Do PhoH2 proteins regulate SigF in mycobacteria?

A thesis

submitted in partial fulfilment

of the requirements for the degree

of

Masters of Research in Biological Sciences

at

The University of Waikato

by

Elizabeth Rzoska-Smith

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2018

Abstract

There is little known about the role of the PhoH2 proteins in biology. PhoH2 consists of two domains, an N-terminal PIN domain with RNase activity and a C-terminal PhoH domain shown to have ATP induced RNA helicase activity. PhoH2 proteins from *Mycobacterium tuberculosis* and *Mycobacterium smegmatis* show ATP and magnesium dependent, sequence-specific RNA unwinding and cleavage activity. It is hypothesised based on RNAseq data that PhoH2 from *M. smegmatis* may be acting as a potential regulator of sigma factor SigF, through its targeted RNA helicase and RNase activity. Regulation may occur directly on the *sigF* mRNA transcript, or indirectly through regulation of anti-sigma factor antagonists. This thesis used purified PhoH2 proteins to examine the activity of PhoH2 against mRNA substrates designed to span regions of the *sigF* gene and the genes co-expressed in the *sigF* operon; *rsbW* and *chaB*. Results show that PhoH2 co-purifies with *E. coli* nucleic acid and that despite this co-purification, activity was observed on *sigF* mRNA, that was improved with increasing amounts of ATP. qPCR experiments were designed to test expression levels of 12 genes, that showed significant up or down regulation expression in *M. smegmatis* mc²155 *ΔphoH2* compared to wild-type *M. smegmatis* mc²155. qPCR analysis revealed a significant upregulation of the genes MSMEG_1804 (*sigF*) and MSMEG_0064 (PPE) in *M. smegmatis* mc²155 *ΔphoH2* compared to *M. smegmatis* mc²155. These results differ from the gene expression results obtained from RNAseq data, yet results still point towards regulation of SigF by PhoH2. Further experiments using qPCR and activity assays with PhoH2 aim to verify the precise role of PhoH2 in the regulation of SigF.

Acknowledgements

There are many people who I would like to give thanks to for all their support, encouragement and knowledge, throughout the duration of my project. Firstly I would like to say thank you to my supervisor; Professor Vic Arcus, for letting me be part of the C.2.10 family, I have really enjoyed my experience in this lab and I really appreciate your guidance and enthusiasm towards my project.

Secondly I would like to thank my two co-supervisors; Dr Joanna Hicks and Dr Emma Andrews. Jo, thank you for all your support and patience when I first came into the lab, I really appropriate all the help and advice you have offered me over the course of my research. Emma, you have provided an enormous amount of energy towards my thesis, I really appreciate your wisdom and continuous encouragement over my research, throughout my time in the lab.

To everyone in the C.2.10 lab, past and present, thank you all for making my time in this lab so enjoyable and a special thanks to Judith, or lab mum, for not only all your yummy baking, but also your kindness and support you have given me from day one. I would also like to thank Dr Greg Jacobson and Dr Steve Bird, from C.2.03, for the knowledge you both provided on qPCR set up and analysis, I really appreciate the time you took to guide me through this part of my research.

And a huge thanks to my very supportive family; mum, dad and Maddy. Mum, thanks for encouraging me to never give up and strive to reach my goals, no matter how daunting they may seem. Dad thanks for always showing an interest towards my research and always being there when I needed help.

To all my friends inside and outside of science, thanks for always been there for talks and coffee and a special thanks to Zoe and Mel for putting up with my continuous rants about proteins and RNA.

And finally Andy, thank you for always being there for me, even if you are six hours away. I really appreciate everything you have done for me this year, it would have been pretty tough without you.

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
AAA	ATPases associated with diverse cellular activities
AAA+	ATPases associated with diverse cellular activities (variation)
AEX	anion exchange chromatography
APS	ammonium persulfate
ATP	adenosine triphosphate
bp	base pair(s)
BLAST	basic local alignment search tool
BS-YlaK	*Bacillus subtilis*-YlaK
cDNA	complementary deoxyribose nucleic acids
Cq	quantification cycle
C-terminal	carboxy terminus
DEPC	diethyl pyruvate carbonate
DNA	deoxyribose nucleic acids
DNase	deoxyribonuclease
dNTP	deoxynucleotide triphosphate
ds	double-stranded
DTT	dithiothreitol
EDTA	ethylene diamine tetraacetic acid (disodium salt)
gDNA	genomic deoxyribose nucleic acids
GITC	guanidium isothiocyanate
GOI	genes of interest
His-tag	poly-histidine tag
HIV	human immunodeficiency virus
HMM	hidden markov model
HPLC	high-performance liquid chromatography
IMAC	immobilised metal affinity chromatography
IPA	isopropanol
IPTG	isopropylthio-β-D-galactosidase
IR	inverted repeat
kb	kilobase
kDa	kilodalton

