chapter Five: Functional Activity of mycobacterial PhoH2 proteins

This chapter will focus on the functional characterisation of three mycobacterial PhoH2 proteins: $PhoH2_{alt\ MTB}$, $PhoH2_{alt\ MSMEG}$ and $PhoH2_{MSMEG}$.

5.1 Introduction

The fusion between PIN and PhoH domains is suggestive of functional linkage (Marcotte et al. 1999). My hypothesis was that given the conserved motifs identified in PhoH2, the biochemical function of the protein is to unwind RNA via the PhoH-domain coupled with the hydrolysis of ATP, and cleave the single stranded RNA product via the PIN-domain.

From the motif analysis (section 1.4.3), it appears that PhoH2 is missing a number of the traditional helicase motifs concerned with nucleic acid binding (Ia, Ib, IV, V, VI) (Table 1.2). However, the highly conserved, arginine rich RGRTL motif, which forms a loop that projects near the P-loop active site, and other conserved motifs, in conjunction with R171 from the adjacent subunit, which bound SO_4 in the PDB 3B85 structure, indicates that PhoH2 may be able to bind RNA (Figure 5.1). R171 (PDB 3B85) could possibly adopt a similar role to motif VI of SF1 and 2, which couples NTP hydrolysis with helicase activity and therefore translocation of the enzyme along the RNA (Bleichert & Baserga 2007). It is also possible with subunit movements during an ATP hydrolysis cycle that the loop may protrude into the central channel.

A number of proteins containing arginine-rich motifs are known to bind RNA (Bayer et al. 2005). The arginine-rich motifs in the PhoH protein structure (PDB 3B85) are positioned close to the central channel (Figure 5.1). In other helicases such as Rho and E1, a series of loops project into the central pore that are responsible for interactions with the incoming substrate (Thomsen & Berger 2009).

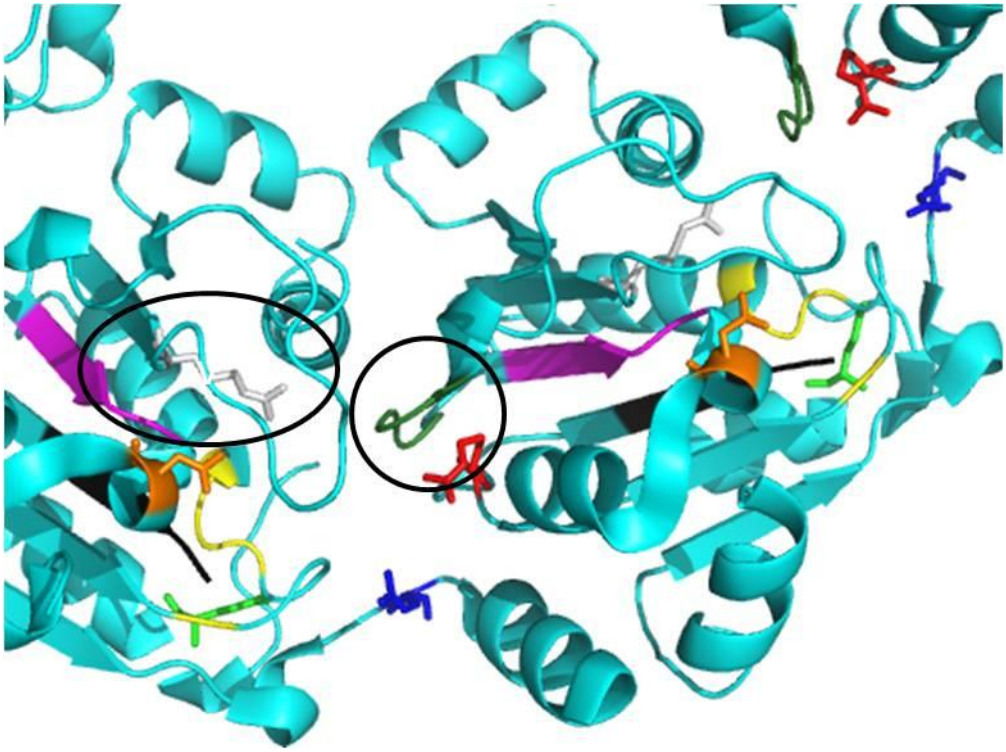

Figure 5.1 Predicted active site of PhoH proteins
Portion of hexameric PhoH protein showing the interface between two adjacent subunits and the positions of conserved residues/motifs Q (green), Walker A (yellow), R residue (light grey), RGRTL (dark green), Walker B (magenta), N residue (SI) (orange), R residue (SRH) (red), III (black) and R residue (SII) (blue). The highly conserved RGRTL motif and R residue (R171) predicted to be involved with RNA recognition are circled.

The RNA binding and substrate specificity of PhoH2 may be determined by the PIN-domain. PIN-domain proteins cleave ssRNA in a sequence specific, Mg^{2+}/Mn^{2+} dependent manner (Arcus et al. 2011). Table 5.1 summarises the sequence specificity of mycobacterial VapC proteins determined to date (Ahidjo et al. 2011; McKenzie et al. 2012a; McKenzie et al. 2012b; Sharp et al. 2012).

Table 5.1 Specificity of VapC proteins from mycobacteria

VapC	Sequence specificity	Reference
MSMEG_1284	A U A U and A U A A	(McKenzie et al. 2012b)
Rv0065/Rv0617	G/C G G/C G/C/A	(Ahidjo et al. 2011)
Rv0595c	A C G C and A C A/U G C	(Sharp et al. 2012)

Alternatively, both the PIN and PhoH-domains may coordinate to confer substrate specificity. For example, RhlB is a helicase that is part of the RNA degradosome that provides substrates suitable for ribonucleases (Chandran et al. 2007). For activity, this protein interacts with RNase E which stimulates

the unwinding and ATPase activity of RhlB, and provides proximal RNA binding sites for RNase E with which it cooperates to unwind RNA. RNase E is a 5'-end-dependent endonuclease (Mackie 1998). The proposed site where interaction between RhlB and RNase E occurs is at a distance from the helicase core (Chandran et al. 2007). It is currently unknown how the PIN and PhoH-domains are structurally arranged. However their arrangement is unlikely to prevent coordination for activity. The interaction between RhlB and RNase E is predicted to partially suppress the RNA binding of the helicase further indicating coordination of both proteins for activity. This strengthens the case that the PIN and PhoH-domain of PhoH2 may work cooperatively to recognise, unwind and cleave substrates as these two domains that make up PhoH2 are fused to one another.

Some helicases show little substrate specificity when analysed *in vitro* and require a large excess of protein over substrate to observe activity. This may be due to the absence of cofactors, low intrinsic affinity for the substrate or partial inactivation of the purified protein (Tanner & Linder 2001).

PhoH proteins have reported ATP binding activity (Kim et al. 1993), however hydrolysis of ATP has not been experimentally verified. ATPase activity in DEAD-box proteins is dependent on RNA (Cordin et al. 2006), although a precise RNA substrate is not always required (Rocak & Linder 2004). ATPase activity can be greatly increased in the presence of specific RNA substrates (Fuller-Pace et al. 1993).

Taken together, the observations made in the literature regarding PIN-domain proteins, PhoH proteins and other helicases led to the hypothesis that the PhoH-domain of PhoH2 has ATPase activity with moderate RNA binding, which relies on the PIN-domain for additional substrate specificity, binding, coordination, and degradation.

5.2 Results

5.2.1 ATPase assays

The first biochemical property of PhoH2 that was investigated was ATPase activity. The hydrolysis of ATP by PhoH2 was measured using a modified malachite green method (Baykov et al. 1988). This test quantifies the green complex that forms between malachite green, molybdate and free orthophosphate, which can be measured using a spectrophotometer at 600-660 nm.

Each of the PhoH2 proteins was screened in a pilot trial and 1.25 µg of protein per 100 µl reaction was suitable for sufficient colour development using substrate concentrations from 0-6 µM ATP in the presence of 1 mM $MgCl_2$ and was incubated for 10 minutes at 37 °C.

Each PhoH2 protein had ATPase activity that was reduced in the presence of EDTA (2 mM), indicating that Mg^{2+} was required for optimal activity (Figure 5.2).

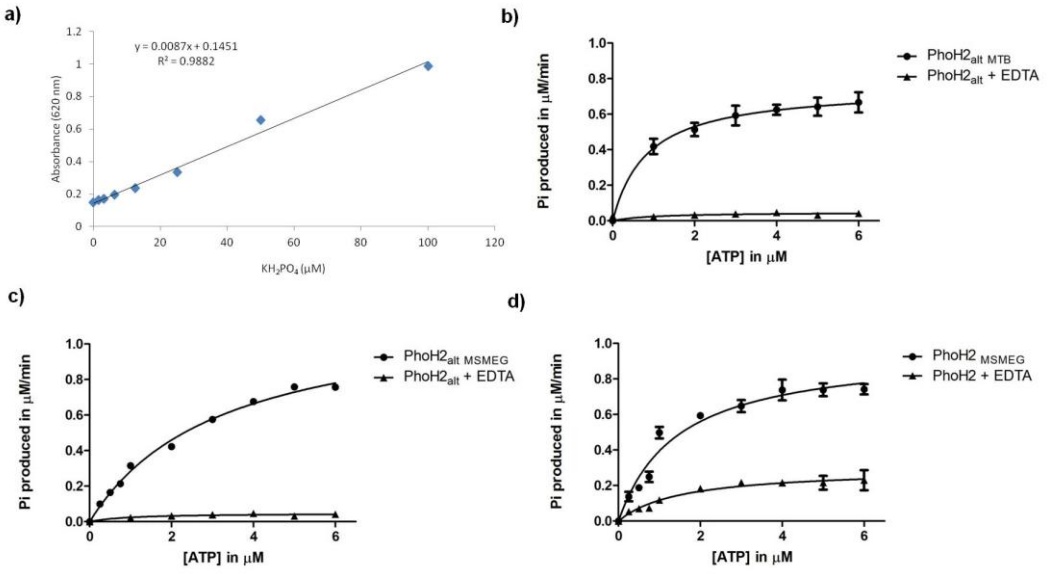

Figure 5.2 ATPase activity of PhoH2$_{alt\ MTB}$, PhoH2$_{alt\ MSMEG}$ and PhoH2$_{MSMEG}$
a) Typical standard curve obtained for 0 µM - 50 µM phosphate and Michaelis Menten plots for 0.25 µM b) PhoH2$_{alt\ MTB}$ c) PhoH2$_{alt\ MSMEG}$ and d) PhoH2$_{MSMEG}$ in the presence of 1 mM $MgCl_2$ and 2 mM EDTA. Error bars represent $^{+/-}$SD.

The kinetic properties calculated for each PhoH2 protein are listed in Table 5.2.

Table 5.2 Michaelis Menten kinetics of PhoH2 proteins

Protein	K_M (μM)	V_{max} (μM/min)	k_{cat} (min^{-1})
$PhoH2_{alt\ MTB}$	0.8	0.8	3.0
$PhoH2_{MSMEG}$	1.4	1.0	3.9
$PhoH2_{alt\ MSMEG}$	3.2	1.2	4.8

The effect of total and specific RNA on ATPase activity was tested by adding 0.0025 μM RNA to each of the reactions containing the same amount of protein as in Figure 5.2 (1:100 molar ratio) and comparing the resulting plots with reactions that did not contain RNA. Initially, a time course assay which included total *M. smegmatis* mc^2155 RNA and 1 μM ATP was performed with $PhoH2_{MSMEG}$ and possibly suggested a slight although not significant, increase of phosphate release in the presence of RNA (Figure 5.3a). ATPase activity was also tested in the presence of a specific RNA oligonucleotide substrate (5'3) selected based on its use in the unwinding and ribonuclease assays (section 5.2.2) with $PhoH2_{MSMEG}$ and $PhoH2_{alt\ MTB}$. Attempts were made to test $PhoH2_{alt\ MSMEG}$ with RNA, however protein purification was unsuccessful. No significant difference was observed in ATPase activity with the addition of RNA (Figure 5.3b and c).

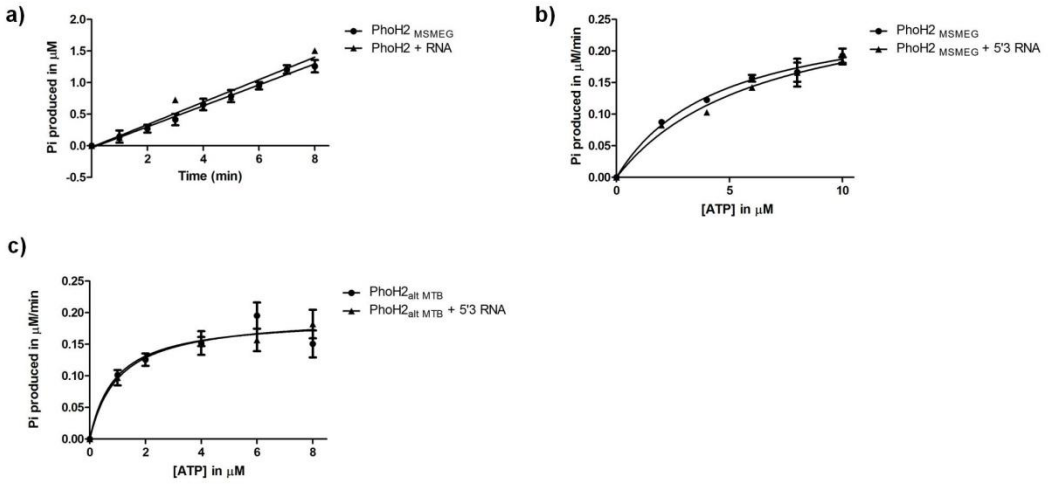

Figure 5.3 The effect of RNA on ATPase activity of $PhoH2_{MSMEG}$ and $PhoH2_{alt\ MTB}$
a) Time course of phosphate released in the presence and absence of total RNA and Michaelis Menten plots of b) $PhoH2_{MSMEG}$ and c) $PhoH2_{alt\ MTB}$ in the presence and absence of specific RNA oligonucleotide (5'3). Error bars represent $^{+/-}$SD.

5.2.2 Unwinding and ribonuclease assays

It is common for helicases to require a ss region for substrate loading, before performing unwinding activity on a duplex region, as a result of the initial low affinity of helicases for dsRNAs (Tanner & Linder 2001). In addition, most helicase proteins also perform helicase activity in specific directions (Singleton et al. 2007). SF1/2 and AAA+ have examples of both 5' to 3' and 3' to 5' directionality, SF3 members demonstrate 3' to 5' activity and SF4/5 demonstrate 5' to 3' activity. Owing to the combination of conserved motifs present in PhoH2 from SFs 1/2/AAA+ and AAA, which adopt both 5' to 3' and 3' to 5' unwinding directionality, the unwinding and ribonuclease activity of PhoH2 was assayed *in vitro* by observing the conversion of a range of RNA duplexes to monomers and degradation. Short self annealing oligomers with identical duplex forming regions and either 5'- or 3'- ss tails and a blunt substrate, were tested in the manner used by Chandran (2007) to characterise ATP-dependent RNA helicase RhlB.

Initially, 5'-tailed, 3'-tailed, ss and blunt RNA and DNA oligonucleotides were tested with each mycobacterial PhoH2 protein at a 1:100 molar ratio. The results of these assays demonstrated that $PhoH2_{alt\ MTB}$, $PhoH2_{alt\ MSMEG}$ and $PhoH2_{MSMEG}$ had ATP-dependent ribonuclease activity with the 5'-tailed and 3'-tailed RNA substrates and perhaps slight activity on ss substrates. This was shown by the degradation of lower molecular weight material and enrichment at the bottom of the gel in the presence of ATP (+ATP) compared with the control reactions (-ATP, +EDTA, AMP-PNP) (Figure 5.4a and b). The control reactions tested for ATP-dependent activity (-ATP), cofactor-dependent activity (+EDTA), and the requirement for ATP hydrolysis (AMP-PNP used in place of ATP).

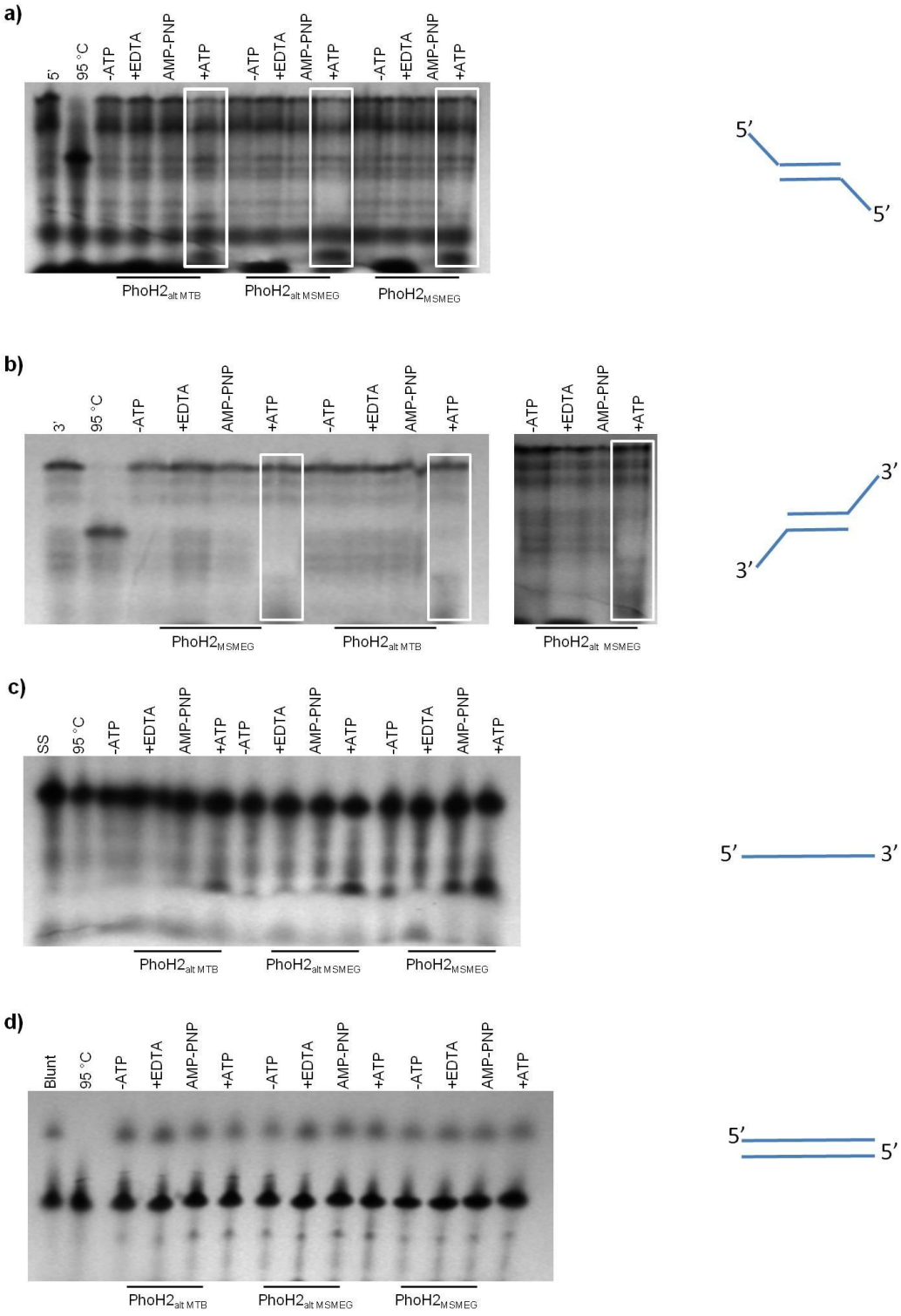

Figure 5.4 Ribonuclease activity of PhoH2$_{\text{alt MTB}}$, PhoH2$_{\text{alt MSMEG}}$ and PhoH2$_{\text{MSMEG}}$ on 5'-tailed, 3'-tailed, ss and blunt RNA substrates

a) 5'-tailed RNA b) 3'-tailed RNA c) ss RNA and d) blunt RNA substrates tested with PhoH2$_{\text{alt MTB}}$, PhoH2$_{\text{alt MSMEG}}$ and PhoH2$_{\text{MSMEG}}$ at 37 °C for 30 minutes. Labels: First column substrates (5', 3' ss or blunt), positive control for unwinding (95 °C), reaction performed in the absence of ATP (-ATP), reaction performed in the presence of EDTA (+EDTA), reaction performed with non-hydrolysable AMP-PNP in place of ATP (AMP-PNP) and reaction performed in the presence of ATP (+ATP). White boxes highlight results of +ATP reactions where activity has occurred.

The hybridisation of the substrates to form 5'-tailed, 3'-tailed or blunt substrates was poor as indicated by the number of hybridisation intermediates present on the gel. Even with the ss oligonucleotide, a mixed population of species was apparent. For accurate analysis of PhoH2 activity on the oligonucleotide substrates, the substrates needed to be more homogeneous. Many hybridisation attempts were made to gain the desired hybridised substrates, these were unsuccessful with these substrates.

There was no activity of any PhoH2 protein on the equivalent DNA substrates (Figure 5.5). The annealing of these substrates was also poor.

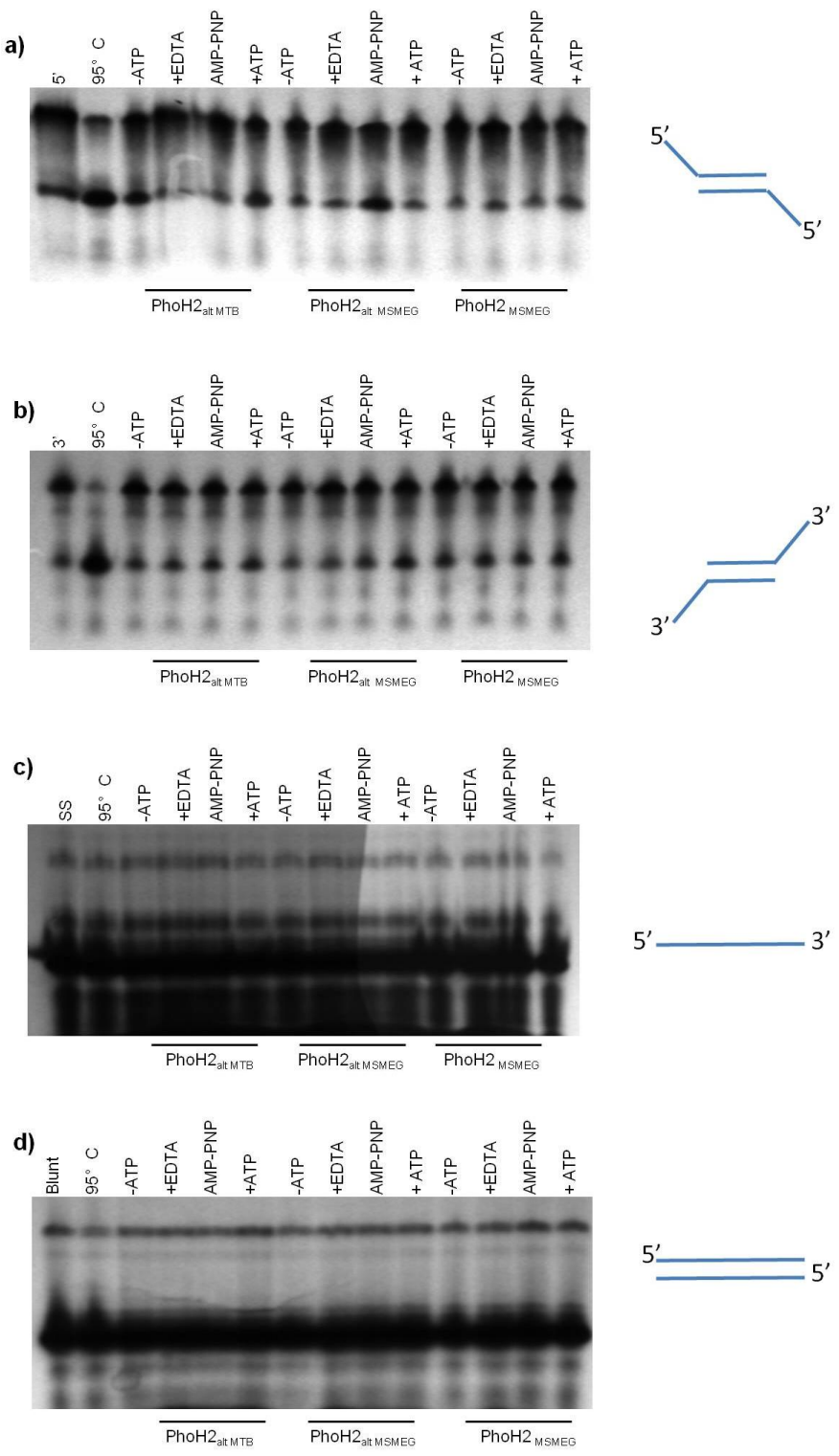

Figure 5.5 Activity of PhoH2$_{alt\ MTB}$, PhoH2$_{alt\ MSMEG}$ and PhoH2$_{MSMEG}$ on 5'-tailed, 3'-tailed, ss and blunt DNA substrates

a) 5'-tailed DNA b) 3'-tailed DNA c) ss DNA and d) blunt DNA substrates tested with PhoH2$_{alt\ MTB}$, PhoH2$_{alt\ MSMEG}$ and PhoH2$_{MSMEG}$ at 37 °C for 30 minutes. Labels: First column substrates (5', 3' ss or blunt), positive control for unwinding (95 °C), reaction performed in the absence of ATP (-ATP), reaction performed in the presence of EDTA (+EDTA), reaction performed with non-hydrolysable AMP-PNP in place of ATP (AMP-PNP) and reaction performed in the presence of ATP (+ATP).

Clearing of lower molecular weight hybridisation intermediates was most apparent with the 3'-tailed RNA oligonucleotide (Figure 5.4b). To obtain a substrate closer to that preferred by PhoH2, two shortened versions of the 3'-tailed RNA oligonucleotide were designed (3'-AU and 3'-UA) that reduced the length of the 3'-tail by either two or five RNA bases respectively (Table 2.12). No activity of any kind was observed by PhoH2 on these substrates and again poor hybridisation was achieved (Figure 5.6).