LB	luria bertani
ko	knock-out
mAU	milli-absorbance units
MDR	multidrug resistant
mg	milligram (10-3 g)
Mic	magnetic induction cycler
MIQE	Minimum Information for Publication of Quantitative Real-Time PCR Experiments
ml	millilitre (10-3 L)
mM	millimolar (10-3 M)
mRNA	messenger RNA
MSMEG	*Mycobacterium smegmatis*
MTB	*Mycobacterium tuberculosis*
MQ	milli Q water – ion exchanged purified water
MW	molecular weight
NCBI	National Center for Biotechnology Information
N-terminal	amino terminus
OD	optical density
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
PDB	protein data bank
pI	isoelectric point
Pi	phosphate
PIN	PilT N-terminal domain
P-loop	Phosphate loop
µg	microgram (10-6 g)
µL	microliter (10-6 L)
µm	micrometer (10-6 m)
µM	micromolar (10-6 M)
rpm	revolutions per minute
RNA	ribonucleic acid
RNAP	RNA polymerase
RNA-Seq	RNA sequencing
RNase	ribonuclease
rNTP	ribonucleotide
rRNA	ribosomal RNA
RT	reverse transcriptase/transcription
RT-qPCR	reverse transcription quantitative real-time polymerase chain reaction

qPCR	quantitative real-time polymerase chain reaction
SDS	sodium dodecyl sulphate
SEC	size exclusion chromatography
ss	single stranded
TA	toxin-antitoxin
TAE	tris-acetate-EDTA
TB	tuberculosis
TE	tris EDTA buffer
TEMED	N, N, N, N,-tetramethylethylenediamine
Tm	melting temperature
Tris	tris-hydroxymethylaminomethane
tRNA	transfer ribonucleic acid
UV	ultraviolet
V	Volts
v/v	volume per volume
WHO	World Health Organisation
wt	wild- type
w/v	weight per volume
w/w	weight per weight
XDR	Extensively Drug Resistance

1 Chapter 1: Introduction

1.1 The persistence of *Mycobacterium tuberculosis*

Mycobacterium tuberculosis is a pathogenic bacilli belonging to the family Mycobacteriaceae, identified as the cause of tuberculosis (TB) in humans, by Robert Koch in 1882 [1, 2, 5]. Tuberculosis is an infectious disease having been prevalent for thousands of years and is one of the top 10 causes of morbidity and mortality around the world [1]. In 2016 there were 1.7 million TB deaths, including 0.4 million TB deaths among people with human immunodeficiency virus (HIV) [2, 5]. The World Health Organisation (WHO) declared a rapid increase in the number of TB infections in both the developed and developing word. This increase has been associated with multidrug resistant (MDR) TB, HIV and global migration [1].

While most TB infections present in the lungs as pulmonary TB, there has also been instances where the disease has been observed in peripheral areas of the body including; bone, the central nervous system and many other organs, this is known as extra-pulmonary TB [2, 84]. Spread of tuberculosis generally occurs when individuals with active pulmonary TB expel the bacteria into the air, usually through coughing, allowing the bacteria to enter new human hosts through the inhalation of infected aerosol droplets [2, 3, 4]. The pathogenic bacilli, contained within these droplets, are no bigger than 2μm and this lets them enter into the alveoli of the respiratory tract in the human host [3]. Alveolar macrophages ingest the *M. tuberculosis* bacilli and enclose them in phagosomes. If macrophages are activated, the mycobacterial-containing phagosomes join with lysosomes, and the bacteria are killed. However if alveolar macrophages are not activated, the bacilli survive and persist within phagosomes by altering their intercellular compartments and forming lesions [7].

The pathogenic nature of *M. tuberculosis* is complex and involves many immune-evading strategies that the bacterium uses to establish primary infection, persistence in the host, and reactivation of pathogenicity under favourable conditions [8, 9]. Primary infection occurs when *M. tuberculosis* overwhelms the host's immune system and enters an active state [6]. This active state allows the

bacteria to rapidly reproduce in the host and eventually spread to other hosts. *M. tuberculosis* can also avoid the host's immune system and enter a dormant state resulting in an asymptomatic and non-infectious host. *M. tuberculosis* is capable of remaining undetected for years, until favourable conditions arise, that allow reactivation of the bacteria and re-emergence of infection [5] (Figure 1.1).

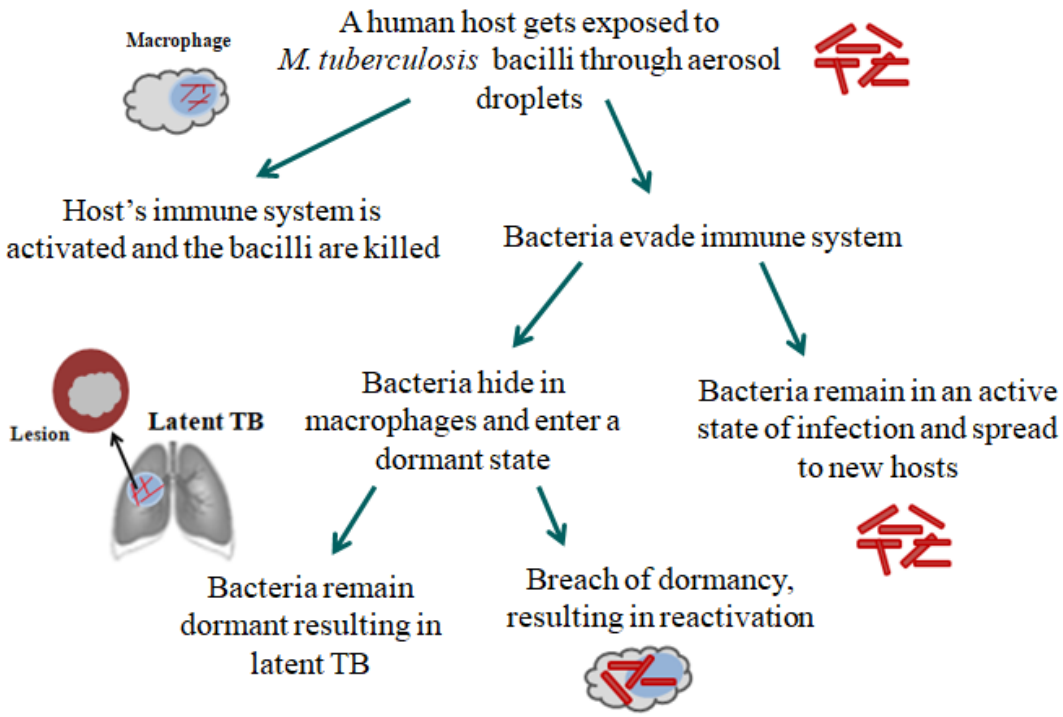

Figure 1. 1: Flow diagram of the progression of tuberculosis disease once *M. tuberculosis* has entered a human host [5, 85].

Generally, only a small number of people infected with *M. tuberculosis* go on to develop active TB disease during their lifetime. Those who do often have HIV, or have a compromised immune system, caused by factors such as; diabetes, smoking, low nutrient diet and alcohol consumption. The WHO states that about one-quarter of the world's population has latent TB [2]. A person with latent TB has been infected with the bacteria, but does not show any symptoms and cannot transmit the disease [3]. This ability of TB bacteria to survive in this latent state, allows them to remain undetectable to the host's immune system, until an opportunity, such as a compromised immune system, arises [3].

In relation to research and development of TB therapies, very few diagnostic technologies were introduced in 2017. There were 17 drugs in phase one, two or three trials, as well as 12 vaccine candidates in clinical trials [120]. New diagnostics, drugs and vaccines are required to achieve the ambitious targets set in

the End TB Strategy that are a; 95% reduction by 2035 in the number of TB deaths compared against 2015, a 90% reduction rate by 2035 in TB incidence compared against 2015; and no major costs to TB affected families by 2035 [2]. To develop new antitubercular agents, it is essential to study the genetics and physiology of *M. tuberculosis* and related mycobacteria to better understand the biology of these organisms. Specific genes and the proteins they encode, involved in bacterial persistence, can provide a suite of novel bacterial targets for vaccine and drug development.

1.2 Toxin-antitoxin systems

Within the human host *M. tuberculosis* must survive in a number of stressful environments, including under hypoxia and nutrient stress in order to persist [11]. The genome of *M. tuberculosis* has 48 proteins residing in the PIN domain family and in 47 cases; these are part of the toxin component of VapBC toxin-antitoxin (TA) systems [18]. Members of this family are hypothesised to be stress response proteins and are thought to be involved in a number of stress related states such as the thermal stress response of *Sulfolobus solfataricus* [36]. They are also involved with pathogen trafficking [49] and metabolic regulation in *Mycoabcterium smegmatis* given their biochemical activity [19]. It is thought by many researchers, that given the collection of PIN domain proteins found within *M. tuberculosis*, these protein may play roles in the persistence and stress response in *M. tuberculosis* [4, 11, 17, 18, 19]. The 48th member belonging to the collection of PIN domains, in *M. tuberculosis*, is located at the N terminus of the single PhoH2 protein encoded in the genome [4].