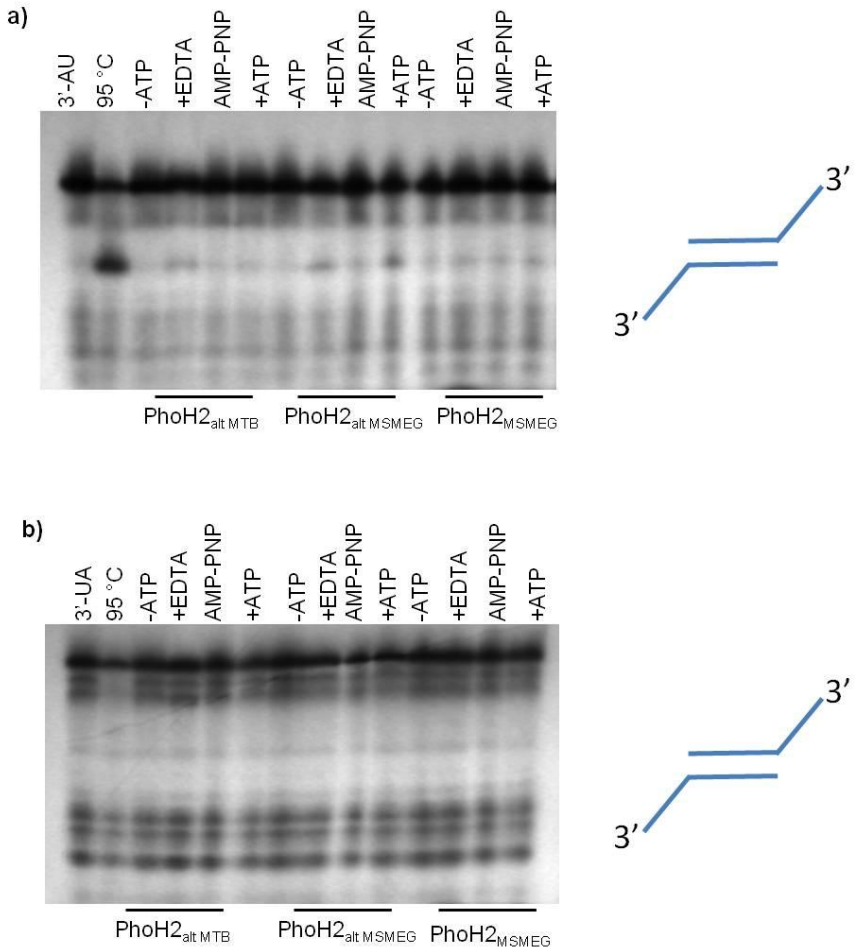

Figure 5.6 Activity of $PhoH2_{alt\ MTB}$, $PhoH2_{alt\ MSMEG}$ and $PhoH2_{MSMEG}$ on 3'-AU and UA RNA substrates

a) 3'AU-tailed RNA b) 3'UA-tailed RNA tested with $PhoH2_{alt\ MTB}$, $PhoH2_{alt\ MSMEG}$ and $PhoH2_{MSMEG}$ at 37 °C for 30 minutes. Labels: First column substrates (3'-AU or 3'-UA), positive control for unwinding (95 °C), reaction performed in the absence of ATP (-ATP), reaction performed in the presence of EDTA (+EDTA), reaction performed with non-hydrolysable AMP-PNP in place of ATP (AMP-PNP) and reaction performed in the presence of ATP (+ATP).

The next substrate designed was 5'-AC. In this substrate, the 5' terminal combination of residues had been changed from GA to AC which aimed to

improve radio labelling with adenosine 5'-triphosphate γ-^{33}P and subsequent hybridisation. This substrate was more homogenous and with each of the PhoH2 proteins ATP-dependent RNA binding of PhoH2 was observed (Figure 5.7a) as shown by the band shift on the gel. Due to this new mode of activity, three further RNA oligonucleotides were screened, 3'-AC which had the 5'-AC overhang positioned in the 3' orientation, a blunt 5'-AC substrate, 5'-ACAA which had the 3rd, 4th and 5th residues changed to their complementary RNA base, and a DNA equivalent of 5'-AC (Figure 5.7).

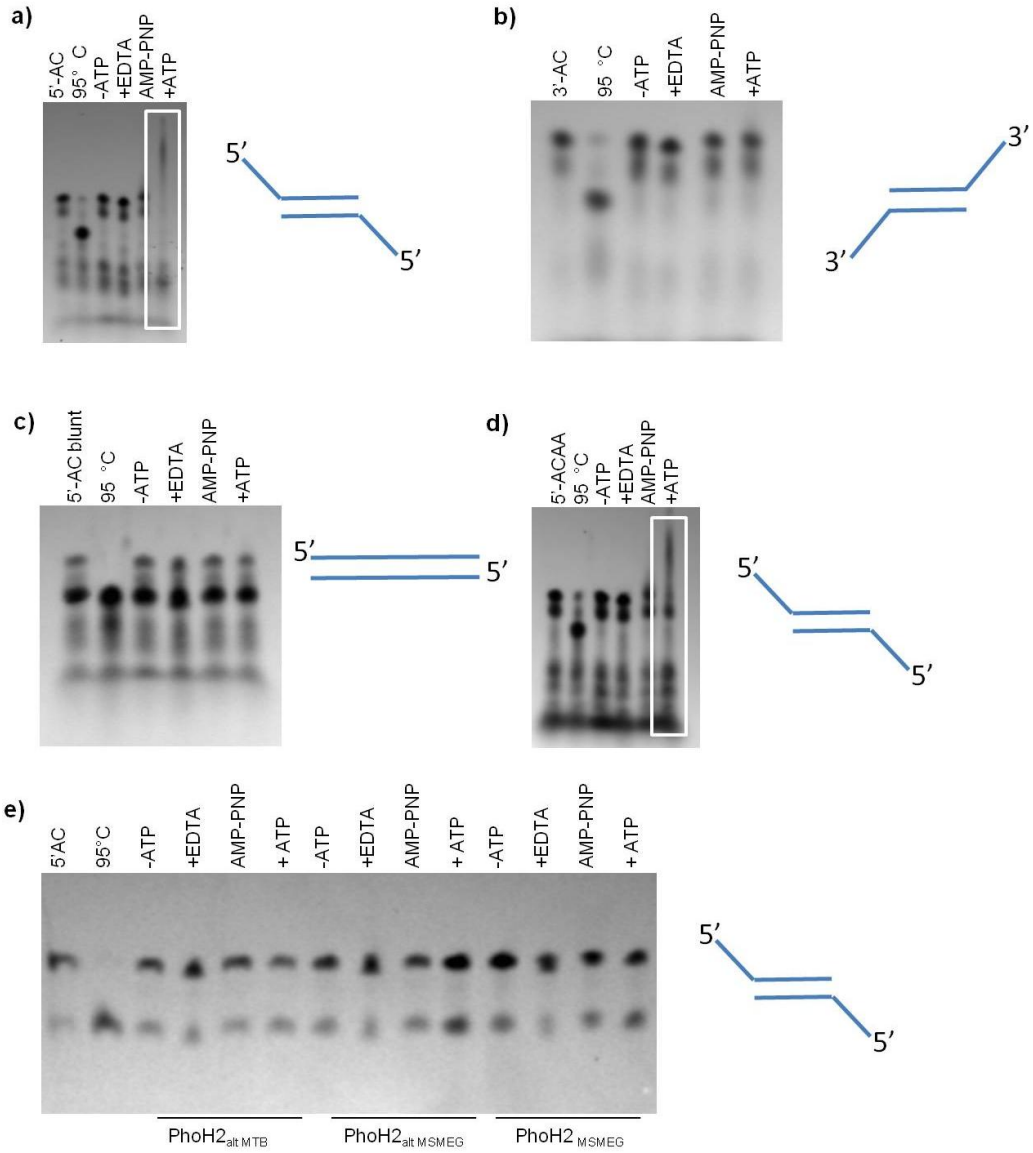

Figure 5.7 Binding and ribonuclease activity of PhoH2$_{MSMEG}$ on 5'-AC, 3'-AC, 5'-AC blunt, 5'-ACAA RNA substrates and 5'AC DNA substrate
a) 5'-AC RNA b) 3'-AC RNA c) 5'AC blunt RNA d) 5'ACAA RNA e) 5'AC DNA substrates tested with PhoH2$_{MSMEG}$ at 37 °C for 30 minutes. The same results were obtained for PhoH2$_{alt\ MTB}$ and PhoH2$_{alt\ MSMEG}$. Labels: First column substrates (5'-AC, 3'-AC, 3'-AC blunt, 5'-ACAA RNA or 5'-AC DNA), positive control for unwinding (95 °C), reaction performed in the absence of ATP (-ATP), reaction performed in the presence of EDTA (+EDTA), reaction performed with non-hydrolysable AMP-PNP in place of ATP (AMP-PNP) and reaction performed in the presence of ATP (+ATP). White boxes highlight results of +ATP reactions where activity has occurred.

Binding was observed with 5'-ACAA in addition to 5'-AC (Figure 5.7d) but no activity of any kind was seen with any of the other substrates tested, likely due to the preference of PhoH2 for 5'-ssRNA overhangs.

A time course assay was undertaken with each PhoH2 protein to look at the rate of binding. The representative results for $PhoH2_{MSMEG}$ (Figure 5.8) show that after 5 minutes the protein has bound up most of the substrate and, over a longer time frame, enrichment at the bottom of the gel increased likely due to the partial cleavage of the substrate. In the presence of AMP-PNP, the substrate was also bound and cleaved, suggesting ATP-independent RNA binding and partial cleavage activity of PhoH2. The results for $PhoH2_{alt\ MTB}$ and $PhoH2_{alt\ MSMEG}$ were essentially the same as for $PhoH2_{MSMEG}$.

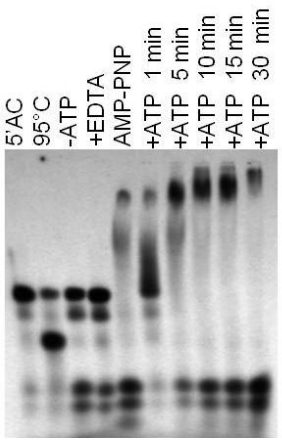

Figure 5.8 Binding and ribonuclease activity of $PhoH2_{MSMEG}$ on 5'AC RNA substrate over time
5'AC RNA time course with $PhoH2_{MSMEG}$. Labels: First column substrate (5'-AC RNA), positive control for unwinding (95 °C), reaction performed in the absence of ATP (-ATP), reaction performed in the presence of EDTA (+EDTA), reaction performed with non-hydrolysable AMP-PNP in place of ATP (AMP-PNP) and reaction performed in the presence of ATP (+ATP).

PIN-domain proteins are known to bind RNA (Lamanna & Karbstein 2009), and binding (in the absence of RNA cleavage) can be an indication that the substrate is still far from optimal. For this reason, a more systematic approach was taken with the design of the next five RNA substrates (5' 1-5). 5'1-5 RNA substrates are modified versions of the 5'AC substrate in which the 1st, 2nd, 3rd, 4th, and 5th residues have been sequentially changed to their complementary RNA base. These substrates were tested with $PhoH2_{MSMEG}$, and showed in the presence of ATP, full degradation of RNA oligonucleotides 5'3, 5'4 and 5'5 (Figure 5.9). It was assumed that the substrate was being unwound in order for the PIN-domain to degrade all the RNA as all PIN-domain proteins characterised thus far cleave only ss material (Arcus et al. 2011). Binding and only partial degradation activity of RNA substrates 5'1 and 5'2 were observed in the presence of ATP (Figure 5.9a and b). With these substrates, binding in the presence of AMP-PNP was not observed.

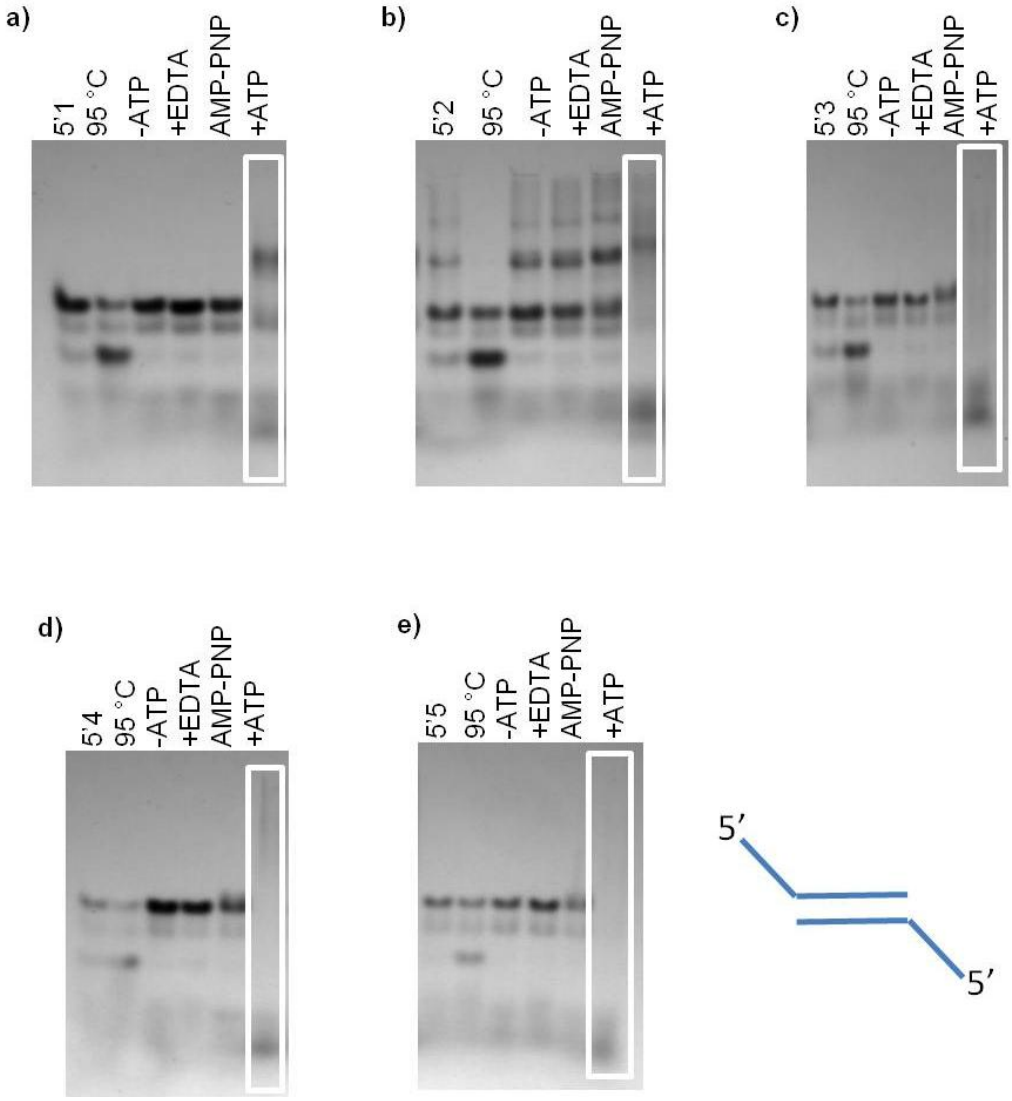

Figure 5.9 Unwinding and ribonuclease activity of PhoH2$_{MSMEG}$ on 5'1, 5'2, 5'3, 5'4 and 5'5 substrates

a) 5'1-tailed RNA b) 5'2-tailed RNA c) 5'3-tailed RNA d) 5'4-tailed RNA e) 5'5-tailed RNA tested with PhoH2$_{MSMEG}$ at 37 °C for 30 minutes. Labels: First column substrate (5'1-5'5 RNA), positive control for unwinding (95 °C), reaction performed in the absence of ATP (-ATP), reaction performed in the presence of EDTA (+EDTA), reaction performed with non-hydrolysable AMP-PNP in place of ATP (AMP-PNP) and reaction performed in the presence of ATP (+ATP). White boxes highlight results of +ATP reactions where activity has occurred.

Table 5.3 summarises the activity of PhoH2 on the 5'-AC derived substrates. The optimal sequence for unwinding and cleavage can be deduced from these results as A C A/U A/U G/C U. A uridine residue at the 6th position is important as with 5'ACAA (ACAACAACAUCA**GAGUGCGCACUC)** an A is present at position 6 and binding activity, but not unwinding and cleavage, was observed despite the ACAA combination of the first four RNA bases (Figure 5.7d).

Table 5.3 Summary of activity on 5'AC derived substrates

Oligonucleotide		Sequence 5'-3 (**duplex**)	Activity
RNA	5'AC	ACAUGUACAUCA**GAGUGCGCACUC**	Bind
	3' AC	**GAGUGCGCACUC**ACUACAUGUACA	X
	5' AC blunt	**ACAUGUACAUCAUGAUGUACAUGU**	X
	5' ACAA	ACAACAACAUCA**GAGUGCGCACUC**	Bind
	5'1	UCAUGUACAUCA**GAGUGCGCACUC**	Bind
	5'2	AGAUGUACAUCA**GAGUGCGCACUC**	Bind
	5'3	ACUUGUACAUCA**GAGUGCGCACUC**	Unwind/Cleave
	5'4	ACAAGUACAUCA**GAGUGCGCACUC**	Unwind/Cleave
	5'5	ACAUCUACAUCA**GAGUGCGCACUC**	Unwind/Cleave

To test the requirement of ATP hydrolysis for unwinding, a non-hydrolysable analogue (AMP-PNP) was used in place of ATP. Unwinding activity was not observed suggesting unwinding is stimulated by ATP hydrolysis (Figure 5.9). Binding activity was observed with some substrates in the presence of AMP-PNP therefore ATP hydrolysis is not required for oligomer binding (Figure 5.8).

5.2.3 UTR assays

RT-PCR revealed that the gene for *phoH2* in *M. tuberculosis* and *M. smegmatis* is part of a longer mRNA transcript (Figure 3.10). As raised in Chapter 3, certain RNase enzymes autoregulate their expression by binding directly to their 5' untranslated region. RNase E is one example which acts on a *cis*-acting stem-loop in the *rne* 5' untranslated region (Schuck et al. 2009). VapBC proteins are also known to autoregulate their expression via the DNA binding activity of VapB (Robson et al. 2009).

As discussed in Chapter 3, and owing to the unwinding and ribonuclease activity of PhoH2, it was possible that the upstream mRNA transcript associated with *phoH2* maybe a biochemical target of the PhoH2 protein. To test this, the alt transcript (from alternative start site to annotated start site) was successfully amplified from vector DNA (Figure 5.10) and transcribed from DNA into RNA.

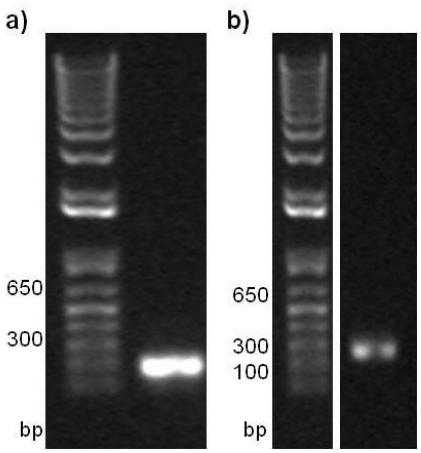

Figure 5.10 Amplified alt UTR substrates
PCR amplified UTR substrates from genomic DNA a) MTB_{alt} (156 bp) and b) $MSMEG_{alt}$ (147 bp).

The alt substrates were radio labelled and $PhoH2_{alt\ MTB}$ was found to have ATP-dependent activity on its own alt substrate, but not the alt substrate of *M. smegmatis* (Figure 5.11a and d). $PhoH2_{MSMEG}$ did not show any activity on either alt substrate (Figure 5.11b and c).

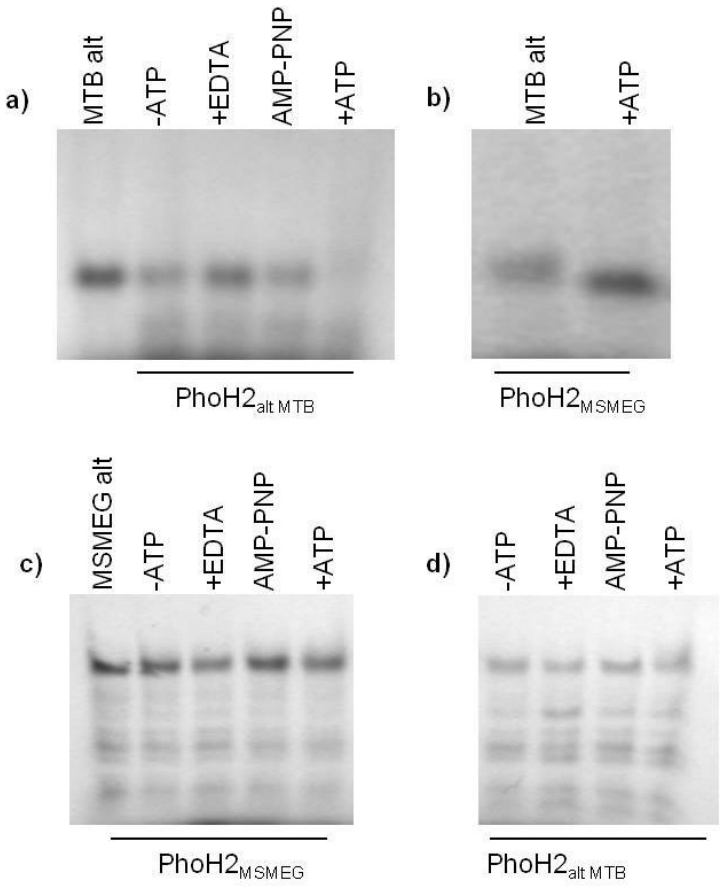

Figure 5.11 Activity of PhoH2$_{alt\ MTB}$ and PhoH2$_{MSMEG}$ on alt UTR substrates
a) PhoH2$_{alt\ MTB}$ activity on MTB alt b) PhoH2$_{MSMEG}$ activity on MTB alt c) PhoH2$_{MSMEG}$ activity on MSMEG alt and d) PhoH2$_{alt\ MTB}$ activity on MSMEG alt. Labels: First column substrate (MTB alt, MSMEG alt), reaction performed in the absence of ATP (-ATP), reaction performed in the presence of EDTA (+EDTA), reaction performed with non-hydrolysable AMP-PNP in place of ATP (AMP-PNP) and reaction performed in the presence of ATP (+ATP).

The optimal sequence deduced from the unwinding and ribonuclease assays was A C A/U A/U G/C U. The sequence ACAUGU is found twice in the alt substrate from *M. tuberculosis* in a ss loop and in a duplex region (Figure 5.12). This combination was not identified in the alt sequence from *M. smegmatis*.

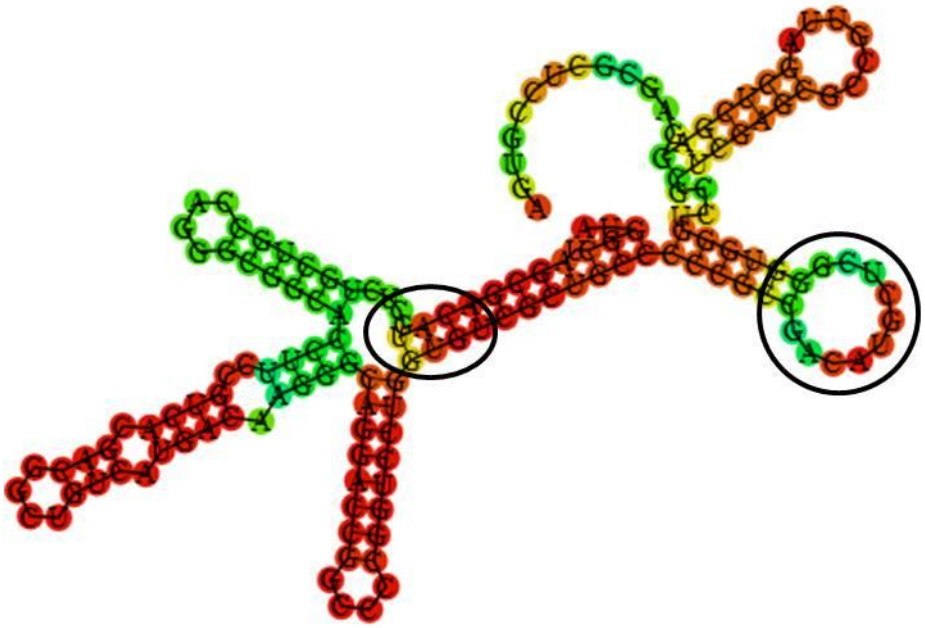

Figure 5.12 Secondary structure prediction of MTB alt sequence
The secondary structure for MTB alt substrate taken from the alternative start site to the end of the overlap at the beginning of the phoH2 gene was predicted using RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi). The black open circles show the positions of the ACAUGU combination. The free energy prediction is - 76.60 kcal/mol.

5.3 Discussion

5.3.1 ATPase activity

In order to perform the ATPase activity assays, each of the proteins had to be purified into a non-phosphate buffer. This proved challenging due to the instability of the protein in non-phosphate buffers, for example in Na-Hepes and Tris buffers the proteins 'crashed out' within four hours of IMAC purification, and if these proteins were further purified with size exclusion chromatography, they appeared to be soluble aggregates as they eluted. Assays were therefore carried out immediately following IMAC purification.

The rate of ATP hydrolysis by PhoH2 proteins was relatively slow. When total RNA or specific RNA were added to the reactions, no significant difference was observed, despite in the literature, there being the general requirement of nucleic acid binding or the association of protein cofactors to stimulate ATP hydrolysis (Singleton et al. 2007). This may be due to trace amounts of residual RNA present following the IMAC purification. The pre-treatment of the proteins with DNase or RNase to remove residual nucleic acids was

unfeasible due to the unstable nature of the proteins. This treatment caused the proteins to precipitate.

Hexameric helicases are expected to be held together in two ways; either by peripheral interactions of the hexameric ring or by interactions with an additional ring stacked on top (Ye et al. 2004). The molecular weights determined by size exclusion for PhoH2 suggest that the proteins are purified in a higher order conformation (Table 4.3). This may explain why ATPase activity is similar in the presence of RNA, as the protein is already in a conformation suitable for the hydrolysis of ATP. This may suggest, that the protein under any circumstances hydrolyses ATP. This is unlikely to occur inside a cell where ATP is a precious resource, therefore it is likely that additional factors contribute to oligomerisation for activity.

$PhoH2_{alt\ MTB}$ was sufficiently able to hydrolyse ATP despite being cloned with a potential regulatory element. $PhoH2_{alt\ MSMEG}$ was able to hydrolyse ATP as well, albeit the K_M was greater than observed for $PhoH2_{MSMEG}$. If this element is acting as an inhibitor like that of VapB, in VapBC systems, then it is likely that the regulatory element would inhibit the PIN domain, thus having potentially little effect on hydrolysis activity. The two proteins are likely to together, therefore an inhibitory effect on the PIN domain should have some inhibitory effect on the PhoH domain.

5.3.2 Unwinding and ribonuclease activity

The unwinding and cleavage activity of PhoH2 in the presence of ATP was specific and relied on 5'-tailed duplex RNA substrates with the 5' terminal combination of A C A/U A/U G/C U. The initial substrates tested with each of the PhoH2 proteins suggested that there was some activity of PhoH2 on substrates that had an overhang. Upon further substrate testing, a difference in activity was observed where the substrate was bound and binding was apparent with each of the PhoH2 proteins. From this point on, the last 5'1-5'5 substrates were tested with $PhoH2_{MSMEG}$ only, owing to the more consistent purification of this protein compared with $PhoH2_{alt\ MTB}$ and $PhoH2_{alt\ MSMEG}$. It is assumed, due to the consistent activity observed with all three proteins,

that PhoH2$_{alt\ MTB}$ and PhoH2$_{alt\ MSMEG}$ would have the same activity on 5'1-5'5. For these substrates very minor changes to the oligonucleotide substrate led to unwinding and RNase activity.

There are many reports of substrate specificity of helicases arising through interactions of flanking sequences and sequences on the external surface of the helicase cores, that serve as sites for interaction with regulatory proteins or cofactors, while the catalytic cores of many have a generic substrate binding affinity (Tanner & Linder 2001). Protein cofactors are reported to promote target recognition and helicase activity, stimulate ATPase activity, confer substrate specificity and increase the affinity of helicases for their substrates or inhibit helicase activity (Cordin et al. 2006). In cases in which a RNA helicase participates in multiple cellular processes, protein cofactors are also likely to act to recruit the helicase to the appropriate complex (Silverman et al. 2003). By testing PhoH2 as a whole, full length protein, it is unknown if one or both domains of the protein are contributing to substrate recognition and binding. Evidence suggests, however that the domains are working together, as for the assays where unwinding and cleavage was observed there were no degradation products in the absence of ATP. Similarly, in the presence of AMP-PNP for some substrates (5'AC), binding activity was observed which did not occur in any of the other reactions (-ATP, +EDTA). If the PIN-domain was working alone then in the absence of ATP (-ATP) activity should have been observed. Further, substrate recognition is likely to be determined by both domains, as with the binding activity of some substrates in the presence of ATP (+ATP) some substrate degradation occurred, suggesting possibly that some of the substrate is being 'held up' by the PhoH-domain allowing for cleavage by the PIN-domain. *In vivo*, PhoH2 is likely to act on a particular set of targets whose specificity is determined by both domains.

5.3.3 UTR assays

The activity of PhoH2$_{alt\ MTB}$ on the MTB alt substrate suggests a possible autoregulatory mechanism of this protein at the mRNA level. There was no cross specificity of either PhoH2 protein for their opposing alt substrate. This

may be, in part, due to the lack of suitable recognition sites in the *M. smegmatis* alt sequence, or the specificity of $PhoH2_{alt\ MTB}$ for its own upstream region. This mechanism of regulation is similar to that of RNase E, which autoregulates its synthesis through binding to a stem-loop in its 5'UTR to prevent the transcription of its gene (Schuck et al. 2009). It is interesting that $PhoH2_{MSMEG}$ did not have an effect on the MTB alt substrate (Figure 5.11b), despite the two PhoH2 proteins showing the same specificity for the shorter oligonucleotide substrates.

As a result of the binding activity observed with some substrates in the unwinding and ribonuclease assays it is plausible that PhoH2 adopts three types of activity on its targets: 1) unwinding and cleavage 2) binding and cleavage and 3) binding. All of these require ATP as a cofactor. Evidence suggests that PhoH2 from *M. tuberculosis* and *M. smegmatis* differ in their modes of regulation and that the role of the upstream sequence in *M. smegmatis* is not as influential as the upstream sequence of *M. tuberculosis*.

5.4 Conclusions

This is the first biochemical characterisation of a PIN-PhoH protein. The results of this biochemical investigation of three mycobacterial PhoH2 proteins can provide data that may be applied to other PhoH2 proteins.

PhoH2 can hydrolyse ATP irrespective of the addition of secondary substrate, suggesting assembly of the protein in a higher order conformation that is suitable for hydrolysis activity. A number of short self annealing oligonucleotides were screened, and the sequence A C A/U A/U G/C U was deduced as a specific target of PhoH2 proteins. UTR assays revealed activity of $PhoH2_{alt\ MTB}$ on its respective upstream sequence which was not observed for $PhoH2_{MSMEG}$ suggesting a possible autoregulatory role of $PhoH2_{alt\ MTB}$. Furthermore, this suggests a critical regulatory role of the upstream region of *phoH2* in *M. tuberculosis* which is consistent with the requirement of this sequence for soluble protein expression and offsetting negative effects of conditional over expression on growth and colony formation.

Chapter Six: Discussion

PhoH2 proteins are abundant across bacterial and archaeal species. Despite this, little work has been done to determine their biological role. Biochemically, PIN-domain proteins are RNases, and as the toxic component of TA systems, those that have been extensively characterised, are sequence specific, and act to tune cellular metabolism (McKenzie et al. 2012b). The VapBC TA proteins are overrepresented in organisms that are pathogenic to mammals (McKenzie et al. 2012b), implying the involvement of these proteins in stress response and persistence (Gerdes 2000; Hayes 2003; Arcus et al. 2005; Buts et al. 2005; Gerdes et al. 2005). It is probable that other PIN-domain variants, by virtue of their biochemical activity, contribute to the adaptive response to environmental stress.

This research has identified that *phoH2* from two mycobacterial species is transcribed as part of an extensively structured mRNA transcript. Long 5'UTRs generally play roles in stability of the mRNA transcript, provide components for the correct intracellular mRNA localisation, and contain *cis*-acting regulatory elements (Irnov et al. 2010). In *M. tuberculosis*, evidence suggests that a ~152 base segment of this transcript closest to the 5'-terminus of *phoH2* is regulatory, shown to be necessary for 'normal' growth, colony formation, and protein overexpression of PhoH2, not otherwise demonstrated by *phoH2* alone.

Most PIN-domain proteins are co-expressed with an inhibitor that forms a complex with the PIN-domain protein (VapBC TA) (Arcus et al. 2011). This inhibitor allows for successful PIN-domain protein overexpression, and offsets the activity of the PIN-domain. These proteins have also been shown to affect growth when conditionally expressed in the absence of their inhibitor (Robson 2010). Evidence presented here suggests that it is likely that the alt peptide interacts with PhoH2. This peptide was required for to alleviate negative growth affects observed with $PhoH2_{MTB}$ alone, and was required for soluble protein expression.

LTQ-Orbitrap MS confirmed that the alt transcript is translated along with PhoH2 and can suggest an interaction between these two proteins. A difference between this peptide and the inhibitors of PIN-domains in the VapBC TA systems is that this peptide appears to not need to be removed in order to test for biochemical activity of PhoH2 proteins. The combined MALDI TOF/TOF MS and LTQ-Orbitrap analyses performed at the University of Otago, suggested that the sample of PhoH2 protein was a pool of the same protein but of different lengths. This may suggest a labile nature of the regulator of PhoH2 which, therefore, would not be able to prevent the activity of the protein.

It is likely transcription and translation of this peptide and of the *phoH2* gene is regulated. This may be controlled by one or more of the predicted transcriptional regulators positioned 1) within the intergenic region of *Rv1094* and *Rv1095* (*Rv1990c*), 2) further upstream near the start of the full length mRNA transcript (*Rv0494*), 3) further downstream of *phoH2* (*Rv3133c*), or 4) by negative feedback of the PhoH2 protein, which demonstrates activity on the alt RNA transcript (Figure 6.1).

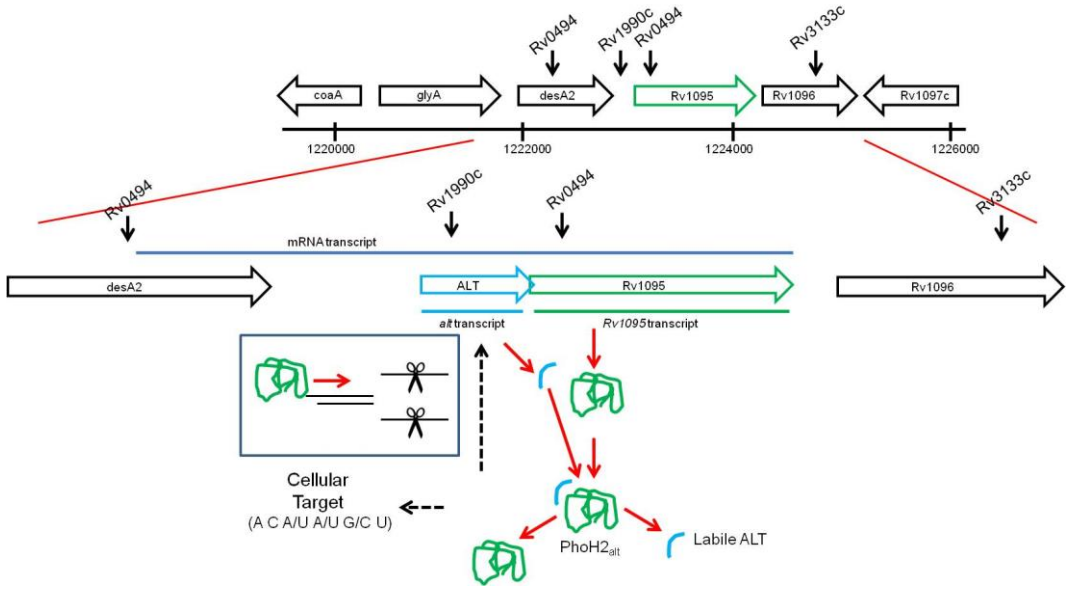

Figure 6.1 Model of proposed *Rv1095* regulation and PhoH2 activity
Rv1095 has four predicted transcriptional regulators (*Rv0494* x2, *Rv1990c*, and *Rv3133c*). Solid blue line (mRNA transcript) shows the length of the mRNA transcript determined by RT-PCR, this would extend at least to the end of the *Rv1095* gene. The target of *Rv0494* is located approximately half way through the *Rv1094* gene. Transcriptional regulation by this transcription factor may influence the stability of the entire mRNA transcript. Similarly, the position of *Rv1990c* within the intergenic region of *Rv1094* and *Rv1095*, may regulate the transcript from the alt site onward. If the mRNA transcript does extend past the 3' end of *Rv1095* then *Rv3133c* may play an additional role in regulation. Once $PhoH2_{alt}$ is transcribed and translated, the lability of the alt peptide would leave a pool of $PhoH2_{alt}$ protein and PhoH2 protein. The PhoH2 protein would be free to act on its cellular targets and negatively regulate its expression by acting on the alt mRNA.

Each of these regulators, either directly or indirectly, respond to oxidative stress and phosphate starvation as described in sections 1.5.2, 3.1.2 and 3.1.3. Homologues to all these regulators are found in *M. smegmatis*, with the exception of *Rv1990c*, show elevated expression when cultured slowly under oxygen limited conditions (0.6 % air), including *MSMEG_5247* (Berney & Cook 2010). The functional partners identified for *phoH2* are similar between *M. tuberculosis* and *M. smegmatis*, and implicate *phoH2* further in the oxidative stress response and may suggest involvement in virulence associated lipid biogenesis.

The incorporation of the alt sequence was critical for 'normal' growth and colony formation, and protein overexpression. It is possible then that *Rv1990c* would have the greatest effect of all the three regulators as its predicted target position is closest to the 5' end of the alt transcript. This

transcriptional regulator is under control of SenX3-RegX3 in *M. tuberculosis*, which controls expression of phosphate specific transport genes under phosphate starvation (Rifat et al. 2009) and contains sensors of oxygen and redox state (Parish et al. 2003b; Rickman et al. 2004; Roberts et al. 2011). It is also possible that the candidate pseudopalindrome sequence identified by Kazakov et al. (2002) located within the alt sequence may be targeted for regulation.

PhoH2 from *M. smegmatis* did not require the cloning of the alt sequence for protein overexpression or for 'normal' growth or colony formation in *M. smegmatis* cells, likely due to endogenous levels of the alt peptide. PhoH2 from *M. smegmatis* in addition may be regulated slightly differently to PhoH2 in *M. tuberculosis*. As mentioned, each of the predicted regulators of PhoH2 in *M. tuberculosis* with the exception of *Rv1990c* (not found in *M. smegmatis*) had elevated expression when cultured in hypoxic conditions. In *M. smegmatis*, SenX3-RegX3 is responsible for expression of genes involved in the response to phosphate starvation (Glover et al. 2007) and this two component system was also up regulated under hypoxic growth conditions (Berney & Cook 2010). Overall, it can be hypothesised that PhoH2 from *M. tuberculosis* and *M. smegmatis* are under the ultimate control of the SenX3-RegX3 or DevRS systems, which respond to phosphate starvation and oxidative stress. A flow diagram of predicted PhoH2 regulation in *M. tuberculosis* and *M. smegmatis* is outlined in Figure 6.2.

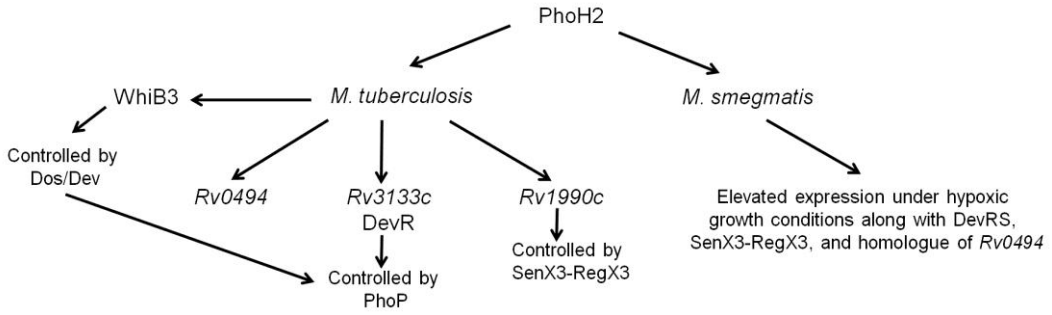

Figure 6.2 Overview of predicted regulation of PhoH2 in *M. tuberculosis* and *M. smegmatis*

Biochemically, each PhoH2 protein tested was able to hydrolyse ATP in the presence of Mg^{2+}, irrespective of the addition of secondary substrate. PhoH2 proteins also demonstrated 1) binding, 2) binding and cleavage and 3) unwinding and cleavage, of RNA substrates in the presence of Mg^{2+} and ATP. In some cases, binding activity occurred in the presence of AMP-PNP suggesting for binding activity the hydrolysis of ATP was not required. For full unwinding and cleavage activity, ATP hydrolysis was required, as unwinding did not occur with AMP-PNP in place of ATP. PhoH2 from *M. tuberculosis* and *M. smegmatis* were found to act on the same short RNA substrates. The optimal substrate had a 5'-ssRNA overhang composed of the sequence A C A/U A/U G/C U. PhoH2 from *M. tuberculosis* was further demonstrated to act on its respective alt transcript, suggesting that it may autoregulate its own expression. Other PhoH2 proteins may be able to act on longer RNA targets as well, and may rely on the presence of specific secondary structural elements within the RNA for activity.

Currently, all characterised PIN-domain proteins, in spite of carrying the same set of conserved residues that make up the active site, target different specific RNA sequences (Ahidjo et al. 2011; McKenzie et al. 2012a; McKenzie et al. 2012b; Sharp et al. 2012). It is possible, given the results presented here, that all PhoH2 proteins target the same short sequences, but *in vivo* their biological targets may differ depending under which conditions PhoH2 is expressed.

The three types of activity displayed by PhoH2 (binding, binding and cleavage, and, unwinding and cleavage) can suggest that these proteins may target substrates within a gene or within the 5' or 3' UTR to destabilise the transcript. This type of model has been proposed for other PIN-domain proteins where the target sequences have been identified either 5' or 3' of the gene of interest (McKenzie 2011). Further, the reaction products of PhoH2 may be targets of other proteins involved with RNA metabolism.

6.1 Recommendations for further work

To elucidate the domain interactions and confirm the higher order conformation of PhoH2, a higher resolution structure would need to be obtained. This may be achieved by further optimising crystal growth for structure determination by crystallography or by further pursuing electron microscopy for image data. For both crystallisation and electron microscopy, the homogeneity of the protein sample needs to be improved. EM image data may be used in conjunction with diffraction data to gain a low resolution structure.

A structure would show whether the PIN-domain, in a PIN-PhoH fusion, forms a ring and tunnel in its higher order conformation in a similar manner to that of PDB 3I8O. Further, whether one or more hexamers work together to perform mechanical work, as suggested by PDB 3B85. A structure may also reveal where the alt peptide is docking, and help confirm the lability of this peptide. The mechanism of activity, may also be uncovered by structures obtained with substrates bound at different stages of the ATPase cycle.

To confirm if a single domain is responsible for binding activity, the PIN and PhoH domains of $PhoH2_{TBIS}$, , could be cloned separately and tested in expression trials. Each domain could then be tested in ATPase/binding/unwinding/ribonuclease assays. These tests would also shed light on whether the PIN-domain is responsible for maintaining the higher order hexameric conformation of the PhoH-domain. Additionally, ATPase assays could be performed in the presence of the PIN-domain to see if ATPase activity is increased or whether ATP turnover is carried out independently of the PIN-domain.

To test the requirement of PhoH2 in the adaptation to phosphate limiting conditions and hypoxia, a qPCR analysis could be performed on samples taken during culture under these two conditions. If expression of *phoH2* was elevated during a shift from phosphate sufficient to phosphate limiting

conditions, this would be consistent with *phoH* (PhoH2) from *C. glutamicum* (Ishige et al. 2003).

To further verify the requirement of PhoH2 for growth under phosphate starvation and hypoxic conditions and to study these effects *in vivo*, a deletion mutant would need to be made. If growth was affected by the absence of PhoH2, then this protein is involved in these responses. Further, mRNA microarray analyses could be performed on the mutant to determine which genes are affected by *phoH2*, and those that are affected, could be tested *in vitro*.

Attempts were made to generate an unmarked deletion of *phoH2* in *M. smegmatis* using the method developed for two-step allelic exchange mutagenesis in *M. tuberculosis* (Parish & Stoker 2000), which is based on the method of Pelicic et al. (1997). These methods rely on the use of a mycobacteria-specific delivery plasmid pX33 (modified pPR23) (Pelicic et al. 1997; Gebhard et al. 2006) that contains a thermosensitive origin of replication and gene *sac*B, which can be used to promote recombination and positive selection of gene replacement events during knockout construction. Two recombination events aim to exchange the gene for an insert (~1650 bp) composed of ~800 bp of the left and right flank of the gene which has been cloned into the delivery vector. The left and right flank PCR products of *phoH2* were relatively straightforward to obtain after gradient cycling parameters were trialled. The overlapping LFRF PCR to gain the LFRF product was very challenging to amplify and was not obtained successfully despite numerous rounds and re-designing of primers to improve the left and right flank used in the second PCR reaction.

To test the hypothesis that PhoH2 from *M. tuberculosis* is autoregulatory, the exact transcriptional start site would need to be obtained for the construction of a transcriptional fusion. The fusion could be introduced into the mutant background to identify any unregulated expression using systems such as lacZ. To verify the exact biochemical target within the alt RNA

sequence, the substrate could be split into shorter sequences to determine which portion is targeted.

6.2 Concluding remarks

The two mycobacterial genes investigated here were both transcribed as part of a longer mRNA transcript that is most likely to contain regulatory elements or sites for regulation. The sequence taken from alternative start site 1 (*phoH2*$_{alt}$) was found to be required to alleviate the negative effect of PhoH2 protein expression otherwise observed with PhoH2$_{MTB}$ alone and was also required to gain soluble protein expression of PhoH2 from *M. tuberculosis*. These results raised the possibility of regulatory elements in this sequence required for expression, regulation or translation or a potential inhibitory role of this sequence on the activity of PhoH2. It was later uncovered that the sequence from the alternative start site was being translated along with PhoH2 and that these two proteins interact with one another. This MS analysis also suggested that the peptide attachment to PhoH2 is labile in nature as a number of intact protein masses were identified suggesting a pool of the same protein but of different lengths. This protein construct and PhoH2$_{alt\ MSMEG}$ did not require the active removal of this peptide in order to test for biochemical activity, this is likely due to the labile nature of the peptide therefore there was enough "free" PhoH2 to perform work. Two hypotheses are presented regarding the translation of the alt peptide along with PhoH2, both of which lead to the production of two polypeptides that interact with one another post-translation. Biologically, PhoH2 proteins are thought to contribute to the adaptive response of organisms to phosphate and oxidative stress where they are likely to act on specific RNA substrates.

References

Agarwal N, Tyagi AK 2006. Mycobacterial transcriptional signals: requirements for recognition by RNA polymerase and optimal transcriptional activity. Nucleic Acids Research 34(15): 4245–4257.

Ahidjo BA, Kuhnert D, McKenzie JL, Machowski EE, Gordhan BG, Arcus VL, Abrahams GL, Mizrahi V 2011. VapC Toxins from *Mycobacterium tuberculosis* Are Ribonucleases that Differentially Inhibit Growth and Are Neutralized by Cognate VapB Antitoxins. PLoS One 6(6): e21738.

Anantharaman V, Aravind L 2003. New connections in the prokaryotic toxin-antitoxin network: relationship with the eukaryotic nonsense-mediated RNA decay system. Genome Biology 4: R81.1-R81.15.

Anantharaman V, Koonin EV, Aravind L 2001. TRAM, a predicted RNA-binding domain, common to tRNA uracil methylation and adenine thiolation enzymes. FEMS Microbiology Letters 197(2): 215-221.

Anantharaman V, Koonin EV, Aravind L 2002. Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Research 30(7): 1427-1464.

Andreeva A, Howorth D, Chandonia J, Brenner SE, Hubbard TJP, Chothia C, Murzin AG 2008. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Research 36: D419-425.

Appaji RN, Ambili M, Jala VR, Subramanya HS, Savithri HS 2003. Structure-function relationship in serine hydroxymethyltransferase. Biochimica et Biophysica Acta 1647(1-2): 24-29.

Arcus VL, Rainey PB, Turner SJ 2005. The PIN-domain toxin-antitoxin array in mycobacteria. Trends in Microbiology 13(8): 360-365.

Arcus VL, McKenzie JL, Robson J, Cook GM 2011. The PIN-domain ribonucleases and the prokaryotic VapBC toxin-antitoxin array. Protein Engineering, Design & Selection: 1-8.

Arcus VL, Bäckbro K, Roos A, Daniel EL, Baker EN 2004. Distant structural homology leads to the functional characterization of an archaeal PIN domain as an exonuclease. Journal of Biological Chemistry 279(16): 16471-16478.

Bashiri G, Squire CJ, Baker EN, Moreland NJ 2007. Expression, purification and crystallization of native and selenomethonine labeled *Mycobacterium tuberculosis* FGD1 (Rv0407) using a *Mycobacterium smegmatis* expression system. Protein Expression and Purification 54: 38-44.

Bashiri G, Rehan AM, Greenwood DR, Dickson JMJ, Baker EN 2010. Metabolic Engineering of Cofactor F_{420} Production in *Mycobacterium smegmatis*. PLoS One 5(12): e15803.

Bashyam MD, Kaushal D, Dasgupta SK, Tyagi AK 1996. A study of Mycobacterial Transcriptional Apparatus: Identification of novel features in promoter elements. Journal of Bacteriology 178 (16): 4847-4853.

Bayer TS, Booth LN, Knudsen SM, Ellington AD 2005. Arginine-rich motifs present multiple interfaces for specific binding by RNA. RNA 11: 1848-1857.

Baykov AA, Evtushenko OA, Avaeva SM 1988. A Malachite Green Procedure for Orthophosphate Determination and Its Use in Alkaline Phosphatase-Based Immunoassay. Analytical Biochemistry 171: 266-270.

Berney M, Cook GM 2010. Unique Flexibility in Energy Metabolism Allows Mycobacteria to Combat Starvation and Hypoxia. PLoS One 5(1): e8614.

Bhamidi S 2009. Mycobacterial Cell Wall Arabinogalactan. Bacterial Polysaccharides: Current Innovations and Future Trends, Caister Academic Press.

Bleichert F, Baserga SJ 2007. The Long Unwinding Road of RNA Helicases. Molecular Cell 27(3): 339-352.

Blower TR, Pei XU, Short FL, Fineran PC, Luisi BF, Salmond GPC 2011. A processed noncoding RNA regulates an altruistic bacterial antiviral system. Nature Structural & Molecular Biology 18: 185-190.

Bodogai M, Ferenczi S, Bashtovyy D, Miclea P, Papp P, Dusha I 2006. The *ntrPR* Operon of *Sinorhizobium meliloti* Is Organized and Functions as a Toxin-Antitoxin Module. Molecular Plant-Microbe Interactions 19(7): 811-822.

Bunker RD, McKenzie JL, Baker EN, Arcus VL 2008. Crystal structure of PAE0151 from *Pyrobaculum aerophilum*, a PIN-domain (VapC) protein from a toxin-antitoxin operon. Proteins-Structure Function and Bioinformatics 72(1): 510-518.

Buts L, Lah J, Dao-Thi MH, Wyns L, Loris R 2005. Toxin-antitoxin modules as bacterial metabolic stress managers. Trends in Biochemical Science 30: 672-679.

Caruthers JM, McKay DB 2002. Helicase structure and mechanism. Current Opinion in Structural Biology 12: 123-133.

Chandran V, Poljak L, Vanzo NF, Leroy A, Miguel RN, Fernandez-Recio J, Parkinson J, Burns C, Carpousis AJ, Luisi BF 2007. Recognition and Cooperation Between the ATP-dependent RNA Helicase RhlB and Ribonuclease RNase E. Journal of Molecular Biology 367: 113-132.

Chen P, Nie S, Mi W, Wang X, Liang S 2004. De novo sequencing of tryptic peptides sulfonated by 4-sulfophenyl isothiocyanate for unambiguous protein identification using post-source decay matrix-assisted laser desorption/ionisation mass spectrometry. Rapid Communications in Mass Spectrometry: 191-198.

Connolly LE, Edelstein PH, Ramakrishnan L 2007. Why Is Long-Term Therapy Required to Cure Tuberculosis? PLoS One Medicine 4(3): 0435-0442.

Cooper CR, Daugherty AJ, Tachdjian S, Blum PH 2009. Role of *vapBC* toxin-antitoxin loci in the thermal stress response of *Sulfolobus solfataricus*. Biochemical Society Transactions 37(pt 1): 123-126.

Cordin O, Banroques J, Tanner NK, Linder P 2006. The DEAD-box protein family of RNA helicases. Gene 367: 17-37.

Daines DA, Wu MH, Yuan SY 2007. VapC-1 of nontypeable *Haemophilus influenzae* is a ribonuclease. Journal of Bacteriology 189(14): 5041-5048.