TA systems are described as small genetic modules that are made up of two elements. The first element is a toxin protein that inhibits cell growth by directly targeting important cellular metabolic pathways and the second element is an unstable antitoxin RNA or protein that binds to and blocks the toxin protein, inhibiting its toxic activity [16, 98, 99]. TA systems can be found in both bacterial and archaeal genomes and allows an immediate response to environmental stresses such as; antibiotics, oxidative stress, nutrient stress, pH, heat stress and the hosts immune response [100, 124]. TA systems have also been suggested to play a role in the virulence of *M. tuberculosis*, through supporting entry into dormancy, in macrophages [135].

Currently, TA systems can be separated into five distinct groups, in regard to their regulation and genetic structure. In type I systems, the antitoxin is a small antisense RNA molecule that inhibits toxin translation, while type II systems encode a protein antitoxin that binds its cognate toxin [137]. Like type I systems in type III TA systems the antitoxins are RNAs that regulate manufacturing of the toxin [123]. In type II, IV and V TA systems, antitoxins are proteins that can either; impound, counterbalance toxin activity or stop toxin synthesis [15]. In type II systems; modules encode two genes in an operon, a protein 'toxin' that stops cell growth and a protein 'antitoxin' that inhibits the activity of the toxin by direct protein-protein interaction [21]. The antitoxins, of these systems, are usually metabolically unstable while the toxins are stable. Therefore regulation of proteolysis on the antitoxin will determine the activity of the partner toxin [22]. In type IV systems the toxin and antitoxin are both made of proteins. Inhibition of the toxin occurs when antitoxin equalise the target proteins, therefore blocking toxin activity. Group V systems have only recently been discovered. In this system, the protein antitoxin is defined as an endoribonuclease that binds and degrades its partner toxin-encoding mRNA [15].

1.2.1 The VapBC toxin-antitoxin family

Type II toxin-antitoxins belonging to the VapBC family, are the most common and widespread TA modules in bacteria and archaea [90]. It appears that there is no particular pattern to the distribution of *vapBC* operons between organisms. In several unrelated organisms, the *vapBC* family of TAs is expanded in number. For example, among mycobacteria the large number of TA operons is reserved to just the mammalian pathogens *M. tuberculosis* and *Mycobacterium bovis*. In the genome of *M. tuberculosis* there are 47 accepted *vapBC* operons. Adversely, *M. smegmatis* and *Mycobacterium avium* subsp. *paratuberculosis* have only a single *vapBC* TA operon [82] and *Mycobacterium leprae* has none [103].

The VapBC family is characterised by a PIN-domain, which is present in the VapC toxin component [18, 96]. The PIN-domain has endoribonuclease activity which is important in pre-rRNA processing, nonsense-mediated mRNA decay [97] and cleavage of RNA molecules, within VapBC systems [90]. The selectivity of the VapC ribonuclease activity is shaped by both the secondary structure motifs

and sequence of the target [92]. These systems produce a response to stress conditions, such as hypoxia and the presence of macrophage enzymes, [101] which can trigger the induction of dormancy and the formation of persisters [102].

VapBC modules are also hypothesised to play roles in controlling intracellular bacterial growth as with the FitAB system of *Neisseria gonorrhoeae* [136] and regulating rates of nitrogen fixation and biomass formation during symbiosis in a host plant seen in the *ntrPR* TA system in *Sinorhizobium meliloti* [138]. VapB proteins consist of two domains: an N-terminal DNA-binding domain and a C-terminal region involved in toxin binding [26]. The VapBC TA complex autoregulates its own expression, through the VapB component that has DNA binding specificity that controls the synthesis of both proteins, by binding to the operon and repressing transcription (Figure 1.2) [27, 91, 97]. Inhibition of the VapC toxin occurs when the VapB antitoxin binds to VapC forming a protein-protein complex, resulting in neutralisation of the toxin [93]. The toxin component of VapBC is very stable while the antitoxin is not and during times of cellular stress, the VapB antitoxin can be easily hydrolysed by proteases belonging to the Clp or Lon family [94, 95, 97]. Upon the degradation of the VapB antitoxin by proteases, the PIN domain of the VapC toxin will be free to inhibit cell growth or viability through RNase activity on the cellular RNA (Figure 1.2) [90].