De Bernardez CE 1998. Refolding of recombinant proteins. Current Opinion in Biotechnology 9: 157-163.

Donmez I, Patel SS 2006. Mechanisms of a ring shaped helicase. Nucleic Acids Research 34(15): 4216-4224.

Edwards TE, Liao R, Phan I, Myler PJ, Grundner C 2012. *Mycobacterium thermoresistibile* as a source of thermostable orthologs of *Mycobacterium tuberculosis* proteins. Protein Science 21(7): 1093-1096.

Erzberger JP, Berger JM 2006. Evolutionary Relationships and Structural Mechanisms of AAA+ Proteins. Annual Review of Biophysics and Biomolecular Structure 35: 93-114.

Fairman-Williams ME, Guenther UP, Jankowsky E 2010. SF1 and SF2 helicases: family matters. Current Opinion in Structural Biology 20(3): 313-324.

Fineran PC, Blower TR, Fouldsa IJ, Humphreys DP, Lilleya KS, Salmond GPC 2009. The phage abortive infection system, ToxIN, functions as a protein–RNA toxin–antitoxin pair. PNAS 106(3): 894-899.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K and others 2010. The PFam protein families database. Nucleic Acids Research 38: D211-222.

Fozo EM, Makarova KS, Shabalina SA, Yutin N, Koonin EV, Storz G 2010. Abundance of type I toxin–antitoxin systems in bacteria: searches for new candidates and discovery of novel families. Nucleic Acids Research 38(11): 3743–3759.

Fuller-Pace FV, Nicol SM, Reid AD, Lane DP 1993. DbpA: a DEAD box protein specifically activated by 23S rRNA. The EMBO Journal 12: 3619-3626.

Gebhard S, Cook GM 2008. Differential Regulation of High-Affinity Phosphate Transport Systems of *Mycobacterium smegmatis*: Identification of PhnF, a Repressor of the phnDCE Operon. Journal of Bacteriology 190(4): 1335–1343.

Gebhard S, Tran SL, Cook GM 2006. The Phn system of *Mycobacterium smegmatis*: a second high-affinity ABC-transporter for phosphate. Microbiology 152: 3453-3465.

Gerdes K 2000. Toxin-Antitoxin Modules May Regulate Synthesis of Macromolecules during Nutritional Stress. Journal of Bacteriology 182(3): 561-572.

Gerdes K, Wagner EGH 2007. RNA Antitoxins. Current Opinion in Microbiology 10: 117-124.

Gerdes K, Rasmussen PB, Molin S 1986. Unique type of plasmid maintenance function: Postsegregational killing of plasmid-free cells. PNAS 83: 3116-3120.

Gerdes K, Christensen SK, Løbner-Olesen A 2005. Prokaryotic toxin-antitoxin stress response loci. Nature 3(5): 371-382.

Glavan F, Behm-Ansmant I, Izaurralde E, Conti E 2006. Structures of the PIN domains of SMG6 and SMG5 reveal a nuclease within the mRNA surveillance complex. The EMBO Journal 25(21): 5117-5125.

Glover RT, Kriakov J, Garforth SJ, Baughn AD, Jacobs WRJ 2007. The Two-Component Regulatory System senX3-regX3 Regulates Phosphate-Dependent Gene Expression in *Mycobacterium smegmatis*. Journal of Bacteriology 189(15): 5495–5503.

Goldstone RM, Moreland NJ, Bashiri G, Baker EN, Lott JS 2008. A new Gateway vector and expression protocol for fast and efficient recombinant protein expression in *Mycobacterium smegmatis*. Protein Expression and Purification 57: 81-87.

Gonzalo-Asensio J, Mostowy S, Harders-Westerveen J, Huygen K, Herna´ndez-Pando R, Thole J, Behr M, Gicque B, Martın C 2008. PhoP: A Missing Piece in the Intricate Puzzle of *Mycobacterium tuberculosis* Virulence. 3 10: e3496.

Gorbalenya AE, Koonin EV 1993. Helicases: amino acid sequence comparisons and structure-function relationships. Current Opinion in Structural Biology 3: 419-429.

Han X, Aslanian A, Yates JR 2008. Mass spectrometry for proteomics. Current Opinion in Chemical Biology 12: 483-490.

Hayes CS, Sauer RT 2003. Toxin-Antitoxin Pairs in Bacteria: Killers of Stress Regulators. Cell 112: 2-4.

Hayes F 2003. Toxins-Antitoxins: Plasmid Maintenance, Programmed Cell Death and Cell Cycle Arrest. Science 301: 1496-1499.

Hickman AB, Dyda F 2005. Binding and unwinding: SF3 viral helicases. Current Opinion in Structural Biology 15: 77-85.

Hopper S, Wilbur JS, Vasquez BL, Larson J, Clary S, Mehr IJ, Seifert HS, So M 2000. Isolation of *Neisseria gonorrhoeae* Mutants That Show Enhanced Trafficking across Polarized T84 Epithelial Monolayers. Infection and Immunity 68(2): 896-905.

Huntzinger E, Kashima I, Fauser M, Sauliere J, Izaurralde E 2008. SMG6 is the catalytic endonuclease that cleaves mRNAs containing nonsense codons in metazoan RNA 14: 2609-2617.

Inouye M 2006. The Discovery of mRNA Interferases: Implication in Bacterial Physiology and Application to Biotechnology. Journal of Cellular Physiology 209: 670-676.

Irnov I, Sharma CM, Vogel J, Winkler WC 2010. Identification of regulatory RNAs in Bacillus subtilis. Nucleic Acids Research 38(19): 6637-6651.

Ishige T, Krause M, Bott M, Wendisch VF, Sahm H 2003. The Phosphate Starvation Stimulon of *Corynebacterium glutamicum* Determined by DNA Microarray Analyses. Journal of Bacteriology 185(15): 4519-4529.

Iyer LM, Leipe DD, Koonin EV, Aravind L 2004. Evolutionary history and higher order classification of AAA+ ATPases. Journal of Structural Biology 146: 11-31.

Jaworski JG, Stumpf PK 1974. Fat metabolism in higher plants. Properties of a soluble stearyl-acyl carrier protein desaturase from maturing Carthamus tinctorius. Archives of Biochemistry and Biophysics 162(1): 158-165.

Kazakov AE, Vassieva O, Gelfand MS, Osterman A, Overbeek R 2003. Bioinformatics classification and functional analysis of PhoH homologs. *In Silico* Biology 3: 3-15.

Kim SK, Makino K, Amemura M, Shinagawa H, Nakata A 1993. Molecular analysis of the phoH gene, belonging to the phosphate regulon in *Escherichia coli*. Journal of Bacteriology 175(5): 1316-1324.

Kocan M, Schaffer S, Ishige T, Sorger-Herrmann U, Wendisch VF, Bott M 2006. Two-component systems of *Corynebacterium glutamicum*: Deletion analysis and involvement of the PhoS-PhoR system in the phosphate starvation response. Journal of Bacteriology 188(2): 724-732.

Koonin EV, Rudd KE 1996. Two domains of superfamily I helicases may exist as separate domains. Protein Science 5: 178-180.

Kozak M 2005. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene 361: 13-37.

Kurz M, Cowieson NP, Robin G, Hume DA, Martin JL, Kobe B, Listwan P 2006. Incorporating a TEV cleavage site reduces the solubility of nine recombinant mouse proteins. Protein Expression and Purification 50: 68-73.

Lamanna AC, Karbstein K 2009. Nob1 binds the single-stranded cleavage site D at the 3' -end of 18S rRNA with its PIN domain. PNAS 106(34): 14259-14264.

Leipe DD, Wolf G, Koonin EV, Aravind L 2002. Classification and Evolution of P-loop GTPases and Related ATPases. Journal of Molecular Biology 317: 41-72.

Liolios K, Skorski J, Jando M, Lapidus A, Copeland A, Glavina Del Rio T, Nolan M, Lucas S, Tice H, Cheng J and others 2010. Complete genome sequence of *Thermobispora bispora* type strain (R51^{T}). Standards in Genomic Sciences 2: 318-326.

Lupas AN, Martin J 2002. AAA proteins. Current Opinion in Structural Biology 12: 746-753.

Lyubimov AY, Strycharska M, Berger JM 2011. The nuts and bolts of ring-translocase structure and mechanism. Current Opinion in Structural Biology 21: 240-248.

Mackie GA 1998. RNase E lacks sequence specificity but is a 5' end endonuclease. Nature 15(395): 720-723.

Makarova KS, Wolf YI, van der Oost J, Koonin EV 2009. Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. Biology Direct 4(29): doi:10.1186/1745-6150-4-29.

Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR and others 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Research 39(D): 225-229.

Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D 1999. A combined algorithm for genomewide prediction of protein function. Nature 402: 83-86.

Mattison K, Wilbur JS, So M, Brennan RG 2006. Structure of FitAB from *Neisseria gonorrhoeae* Bound to DNA Reveals a Tetramer of Toxin-Antitoxin Heterodimers Containing Pin Domains and Ribbon-Helix-Helix Motifs. Journal of Biological Chemistry 281(49): 37942-37951.

McKenzie JL 2011. The Biochemistry of VapBC Toxin-Antitoxins. Unpublished thesis, University of Waikato, Hamilton

McKenzie JL, Duyvestyn JM, Smith T, Bendak K, Mackay J, Cursons R, Cook GM, Arcus VL 2012a. Determination of ribonuclease sequence-specificity using Pentaprobes and mass spectrometry RNA 18: 1267-1278.

McKenzie JL, Robson J, Berney M, Smith TC, Ruthe A, Gardner PP, Arcus VL, Cook GM 2012b. A VapBC Toxin-Antitoxin Module is a Post-Transcriptional Regulator of Metabolic Flux in Mycobacteria. Journal of Bacteriology 194: 2189-2204.

Metcalf WW, Steed PM, Wanner BL 1990. Identification of phosphate starvation-inducible genes in Escherichia coli K-12 by DNA sequence analysis of psi::lacZ(Mu d1) transcriptional fusions. Journal of Bacteriology 172(6): 3191-3200.

Moreland N, Ashton R, Baker HM, Ivanovic I, Patterson S, Arcus VL, Baker EN, Lott JS 2005. A flexible and economical medium-throughput strategy for protein production and crystallisation. Acta Crystallographica Section D Biological Crystallography 61(Pt 10): 1378-1385.

Musco G, Kharrat A, Steier G, Fraternali F, Gibson TJ, Pastore A 1997. The solution structure of the first KH domain of FMR1, the protein responsible for the fragile X syndrome. Nature Structural Biology 4: 712-716.

Nagai J, Bloch K 1968. Enzymatic Desaturation of Stearyl Acyl Carrier Protein. Journal of Biological Chemistry 243(17): 4626-4633.

O'Toole R, Smeulders MJ, Blokpoel MC, Kay EJ, Lougheed K, Williams HD 2002. A Two-Component Regulator of Universal Stress Protein Expression and Adaptation to Oxygen Starvation in Mycobacterium smegmatis. Journal of Bacteriology 185(5): 1543-1554.

Olsen DB, Carroll SS, Culberson JC, Shafer JA, Kuo LC 1994. Effect of template secondary structure on the inhibition of HIV-1 reverse transcriptase by a pyridinone non-nucleoside inhibitor. Nucleic Acids Research 22(8): 1437-1443.

Pandey DP, Gerdes K 2005. Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. Nucleic Acids Research 33(5): 966-976.

Parish T, Stoker NG 2000. *glnE* Is an Essential Gene in *Mycobacterium tuberculosis*. Journal of Bacteriology 182(20): 5715-5720.

Parish T, Smith DA, Roberts G, Betts J, Stoker NG 2003a. The senX3–regX3 two-component regulatory system of *Mycobacterium tuberculosis* is required for virulence. Microbiology 149: 1423-1435.

Parish T, Smith DA, Kendall S, Casali N, G.J. B, Stoker NG 2003b. Deletion of Two-Component Regulatory Systems Increases the Virulence of *Mycobacterium tuberculosis*. Infection and Immunity 71(3): 1134-1140.

Park DW, Kim SS, Nam MK, Kim GY, Kim J, Rhim H 2011. Improved recovery of active GST-fusion proteins from insoluble aggregates: solubilization and purification conditions using PKM2 and HtrA2 as model proteins. BMB reports 44(4): 279-284.

Pelicic V, Jackson MP, Reyrat J, Jacobs WR, Gicquel B, Guilhot C 1997. Efficient allelic exchange and transposon mutagenesis in *Mycobacterium tuberculosis*. PNAS 94: 10955-10960.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J and others 2012. The Pfam protein families database. PNAS 40: doi:10.1093/nar/gkr1065.

Puskas LG, Nagy ZB, Kelemen JZ, Ruberg S, Bodogai M, Becker A, Dusha I 2004. Wide-range transcriptional modulating effect of *ntrR* under microaerobiosis in *Sinorhizobium meliloti*. Molecular Genetics and Genomics 272: 275-289.

Rabhi M, Tuma R, Boudvillain M 2010. RNA remodelling by hexameric RNA helicases. RNA Biology 7(6): 655-666.

Ramage HR, Connolly LE, Cox JS 2009. Comprehensive Functional Analysis of *Mycobacterium tuberculosis* Toxin-Antitoxin Systems: Implications for Pathogenesis, Stress Responses, and Evolution. PLoS One 5(12): e100767.

Rickman L, Saldanha JW, Hunt DM, Hoar DN, Colston MJ, Millar JBA, Buxton RS 2004. A two-component signal transduction system with a PAS domain-containing sensor is required for virulence of *Mycobacterium tuberculosis* in mice. Biochemical and Biophysical Research Communications 314: 259-267.

Rifat D, Bishai WR, Karakousis PC 2009. Phosphate Depletion: A Novel Triger for *Mycobacterium tuberculosis* Persistence. Journal of Infectious Diseases 200: 1126-1135.

Roberts G, Vadrevu IS, Madiraju MV, Parish T 2011. Control of CydB and GltA1 Expression by the SenX3 RegX3 Two Component Regulatory System of *Mycobacterium tuberculosis*. PLoS One 6(6): e21090.

Robson J 2010. Characterisation of the *vapBC* operon of *Mycobacterium smegmatis*. Unpublished thesis, University of Otago, Dunedin.

Robson J, McKenzie JL, Cursons R, Cook GM, Arcus VL 2009. The vapBC operon from *Mycobacterium smegmatis* is an autoregulated toxin-antitoxin module that controls growth via inhibition of translation. Journal of Molecular Biology 390(3): 353-367.

Rocak S, Linder P 2004. Dead-box proteins: The driving forces behind RNA metabolism. Nature Reviews Molecular Cell Biology 5(3): 232-241.

Rodrigue S, Provvedi R, Jacques PE, Gaudreau L, Manganelli R 2006. The sigma factors of *Mycobacterium tuberculosis*. FEMS Microbiology Reviews 30: 926-941.

Ryndak M, Wang S, Smith I 2008. PhoP, a key player in *Mycobacterium tuberculosis* virulence. Trends in Microbiology 16(11): 528-534.

Sambrook J, Russell DW 2001. Molecular Cloning A Laboratory Manual. 3 ed, Cold Spring Harbor Laboratory Press, Cold Spring Habor, New York.

Savvides SN, Yeo H, Beck MR, Blaesing F, Lurz R, Lanka E, Buhrdorf R, Fischer W, Hass R, Waksman G 2003. VirB11 ATPases are dynamic hexameric assemblies: new insights into bacterial type IV secretion. The EMBO Journal 22(9): 1969-1980.

Schuck A, Diwa A, Belasco JG 2009. RNase E autoregulates its synthesis in *Escherichia coli* by binding directly to a stem-loop in the rne 5′ untranslated region. Molecular Microbiology 2(72): 470–478.

Sevin EW, Barloy-Hubler F 2007. RASTA-Bacteria: a web-based tool for identifying toxin-antitoxin loci in prokaryotes. Genome Biology 8: R155.

Sharp JD, Cruz JW, Raman S, Inouye M, Husson RN, Woychik NA 2012. Growth and translation inhibition through sequence-specific RNA binding by a *Mycobacterium tuberculosis* VapC toxin. Journal of Biological Chemistry 287: 12835-12847.

Shevchenko A, Wilm M, Vorm O, Mann M 1996. Mass Spectrometric Sequencing of Proteins from Silver-Stained Polyacrylamide Gels. Analytical Chemistry 68: 850-858.

Silverman E, Edwalds-Gilbert G, Lin RJ 2003. DExD/H-box proteins and their partners: helping RNA helicases unwind. Gene 312: 1-16.

Singh A, Crossman DK, Mai D, Guidry L, Voskuil MI, Renfrow MB, Steyn AJC 2009. *Mycobacterium tuberculosis* WhiB3 Maintains Redox Homeostasis by Regulating Virulence Lipid Anabolism to Modulate Macrophage Response. PLoS Pathogens 5 (8): e1000545.

Singleton MR, Dillingham MS, Wigley DB 2007. Structure and Mechanism of Helicases and Nucleic Acid Translocases. Annual Review of Biochemistry 76: 23-50.

Snapper SB, Melton RE, Mustafa S, Kieser T, Jacobs WRJ 1990. Isolation and characterization of efficient plasmid transformation mutants of *Mycobaterium smegmatis*. Molecular Microbiology 4(11): 1911-9.

Stover P, Schirch V 1990. Serine hydroxymethyltransferase catalyzes the hydrolysis of 5,10-methenyltetrahydrofolate to 5-formyltetrahydrofolate. Journal of Biological Chemistry 265(24): 14227-14233.

Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P and others 2011. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Research 39 (Database issue): D561-8. doi: 10.1093/nar/gkq973

Takeshita D, Zenno S, Lee WC, Saigo K, Tanokura M 2006. Crystallization and preliminary X-ray analysis of the PIN domain of human EST1A. Acta Crystallographica Section F-Structural Biology and Crystallization Communications 62: 656-658.

Takeshita D, Zenno S, Lee WC, Saigo K, Tanokura M 2007. Crystal structure of the PIN domain of human telomerase-associated protein EST1A. Proteins-Structure Function and Bioinformatics 68(4): 980-989.

Tanner NK, Linder P 2001. DExD/H box RNA helicases: From generic motors to specific dissociation functions. Molecular Cell 8(2): 251-262.

Thomsen ND, Berger JM 2009. Running in Reverse: The Structural Basis for Translocation Polarity in Hexameric Helicases. Cell 139: 523-534.

Tsumoto K, Ejima D, Kumagai I, Arakawa T 2003. Practical considerations in refolding proteins from inclusion bodies. Protein Expression and Purification 28: 1-8.

Tuerk C, MacDougal S, Gold L 1992. RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. PNAS 89: 6988-6992.

Walker JE, Saraste M, Runswick MJ, Gay NJ 1982. Distantly related sequences in the α- and β-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold The EMBO Journal 1(8): 945-951.

Wall D, Kaiser D 1999. Type IV pili and cell motility. Molecular Microbiology 32(1): 1-10.

Walters SB, Dubnau E, Kolesnikova I, Laval F, Daffe M, Smith I 2006. The *Mycobacterium tuberculosis* PhoPR two-component system regulates genes essential for virulence and complex lipid biosynthesis. Molecular Microbiology 60(2): 312-330.

WHO 2007. Fact sheet No. 104. World Health Organisation.

WHO 2012. Global Tuberculosis report 2012. World Health Organisation.

Winther KS, Gerdes K 2009. Ectopic production of VapCs from Enterobacteria inhibits translation and trans-activates YoeB mRNA interferase. Molecular Microbiology 74(2): 918–930.

Winther KS, Gerdes K 2011. Enteric virulence associated protein VapC inhibits translation by cleavage of initiator tRNA. PNAS 108(18): 7403-7407.

Woestenenk EA, Hammarstrom M, Van den Berg S, Hard T, Berglund H 2004. His tag effect on solubility of human proteins produced in *Escherichia coli*: a comparison between four expression vectors. Journal of Structural and Functional Genomics 5(3): 217-29

Wong C, Sridhara S, Bardwell JCA, Jakob U 2000. Heating Greatly Speeds Coomassie Blue Staining and Destaining BioTechniques 28: 426-432.

Ye J, Osborne AR, Groll M, Rapoport TA 2004. RecA-like motor ATPases-lessons from structures. Biochemica et Biophysica Acta 1659: 1-18.

Yu X, Horiguchi T, Shigesada K, Egelman EH 2000. Three dimensional reconstruction of transcription termination factor rho: orientation of the N-terminal domain and visualization of an RNA binding site. Journal of Molecular Biology 299 1279-1287.

Zhang YX, Guo XK, Wu C, Bi B, Ren SX, Wu CF, Zhao GP 2004. Characterisation of a novel toxin-antitoxin module, VapBC, encoded my *Leptospira interrogans* chromosome. Cell Research 14(3): 208-216.

Zou J, Xu J, Liu L, Li S, Wu C, Du G 2009. Solubilization and purification of Escherichia coli expressed GST-fusion human vascular endothelial growth factors with N-Lauroylsarcosine. African Journal od Biotechnology 8(10): 2362-2366.

Appendices

Appendix A: Bacterial strains, vectors and media

Table A.1 Bacterial strains and vectors used in this study

Bacterial strain	Description
E. coli	
DB10B (TOP10)	F-mcrA Δ(mrr-hsdRMS-mcrBC) Φ80 dlacZ ΔM15 Δ lacX74 deoR recA1 araD139 Δ(ara leu)7697 galU galK rpsL endA1 nupG
DH5α (electrocompentent)	fhuA2 Δ(argF-lacZ)U169 phoA glnV44 Φ80 Δ(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17
DH5α (chemically compentent)	F- φ80lacZΔM15 Δ(lacZYA-argF)U169 deoR recA1 endA1 hsdR17(rk -, mk+) phoA supE44 thi-1 gyrA96 relA1 λ-
BL21 (DE3)	F- ompT hsdSB (rB-mB-) gal dcm (DE3)
M. smegmatis	
mc^2155	Electrocompetent wild-type strain of *M. smegmatis*
$mc^2 4517$	*M. smegmatis* expression strain with T7 RNA polymerase, Km^r
Vectors	
pYUB1049	*E. coli* mycobacterium shuttle vector with T7 promoter and encoding both C- and N-terminal His-tags; Hyg^r
$pDONR_{221}$	Donor vector; ColE1 (pUC-type) origin of replication; rrnB T1 transcription termination sequence; *att*P recombination site; *att*L recombination site, ccdB negative selection gene; chloramphenicol resistance gene; Km^r
$pDEST_{SMG}$	Modified pYUB1049 *E. coli* mycobacterium shuttle vector *att* recombination recognition sequences; Hyg^r
$pDEST_{17}$	bla promoter; T7 promoter; attR1 recombination site; attR2 recombination site; chlorampenicol resistance gene; 6xHis tag N-terminal; Amp^r
pYUB28b	*E. coli* mycobacterium shuttle vector with T7 promoter, MCS pET28b, and encoding both C- and N-terminal His-tags; Hyg^r
pET28b-*Pst*I	Modified pET28b where the *Nco*I sites has been replaced with *Pst*I site; C- or N-terminal His-tags; Km^r
pMIND	Tetracycline inducible expression vector; Km^r Hyg^r

Table A.2 Growth media used in this study

Solid	Description
LB agar	1 % (w/v) bactotryptone, 0.5 % (w/v) yeast extract, 1 % (w/v) NaCl, 15 g/L agar pH 8
Low salt LB agar	1 % (w/v) bactotryptone, 0.5 % (w/v) yeast extract, 0.5 % (w/v) NaCl, 15g/L agar pH 8
7H10	1.9 g 7H10 powder, 0.5 % (w/v) glycerol in 90 ml H_2O, autoclaved then 10 ml ADC enrichment and 0.05 % (v/v) Tween-80 added at 50 °C
Liquid	
LB	1 % (w/v) bactotryptone, 0.5 % (w/v) yeast extract, 1 % (w/v) NaCl, pH 8
Low salt LB	1 % (w/v) bactotryptone, 0.5 % (w/v) yeast extract, 0.5 % (w/v) NaCl
LBT	1 % (w/v) bactotryptone, 0.5 % (w/v) yeast extract, 1 % (w/v) NaCl 0.05 % (v/v) Tween-80, pH 8
7H9	0.47 g 7H9 powder, 0.2 % (w/v) glycerol in 90 ml H_2O, autoclaved then 10 ml ADC enrichment and 0.05 % (v/v) Tween-80 added at 50 °C
PA-0.5G	50 mM Na_2HPO_4, 50 mM KH_2PO_4, 25 mM $(NH_4)_2SO_4$, 1 mM $MgSO_4$, 0.5 % (w/v) glucose, 0.1 x metals mix**, 200 µg/ml each of 17 amino acids (no Cys, Tyr or Met). Individual components autoclaved or sterile filtered before adding to sterile H_2O
ZYP-5052	1 % (w/v) bactotryptone, 0.5% (w/v) yeast extract, 50 mM Na_2HPO_4, 50 mM KH_2PO_4, 25 mM $(NH_4)_2SO_4$, 1 mM $MgSO_4$, 0.5 % (w/v) glycerol, 0.05 % (w/v) glucose, 0.2 % (w/v) α-lactose, 1 x metals mix *
SOC	2 % (w/v) bactotryptone or bactopeptone, 0.55 % (w/v) yeast extract, 10 mM NaCl, 2.5 mM KCl 10 mM $MgCl_2$, 10 mM $MgSO_4$, 20 mM glucose
Hartmans de Bont	Components added in order given: 1 x trace metal stock**, 15 mM (NH4)2SO4, 0.05 % (v/v) Tween-80, 0.2 % (v/v) glycerol, 50 mM MES. Autoclave then add 8.9 mM K_2HPO_4 and 7.08 mM NaH_2PO_4

*Metals mix 1000x, made up from the sterile stock solutions of each component to give the following concentrations: 50 µM $FeCl_3$ in 0.12 M HCl (filter sterile), 20 µM $CaCl_2$, 10 µM $MnCl_2$, 10 µM $ZnSO_4$, 2 µM $CoCl_2$, 2 µM $CuCl_2$, 2 µM $NiCl_2$, 2 µM Na_2MoO_4, 2 µM Na_2SeO_3, 2 µM H_3BO_3

** Trace metals stock 100x for HdB media: EDTA (0.1 g), $MgCl_2.6H_2O$ (1 g), $CaCl_2.2H_2O$ (10 mg), $NaMoO_4.2H_2O$ (2 mg), $CoCl_2.6H2O$ (4 mg), $MnCl_2.2H_2O$ (10 mg), $ZnSO_4.7H_2O$ (20 mg), $FeSO_4.7H_2O$ (50 mg), $CuSO_4.5H_2O$ to 100 ml

Refolding screen

	Buffer	Salt	Cation 2+	PEG 3350 (% w/v)	Detergent*	Chaotrope*[1]	Polar (P)/ Non-polar (NP) additives	Oxidising/ Reducing agents
1	50 mM tris pH 8.2	250 mM NaCl	1 mM EDTA	0.05 %	0.0 mM	0 mM	0 mM	1 mM DTT
2	50 mM MES pH 6.5	10 mM NaCl	2 mM $MgCl_2$ 2mM $CaCl_2$	0.00 %	0.3 mM	500 mM	0 mM	1 mM GSH 0.1 mM GSSG
3	50 mM MES pH 6.5	10 mM NaCl	1 mM EDTA	0.05 %	0.0 mM	500 mM	200 mM sucrose (NP) 200 mM L-arginine (P)	1 mM GSH 0.1 mM GSSG
4	50 mM tris pH 8.2	250 mM NaCl	2 mM $MgCl_2$ 2mM $CaCl_2$	0.00 %	0.3 mM	0 mM	200 mM sucrose (NP) 200 mM L-arginine (P)	1 mM DTT
5	50 mM MES pH 6.5	250 mM NaCl	2 mM $MgCl_2$ 2 mM $CaCl_2$	0.00 %	0.0 mM	0 mM	200 mM sucrose (NP)	1 mM GSH 0.1 mM GSSG
6	50 mM tris pH 8.2	10 mM NaCl	1 mM EDTA	0.05 %	0.3 mM	500 mM	200 mM sucrose (NP)	1 mM DTT
7	50 mM tris pH 8.2	10 mM NaCl	2 mM $MgCl_2$ 2mM $CaCl_2$	0.00 %	0.0 mM	500 mM	200 mM L-arginine (P)	1 mM DTT
8	50 mM MES pH 6.5	250 mM NaCl	1 mM EDTA	0.05 %	0.3 mM	0 mM	200 mM L-arginine (P)	1 mM GSH 0.1 mM GSSG
9	50 mM MES pH 6.5	250 mM NaCl	2 mM $MgCl_2$ 2mM $CaCl_2$	0.05 %	0.0 mM	500 mM	200 mM sucrose (NP)	1 mM DTT
10	50 mM tris pH 8.2	10 mM NaCl	1 mM EDTA	0.00 %	0.3 mM	0 mM	200 mM sucrose (NP)	1 mM GSH 0.1 mM GSSG
11	50 mM tris pH 8.2	10 mM NaCl	2 mM $MgCl_2$ 2mM $CaCl_2$	0.05 %	0.0 mM	0 mM	200 mM L-arginine (P)	1 mM GSH 0.1 mM GSSG
12	50 mM MES pH 6.5	250 mM NaCl	1 mM EDTA	0.00 %	0.3 mM	500 mM	200 mM L-arginine (P)	1 mM DTT
13	50 mM tris pH 8.2	250 mM NaCl	1 mM EDTA	0.00 %	0.0 mM	500 mM	0 mM	1 mM GSH 0.1 mM GSSG
14	50 mM MES pH 6.5	10 mM NaCl	2 mM $MgCl_2$ 2mM $CaCl_2$	0.05 %	0.3 mM	0 mM	0 mM	1 mM DTT
15	50 mM MES pH 6.5	10 mM NaCl	1 mM EDTA	0.00 %	0.0 mM	0 mM	200 mM sucrose (NP) 200 mM L-arginine (P)	1 mM DTT
16	50 mM tris pH 8.2	250 mM NaCl	2 mM $MgCl_2$ 2mM $CaCl_2$	0.05 %	0.3 mM	500 mM	200 mM sucrose (NP) 200 mM L-arginine (P)	1 mM GSH 0.1 mM GSSG

* Dodecyl maltoside *[1] Guanidine HCl

Appendix B: Gene and protein Information

Rv1095/BCG_1155 (1302 bp)

GTGACCGATACCCGCACGTACGTGCTCGACACCTCTGTGCTGCTGTCCGATCCGTGGGCGTG
CAGCCGGTTCGCCGAACACGATGTGGTGGTTCCGTTGGTGGTGATCAGCGAGCTAGAAGCCA
AGCGCCACCACCACGAGCTGGGATGGTTCGCCCGCCAGGCGTTGCGTCTGTTCGACGATCTG
CGCCTAGAACACGGGCGGTTGGATCAGCCGATTCCGGTTGGCACCCAAGGCGGTACGCTGCA
CGTCGAACTCAATCACACCGACCCGGCGGTGCTGCCCGCAGGCTTTCGCACCGACAGCAACG
ACTCGAGGATCTTGAGTTGCGCCGCCAACCTCGCCGCCGAGGGCAAGCGGGTCACGTTGGTC
AGCAAGGACATTCCGCTGCGCGTTAAGGCCGCCGCGGTGGGGCTGGCCGCCGACGAGTACCA
CGCGCAGGACGTCGTTGTGTCCGGATGGTCGGGGATGCACGAGCTCGAGACCGCTTCCGCGG
ATATCGATGCGTTGTTCGCCGATGGCGAGATCGACCTGGTCGAAGCCCGGGACCTACCGTGT
CACACCGGGATTCGGTTGCTGGGCGGCGGTTCCCACGCGCTGGGCCGGGTCAATGCGCATAA
ACGTGTTCAGCTGGTGCGAGGTGACCGTGAGGCGTTCGGTCTGCGTGGCCGCTCCGCCGAGC
AGCGGGTGGCGCTGGATTTGCTGCTCGATGAGTCGGTGGGCATCGTGTCGCTGGGCGGCAAA
GCCGGCACGGGCAAGTCCGCTTTGGCGTTGTGTGCGGGTCTGGAAGCCGTGCTGGAGCGACG
CACCCACCGCAAGGTGGTGGTCTTCCGCCCGCTGTACGCGGTCGGCGGCCAGGAGCTGGGCT
ACCTGCCCGGTAGCGAGAGCGAGAAGATGGGCCCGTGGGCGCAGGCGGTCTTCGACACCCTC
GAGGGGCTGGCCAGCCCGGCGGTGCTCGAGGAAGTGCTGTCCCGTGGCATGCTCGAGGTGCT
GCCGCTGACCCACATCCGGGGCCGCTCGTTGCATGACTCGTTCGTCATCGTCGACGAGGCAC
AGTCGCTGGAGCGCAATGTGTTGCTGACCGTGCTGTCCCGGTTGGGGACCGGTTCCCGGGTG
GTGTTGACCCACGACATCGCCCAGCGCGACAACCTGCGGGTCGGCCGCCACGACGGGGTCGC
CGCGGTGATCGAGAAGCTCAAAGGTCATCCGTTGTTCGCCCACATCACCTTGCTGCGCAGTG
AGCGCTCGCCGATCGCCGCGCTGGTCACCGAGATGCTCGAGGAGATCACCGGGCCGCGCTGA

Rv1095/BCG_1155 (433 aa)

MW 46881.5 Da pI 5.96

VTDTRTYVLDTSVLLSDPWACSRFAEHDVVVPLVVISELEAKRHHHELGWFARQALRLFDDL
RLEHGRLDQPIPVGTQGGTLHVELNHTDPAVLPAGFRTDSNDSRILSCAANLAAEGKRVTLV
SKDIPLRVKAAAVGLAADEYHAQDVVVSGWSGMHELETASADIDALFADGEIDLVEARDLPC
HTGIRLLGGGSHALGRVNAHKRVQLVRGDREAFGLRGRSAEQRVALDLLLDESVGIVSLGGK
AGTGKSALALCAGLEAVLERRTHRKVVVFRPLYAVGGQELGYLPGSESEKMGPWAQAVFDTL
EGLASPAVLEEVLSRGMLEVLPLTHIRGRSLHDSFVIVDEAQSLERNVLLTVLSRLGTGSRV
VLTHDIAQRDNLRVGRHDGVAAVIEKLKGHPLFAHITLLRSERSPIAALVTEMLEEITGPR*

Rv1095/BCG_1155 pYUB1049 C-terminal fusion tag (1368 bp)

GTGACCGATACCCGCACGTACGTGCTCGACACCTCTGTGCTGCTGTCCGATCCGTGGGCGTG
CAGCCGGTTCGCCGAACACGATGTGGTGGTTCCGTTGGTGGTGATCAGCGAGCTAGAAGCCA
AGCGCCACCACCACGAGCTGGGATGGTTCGCCCGCCAGGCGTTGCGTCTGTTCGACGATCTG
CGCCTAGAACACGGGCGGTTGGATCAGCCGATTCCGGTTGGCACCCAAGGCGGTACGCTGCA
CGTCGAACTCAATCACACCGACCCGGCGGTGCTGCCCGCAGGCTTTCGCACCGACAGCAACG
ACTCGAGGATCTTGAGTTGCGCCGCCAACCTCGCCGCCGAGGGCAAGCGGGTCACGTTGGTC
AGCAAGGACATTCCGCTGCGCGTTAAGGCCGCCGCGGTGGGGCTGGCCGCCGACGAGTACCA
CGCGCAGGACGTCGTTGTGTCCGGATGGTCGGGGATGCACGAGCTCGAGACCGCTTCCGCGG
ATATCGATGCGTTGTTCGCCGATGGCGAGATCGACCTGGTCGAAGCCCGGGACCTACCGTGT
CACACCGGGATTCGGTTGCTGGGCGGCGGTTCCCACGCGCTGGGCCGGGTCAATGCGCATAA
ACGTGTTCAGCTGGTGCGAGGTGACCGTGAGGCGTTCGGTCTGCGTGGCCGCTCCGCCGAGC
AGCGGGTGGCGCTGGATTTGCTGCTCGATGAGTCGGTGGGCATCGTGTCGCTGGGCGGCAAA
GCCGGCACGGGCAAGTCCGCTTTGGCGTTGTGTGCGGGTCTGGAAGCCGTGCTGGAGCGACG
CACCCACCGCAAGGTGGTGGTCTTCCGCCCGCTGTACGCGGTCGGCGGCCAGGAGCTGGGCT
ACCTGCCCGGTAGCGAGAGCGAGAAGATGGGCCCGTGGGCGCAGGCGGTCTTCGACACCCTC
GAGGGGCTGGCCAGCCCGGCGGTGCTCGAGGAAGTGCTGTCCCGTGGCATGCTCGAGGTGCT
GCCGCTGACCCACATCCGGGGCCGCTCGTTGCATGACTCGTTCGTCATCGTCGACGAGGCAC
AGTCGCTGGAGCGCAATGTGTTGCTGACCGTGCTGTCCCGGTTGGGGACCGGTTCCCGGGTG
GTGTTGACCCACGACATCGCCCAGCGCGACAACCTGCGGGTCGGCCGCCACGACGGGGTCGC
CGCGGTGATCGAGAAGCTCAAAGGTCATCCGTTGTTCGCCCACATCACCTTGCTGCGCAGTG
AGCGCTCGCCGATCGCCGCGCTGGTCACCGAGATGCTCGAGGAGATCACCGGGCCGCGCGCG

GATCCGAATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCA
CTGA

Rv1095/BCG_1155 pYUB1049 C-terminal fusion tag (456 aa)
MW 49274.0 Da pI 6.02
VTDTRTYVLDTSVLLSDPWACSRFAEHDVVVPLVVISELEAKRHHHELGWFARQALRLFDDL
RLEHGRLDQPIPVGTQGGTLHVELNHTDPAVLPAGFRTDSNDSRILSCAANLAAEGKRVTLV
SKDIPLRVKAAAVGLAADEYHAQDVVVSGWSGMHELETASADIDALFADGEIDLVEARDLPC
HTGIRLLGGGSHALGRVNAHKRVQLVRGDREAFGLRGRSAEQRVALDLLLDESVGIVSLGGK
AGTGKSALALCAGLEAVLERRTHRKVVVFRPLYAVGGQELGYLPGSESEKMGPWAQAVFDTL
EGLASPAVLEEVLSRGMLEVLPLTHIRGRSLHDSFVIVDEAQSLERNVLLTVLSRLGTGSRV
VLTHDIAQRDNLRVGRHDGVAAVIEKLKGHPLFAHITLLRSERSPIAALVTEMLEEITGPRA
DPNSSSVDKLAAALEHHHHHH*

Rv1095/BCG_1155 pYUB1049 N-terminal fusion tag (1365 bp)
ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATAT
GGTGACCGATACCCGCACGTACGTGCTCGACACCTCTGTGCTGCTGTCCGATCCGTGGGCGT
GCAGCCGGTTCGCCGAACACGATGTGGTGGTTCCGTTGGTGGTGATCAGCGAGCTAGAAGCC
AAGCGCCACCACCACGAGCTGGGATGGTTCGCCCGCCAGGCGTTGCGTCTGTTCGACGATCT
GCGCCTAGAACACGGGCGGTTGGATCAGCCGATTCCGGTTGGCACCCAAGGCGGTACGCTGC
ACGTCGAACTCAATCACACCGACCCGGCGGTGCTGCCCGCAGGCTTTCGCACCGACAGCAAC
GACTCGAGGATCTTGAGTTGCGCCGCCAACCTCGCCGCCGAGGGCAAGCGGGTCACGTTGGT
CAGCAAGGACATTCCGCTGCGCGTTAAGGCCGCCGCGGTGGGGCTGGCCGCCGACGAGTACC
ACGCGCAGGACGTCGTTGTGTCCGGATGGTCGGGGATGCACGAGCTCGAGACCGCTTCCGCG
GATATCGATGCGTTGTTCGCCGATGGCGAGATCGACCTGGTCGAAGCCCGGGACCTACCGTG
TCACACCGGGATTCGGTTGCTGGGCGGCGGTTCCCACGCGCTGGGCCGGGTCAATGCGCATA
AACGTGTTCAGCTGGTGCGAGGTGACCGTGAGGCGTTCGGTCTGCGTGGCCGCTCCGCCGAG
CAGCGGGTGGCGCTGGATTTGCTGCTCGATGAGTCGGTGGGCATCGTGTCGCTGGGCGGCAA
AGCCGGCACGGGCAAGTCCGCTTTGGCGTTGTGTGCGGGTCTGGAAGCCGTGCTGGAGCGAC
GCACCCACCGCAAGGTGGTGGTCTTCCGCCCGCTGTACGCGGTCGGCGGCCAGGAGCTGGGC
TACCTGCCCGGTAGCGAGAGCGAGAAGATGGGCCCGTGGGCGCAGGCGGTCTTCGACACCCT
CGAGGGGCTGGCCAGCCCGGCGGTGCTCGAGGAAGTGCTGTCCCGTGGCATGCTCGAGGTGC
TGCCGCTGACCCACATCCGGGGCCGCTCGTTGCATGACTCGTTCGTCATCGTCGACGAGGCA
CAGTCGCTGGAGCGCAATGTGTTGCTGACCGTGCTGTCCCGGTTGGGGACCGGTTCCCGGGT
GGTGTTGACCCACGACATCGCCCAGCGCGACAACCTGCGGGTCGGCCGCCACGACGGGGTCG
CCGCGGTGATCGAGAAGCTCAAAGGTCATCCGTTGTTCGCCCACATCACCTTGCTGCGCAGT
GAGCGCTCGCCGATCGCCGCGCTGGTCACCGAGATGCTCGAGGAGATCACCGGGCCGCGCTG
A

Rv1095/BCG_1155 pYUB1049 N-terminal fusion tag (455 aa)
MW 49176.0 Da pI 6.22
MGSSHHHHHHSSGLVPRGSHMVTDTRTYVLDTSVLLSDPWACSRFAEHDVVVPLVVISELEA
KRHHHELGWFARQALRLFDDLRLEHGRLDQPIPVGTQGGTLHVELNHTDPAVLPAGFRTDSN
DSRILSCAANLAAEGKRVTLVSKDIPLRVKAAAVGLAADEYHAQDVVVSGWSGMHELETASA
DIDALFADGEIDLVEARDLPCHTGIRLLGGGSHALGRVNAHKRVQLVRGDREAFGLRGRSAE
QRVALDLLLDESVGIVSLGGKAGTGKSALALCAGLEAVLERRTHRKVVVFRPLYAVGGQELG
YLPGSESEKMGPWAQAVFDTLEGLASPAVLEEVLSRGMLEVLPLTHIRGRSLHDSFVIVDEA
QSLERNVLLTVLSRLGTGSRVVLTHDIAQRDNLRVGRHDGVAAVIEKLKGHPLFAHITLLRS
ERSPIAALVTEMLEEITGPR*

Rv1095/BCG_1155 pDEST$_{SMG}$ N-terminal fusion tag (1368 bp)
ATGTCGCATCACCATCACCATCACCTCGAATCACCATCAACAAGTTTGTACAAAAAAGCAGG
CTTCGTGACCGATACCCGCACGTACGTGCTCGACACCTCTGTGCTGCTGTCCGATCCGTGGG
CGTGCAGCCGGTTCGCCGAACACGATGTGGTGGTTCCGTTGGTGGTGATCAGCGAGCTAGAA
GCCAAGCGCCACCACCACGAGCTGGGATGGTTCGCCCGCCAGGCGTTGCGTCTGTTCGACGA
TCTGCGCCTAGAACACGGGCGGTTGGATCAGCCGATTCCGGTTGGCACCCAAGGCGGTACGC

TGCACGTCGAACTCAATCACACCGACCCGGCGGTGCTGCCCGCAGGCTTTCGCACCGACAGC
AACGACTCGAGGATCTTGAGTTGCGCCGCCAACCTCGCCGCCGAGGGCAAGCGGGTCACGTT
GGTCAGCAAGGACATTCCGCTGCGCGTTAAGGCCGCCGCGGTGGGGCTGGCCGCCGACGAGT
ACCACGCGCAGGACGTCGTTGTGTCCGGATGGTCGGGGATGCACGAGCTCGAGACCGCTTCC
GCGGATATCGATGCGTTGTTCGCCGATGGCGAGATCGACCTGGTCGAAGCCCGGGACCTACC
GTGTCACACCGGGATTCGGTTGCTGGGCGGCGGTTCCCACGCGCTGGGCCGGGTCAATGCGC
ATAAACGTGTTCAGCTGGTGCGAGGTGACCGTGAGGCGTTCGGTCTGCGTGGCCGCTCCGCC
GAGCAGCGGGTGGCGCTGGATTTGCTGCTCGATGAGTCGGTGGGCATCGTGTCGCTGGGCGG
CAAAGCCGGCACGGGCAAGTCCGCTTTGGCGTTGTGTGCGGGTCTGGAAGCCGTGCTGGAGC
GACGCACCCACCGCAAGGTGGTGGTCTTCCGCCCGCTGTACGCGGTCGGCGGCCAGGAGCTG
GGCTACCTGCCCGGTAGCGAGAGCGAGAAGATGGGCCCGTGGGCGCAGGCGGTCTTCGACAC
CCTCGAGGGGCTGGCCAGCCCGGCGGTGCTCGAGGAAGTGCTGTCCCGTGGCATGCTCGAGG
TGCTGCCGCTGACCCACATCCGGGGCCGCTCGTTGCATGACTCGTTCGTCATCGTCGACGAG
GCACAGTCGCTGGAGCGCAATGTGTTGCTGACCGTGCTGTCCCGGTTGGGGACCGGTTCCCG
GGTGGTGTTGACCCACGACATCGCCCAGCGCGACAACCTGCGGGTCGGCCGCCACGACGGGG
TCGCCGCGGTGATCGAGAAGCTCAAAGGTCATCCGTTGTTCGCCCACATCACCTTGCTGCGC
AGTGAGCGCTCGCCGATCGCCGCGCTGGTCACCGAGATGCTCGAGGAGATCACCGGGCCGCG
CTGA

Rv1095/BCG_1155 pDEST$_{SMG}$ N-terminal fusion tag (456 aa)
MW 49432.3 Da pI 6.2
MSHHHHHHLESPSTSLYKKAGFVTDTRTYVLDTSVLLSDPWACSRFAEHDVVVPLVVISELE
AKRHHHELGWFARQALRLFDDLRLEHGRLDQPIPVGTQGGTLHVELNHTDPAVLPAGFRTDS
NDSRILSCAANLAAEGKRVTLVSKDIPLRVKAAAVGLAADEYHAQDVVVSGWSGMHELETAS
ADIDALFADGEIDLVEARDLPCHTGIRLLGGGSHALGRVNAHKRVQLVRGDREAFGLRGRSA
EQRVALDLLLDESVGIVSLGGKAGTGKSALALCAGLEAVLERRTHRKVVVFRPLYAVGGQEL
GYLPGSESEKMGPWAQAVFDTLEGLASPAVLEEVLSRGMLEVLPLTHIRGRSLHDSFVIVDE
AQSLERNVLLTVLSRLGTGSRVVLTHDIAQRDNLRVGRHDGVAAVIEKLKGHPLFAHITLLR
SERSPIAALVTEMLEEITGPR*

Rv1095/BCG_1155 pDEST$_{17}$ N-terminal fusion tag (1368 bp)
ATGTCGTACTACCATCACCATCACCATCACCTCGAATCAACAAGTTTGTACAAAAAAGCAGG
CTTCGTGACCGATACCCGCACGTACGTGCTCGACACCTCTGTGCTGCTGTCCGATCCGTGGG
CGTGCAGCCGGTTCGCCGAACACGATGTGGTGGTTCCGTTGGTGGTGATCAGCGAGCTAGAA
GCCAAGCGCCACCACCACGAGCTGGGATGGTTCGCCCGCCAGGCGTTGCGTCTGTTCGACGA
TCTGCGCCTAGAACACGGGCGGTTGGATCAGCCGATTCCGGTTGGCACCCAAGGCGGTACGC
TGCACGTCGAACTCAATCACACCGACCCGGCGGTGCTGCCCGCAGGCTTTCGCACCGACAGC
AACGACTCGAGGATCTTGAGTTGCGCCGCCAACCTCGCCGCCGAGGGCAAGCGGGTCACGTT
GGTCAGCAAGGACATTCCGCTGCGCGTTAAGGCCGCCGCGGTGGGGCTGGCCGCCGACGAGT
ACCACGCGCAGGACGTCGTTGTGTCCGGATGGTCGGGGATGCACGAGCTCGAGACCGCTTCC
GCGGATATCGATGCGTTGTTCGCCGATGGCGAGATCGACCTGGTCGAAGCCCGGGACCTACC
GTGTCACACCGGGATTCGGTTGCTGGGCGGCGGTTCCCACGCGCTGGGCCGGGTCAATGCGC
ATAAACGTGTTCAGCTGGTGCGAGGTGACCGTGAGGCGTTCGGTCTGCGTGGCCGCTCCGCC
GAGCAGCGGGTGGCGCTGGATTTGCTGCTCGATGAGTCGGTGGGCATCGTGTCGCTGGGCGG
CAAAGCCGGCACGGGCAAGTCCGCTTTGGCGTTGTGTGCGGGTCTGGAAGCCGTGCTGGAGC
GACGCACCCACCGCAAGGTGGTGGTCTTCCGCCCGCTGTACGCGGTCGGCGGCCAGGAGCTG
GGCTACCTGCCCGGTAGCGAGAGCGAGAAGATGGGCCCGTGGGCGCAGGCGGTCTTCGACAC
CCTCGAGGGGCTGGCCAGCCCGGCGGTGCTCGAGGAAGTGCTGTCCCGTGGCATGCTCGAGG
TGCTGCCGCTGACCCACATCCGGGGCCGCTCGTTGCATGACTCGTTCGTCATCGTCGACGAG
GCACAGTCGCTGGAGCGCAATGTGTTGCTGACCGTGCTGTCCCGGTTGGGGACCGGTTCCCG
GGTGGTGTTGACCCACGACATCGCCCAGCGCGACAACCTGCGGGTCGGCCGCCACGACGGGG
TCGCCGCGGTGATCGAGAAGCTCAAAGGTCATCCGTTGTTCGCCCACATCACCTTGCTGCGC
AGTGAGCGCTCGCCGATCGCCGCGCTGGTCACCGAGATGCTCGAGGAGATCACCGGGCCGCG
CTGA

Rv1095/BCG_1155 pDEST$_{17}$ N-terminal fusion tag (456 aa)
MW 49574.5 Da pI 6.2

MSYYHHHHHHLESTSLYKKAGFVTDTRTYVLDTSVLLSDPWACSRFAEHDVVVPLVVISELE
AKRHHHELGWFARQALRLFDDLRLEHGRLDQPIPVGTQGGTLHVELNHTDPAVLPAGFRTDS
NDSRILSCAANLAAEGKRVTLVSKDIPLRVKAAAVGLAADEYHAQDVVVSGWSGMHELETAS
ADIDALFADGEIDLVEARDLPCHTGIRLLGGGSHALGRVNAHKRVQLVRGDREAFGLRGRSA
EQRVALDLLLDESVGIVSLGGKAGTGKSALALCAGLEAVLERRTHRKVVVFRPLYAVGGQEL
GYLPGSESEKMGPWAQAVFDTLEGLASPAVLEEVLSRGMLEVLPLTHIRGRSLHDSFVIVDE
AQSLERNVLLTVLSRLGTGSRVVLTHDIAQRDNLRVGRHDGVAAVIEKLKGHPLFAHITLLR
SERSPIAALVTEMLEEITGPR*

Rv1095/BCG_1155 pYUB28b C-terminal fusion tag (1344 bp)

ATGGTGACCGATACCCGCACGTACGTGCTCGACACCTCTGTGCTGCTGTCCGATCCGTGGGC
GTGCAGCCGGTTCGCCGAACACGATGTGGTGGTTCCGTTGGTGGTGATCAGCGAGCTAGAAG
CCAAGCGCCACCACCACGAGCTGGGATGGTTCGCCCGCCAGGCGTTGCGTCTGTTCGACGAT
CTGCGCCTAGAACACGGGCGGTTGGATCAGCCGATTCCGGTTGGCACCCAAGGCGGTACGCT
GCACGTCGAACTCAATCACACCGACCCGGCGGTGCTGCCCGCAGGCTTTCGCACCGACAGCA
ACGACTCGAGGATCTTGAGTTGCGCCGCCAACCTCGCCGCCGAGGGCAAGCGGGTCACGTTG
GTCAGCAAGGACATTCCGCTGCGCGTTAAGGCCGCCGCGGTGGGGCTGGCCGCCGACGAGTA
CCACGCGCAGGACGTCGTTGTGTCCGGATGGTCGGGGATGCACGAGCTCGAGACCGCTTCCG
CGGATATCGATGCGTTGTTCGCCGATGGCGAGATCGACCTGGTCGAAGCCCGGGACCTACCG
TGTCACACCGGGATTCGGTTGCTGGGCGGCGGTTCCCACGCGCTGGGCCGGGTCAATGCGCA
TAAACGTGTTCAGCTGGTGCGAGGTGACCGTGAGGCGTTCGGTCTGCGTGGCCGCTCCGCCG
AGCAGCGGGTGGCGCTGGATTTGCTGCTCGATGAGTCGGTGGGCATCGTGTCGCTGGGCGGC
AAAGCCGGCACGGGCAAGTCCGCTTTGGCGTTGTGTGCGGGTCTGGAAGCCGTGCTGGAGCG
ACGCACCCACCGCAAGGTGGTGGTCTTCCGCCCGCTGTACGCGGTCGGCGGCCAGGAGCTGG
GCTACCTGCCCGGTAGCGAGAGCGAGAAGATGGGCCCGTGGGCGCAGGCGGTCTTCGACACC
CTCGAGGGGCTGGCCAGCCCGGCGGTGCTCGAGGAAGTGCTGTCCCGTGGCATGCTCGAGGT
GCTGCCGCTGACCCACATCCGGGGCCGCTCGTTGCATGACTCGTTCGTCATCGTCGACGAGG
CACAGTCGCTGGAGCGCAATGTGTTGCTGACCGTGCTGTCCCGGTTGGGGACCGGTTCCCGG
GTGGTGTTGACCCACGACATCGCCCAGCGCGACAACCTGCGGGTCGGCCGCCACGACGGGGT
CGCCGCGGTGATCGAGAAGCTCAAAGGTCATCCGTTGTTCGCCCACATCACCTTGCTGCGCA
GTGAGCGCTCGCCGATCGCCGCGCTGGTCACCGAGATGCTCGAGGAGATCACCGGGCCGCGC
AAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGA