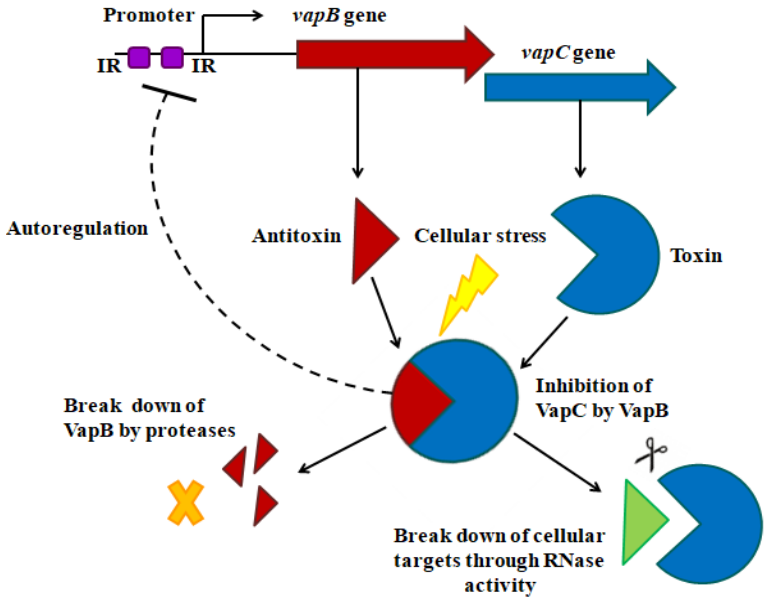

Figure 1. 2: Schematic of a VapBC family from a type II toxin-antitoxin system. The *vapB* (red arrow) and *vapC* (blue arrow) genes overlap in an operon. *vapB* encodes for the VapB antitoxin which binds to and inhibits the VapC toxin, encoded by *vapC*. Cellular stress can result in the breakdown of the unstable VapB antitoxin by cellular proteases releasing the toxin. Free VapC can now exhibit its RNase activity on cellular targets like mRNA. The VapBC system auto-regulates its own transcription by binding to inverted repeats (IR) in the promoter region. Image is adapted from [18, 114, 115].

1.3 The PIN domain

The PIN domain is found in a diverse range of organisms. They 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* [25]. Sequences of PIN domains show sequence homology to many proteins that are involved in eukaryotic nonsense-mediated decay, as well as the 5`-3` nucleases, T4 RNase H, flap endonucleases and exonuclease domains of DNA polymerases. These similarities in sequence homology further suggest that PIN domains bind RNA and have a broad range of nuclease activity [30, 120]. In prokaryotes, the majority of PIN domain proteins are part of the toxin component of toxin-antitoxins, as described in section 1.2 [30].

Structures and biochemical studies of PIN domain family members show low identity across proteins, aside from four well-conserved acidic amino acid residues, (Asp, Glu, Asp & Asp) (Figures 1.3 & 1.4 B), that constitute the active site [11, 104]. Another sequence feature common to this family is the presence of a polar residue (Thr, Ser or Asn) at the position $i + 1$ or $i + 2$ following the first

conserved aspartic acid, that play a structural role within this family (positions 5 and 6 in Figure 1.3) [30].