Rv1095/BCG_1155 pYUB28b C-terminal fusion tag (447 aa)

MW 48532.4 pI 6.14

MVTDTRTYVLDTSVLLSDPWACSRFAEHDVVVPLVVISELEAKRHHHELGWFARQALRLFDD
LRLEHGRLDQPIPVGTQGGTLHVELNHTDPAVLPAGFRTDSNDSRILSCAANLAAEGKRVTL
VSKDIPLRVKAAAVGLAADEYHAQDVVVSGWSGMHELETASADIDALFADGEIDLVEARDLP
CHTGIRLLGGGSHALGRVNAHKRVQLVRGDREAFGLRGRSAEQRVALDLLLDESVGIVSLGG
KAGTGKSALALCAGLEAVLERRTHRKVVVFRPLYAVGGQELGYLPGSESEKMGPWAQAVFDT
LEGLASPAVLEEVLSRGMLEVLPLTHIRGRSLHDSFVIVDEAQSLERNVLLTVLSRLGTGSR
VVLTHDIAQRDNLRVGRHDGVAAVIEKLKGHPLFAHITLLRSERSPIAALVTEMLEEITGPR
KLAAALEHHHHHH*

Rv1095alt (1454 bp)

ATGGCTAGCGACATGCTCTGCTGCCAGGGCGGCACCTTCCGTCACGACGGCTGTCATGACAA
GGGCAGGACCGGCCCCGGTCCTGGTGTCGCTGCCCCCGCCGACATGCTCGGGTGGGTCCGCT
CGAGCGCCGTTAGCTCGAGGAGCGCTCCGTGACCGATACCCGCACGTACGTGCTCGACACCT
CTGTGCTGCTGTCCGATCCGTGGGCGTGCAGCCGGTTCGCCGAACACGATGTGGTGGTTCCG
TTGGTGGTGATCAGCGAGCTAGAAGCCAAGCGCCACCACCACGAGCTGGGATGGTTCGCCCG
CCAGGCGTTGCGTCTGTTCGACGATCTGCGCCTAGAACACGGGCGGTTGGATCAGCCGATTC
CGGTTGGCACCCAAGGCGGTACGCTGCACGTCGAACTCAATCACACCGACCCGGCGGTGCTG
CCCGCAGGCTTTCGCACCGACAGCAACGACTCGAGGATCTTGAGTTGCGCCGCCAACCTCGC
CGCCGAGGGCAAGCGGGTCACGTTGGTCAGCAAGGACATTCCGCTGCGCGTTAAGGCCGCCG
CGGTGGGGCTGGCCGCCGACGAGTACCACGCGCAGGACGTCGTTGTGTCCGGATGGTCGGGG

ATGCACGAGCTCGAGACCGCTTCCGCGGATATCGATGCGTTGTTCGCCGATGGCGAGATCGA
CCTGGTCGAAGCCCGGGACCTACCGTGTCACACCGGGATTCGGTTGCTGGGCGGCGGTTCCC
ACGCGCTGGGCCGGGTCAATGCGCATAAACGTGTTCAGCTGGTGCGAGGTGACCGTGAGGCG
TTCGGTCTGCGTGGCCGCTCCGCCGAGCAGCGGGTGGCGCTGGATTTGCTGCTCGATGAGTC
GGTGGGCATCGTGTCGCTGGGCGGCAAAGCCGGCACGGGCAAGTCCGCTTTGGCGTTGTGTG
CGGGTCTGGAAGCCGTGCTGGAGCGACGCACCCACCGCAAGGTGGTGGTCTTCCGCCCGCTG
TACGCGGTCGGCGGCCAGGAGCTGGGCTACCTGCCCGGTAGCGAGAGCGAGAAGATGGGCCC
GTGGGCGCAGGCGGTCTTCGACACCCTCGAGGGGCTGGCCAGCCCGGCGGTGCTCGAGGAAG
TGCTGTCCCGTGGCATGCTCGAGGTGCTGCCGCTGACCCACATCCGGGGCCGCTCGTTGCAT
GACTCGTTCGTCATCGTCGACGAGGCACAGTCGCTGGAGCGCAATGTGTTGCTGACCGTGCT
GTCCCGGTTGGGGACCGGTTCCCGGGTGGTGTTGACCCACGACATCGCCCAGCGCGACAACC
TGCGGGTCGGCCGCCACGACGGGGTCGCCGCGGTGATCGAGAAGCTCAAAGGTCATCCGTTG
TTCGCCCACATCACCTTGCTGCGCAGTGAGCGCTCGCCGATCGCCGCGCTGGTCACCGAGAT
GCTCGAGGAGATCACCGGGCCGCGCTGA

Rv1095alt (486 aa)
MW 52053.3 Da pI 6.09
MASDMLCCQGGTFRHDGCHDKGRTGPGPGVAAPADMLGWVRSSAVSSRSAP*VTDTRTYVLD
TSVLLSDPWACSRFAEHDVVVPLVVISELEAKRHHHELGWFARQALRLFDDLRLEHGRLDQP
IPVGTQGGTLHVELNHTDPAVLPAGFRTDSNDSRILSCAANLAAEGKRVTLVSKDIPLRVKA
AAVGLAADEYHAQDVVVSGWSGMHELETASADIDALFADGEIDLVEARDLPCHTGIRLLGGG
SHALGRVNAHKRVQLVRGDREAFGLRGRSAEQRVALDLLLDESVGIVSLGGKAGTGKSALAL
CAGLEAVLERRTHRKVVVFRPLYAVGGQELGYLPGSESEKMGPWAQAVFDTLEGLASPAVLE
EVLSRGMLEVLPLTHIRGRSLHDSFVIVDEAQSLERNVLLTVLSRLGTGSRVVLTHDIAQRD
NLRVGRHDGVAAVIEKLKGHPLFAHITLLRSERSPIAALVTEMLEEITGPR*

Rv1095alt pYUB1049 C-terminal fusion tag (1524 bp)
ATGGCTAGCGACATGCTCTGCTGCCAGGGCGGCACCTTCCGTCACGACGGCTGTCATGACAA
GGGCAGGACCGGCCCCGGTCCTGGTGTCGCTGCCCCCGCCGACATGCTCGGGTGGGTCCGCT
CGAGCGCCGTTAGCTCGAGGAGCGCTCCGTGAGTGACCGATACCCGCACGTACGTGCTCGAC
ACCTCTGTGCTGCTGTCCGATCCGTGGGCGTGCAGCCGGTTCGCCGAACACGATGTGGTGGT
TCCGTTGGTGGTGATCAGCGAGCTAGAAGCCAAGCGCCACCACCACGAGCTGGGATGGTTCG
CCCGCCAGGCGTTGCGTCTGTTCGACGATCTGCGCCTAGAACACGGGCGGTTGGATCAGCCG
ATTCCGGTTGGCACCCAAGGCGGTACGCTGCACGTCGAACTCAATCACACCGACCCGGCGGT
GCTGCCCGCAGGCTTTCGCACCGACAGCAACGACTCGAGGATCTTGAGTTGCGCCGCCAACC
TCGCCGCCGAGGGCAAGCGGGTCACGTTGGTCAGCAAGGACATTCCGCTGCGCGTTAAGGCC
GCCGCGGTGGGGCTGGCCGCCGACGAGTACCACGCGCAGGACGTCGTTGTGTCCGGATGGTC
GGGGATGCACGAGCTCGAGACCGCTTCCGCGGATATCGATGCGTTGTTCGCCGATGGCGAGA
TCGACCTGGTCGAAGCCCGGGACCTACCGTGTCACACCGGGATTCGGTTGCTGGGCGGCGGT
TCCCACGCGCTGGGCCGGGTCAATGCGCATAAACGTGTTCAGCTGGTGCGAGGTGACCGTGA
GGCGTTCGGTCTGCGTGGCCGCTCCGCCGAGCAGCGGGTGGCGCTGGATTTGCTGCTCGATG
AGTCGGTGGGCATCGTGTCGCTGGGCGGCAAAGCCGGCACGGGCAAGTCCGCTTTGGCGTTG
TGTGCGGGTCTGGAAGCCGTGCTGGAGCGACGCACCCACCGCAAGGTGGTGGTCTTCCGCCC
GCTGTACGCGGTCGGCGGCCAGGAGCTGGGCTACCTGCCCGGTAGCGAGAGCGAGAAGATGG
GCCCGTGGGCGCAGGCGGTCTTCGACACCCTCGAGGGGCTGGCCAGCCCGGCGGTGCTCGAG
GAAGTGCTGTCCCGTGGCATGCTCGAGGTGCTGCCGCTGACCCACATCCGGGGCCGCTCGTT
GCATGACTCGTTCGTCATCGTCGACGAGGCACAGTCGCTGGAGCGCAATGTGTTGCTGACCG
TGCTGTCCCGGTTGGGGACCGGTTCCCGGGTGGTGTTGACCCACGACATCGCCCAGCGCGAC
AACCTGCGGGTCGGCCGCCACGACGGGGTCGCCGCGGTGATCGAGAAGCTCAAAGGTCATCC
GTTGTTCGCCCACATCACCTTGCTGCGCAGTGAGCGCTCGCCGATCGCCGCGCTGGTCACCG
AGATGCTCGAGGAGATCACCGGGCCGCGCGCGGATCCGAATTCGAGCTCCGTCGACAAGCTT
GCGGCCGCACTCGAGCACCACCACCACCACCACTGA

Rv1095alt pYUB1049 C-terminal fusion tag (508 aa)
MW 54445.9 Da pI 6.13
MASDMLCCQGGTFRHDGCHDKGRTGPGPGVAAPADMLGWVRSSAVSSRSAP*VTDTRTYVLD
TSVLLSDPWACSRFAEHDVVVPLVVISELEAKRHHHELGWFARQALRLFDDLRLEHGRLDQP

IPVGTQGGTLHVELNHTDPAVLPAGFRTDSNDSRILSCAANLAAEGKRVTLVSKDIPLRVKA
AAVGLAADEYHAQDVVVSGWSGMHELETASADIDALFADGEIDLVEARDLPCHTGIRLLGGG
SHALGRVNAHKRVQLVRGDREAFGLRGRSAEQRVALDLLLDESVGIVSLGGKAGTGKSALAL
CAGLEAVLERRTHRKVVVFRPLYAVGGQELGYLPGSESEKMGPWAQAVFDTLEGLASPAVLE
EVLSRGMLEVLPLTHIRGRSLHDSFVIVDEAQSLERNVLLTVLSRLGTGSRVVLTHDIAQRD
NLRVGRHDGVAAVIEKLKGHPLFAHITLLRSERSPIAALVTEMLEEITGPRADPNSSSVDKL
AAALEHHHHHH*

Rv1095alt pYUB1049 N-terminal fusion tag (1514 bp)

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATAT
GGCTAGCGACATGCTCTGCTGCCAGGGCGGCACCTTCCGTCACGACGGCTGTCATGACAAGG
GCAGGACCGGCCCCGGTCCTGGTGTCGCTGCCCCCGCCGACATGCTCGGGTGGGTCCGCTCG
AGCGCCGTTAGCTCGAGGAGCGCTCCGTGACCGATACCCGCACGTACGTGCTCGACACCTCT
GTGCTGCTGTCCGATCCGTGGGCGTGCAGCCGGTTCGCCGAACACGATGTGGTGGTTCCGTT
GGTGGTGATCAGCGAGCTAGAAGCCAAGCGCCACCACCACGAGCTGGGATGGTTCGCCCGCC
AGGCGTTGCGTCTGTTCGACGATCTGCGCCTAGAACACGGGCGGTTGGATCAGCCGATTCCG
GTTGGCACCCAAGGCGGTACGCTGCACGTCGAACTCAATCACACCGACCCGGCGGTGCTGCC
CGCAGGCTTTCGCACCGACAGCAACGACTCGAGGATCTTGAGTTGCGCCGCCAACCTCGCCG
CCGAGGGCAAGCGGGTCACGTTGGTCAGCAAGGACATTCCGCTGCGCGTTAAGGCCGCCGCG
GTGGGGCTGGCCGCCGACGAGTACCACGCGCAGGACGTCGTTGTGTCCGGATGGTCGGGGAT
GCACGAGCTCGAGACCGCTTCCGCGGATATCGATGCGTTGTTCGCCGATGGCGAGATCGACC
TGGTCGAAGCCCGGGACCTACCGTGTCACACCGGGATTCGGTTGCTGGGCGGCGGTTCCCAC
GCGCTGGGCCGGGTCAATGCGCATAAACGTGTTCAGCTGGTGCGAGGTGACCGTGAGGCGTT
CGGTCTGCGTGGCCGCTCCGCCGAGCAGCGGGTGGCGCTGGATTTGCTGCTCGATGAGTCGG
TGGGCATCGTGTCGCTGGGCGGCAAAGCCGGCACGGGCAAGTCCGCTTTGGCGTTGTGTGCG
GGTCTGGAAGCCGTGCTGGAGCGACGCACCCACCGCAAGGTGGTGGTCTTCCGCCCGCTGTA
CGCGGTCGGCGGCCAGGAGCTGGGCTACCTGCCCGGTAGCGAGAGCGAGAAGATGGGCCCGT
GGGCGCAGGCGGTCTTCGACACCCTCGAGGGGCTGGCCAGCCCGGCGGTGCTCGAGGAAGTG
CTGTCCCGTGGCATGCTCGAGGTGCTGCCGCTGACCCACATCCGGGGCCGCTCGTTGCATGA
CTCGTTCGTCATCGTCGACGAGGCACAGTCGCTGGAGCGCAATGTGTTGCTGACCGTGCTGT
CCCGGTTGGGGACCGGTTCCCGGGTGGTGTTGACCCACGACATCGCCCAGCGCGACAACCTG
CGGGTCGGCCGCCACGACGGGGTCGCCGCGGTGATCGAGAAGCTCAAAGGTCATCCGTTGTT
CGCCCACATCACCTTGCTGCGCAGTGAGCGCTCGCCGATCGCCGCGCTGGTCACCGAGATGC
TCGAGGAGATCACCGGGCCGCGCTGA

Rv1095alt pYUB1049 N-terminal fusion tag (506 aa)

MW 54216.6 Da pI 6.33

MGSSHHHHHHSSGLVPRGSHMASDMLCCQGGTFRHDGCHDKGRTGPGPGVAAPADMLGWVRS
SAVSSRSAP*VTDTRTYVLDTSVLLSDPWACSRFAEHDVVVPLVVISELEAKRHHHELGWFA
RQALRLFDDLRLEHGRLDQPIPVGTQGGTLHVELNHTDPAVLPAGFRTDSNDSRILSCAANL
AAEGKRVTLVSKDIPLRVKAAAVGLAADEYHAQDVVVSGWSGMHELETASADIDALFADGEI
DLVEARDLPCHTGIRLLGGGSHALGRVNAHKRVQLVRGDREAFGLRGRSAEQRVALDLLLDE
SVGIVSLGGKAGTGKSALALCAGLEAVLERRTHRKVVVFRPLYAVGGQELGYLPGSESEKMG
PWAQAVFDTLEGLASPAVLEEVLSRGMLEVLPLTHIRGRSLHDSFVIVDEAQSLERNVLLTV
LSRLGTGSRVVLTHDIAQRDNLRVGRHDGVAAVIEKLKGHPLFAHITLLRSERSPIAALVTE
MLEEITGPR*

Rv1095alt pDEST$_{SMG}$ N-terminal fusion tag (1520 bp)

ATGTCGCATCACCATCACCATCACCTCGAATCACCATCAACAAGTTTGTACAAAAAAGCAGG
CTTCATGGCTAGCGACATGCTCTGCTGCCAGGGCGGCACCTTCCGTCACGACGGCTGTCATG
ACAAGGGCAGGACCGGCCCCGGTCCTGGTGTCGCTGCCCCCGCCGACATGCTCGGGTGGGTC
CGCTCGAGCGCCGTTAGCTCGAGGAGCGCTCCGTGACCGATACCCGCACGTACGTGCTCGAC
ACCTCTGTGCTGCTGTCCGATCCGTGGGCGTGCAGCCGGTTCGCCGAACACGATGTGGTGGT

TCCGTTGGTGGTGATCAGCGAGCTAGAAGCCAAGCGCCACCACCACGAGCTGGGATGGTTCG
CCCGCCAGGCGTTGCGTCTGTTCGACGATCTGCGCCTAGAACACGGGCGGTTGGATCAGCCG
ATTCCGGTTGGCACCCAAGGCGGTACGCTGCACGTCGAACTCAATCACACCGACCCGGCGGT
GCTGCCCGCAGGCTTTCGCACCGACAGCAACGACTCGAGGATCTTGAGTTGCGCCGCCAACC
TCGCCGCCGAGGGCAAGCGGGTCACGTTGGTCAGCAAGGACATTCCGCTGCGCGTTAAGGCC
GCCGCGGTGGGGCTGGCCGCCGACGAGTACCACGCGCAGGACGTCGTTGTGTCCGGATGGTC
GGGGATGCACGAGCTCGAGACCGCTTCCGCGGATATCGATGCGTTGTTCGCCGATGGCGAGA
TCGACCTGGTCGAAGCCCGGGACCTACCGTGTCACACCGGGATTCGGTTGCTGGGCGGCGGT
TCCCACGCGCTGGGCCGGGTCAATGCGCATAAACGTGTTCAGCTGGTGCGAGGTGACCGTGA
GGCGTTCGGTCTGCGTGGCCGCTCCGCCGAGCAGCGGGTGGCGCTGGATTTGCTGCTCGATG
AGTCGGTGGGCATCGTGTCGCTGGGCGGCAAAGCCGGCACGGGCAAGTCCGCTTTGGCGTTG
TGTGCGGGTCTGGAAGCCGTGCTGGAGCGACGCACCCACCGCAAGGTGGTGGTCTTCCGCCC
GCTGTACGCGGTCGGCGGCCAGGAGCTGGGCTACCTGCCCGGTAGCGAGAGCGAGAAGATGG
GCCCGTGGGCGCAGGCGGTCTTCGACACCCTCGAGGGGCTGGCCAGCCCGGCGGTGCTCGAG
GAAGTGCTGTCCCGTGGCATGCTCGAGGTGCTGCCGCTGACCCACATCCGGGGCCGCTCGTT
GCATGACTCGTTCGTCATCGTCGACGAGGCACAGTCGCTGGAGCGCAATGTGTTGCTGACCG
TGCTGTCCCGGTTGGGGACCGGTTCCCGGGTGGTGTTGACCCACGACATCGCCCAGCGCGAC
AACCTGCGGGTCGGCCGCCACGACGGGGTCGCCGCGGTGATCGAGAAGCTCAAAGGTCATCC
GTTGTTCGCCCACATCACCTTGCTGCGCAGTGAGCGCTCGCCGATCGCCGCGCTGGTCACCG
AGATGCTCGAGGAGATCACCGGGCCGCGCTGA

Rv1095alt pDEST$_{SMG}$ N-terminal fusion tag (506 aa)
MW 54604.2 pI 6.31
MSHHHHHHLESPSTSLYKKAGFMASDMLCCQGGTFRHDGCHDKGRTGPGPGVAAPADMLGWV
RSSAVSSRSAP*VTDTRTYVLDTSVLLSDPWACSRFAEHDVVVPLVVISELEAKRHHHELGW
FARQALRLFDDLRLEHGRLDQPIPVGTQGGTLHVELNHTDPAVLPAGFRTDSNDSRILSCAA
NLAAEGKRVTLVSKDIPLRVKAAAVGLAADEYHAQDVVVSGWSGMHELETASADIDALFADG
EIDLVEARDLPCHTGIRLLGGGSHALGRVNAHKRVQLVRGDREAFGLRGRSAEQRVALDLLL
DESVGIVSLGGKAGTGKSALALCAGLEAVLERRTHRKVVVFRPLYAVGGQELGYLPGSESEK
MGPWAQAVFDTLEGLASPAVLEEVLSRGMLEVLPLTHIRGRSLHDSFVIVDEAQSLERNVLL
TVLSRLGTGSRVVLTHDIAQRDNLRVGRHDGVAAVIEKLKGHPLFAHITLLRSERSPIAALV
TEMLEEITGPR*

Rv1095alt alt (165 bp)
ATGGCTAGCGACATGCTCTGCTGCCAGGGCGGCACCTTCCGTCACGACGGCTGTCATGACAA
GGGCAGGACCGGCCCCGGTCCTGGTGTCGCTGCCCCCGCCGACATGCTCGGGTGGGTCCGCT
CGAGCGCCGTTAGCTCGAGGAGCGCTCCGTGA

Rv1095alt alt (52 aa)
MW 5189.8 pI 7.78
MASDMLCCQGGTFRHDGCHDKGRTGPGPGVAAPADMLGWVRSSAVSSRSAP*****

Rv1095alt alt pYUB28b C-terminal fusion tag (95 bp)
ATGGCTAGCGACATGCTCTGCTGCCAGGGCGGCACCTTCCGTCACGACGGCTGTCATGACAA
GGGCAGGACCGGCCCCGGTCCTGGTGTCGCTGCCCCCGCCGACATGCTCGGGTGGGTCCGCT
CGAGCGCCGTTAGCTCGAGGAGCGCTCCGAAGCTTGCGGCCGCACTCGAGCACCACCACCAC
CACCACTGA

Rv1095alt alt pYUB28b C-terminal fusion tag (65 aa)
MW 6709.5 Da pI 7.87
MASDMLCCQGGTFRHDGCHDKGRTGPGPGVAAPADMLGWVRSSAVSSRSAPKLAAALEHHHH
HH*

Rv1095 PIN-domain (396 bp)

GTGACCGATACCCGCACGTACGTGCTCGACACCTCTGTGCTGCTGTCCGATCCGTGGGCGTG
CAGCCGGTTCGCCGAACACGATGTGGTGGTTCCGTTGGTGGTGATCAGCGAGCTAGAAGCCA
AGCGCCACCACCACGAGCTGGGATGGTTCGCCCGCCAGGCGTTGCGTCTGTTCGACGATCTG
CGCCTAGAACACGGGCGGTTGGATCAGCCGATTCCGGTTGGCACCCAAGGCGGTACGCTGCA
CGTCGAACTCAATCACACCGACCCGGCGGTGCTGCCCGCAGGCTTTCGCACCGACAGCAACG
ACTCGAGGATCTTGAGTTGCGCCGCCAACCTCGCCGCCGAGGGCAAGCGGGTCACGTTGGTC
AGCAAGGACATTCCGCTGCGCGTT

Rv1095 PIN-domain (132 aa)

MW 14652.6 Da pI 5.94

VTDTRTYVLDTSVLLSDPWACSRFAEHDVVVPLVVISELEAKRHHHELGWFARQALRLFDDL
RLEHGRLDQPIPVGTQGGTLHVELNHTDPAVLPAGFRTDSNDSRILSCAANLAAEGKRVTLV
SKDIPLRV

Rv1095 PIN-domain pDEST$_{SMG}$ N-terminal fusion tag (465 bp)

ATGTCGCATCACCATCACCATCACCTCGAATCACCATCAACAAGTTTGTACAAAAAAGCAGG
CTTCGTGACCGATACCCGCACGTACGTGCTCGACACCTCTGTGCTGCTGTCCGATCCGTGGG
CGTGCAGCCGGTTCGCCGAACACGATGTGGTGGTTCCGTTGGTGGTGATCAGCGAGCTAGAA
GCCAAGCGCCACCACCACGAGCTGGGATGGTTCGCCCGCCAGGCGTTGCGTCTGTTCGACGA
TCTGCGCCTAGAACACGGGCGGTTGGATCAGCCGATTCCGGTTGGCACCCAAGGCGGTACGC
TGCACGTCGAACTCAATCACACCGACCCGGCGGTGCTGCCCGCAGGCTTTCGCACCGACAGC
AACGACTCGAGGATCTTGAGTTGCGCCGCCAACCTCGCCGCCGAGGGCAAGCGGGTCACGTT
GGTCAGCAAGGACATTCCGCTGCGCGTTTAG

Rv1095 PIN-domain pDEST$_{SMG}$ N-terminal fusion tag (154 aa)

MW 17203.4 Da pI 6.48

MSHHHHHHLESPSTSLYKKAGFVTDTRTYVLDTSVLLSDPWACSRFAEHDVVVPLVVISELE
AKRHHHELGWFARQALRLFDDLRLEHGRLDQPIPVGTQGGTLHVELNHTDPAVLPAGFRTDS
NDSRILSCAANLAAEGKRVTLVSKDIPLRV*

Rv1095 PIN-domain pDEST$_{17}$ N-terminal fusion tag (465 bp)

ATGTCGTACTACCATCACCATCACCATCACCTCGAATCAACAAGTTTGTACAAAAAAGCAGG
CTTCGTGACCGATACCCGCACGTACGTGCTCGACACCTCTGTGCTGCTGTCCGATCCGTGGG
CGTGCAGCCGGTTCGCCGAACACGATGTGGTGGTTCCGTTGGTGGTGATCAGCGAGCTAGAA
GCCAAGCGCCACCACCACGAGCTGGGATGGTTCGCCCGCCAGGCGTTGCGTCTGTTCGACGA
TCTGCGCCTAGAACACGGGCGGTTGGATCAGCCGATTCCGGTTGGCACCCAAGGCGGTACGC
TGCACGTCGAACTCAATCACACCGACCCGGCGGTGCTGCCCGCAGGCTTTCGCACCGACAGC
AACGACTCGAGGATCTTGAGTTGCGCCGCCAACCTCGCCGCCGAGGGCAAGCGGGTCACGTT
GGTCAGCAAGGACATTCCGCTGCGCGTTTAG