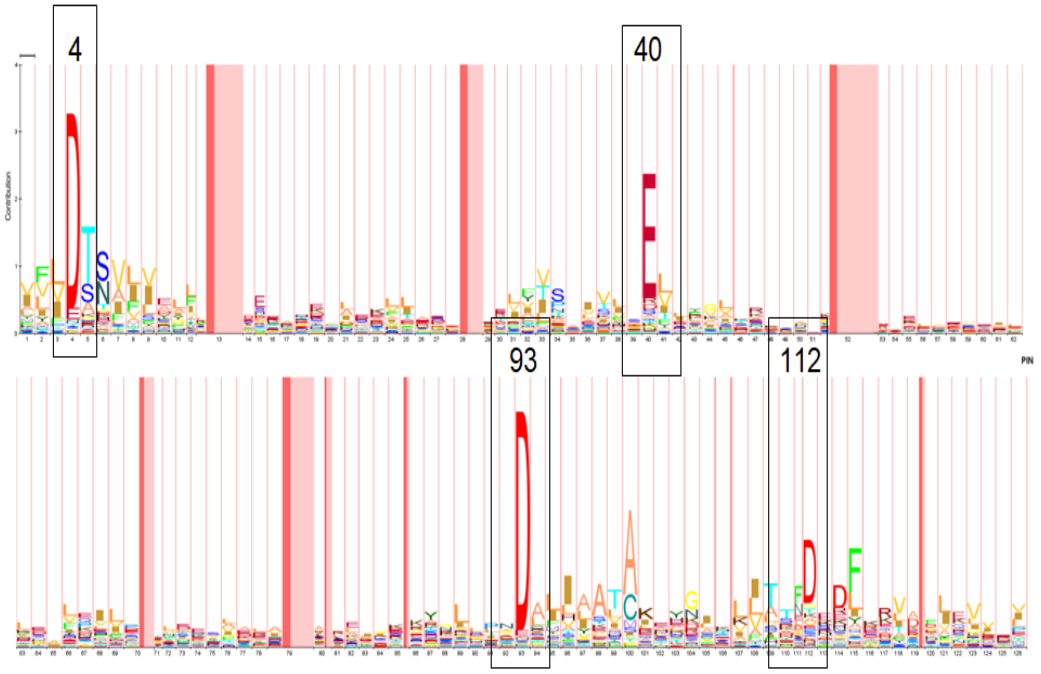

Figure 1. 3: Hidden Markov Model (HMM) that defines PIN domain proteins. This model is based on a multiple sequence alignment and shows four important conserved amino acids (inside black boxes) located at positions 4 (D), 40 (E), 93 (D) and 112 (D), which define the PIN domain family. The height of each letter describes the significance of the amino acid at that position within the PIN family. While the width of the letters defines the importance of this position in the PIN domain family. The light and dark pink areas show where insertions have occurred [4, 10, 11]. Generated from pFam, [104].

The active site residues are located at the C-terminal region of the core β-strands [24]. The structures of PIN domain-like super families can be characterised by a common fold that consists of a 3-layer α/β/α sandwich with a 5-stranded parallel β sheet, located in the middle of the structure, with α helices at both sides (Figure 1.4 A) [29]. While the four conserved residues vary in sequence distance, the arrangement of the fold causes the conserved acidic residues to form the active site [18] which binds Mg^{2+} or Mn^{2+} (Figure 1.4 B) and promotes cleavage of RNA [120, 28]. Prokaryotic PIN domains commonly form dimers (Figure 1.4 C), this arrangement allows the active sites to position into the groove along the long axis of the structure [120].

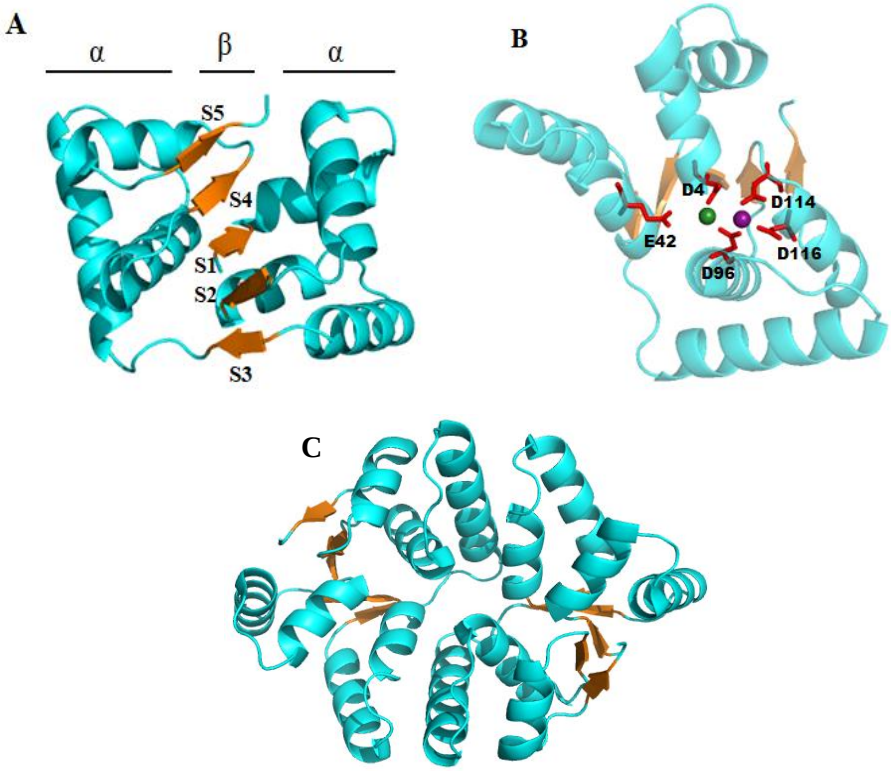

Figure 1. 4: PIN domain VapC-15 from *M. tuberculosis* (PDB 4CHG). Figure A shows the PIN domain displaying a fold consisting of a 3-layer α/β/α formation with 5 β sheets (orange) in the middle of α helices either side (blue). The 5 β sheets are labelled from S1 to S5. Figure B shows conserved residues (Asp, Glu, Asp & Asp) located in the active site of the PIN domain. Manganese (purple sphere) and magnesium (green sphere) are also present in the active site. Figure C shows the PIN domain existing as a dimmer formation in nature. Structures generated using PyMol V1.6.0.0.

Around 95% of defined PIN domain proteins are part of toxin-antitoxin systems. However, in the remaining 5% they are joined with TRAM domains, KH domains and AAA+ ATPase domains [11, 18]. TRAM and KH domains are described as RNA-binding domains [40] and the AAA+ ATPase domains are part of the AAA+ superfamily of ring-shaped P-loop NTPases [122]. These domains can display hexameric ring-like structures and take part in the remodelling and translocating of macromolecules in reactions requiring ATP [11].

In the genome of *M. tuberculosis* there are 48 known proteins that belong to the PIN domain family. Forty seven of these proteins are part of a VapBC toxin-antitoxin family [18]. The last PIN domain belongs to the PhoH2 super family [4].

1.4 PhoH2 proteins

The protein PhoH2 consists of two domains that are fused together; an N-terminal PIN domain fused to a C-terminal PhoH domain [4]. PIN domains are predicted RNases [30] and PhoH domains are predicted ATPases [13]. The two domains are joined by a long flexible linker region that does not correspond to any known conserved domain (Figure 1.5) [10].

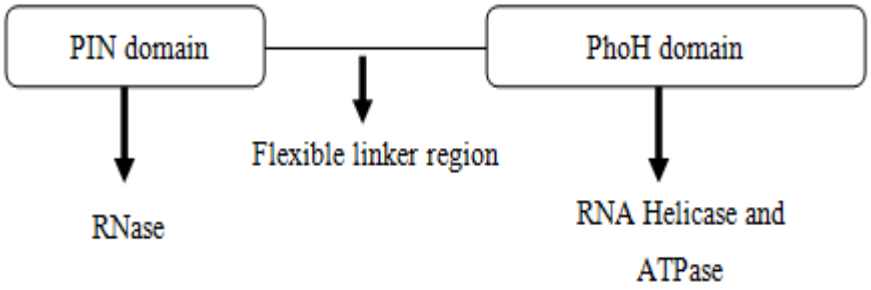

Figure 1. 5: Schematic arrangement of the PhoH and PIN domains that make up a PhoH2 protein. The N-terminal PIN domain is fused to a C-terminal PhoH domain, which are joined together by a flexible linker region with unknown function.

While PhoH2 proteins are found across a diverse range of prokaryotes, they have not been identified in any eukaryotes to date. pFam lists 1605 sequences with PIN-PhoH architecture [13] these sequences are annotated as ribonucleases fused with an ATPase domain belonging to the PhoH super family [11]. PhoH2 proteins are found in a wide variety of different organisms and are very common among mycobacterial species. A BLAST search shows that the PhoH2 protein from *M. tuberculosis* has 100% identify to PhoH2 homologues from *M. bovis*, *Mycobacterium pinnipedii*, *Mycobacterium canettii* and *Mycobacterium microti*. The PhoH2 homologue from *M. smegmatis* has 90.7% sequence identity to PhoH2 from *M. tuberculosis*.

1.5 Is PhoH2 posing as a toxin-antitoxin system?

One of the first examples of a variation of a type II toxin is encoded by the *phoH2* locus found in mycobacterial genomes including *M. tuberculosis* and *M. smegmatis* [4]. PhoH2 is co-transcribed from a bicistronic operon, together with the open reading frame of the upstream PhoAT component, in the same 5' to 3' orientation as the PhoH2 gene. The existence of antitoxins upstream of the PhoH2 gene is limited to mycobacteria, there are no other reports that this occurs in any other organisms [10]. Although much shorter in length PhoAT is thought to