Rv1095 PIN-domain pDEST$_{17}$ N-terminal fusion tag (155 aa)

MW 17345.6 Da pI 6.48

MSYYHHHHHHLESTSLYKKAGFVTDTRTYVLDTSVLLSDPWACSRFAEHDVVVPLVVISELE
AKRHHHELGWFARQALRLFDDLRLEHGRLDQPIPVGTQGGTLHVELNHTDPAVLPAGFRTDS
NDSRILSCAANLAAEGKRVTLVSKDIPLRV*

Rv1095 PIN-domain pYUB28b C-terminal fusion tag (438 bp)

GTGACCGATACCCGCACGTACGTGCTCGACACCTCTGTGCTGCTGTCCGATCCGTGGGCGTG
CAGCCGGTTCGCCGAACACGATGTGGTGGTTCCGTTGGTGGTGATCAGCGAGCTAGAAGCCA
AGCGCCACCACCACGAGCTGGGATGGTTCGCCCGCCAGGCGTTGCGTCTGTTCGACGATCTG
CGCCTAGAACACGGGCGGTTGGATCAGCCGATTCCGGTTGGCACCCAAGGCGGTACGCTGCA
CGTCGAACTCAATCACACCGACCCGGCGGTGCTGCCCGCAGGCTTTCGCACCGACAGCAACG

ACTCGAGGATCTTGAGTTGCGCCGCCAACCTCGCCGCCGAGGGCAAGCGGGTCACGTTGGTC
AGCAAGGACATTCCGCTGCGCGTTAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCA
CTGA

Rv1095 PIN-domain pYUB28b C-terminal fusion tag (146 aa)
MW 16303.5 Da pI 6.33
MVTDTRTYVLDTSVLLSDPWACSRFAEHDVVVPLVVISELEAKRHHHELGWFARQALRLFDD
LRLEHGRLDQPIPVGTQGGTLHVELNHTDPAVLPAGFRTDSNDSRILSCAANLAAEGKRVTL
VSKDIPLRVKLAAALEHHHHHH*

Rv1095 PhoH-domain (654 bp)
GAGGCGTTCGGTCTGCGTGGCCGCTCCGCCGAGCAGCGGGTGGCGCTGGATTTGCTGCTCGA
TGAGTCGGTGGGCATCGTGTCGCTGGGCGGCAAAGCCGGCACGGGCAAGTCCGCTTTGGCGT
TGTGTGCGGGTCTGGAAGCCGTGCTGGAGCGACGCACCCACCGCAAGGTGGTGGTCTTCCGC
CCGCTGTACGCGGTCGGCGGCCAGGAGCTGGGCTACCTGCCCGGTAGCGAGAGCGAGAAGAT
GGGCCCGTGGGCGCAGGCGGTCTTCGACACCCTCGAGGGGCTGGCCAGCCCGGCGGTGCTCG
AGGAAGTGCTGTCCCGTGGCATGCTCGAGGTGCTGCCGCTGACCCACATCCGGGGCCGCTCG
TTGCATGACTCGTTCGTCATCGTCGACGAGGCACAGTCGCTGGAGCGCAATGTGTTGCTGAC
CGTGCTGTCCCGGTTGGGGACCGGTTCCCGGGTGGTGTTGACCCACGACATCGCCCAGCGCG
ACAACCTGCGGGTCGGCCGCCACGACGGGGTCGCCGCGGTGATCGAGAAGCTCAAAGGTCAT
CCGTTGTTCGCCCACATCACCTTGCTGCGCAGTGAGCGCTCGCCGATCGCCGCGCTGGTCAC
CGAGATGCTCGAGGAGATCACCGGGCCGCGCTGA

Rv1095 PhoH-domain (218 aa)
MW 23407.0 Da pI 6.43
EAFGLRGRSAEQRVALDLLLDESVGIVSLGGKAGTGKSALALCAGLEAVLERRTHRKVVVFR
PLYAVGGQELGYLPGSESEKMGPWAQAVFDTLEGLASPAVLEEVLSRGMLEVLPLTHIRGRS
LHDSFVIVDEAQSLERNVLLTVLSRLGTGSRVVLTHDIAQRDNLRVGRHDGVAAVIEKLKGH
PLFAHITLLRSERSPIAALVTEMLEEITGPR*

Rv1095 PhoH-domain pDEST$_{SMG}$ N-terminal fusion tag (720 bp)
ATGTCGCATCACCATCACCATCACCTCGAATCACCATCAACAAGTTTGTACAAAAAAGCAGG
CTTCGAGGCGTTCGGTCTGCGTGGCCGCTCCGCCGAGCAGCGGGTGGCGCTGGATTTGCTGC
TCGATGAGTCGGTGGGCATCGTGTCGCTGGGCGGCAAAGCCGGCACGGGCAAGTCCGCTTTG
GCGTTGTGTGCGGGTCTGGAAGCCGTGCTGGAGCGACGCACCCACCGCAAGGTGGTGGTCTT
CCGCCCGCTGTACGCGGTCGGCGGCCAGGAGCTGGGCTACCTGCCCGGTAGCGAGAGCGAGA
AGATGGGCCCGTGGGCGCAGGCGGTCTTCGACACCCTCGAGGGGCTGGCCAGCCCGGCGGTG
CTCGAGGAAGTGCTGTCCCGTGGCATGCTCGAGGTGCTGCCGCTGACCCACATCCGGGGCCG
CTCGTTGCATGACTCGTTCGTCATCGTCGACGAGGCACAGTCGCTGGAGCGCAATGTGTTGC
TGACCGTGCTGTCCCGGTTGGGGACCGGTTCCCGGGTGGTGTTGACCCACGACATCGCCCAG
CGCGACAACCTGCGGGTCGGCCGCCACGACGGGGTCGCCGCGGTGATCGAGAAGCTCAAAGG
TCATCCGTTGTTCGCCCACATCACCTTGCTGCGCAGTGAGCGCTCGCCGATCGCCGCGCTGG
TCACCGAGATGCTCGAGGAGATCACCGGGCCGCGCTGA

Rv1095 PhoH-domain pDEST$_{SMG}$ N-terminal fusion tag (240 aa)
MW 25957.8 Da pI 6.9
MSHHHHHHLESPSTSLYKKAGFEAFGLRGRSAEQRVALDLLLDESVGIVSLGGKAGTGKSAL
ALCAGLEAVLERRTHRKVVVFRPLYAVGGQELGYLPGSESEKMGPWAQAVFDTLEGLASPAV
LEEVLSRGMLEVLPLTHIRGRSLHDSFVIVDEAQSLERNVLLTVLSRLGTGSRVVLTHDIAQ
RDNLRVGRHDGVAAVIEKLKGHPLFAHITLLRSERSPIAALVTEMLEEITGPR*

Rv1095 PhoH-domain pDEST$_{17}$ N-terminal fusion tag (720 bp)
ATGTCGTACTACCATCACCATCACCATCACCTCGAATCAACAAGTTTGTACAAAAAAGCAGG
CTTCGAGGCGTTCGGTCTGCGTGGCCGCTCCGCCGAGCAGCGGGTGGCGCTGGATTTGCTGC
TCGATGAGTCGGTGGGCATCGTGTCGCTGGGCGGCAAAGCCGGCACGGGCAAGTCCGCTTTG

GCGTTGTGTGCGGGTCTGGAAGCCGTGCTGGAGCGACGCACCCACCGCAAGGTGGTGGTCTT
CCGCCCGCTGTACGCGGTCGGCGGCCAGGAGCTGGGCTACCTGCCCGGTAGCGAGAGCGAGA
AGATGGGCCCGTGGGCGCAGGCGGTCTTCGACACCCTCGAGGGGCTGGCCAGCCCGGCGGTG
CTCGAGGAAGTGCTGTCCCGTGGCATGCTCGAGGTGCTGCCGCTGACCCACATCCGGGGCCG
CTCGTTGCATGACTCGTTCGTCATCGTCGACGAGGCACAGTCGCTGGAGCGCAATGTGTTGC
TGACCGTGCTGTCCCGGTTGGGGACCGGTTCCCGGGTGGTGTTGACCCACGACATCGCCCAG
CGCGACAACCTGCGGGTCGGCCGCCACGACGGGGTCGCCGCGGTGATCGAGAAGCTCAAAGG
TCATCCGTTGTTCGCCCACATCACCTTGCTGCGCAGTGAGCGCTCGCCGATCGCCGCGCTGG
TCACCGAGATGCTCGAGGAGATCACCGGGCCGCGCTGA

Rv1095 PhoH-domain pDEST$_{17}$ N-terminal fusion tag (240 aa)
MW 26100.0 Da pI 6.9
MSYYHHHHHHLESTSLYKKAGFEAFGLRGRSAEQRVALDLLLDESVGIVSLGGKAGTGKSAL
ALCAGLEAVLERRTHRKVVVFRPLYAVGGQELGYLPGSESEKMGPWAQAVFDTLEGLASPAV
LEEVLSRGMLEVLPLTHIRGRSLHDSFVIVDEAQSLERNVLLTVLSRLGTGSRVVLTHDIAQ
RDNLRVGRHDGVAAVIEKLKGHPLFAHITLLRSERSPIAALVTEMLEEITGPR*

Rv1095 PhoH-domain pYUB28b C-terminal fusion tag (696 bp)
ATGGAGGCGTTCGGTCTGCGTGGCCGCTCCGCCGAGCAGCGGGTGGCGCTGGATTTGCTGCT
CGATGAGTCGGTGGGCATCGTGTCGCTGGGCGGCAAAGCCGGCACGGGCAAGTCCGCTTTGG
CGTTGTGTGCGGGTCTGGAAGCCGTGCTGGAGCGACGCACCCACCGCAAGGTGGTGGTCTTC
CGCCCGCTGTACGCGGTCGGCGGCCAGGAGCTGGGCTACCTGCCCGGTAGCGAGAGCGAGAA
GATGGGCCCGTGGGCGCAGGCGGTCTTCGACACCCTCGAGGGGCTGGCCAGCCCGGCGGTGC
TCGAGGAAGTGCTGTCCCGTGGCATGCTCGAGGTGCTGCCGCTGACCCACATCCGGGGCCGC
TCGTTGCATGACTCGTTCGTCATCGTCGACGAGGCACAGTCGCTGGAGCGCAATGTGTTGCT
GACCGTGCTGTCCCGGTTGGGGACCGGTTCCCGGGTGGTGTTGACCCACGACATCGCCCAGC
GCGACAACCTGCGGGTCGGCCGCCACGACGGGGTCGCCGCGGTGATCGAGAAGCTCAAAGGT
CATCCGTTGTTCGCCCACATCACCTTGCTGCGCAGTGAGCGCTCGCCGATCGCCGCGCTGGT
CACCGAGATGCTCGAGGAGATCACCGGGCCGCGCAAGCTTGCGGCCGCACTCGAGCACCACC
ACCACCACCACTGA

Rv1095 PhoH-domain pYUB28b C-terminal fusion tag (232 aa)
MW 25057.9 Da pI 6.67
MEAFGLRGRSAEQRVALDLLLDESVGIVSLGGKAGTGKSALALCAGLEAVLERRTHRKVVVF
RPLYAVGGQELGYLPGSESEKMGPWAQAVFDTLEGLASPAVLEEVLSRGMLEVLPLTHIRGR
SLHDSFVIVDEAQSLERNVLLTVLSRLGTGSRVVLTHDIAQRDNLRVGRHDGVAAVIEKLKG
HPLFAHITLLRSERSPIAALVTEMLEEITGPRKLAAALEHHHHHH*

Rv1095 PIN-domainalt (548 bp)
ATGGCTAGCGACATGCTCTGCTGCCAGGGCGGCACCTTCCGTCACGACGGCTGTCATGACAA
GGGCAGGACCGGCCCCGGTCCTGGTGTCGCTGCCCCCGCCGACATGCTCGGGTGGGTCCGCT
CGAGCGCCGTTAGCTCGAGGAGCGCTCCGTGACCGATACCCGCACGTACGTGCTCGACACCT
CTGTGCTGCTGTCCGATCCGTGGGCGTGCAGCCGGTTCGCCGAACACGATGTGGTGGTTCCG
TTGGTGGTGATCAGCGAGCTAGAAGCCAAGCGCCACCACCACGAGCTGGGATGGTTCGCCCG
CCAGGCGTTGCGTCTGTTCGACGATCTGCGCCTAGAACACGGGCGGTTGGATCAGCCGATTC
CGGTTGGCACCCAAGGCGGTACGCTGCACGTCGAACTCAATCACACCGACCCGGCGGTGCTG
CCCGCAGGCTTTCGCACCGACAGCAACGACTCGAGGATCTTGAGTTGCGCCGCCAACCTCGC
CGCCGAGGGCAAGCGGGTCACGTTGGTCAGCAAGGACATTCCGCTGCGCGTT

Rv1095 PIN-domain$_{alt}$ (183 aa)
MW 19824.4 Da pI 6.28
MASDMLCCQGGTFRHDGCHDKGRTGPGPGVAAPADMLGWVRSSAVSSRSAP*VTDTRTYVLD
TSVLLSDPWACSRFAEHDVVVPLVVISELEAKRHHHELGWFARQALRLFDDLRLEHGRLDQP
IPVGTQGGTLHVELNHTDPAVLPAGFRTDSNDSRILSCAANLAAEGKRVTLVSKDIPLRV

Rv1095 PIN-domain$_{alt}$ pYUB28b C-terminal fusion tag (594 bp)

```
ATGGCTAGCGACATGCTCTGCTGCCAGGGCGGCACCTTCCGTCACGACGGCTGTCATGACAA
GGGCAGGACCGGCCCCGGTCCTGGTGTCGCTGCCCCCGCCGACATGCTCGGGTGGGTCCGCT
CGAGCGCCGTTAGCTCGAGGAGCGCTCCGTGAGTGACCGATACCCGCACGTACGTGCTCGAC
ACCTCTGTGCTGCTGTCCGATCCGTGGGCGTGCAGCCGGTTCGCCGAACACGATGTGGTGGT
TCCGTTGGTGGTGATCAGCGAGCTAGAAGCCAAGCGCCACCACCACGAGCTGGGATGGTTCG
CCCGCCAGGCGTTGCGTCTGTTCGACGATCTGCGCCTAGAACACGGGCGGTTGGATCAGCCG
ATTCCGGTTGGCACCCAAGGCGGTACGCTGCACGTCGAACTCAATCACACCGACCCGGCGGT
GCTGCCCGCAGGCTTTCGCACCGACAGCAACGACTCGAGGATCTTGAGTTGCGCCGCCAACC
TCGCCGCCGAGGGCAAGCGGGTCACGTTGGTCAGCAAGGACATTCCGCTGCGCGTTAAGCTT
GCGGCCGCACTCGAGCACCACCACCACCACCACTGA
```

Rv1095 PIN-domain$_{alt}$ pYUB28b C-terminal fusion tag (198 aa)

MW 21344.1 Da pI 6.55

```
MASDMLCCQGGTFRHDGCHDKGRTGPGPGVAAPADMLGWVRSSAVSSRSAP*VTDTRTYVLD
TSVLLSDPWACSRFAEHDVVVPLVVISELEAKRHHHELGWFARQALRLFDDLRLEHGRLDQP
IPVGTQGGTLHVELNHTDPAVLPAGFRTDSNDSRILSCAANLAAEGKRVTLVSKDIPLRVKL
AAALEHHHHHH*
```

MSMEG_5247 (1314 bp)

```
GTGACTGAGCAAGCTGTCCGTACCTATGTGCTCGACACCTCGGTGTTGCTGTCAGATCCCTG
GGCATGCACCCGGTTCGCCGAGCACGAGGTGGTGGTCCCGCTGGTCGTCATCAGTGAGTTGG
AGGCCAAACGGCACCATCACGAACTCGGCTGGTTCGCGCGGCAGGCCTTGCGGATGTTCGAC
GACATGCGACTCGAACACGGCCGCCTGGATCAACCGGTCCCGGTCGGAACACAGGGCGGCAC
ACTGCACGTCGAGCTGAACCACAGCGACCCGTCGGTGCTGCCCGCCGGGTTCCGCAACGACA
GCAACGACGCCAGGATCCTCACGGTCGCGGCCAATCTCGCCGCCGAGGGCAAGCACGTCACG
TTGGTGAGCAAGGACATCCCGCTGCGCGTCAAGGCCGGTGCGGTGGGCCTGGCGGCCGACGA
GTACCACGCACAGGACGTCGTGGTGTCCGGGTGGACGGGCATGACCGAGATGGACGTCGCAG
GCGAGGACATCGACACGCTGTTCGCCGACGGTGAGATCGACCTGGCCGAGGCGCGGGATCTG
CCGTGCCACACAGGCATTCGGTTGCTCGGTGGCACCTCGCACGCGCTGGGACGGGTGAACGC
GGCCAAGAAGGTGCAGCTGGTCCGTGGTGATCGCGAAGTGTTCGGCCTCCGGGGAAGGTCAG
CCGAACAGCGCGTCGCCCTCGACCTGCTGCTCGACGAGTCCGTCGGCATCGTGTCACTCGGC
GGCAAGGCCGGCACCGGCAAGTCGGCGCTCGCGCTGTGCGCGGGCCTGGAGGCCGTGCTGGA
GCGCCGCACGCAGCGCAAGGTCGTGGTGTTCCGCCCGCTGTATGCCGTGGGCGGTCAGGATC
TCGGCTACCTGCCCGGTAGCGAGAGCGAGAAGATGGGCCCGTGGGCGCAGGCGGTGTTCGAC
ACCCTCGAAGGGCTCGCGAGCCCCGCGGTGCTCGACGAGGTGCTCTCCCGCGGCATGCTCGA
AGTGCTGCCCCTGACCCACATCCGGGGCCGCTCACTGCACGACTCGTTCGTGATCGTCGACG
AGGCGCAGTCGCTCGAACGCAACGTGCTGCTGACGGTGCTGTCCCGGCTGGGTGCCGGGTCC
CGCGTGGTGCTCACCCACGACGTCGCCCAGCGCGACAACCTGCGGGTGGGACGCCATGACGG
CGTGGCCGCGGTGATCGAGAAGCTCAAGGGGCACCCGCTGTTCGCCCACGTGACGTTGCAGC
GCAGCGAACGCTCGCCGATCGCCGCGCTGGTCACCGAGATGCTCGAGGAGATCAGCCCCGGC
GCCCTGCCCTGA
```

MSMEG_5247 (438 aa)

MW 47137.8 Da pI 5.66

```
VTEQAVRTYVLDTSVLLSDPWACTRFAEHEVVVPLVVISELEAKRHHHELGWFARQALRMFD
DMRLEHGRLDQPVPVGTQGGTLHVELNHSDPSVLPAGFRNDSNDARILTVAANLAAEGKHVT
LVSKDIPLRVKAGAVGLAADEYHAQDVVVSGWTGMTEMDVAGEDIDTLFADGEIDLAEARDL
PCHTGIRLLGGTSHALGRVNAAKKVQLVRGDREVFGLRGRSAEQRVALDLLLDESVGIVSLG
GKAGTGKSALALCAGLEAVLERRTQRKVVVFRPLYAVGGQDLGYLPGSESEKMGPWAQAVFD
TLEGLASPAVLDEVLSRGMLEVLPLTHIRGRSLHDSFVIVDEAQSLERNVLLTVLSRLGAGS
RVVLTHDVAQRDNLRVGRHDGVAAVIEKLKGHPLFAHVTLQRSERSPIAALVTEMLEEISPG
ALP*
```

MSMEG_5247 pYUB28b C-terminal fusion tag (1353 bp)

GTGACTGAGCAAGCTGTCCGTACCTATGTGCTCGACACCTCGGTGTTGCTGTCAGATCCCTG
GGCATGCACCCGGTTCGCCGAGCACGAGGTGGTGGTCCCGCTGGTCGTCATCAGTGAGTTGG
AGGCCAAACGGCACCATCACGAACTCGGCTGGTTCGCGCGGCAGGCCTTGCGGATGTTCGAC
GACATGCGACTCGAACACGGCCGCCTGGATCAACCGGTCCCGGTCGGAACACAGGGCGGCAC
ACTGCACGTCGAGCTGAACCACAGCGACCCGTCGGTGCTGCCCGCCGGGTTCCGCAACGACA
GCAACGACGCCAGGATCCTCACGGTCGCGGCCAATCTCGCCGCCGAGGGCAAGCACGTCACG
TTGGTGAGCAAGGACATCCCGCTGCGCGTCAAGGCCGGTGCGGTGGGCCTGGCGGCCGACGA
GTACCACGCACAGGACGTCGTGGTGTCCGGGTGGACGGGCATGACCGAGATGGACGTCGCAG
GCGAGGACATCGACACGCTGTTCGCCGACGGTGAGATCGACCTGGCCGAGGCGCGGGATCTG
CCGTGCCACACAGGCATTCGGTTGCTCGGTGGCACCTCGCACGCGCTGGGACGGGTGAACGC
GGCCAAGAAGGTGCAGCTGGTCCGTGGTGATCGCGAAGTGTTCGGCCTCCGGGGAAGGTCAG
CCGAACAGCGCGTCGCCCTCGACCTGCTGCTCGACGAGTCCGTCGGCATCGTGTCACTCGGC
GGCAAGGCCGGCACCGGCAAGTCGGCGCTCGCGCTGTGCGCGGGCCTGGAGGCCGTGCTGGA
GCGCCGCACGCAGCGCAAGGTCGTGGTGTTCCGCCCGCTGTATGCCGTGGGCGGTCAGGATC
TCGGCTACCTGCCCGGTAGCGAGAGCGAGAAGATGGGCCCGTGGGCGCAGGCGGTGTTCGAC
ACCCTCGAAGGGCTCGCGAGCCCCGCGGTGCTCGACGAGGTGCTCTCCCGCGGCATGCTCGA
AGTGCTGCCCCTGACCCACATCCGGGGCCGCTCACTGCACGACTCGTTCGTGATCGTCGACG
AGGCGCAGTCGCTCGAACGCAACGTGCTGCTGACGGTGCTGTCCCGGCTGGGTGCCGGGTCC
CGCGTGGTGCTCACCCACGACGTCGCCCAGCGCGACAACCTGCGGGTGGGACGCCATGACGG
CGTGGCCGCGGTGATCGAGAAGCTCAAGGGGCACCCGCTGTTCGCCCACGTGACGTTGCAGC
GCAGCGAACGCTCGCCGATCGCCGCGCTGGTCACCGAGATGCTCGAGGAGATCAGCCCCGGC
GCCCTGCCCAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGA

MSMEG_5247 pYUB28b C-terminal fusion tag (451 aa)
MW 48657.5 Da pI 5.9
VTEQAVRTYVLDTSVLLSDPWACTRFAEHEVVVPLVVISELEAKRHHHELGWFARQALRMFD
DMRLEHGRLDQPVPVGTQGGTLHVELNHSDPSVLPAGFRNDSNDARILTVAANLAAEGKHVT
LVSKDIPLRVKAGAVGLAADEYHAQDVVVSGWTGMTEMDVAGEDIDTLFADGEIDLAEARDL
PCHTGIRLLGGTSHALGRVNAAKKVQLVRGDREVFGLRGRSAEQRVALDLLLDESVGIVSLG
GKAGTGKSALALCAGLEAVLERRTQRKVVVFRPLYAVGGQDLGYLPGSESEKMGPWAQAVFD
TLEGLASPAVLDEVLSRGMLEVLPLTHIRGRSLHDSFVIVDEAQSLERNVLLTVLSRLGAGS
RVVLTHDVAQRDNLRVGRHDGVAAVIEKLKGHPLFAHVTLQRSERSPIAALVTEMLEEISPG
ALPKLAAALEHHHHHH*

MSMEG_5247alt (1453 bp)
ATGGCTAGCGACCTGCTCTGCTGTCCGGGCGGTACCGATCGTTTCGATCACGAAAGGACCGG
CCCCGGTCCTGCCGCAGTGTCCACCGAGCAATTCGTGTGGGGCCGCATGAAGAGCGCCGTTC
CCCTAGGAGCGCCACGTGACTGAGCAAGCTGTCCGTACCTATGTGCTCGACACCTCGGTGTT
GCTGTCAGATCCCTGGGCATGCACCCGGTTCGCCGAGCACGAGGTGGTGGTCCCGCTGGTCG
TCATCAGTGAGTTGGAGGCCAAACGGCACCATCACGAACTCGGCTGGTTCGCGCGGCAGGCC
TTGCGGATGTTCGACGACATGCGACTCGAACACGGCCGCCTGGATCAACCGGTCCCGGTCGG
AACACAGGGCGGCACACTGCACGTCGAGCTGAACCACAGCGACCCGTCGGTGCTGCCCGCCG
GGTTCCGCAACGACAGCAACGACGCCAGGATCCTCACGGTCGCGGCCAATCTCGCCGCCGAG
GGCAAGCACGTCACGTTGGTGAGCAAGGACATCCCGCTGCGCGTCAAGGCCGGTGCGGTGGG
CCTGGCGGCCGACGAGTACCACGCACAGGACGTCGTGGTGTCCGGGTGGACGGGCATGACCG
AGATGGACGTCGCAGGCGAGGACATCGACACGCTGTTCGCCGACGGTGAGATCGACCTGGCC
GAGGCGCGGGATCTGCCGTGCCACACAGGCATTCGGTTGCTCGGTGGCACCTCGCACGCGCT
GGGACGGGTGAACGCGGCCAAGAAGGTGCAGCTGGTCCGTGGTGATCGCGAAGTGTTCGGCC
TCCGGGGAAGGTCAGCCGAACAGCGCGTCGCCCTCGACCTGCTGCTCGACGAGTCCGTCGGC
ATCGTGTCACTCGGCGGCAAGGCCGGCACCGGCAAGTCGGCGCTCGCGCTGTGCGCGGGCCT
GGAGGCCGTGCTGGAGCGCCGCACGCAGCGCAAGGTCGTGGTGTTCCGCCCGCTGTATGCCG
TGGGCGGTCAGGATCTCGGCTACCTGCCCGGTAGCGAGAGCGAGAAGATGGGCCCGTGGGCG
CAGGCGGTGTTCGACACCCTCGAAGGGCTCGCGAGCCCCGCGGTGCTCGACGAGGTGCTCTC
CCGCGGCATGCTCGAAGTGCTGCCCCTGACCCACATCCGGGGCCGCTCACTGCACGACTCGT
TCGTGATCGTCGACGAGGCGCAGTCGCTCGAACGCAACGTGCTGCTGACGGTGCTGTCCCGG
CTGGGTGCCGGGTCCCGCGTGGTGCTCACCCACGACGTCGCCCAGCGCGACAACCTGCGGGT
GGGACGCCATGACGGCGTGGCCGCGGTGATCGAGAAGCTCAAGGGGCACCCGCTGTTCGCCC

ACGTGACGTTGCAGCGCAGCGAACGCTCGCCGATCGCCGCGCTGGTCACCGAGATGCTCGAG
GAGATCAGCCCCGGCGCCCTGCCCTGA

MSMEG_5247alt (450 aa)
MW 52208.5 Da pI 5.65
MASDLLCCPGGTDRFDHERTGPGPAAVSTEQFVWGRMKSAVPLGAPRD*VTEQAVRTYVLDT
SVLLSDPWACTRFAEHEVVVPLVVISELEAKRHHHELGWFARQALRMFDDMRLEHGRLDQPV
PVGTQGGTLHVELNHSDPSVLPAGFRNDSNDARILTVAANLAAEGKHVTLVSKDIPLRVKAG
AVGLAADEYHAQDVVVSGWTGMTEMDVAGEDIDTLFADGEIDLAEARDLPCHTGIRLLGGTS
HALGRVNAAKKVQLVRGDREVFGLRGRSAEQRVALDLLLDESVGIVSLGGKAGTGKSALALC
AGLEAVLERRTQRKVVVFRPLYAVGGQDLGYLPGSESEKMGPWAQAVFDTLEGLASPAVLDE
VLSRGMLEVLPLTHIRGRSLHDSFVIVDEAQSLERNVLLTVLSRLGAGSRVVLTHDVAQRDN
LRVGRHDGVAAVIEKLKGHPLFAHVTLQRSERSPIAALVTEMLEEISPGALP*

MSMEG_5247alt pYUB28b C-terminal fusion tag (1492 bp)
ATGGCTAGCGACCTGCTCTGCTGTCCGGGCGGTACCGATCGTTTCGATCACGAAAGGACCGG
CCCCGGTCCTGCCGCAGTGTCCACCGAGCAATTCGTGTGGGGCCGCATGAAGAGCGCCGTTC
CCCTAGGAGCGCCACGTGACTGAGCAAGCTGTCCGTACCTATGTGCTCGACACCTCGGTGTT
GCTGTCAGATCCCTGGGCATGCACCCGGTTCGCCGAGCACGAGGTGGTGGTCCCGCTGGTCG
TCATCAGTGAGTTGGAGGCCAAACGGCACCATCACGAACTCGGCTGGTTCGCGCGGCAGGCC
TTGCGGATGTTCGACGACATGCGACTCGAACACGGCCGCCTGGATCAACCGGTCCCGGTCGG
AACACAGGGCGGCACACTGCACGTCGAGCTGAACCACAGCGACCCGTCGGTGCTGCCCGCCG
GGTTCCGCAACGACAGCAACGACGCCAGGATCCTCACGGTCGCGGCCAATCTCGCCGCCGAG
GGCAAGCACGTCACGTTGGTGAGCAAGGACATCCCGCTGCGCGTCAAGGCCGGTGCGGTGGG
CCTGGCGGCCGACGAGTACCACGCACAGGACGTCGTGGTGTCCGGGTGGACGGGCATGACCG
AGATGGACGTCGCAGGCGAGGACATCGACACGCTGTTCGCCGACGGTGAGATCGACCTGGCC
GAGGCGCGGGATCTGCCGTGCCACACAGGCATTCGGTTGCTCGGTGGCACCTCGCACGCGCT
GGGACGGGTGAACGCGGCCAAGAAGGTGCAGCTGGTCCGTGGTGATCGCGAAGTGTTCGGCC
TCCGGGGAAGGTCAGCCGAACAGCGCGTCGCCCTCGACCTGCTGCTCGACGAGTCCGTCGGC
ATCGTGTCACTCGGCGGCAAGGCCGGCACCGGCAAGTCGGCGCTCGCGCTGTGCGCGGGCCT
GGAGGCCGTGCTGGAGCGCCGCACGCAGCGCAAGGTCGTGGTGTTCCGCCCGCTGTATGCCG
TGGGCGGTCAGGATCTCGGCTACCTGCCCGGTAGCGAGAGCGAGAAGATGGGCCCGTGGGCG
CAGGCGGTGTTCGACACCCTCGAAGGGCTCGCGAGCCCCGCGGTGCTCGACGAGGTGCTCTC
CCGCGGCATGCTCGAAGTGCTGCCCCTGACCCACATCCGGGGCCGCTCACTGCACGACTCGT
TCGTGATCGTCGACGAGGCGCAGTCGCTCGAACGCAACGTGCTGCTGACGGTGCTGTCCCGG
CTGGGTGCCGGGTCCCGCGTGGTGCTCACCCACGACGTCGCCCAGCGCGACAACCTGCGGGT
GGGACGCCATGACGGCGTGGCCGCGGTGATCGAGAAGCTCAAGGGGCACCCGCTGTTCGCCC
ACGTGACGTTGCAGCGCAGCGAACGCTCGCCGATCGCCGCGCTGGTCACCGAGATGCTCGAG
GAGATCAGCCCCGGCGCCCTGCCCAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCA
CTGA

MSMEG_5247alt pYUB28b C-terminal fusion tag (498 aa)
MW 53728.2 Da pI 5.88
MASDLLCCPGGTDRFDHERTGPGPAAVSTEQFVWGRMKSAVPLGAPRD*VTEQAVRTYVLDT
SVLLSDPWACTRFAEHEVVVPLVVISELEAKRHHHELGWFARQALRMFDDMRLEHGRLDQPV
PVGTQGGTLHVELNHSDPSVLPAGFRNDSNDARILTVAANLAAEGKHVTLVSKDIPLRVKAG
AVGLAADEYHAQDVVVSGWTGMTEMDVAGEDIDTLFADGEIDLAEARDLPCHTGIRLLGGTS
HALGRVNAAKKVQLVRGDREVFGLRGRSAEQRVALDLLLDESVGIVSLGGKAGTGKSALALC
AGLEAVLERRTQRKVVVFRPLYAVGGQDLGYLPGSESEKMGPWAQAVFDTLEGLASPAVLDE
VLSRGMLEVLPLTHIRGRSLHDSFVIVDEAQSLERNVLLTVLSRLGAGSRVVLTHDVAQRDN
LRVGRHDGVAAVIEKLKGHPLFAHVTLQRSERSPIAALVTEMLEEISPGALPKLAAALEHHH
HHH*

MSMEG_5247 PIN-domain (402 bp)

GTGACTGAGCAAGCTGTCCGTACCTATGTGCTCGACACCTCGGTGTTGCTGTCAGATCCCTG
GGCATGCACCCGGTTCGCCGAGCACGAGGTGGTGGTCCCGCTGGTCGTCATCAGTGAGTTGG
AGGCCAAACGGCACCATCACGAACTCGGCTGGTTCGCGCGGCAGGCCTTGCGGATGTTCGAC
GACATGCGACTCGAACACGGCCGCCTGGATCAACCGGTCCCGGTCGGAACACAGGGCGGCAC
ACTGCACGTCGAGCTGAACCACAGCGACCCGTCGGTGCTGCCCGCCGGGTTCCGCAACGACA
GCAACGACGCCAGGATCCTCACGGTCGCGGCCAATCTCGCCGCCGAGGGCAAGCACGTCACG
TTGGTGAGCAAGGACATCCCGCTGCGCGTC

MSMEG_5247 PIN-domain (134 aa)
MW 14903.9 Da pI 5.86
VTEQAVRTYVLDTSVLLSDPWACTRFAEHEVVVPLVVISELEAKRHHHELGWFARQALRMFD
DMRLEHGRLDQPVPVGTQGGTLHVELNHSDPSVLPAGFRNDSNDARILTVAANLAAEGKHVT
LVSKDIPLRV

MSMEG_5247 PIN-domain pYUB28b C-terminal fusion tag (447 bp)
ATGGTGACTGAGCAAGCTGTCCGTACCTATGTGCTCGACACCTCGGTGTTGCTGTCAGATCC
CTGGGCATGCACCCGGTTCGCCGAGCACGAGGTGGTGGTCCCGCTGGTCGTCATCAGTGAGT
TGGAGGCCAAACGGCACCATCACGAACTCGGCTGGTTCGCGCGGCAGGCCTTGCGGATGTTC
GACGACATGCGACTCGAACACGGCCGCCTGGATCAACCGGTCCCGGTCGGAACACAGGGCGG
CACACTGCACGTCGAGCTGAACCACAGCGACCCGTCGGTGCTGCCCGCCGGGTTCCGCAACG
ACAGCAACGACGCCAGGATCCTCACGGTCGCGGCCAATCTCGCCGCCGAGGGCAAGCACGTC
ACGTTGGTGAGCAAGGACATCCCGCTGCGCGTCAAGCTTGCGGCCGCACTCGAGCACCACCA
CCACCACCACTGA

MSMEG_5247 PIN-domain pYUB28b C-terminal fusion tag (149 aa)
MW 16554.8 Da pI 6.24
MVTEQAVRTYVLDTSVLLSDPWACTRFAEHEVVVPLVVISELEAKRHHHELGWFARQALRMF
DDMRLEHGRLDQPVPVGTQGGTLHVELNHSDPSVLPAGFRNDSNDARILTVAANLAAEGKHV
TLVSKDIPLRVKLAAALEHHHHHH*

MSMEG_5247 PhoH-domain +1 (660 bp)
GAAGTGTTCGGCCTCCGGGGAAGGTCAGCCGAACAGCGCGTCGCCCTCGACCTGCTGCTCGA
CGAGTCCGTCGGCATCGTGTCACTCGGCGGCAAGGCCGGCACCGGCAAGTCGGCGCTCGCGC
TGTGCGCGGGCCTGGAGGCCGTGCTGGAGCGCCGCACGCAGCGCAAGGTCGTGGTGTTCCGC
CCGCTGTATGCCGTGGGCGGTCAGGATCTCGGCTACCTGCCCGGTAGCGAGAGCGAGAAGAT
GGGCCCGTGGGCGCAGGCGGTGTTCGACACCCTCGAAGGGCTCGCGAGCCCCGCGGTGCTCG
ACGAGGTGCTCTCCCGCGGCATGCTCGAAGTGCTGCCCCTGACCCACATCCGGGGCCGCTCA
CTGCACGACTCGTTCGTGATCGTCGACGAGGCGCAGTCGCTCGAACGCAACGTGCTGCTGAC
GGTGCTGTCCCGGCTGGGTGCCGGGTCCCGCGTGGTGCTCACCCACGACGTCGCCCAGCGCG
ACAACCTGCGGGTGGGACGCCATGACGGCGTGGCCGCGGTGATCGAGAAGCTCAAGGGGCAC
CCGCTGTTCGCCCACGTGACGTTGCAGCGCAGCGAACGCTCGCCGATCGCCGCGCTGGTCAC
CGAGATGCTCGAGGAGATCAGCCCCGGCGCCCTGCCCTGA

MSMEG_5247 PhoH-domain +1 (219 aa)
MW 23466.0 DapI 6.11
EVFGLRGRSAEQRVALDLLLDESVGIVSLGGKAGTGKSALALCAGLEAVLERRTQRKVVVFR
PLYAVGGQDLGYLPGSESEKMGPWAQAVFDTLEGLASPAVLDEVLSRGMLEVLPLTHIRGRS
LHDSFVIVDEAQSLERNVLLTVLSRLGAGSRVVLTHDVAQRDNLRVGRHDGVAAVIEKLKGH
PLFAHVTLQRSERSPIAALVTEMLEEISPGALP*

MSMEG_5247 PhoH-domain +1 pYUB28b C-terminal fusion tag (702 bp)
ATGGAAGTGTTCGGCCTCCGGGGAAGGTCAGCCGAACAGCGCGTCGCCCTCGACCTGCTGCT
CGACGAGTCCGTCGGCATCGTGTCACTCGGCGGCAAGGCCGGCACCGGCAAGTCGGCGCTCG
CGCTGTGCGCGGGCCTGGAGGCCGTGCTGGAGCGCCGCACGCAGCGCAAGGTCGTGGTGTTC

CGCCCGCTGTATGCCGTGGGCGGTCAGGATCTCGGCTACCTGCCCGGTAGCGAGAGCGAGAA
GATGGGCCCGTGGGCGCAGGCGGTGTTCGACACCCTCGAAGGGCTCGCGAGCCCCGCGGTGC
TCGACGAGGTGCTCTCCCGCGGCATGCTCGAAGTGCTGCCCCTGACCCACATCCGGGGCCGC
TCACTGCACGACTCGTTCGTGATCGTCGACGAGGCGCAGTCGCTCGAACGCAACGTGCTGCT
GACGGTGCTGTCCCGGCTGGGTGCCGGGTCCCGCGTGGTGCTCACCCACGACGTCGCCCAGC
GCGACAACCTGCGGGTGGGACGCCATGACGGCGTGGCCGCGGTGATCGAGAAGCTCAAGGGG
CACCCGCTGTTCGCCCACGTGACGTTGCAGCGCAGCGAACGCTCGCCGATCGCCGCGCTGGT
CACCGAGATGCTCGAGGAGATCAGCCCCGGCGCCCTGCCCAAGCTTGCGGCCGCACTCGAGC
ACCACCACCACCACCACTGA

MSMEG_5247 PhoH-domain +1 pYUB28b C-terminal fusion tag (233 aa)
MW 25116.9 Da pI 6.46
MEVFGLRGRSAEQRVALDLLLDESVGIVSLGGKAGTGKSALALCAGLEAVLERRTQRKVVVF
RPLYAVGGQDLGYLPGSESEKMGPWAQAVFDTLEGLASPAVLDEVLSRGMLEVLPLTHIRGR
SLHDSFVIVDEAQSLERNVLLTVLSRLGAGSRVVLTHDVAQRDNLRVGRHDGVAAVIEKLKG
HPLFAHVTLQRSERSPIAALVTEMLEEISPGALPKLAAALEHHHHHH*
MSMEG_5247 PhoH-domain +2 (666 bp)
GATCGCGAAGTGTTCGGCCTCCGGGGAAGGTCAGCCGAACAGCGCGTCGCCCTCGACCTGCT
GCTCGACGAGTCCGTCGGCATCGTGTCACTCGGCGGCAAGGCCGGCACCGGCAAGTCGGCGC
TCGCGCTGTGCGCGGGCCTGGAGGCCGTGCTGGAGCGCCGCACGCAGCGCAAGGTCGTGGTG
TTCCGCCCGCTGTATGCCGTGGGCGGTCAGGATCTCGGCTACCTGCCCGGTAGCGAGAGCGA
GAAGATGGGCCCGTGGGCGCAGGCGGTGTTCGACACCCTCGAAGGGCTCGCGAGCCCCGCGG
TGCTCGACGAGGTGCTCTCCCGCGGCATGCTCGAAGTGCTGCCCCTGACCCACATCCGGGGC
CGCTCACTGCACGACTCGTTCGTGATCGTCGACGAGGCGCAGTCGCTCGAACGCAACGTGCT
GCTGACGGTGCTGTCCCGGCTGGGTGCCGGGTCCCGCGTGGTGCTCACCCACGACGTCGCCC
AGCGCGACAACCTGCGGGTGGGACGCCATGACGGCGTGGCCGCGGTGATCGAGAAGCTCAAG
GGGCACCCGCTGTTCGCCCACGTGACGTTGCAGCGCAGCGAACGCTCGCCGATCGCCGCGCT
GGTCACCGAGATGCTCGAGGAGATCAGCCCCGGCGCCCTGCCCTGA

MSMEG_5247 PhoH-domain +2 (221 aa)
MW 23737.3 Da pI 6.1
DREVFGLRGRSAEQRVALDLLLDESVGIVSLGGKAGTGKSALALCAGLEAVLERRTQRKVVV
FRPLYAVGGQDLGYLPGSESEKMGPWAQAVFDTLEGLASPAVLDEVLSRGMLEVLPLTHIRG
RSLHDSFVIVDEAQSLERNVLLTVLSRLGAGSRVVLTHDVAQRDNLRVGRHDGVAAVIEKLK
GHPLFAHVTLQRSERSPIAALVTEMLEEISPGALP*

MSMEG_5247 PhoH-domain +2 pYUB28b C-terminal fusion tag (708 bp)
ATGGATCGCGAAGTGTTCGGCCTCCGGGGAAGGTCAGCCGAACAGCGCGTCGCCCTCGACCT
GCTGCTCGACGAGTCCGTCGGCATCGTGTCACTCGGCGGCAAGGCCGGCACCGGCAAGTCGG
CGCTCGCGCTGTGCGCGGGCCTGGAGGCCGTGCTGGAGCGCCGCACGCAGCGCAAGGTCGTG
GTGTTCCGCCCGCTGTATGCCGTGGGCGGTCAGGATCTCGGCTACCTGCCCGGTAGCGAGAG
CGAGAAGATGGGCCCGTGGGCGCAGGCGGTGTTCGACACCCTCGAAGGGCTCGCGAGCCCCG
CGGTGCTCGACGAGGTGCTCTCCCGCGGCATGCTCGAAGTGCTGCCCCTGACCCACATCCGG
GGCCGCTCACTGCACGACTCGTTCGTGATCGTCGACGAGGCGCAGTCGCTCGAACGCAACGT
GCTGCTGACGGTGCTGTCCCGGCTGGGTGCCGGGTCCCGCGTGGTGCTCACCCACGACGTCG
CCCAGCGCGACAACCTGCGGGTGGGACGCCATGACGGCGTGGCCGCGGTGATCGAGAAGCTC
AAGGGGCACCCGCTGTTCGCCCACGTGACGTTGCAGCGCAGCGAACGCTCGCCGATCGCCGC
GCTGGTCACCGAGATGCTCGAGGAGATCAGCCCCGGCGCCCTGCCCAAGCTTGCGGCCGCAC
TCGAGCACCACCACCACCACCACTGA

MSMEG_5247 PhoH-domain +2 pYUB28b C-terminal fusion tag (235 bp)
MW 25388.2 Da pI 6.46
MDREVFGLRGRSAEQRVALDLLLDESVGIVSLGGKAGTGKSALALCAGLEAVLERRTQRKVV
VFRPLYAVGGQDLGYLPGSESEKMGPWAQAVFDTLEGLASPAVLDEVLSRGMLEVLPLTHIR
GRSLHDSFVIVDEAQSLERNVLLTVLSRLGAGSRVVLTHDVAQRDNLRVGRHDGVAAVIEKL
KGHPLFAHVTLQRSERSPIAALVTEMLEEISPGALPKLAAALEHHHHHH*

TBIS_3092 optimised for *E. coli* protein expression (1323 bp)

CTGCAGATGGTTGCAGTTAGCAGCAGCAATCCGACCCAGACCCGTACCTATGTTCTGGATAC
CAGCGTTCTGCTGGCAGATCCGGCAAGCATGAGCCGTTTTGCAGAACATGAAGTTGTTATTC
CGATTGTGGTGATCAATGAACTGGAAGCAAAACGTCATCATCCGGAACTGGGTTATTTTGCA
CGTGAAGCACTGCGTTTTCTGGATGATCTGCGTGTTCGTCATGGTCGTCTGGATCAGCCGGT
TCCGATTGGTGAAGGCACCATTCGTGTTGAACTGAATCATAGCGATCCGGCAGTTCTGCCTG
CCGGTCTGCGTAGCGGTGATAATGATAGCCGTATTCTGACCGTTGCACAGAATCTGGCAGCC
GAAGGTCGTGATGTTGTTCTGGTTAGCAAAGATCTGCCGATGCGTCTGAAAGCAGCAAGCCT
GGGTCTGAATGCCGAAGAATATCGTGCAGGTATGGTTATTGAAAGCGGTTGGACCGGTATGG
CCGAACTGCAAGTTACCGATGATGACCTGCGTATGCTGTTTGAACATGGCACCATTGAACTG
GCCGAAGCACGTGATCTGCCGTGTCATACAGGTCTGCGTCTGCTGAGCACCCGTGGTAGCGC
ACTGGGTCGTGTTACACCGGATAAAAGCGTTCGTCTGGTTCGTGGTGATCGTGAAGTTTTTG
GTCTGCGTGGTCGTAGCGCAGAACAGCGTATTGCACTGGATCTGCTGATGGATCCGGAAATT
GGTATTGTTAGCATTGGTGGTCGTGCAGGCACCGGTAAAAGTGCACTGGCACTGTGTGCAGG
TCTGGAAGCAGTTCTGGAACGTCGTCAGCATCGTAAAATCATTGTTTTTCGTCCGCTGTATG
CAGTTGGTGGCCAAGAACTGGGCTATCTGCCTGGCACCGAAAATGATAAAATGGGTCCGTGG
GCACAGGCAGTTTATGATACCCTGGGTGCAGTTACCAGTCCGGAAGTTATTGAAGAAGTGCT
GGATCGCGGTATGCTGGAAGTTCTGCCGCTGACCCATATTCGTGGTCGTTCACTGCATGATG
CATTTGTTATTGTTGATGAAGCACAGAGCCTGGAACGTGGTGTTCTGCTGACCGTTCTGAGC
CGTATTGGTAGCAATAGCCGTGTGGTTCTGACCCATGATGTTGCACAGCGTGATAATCTGCG
CGTTGGTCGTCATGATGGTGTTGTTGCAGTTGTTGAAAAACTGAAAGGTCATCCGCTGTTTG
CACATATTACCCTGACCCGTAGCGAACGTAGCCCGATTGCAGCACTGGTTACCGAAATGCTG
CAAGATATTACCATCAAGCTT

TBIS_3092 optimised for *E. coli* protein expression (441 aa)

MW 48248.4 pI 6

LQMVAVSSSNPTQTRTYVLDTSVLLADPASMSRFAEHEVVIPIVVINELEAKRHHPELGYFA
REALRFLDDLRVRHGRLDQPVPIGEGTIRVELNHSDPAVLPAGLRSGDNDSRILTVAQNLAA
EGRDVVLVSKDLPMRLKAASLGLNAEEYRAGMVIESGWTGMAELQVTDDDLRMLFEHGTIEL
AEARDLPCHTGLRLLSTRGSALGRVTPDKSVRLVRGDREVFGLRGRSAEQRIALDLLMDPEI
GIVSIGGRAGTGKSALALCAGLEAVLERRQHRKIIVFRPLYAVGGQELGYLPGTENDKMGPW
AQAVYDTLGAVTSPEVIEEVLDRGMLEVLPLTHIRGRSLHDAFVIVDEAQSLERGVLLTVLS
RIGSNSRVVLTHDVAQRDNLRVGRHDGVVAVVEKLKGHPLFAHITLTRSERSPIAALVTEML
QDITIKL

TBIS_3092 pET28b-*Pst*I C-terminal fusion tag (1353 bp)

ATGGTTGCAGTTAGCAGCAGCAATCCGACCCAGACCCGTACCTATGTTCTGGATACCAGCGT
TCTGCTGGCAGATCCGGCAAGCATGAGCCGTTTTGCAGAACATGAAGTTGTTATTCCGATTG
TGGTGATCAATGAACTGGAAGCAAAACGTCATCATCCGGAACTGGGTTATTTTGCACGTGAA
GCACTGCGTTTTCTGGATGATCTGCGTGTTCGTCATGGTCGTCTGGATCAGCCGGTTCCGAT
TGGTGAAGGCACCATTCGTGTTGAACTGAATCATAGCGATCCGGCAGTTCTGCCTGCCGGTC
TGCGTAGCGGTGATAATGATAGCCGTATTCTGACCGTTGCACAGAATCTGGCAGCCGAAGGT
CGTGATGTTGTTCTGGTTAGCAAAGATCTGCCGATGCGTCTGAAAGCAGCAAGCCTGGGTCT
GAATGCCGAAGAATATCGTGCAGGTATGGTTATTGAAAGCGGTTGGACCGGTATGGCCGAAC
TGCAAGTTACCGATGATGACCTGCGTATGCTGTTTGAACATGGCACCATTGAACTGGCCGAA
GCACGTGATCTGCCGTGTCATACAGGTCTGCGTCTGCTGAGCACCCGTGGTAGCGCACTGGG
TCGTGTTACACCGGATAAAAGCGTTCGTCTGGTTCGTGGTGATCGTGAAGTTTTTGGTCTGC
GTGGTCGTAGCGCAGAACAGCGTATTGCACTGGATCTGCTGATGGATCCGGAAATTGGTATT
GTTAGCATTGGTGGTCGTGCAGGCACCGGTAAAAGTGCACTGGCACTGTGTGCAGGTCTGGA
AGCAGTTCTGGAACGTCGTCAGCATCGTAAAATCATTGTTTTTCGTCCGCTGTATGCAGTTG
GTGGCCAAGAACTGGGCTATCTGCCTGGCACCGAAAATGATAAAATGGGTCCGTGGGCACAG
GCAGTTTATGATACCCTGGGTGCAGTTACCAGTCCGGAAGTTATTGAAGAAGTGCTGGATCG
CGGTATGCTGGAAGTTCTGCCGCTGACCCATATTCGTGGTCGTTCACTGCATGATGCATTTG
TTATTGTTGATGAAGCACAGAGCCTGGAACGTGGTGTTCTGCTGACCGTTCTGAGCCGTATT
GGTAGCAATAGCCGTGTGGTTCTGACCCATGATGTTGCACAGCGTGATAATCTGCGCGTTGG
TCGTCATGATGGTGTTGTTGCAGTTGTTGAAAAACTGAAAGGTCATCCGCTGTTTGCACATA

TTACCCTGACCCGTAGCGAACGTAGCCCGATTGCAGCACTGGTTACCGAAATGCTGCAAGAT
ATTACCATCAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGA

TBIS_3092 pET28b-*Pst*I C-terminal fusion tag (451 aa)
MW 49285.5 Da pI 6.13
MVAVSSSNPTQTRTYVLDTSVLLADPASMSRFAEHEVVIPIVVINELEAKRHHPELGYFARE
ALRFLDDLRVRHGRLDQPVPIGEGTIRVELNHSDPAVLPAGLRSGDNDSRILTVAQNLAAEG
RDVVLVSKDLPMRLKAASLGLNAEEYRAGMVIESGWTGMAELQVTDDDLRMLFEHGTIELAE
ARDLPCHTGLRLLSTRGSALGRVTPDKSVRLVRGDREVFGLRGRSAEQRIALDLLMDPEIGI
VSIGGRAGTGKSALALCAGLEAVLERRQHRKIIVFRPLYAVGGQELGYLPGTENDKMGPWAQ
AVYDTLGAVTSPEVIEEVLDRGMLEVLPLTHIRGRSLHDAFVIVDEAQSLERGVLLTVLSRI
GSNSRVVLTHDVAQRDNLRVGRHDGVVAVVEKLKGHPLFAHITLTRSERSPIAALVTEMLQD
ITIKLAAALEHHHHHH*