be analogous to VapB of the VapBC toxin-antitoxins [11]. PhoH2 and PhoAT have been proposed to have a similar association to type II TA systems where both components of the TA system are proteins. Andrew and Arcus, 2015 [4], showed that the expression of *M. tuberculosis* PhoH2 in *M. smegmatis* was reported to have a negative effect on growth, which was alleviated by the presence of the adjacently encoded PhoAT antitoxin, which is capable of binding to PhoH2 proteins. They also speculate that the acquisition of additional domains expands the versatility of the mycobacterial VapBC TA systems, which function as post-transcriptional regulators in response to environmental stress [4]. PhoH2 has a specific RNA targeted activity that is similar to the targeted activity of other VapC toxin proteins which play a role in the post-transcriptional regulation of RNA. It is due to this similarity, PhoH2 proteins have been proposed to contribute to the adaptive reaction of organisms to stress, through the targeting of mRNA transcripts [10]. PhoAT, the upstream antitoxin is thought, like other antitoxin proteins to inhibit the toxic affect of PhoH2, the toxin, under normal conditions. Then during times of stress PhoAT would be degraded and PhoH2 would be able to act on its cellular RNA targets. The binding of PhoAT to PhoH2 is predicted to either block the active site of the toxin, or cause steric changes to PhoH2 resulting in the loss of activity and a non-functional protein, therefore being incapable of causing toxicity to the cell. It is possible that PhoAT-PhoH2 works alongside the large number of VapBCs present within *M. tuberculosis* genome [4, 10]. Toxin activation then leads to growth arrest, multidrug tolerance, and persistence [20, 23].

In contrast to all previously reported type II toxins that are small, single-domain proteins, the toxin PhoH2 was shown to be a two-domain protein consisting of an N-terminal VapC toxin-like PIN domain linked to a C-terminal PhoH helicase domain adding up to a total molecular weight of approximately 48.7 kDa [4]. The latter ATPase domain was shown to have RNA binding and unwinding activity, complementing the RNase activity of the N-terminal VapC toxin domain.

1.6 Helicase and PhoH domains

1.6.1 Helicase domains

Helicases are defined as a class of enzymes that catalyse the separation of paired RNA or DNA duplexes leading to exposure of the single nucleic acid strands [11]. This exposure of nucleic acids allows proteins like PIN domains access to RNA sequences, where they can execute their cleavage on specific targets [40]. Helicases require the hydrolysis of ATP for energy to unwind the strands [54]. Helicase enzymes share many common biochemical properties involved with the modification and processing of nucleic acids [55], this includes replication, repair, recombination, transcription, translation and many other nucleic acid related processes [10, 32]. These enzymes have been identified in various organisms, ranging from viruses to eukaryotes and because of their ubiquitous nature have evolved various mechanisms to support specific roles in various biological processes. They have also evolved various structural formations; monomer, dimer or hexamer, to accommodate for different active assembly states [56].

The early classification of helicases used amino acid sequence comparisons to reveal three distinct and diverse superfamilies (SF1, SF2, and SF3) and two smaller families (SF4 and SF5) [11, 64]. Singleton (2007) proposed the addition of the AAA+ class as SF6 to encompass the nucleic acid motors that are classified in this family [59]. Current reviews have confirmed the identity of six helicase superfamilies, based upon primary structure and conserved sequence motifs [57, 58], with all family members possessing Walker A and B motifs that are common to many NTPase enzymes [128]. Superfamily 1 (SF1) and superfamily 2 (SF2) helicases are common and generally function as monomers or dimers in diverse RNA or DNA manipulations [58, 59]. Most of the other helicase superfamilies (SF3-SF6) form ring shaped hexameric structures, demonstrated by biochemical and electron microscopy studies [60-63]. These superfamilies generally encircle the nucleic acid and participate at the replication fork [58]. All of these helicases bind and hydrolyse NTP at the interface between two recA-like domains [57]. The NTP binding site consists of a Walker A (P-loop) and a Walker B motif from the first domain and other elements such as an arginine finger from the other domain [57, 11]. The SF1 and SF2 helicases are the largest of the superfamilies and contain similar sets of seven conserved motifs and two recA-like domains that are

coupled by a short linker [59]. The ATP-binding and hydrolysis site is located at the interface of these two domains [57]. SF2 is formed by the DEAD-box and related DEAH, DExH and DExD RNA helicases [35]. In a majority of the hexameric helicases (SF3-SF6), the ATP-site consists of elements derived from adjacent monomers in the complex. The third superfamily (SF3) is made up of viral helicases that all contain four conserved motifs. The DnaB-like family (SF4) contains helicases that bear the Walker A and B motifs as well as three distinct conserved motifs. The Rho-like family (SF5) is categorised as a separate family, due to its sequence, despite its similarity to SF4 helicases, [59, 11]. The SF3 and SF6 helicases belong to the AAA+ family of ATPases. The AAA+ group consists of a large class of ATPases that include several other complexes that are involved in DNA replication [57]. Some proteins in the SF6 groups are; mini chromosome maintenance (MCM), RuvB, RuvA and RuvC [59].

1.6.2 RNA helicases

RNA molecules are normally single stranded, but have the ability to fold into complex structures or to join with complementary RNAs that can make it difficult for nucleases to process or degrade target RNA's [35]. Proteins that act as chaperones or helicases are capable of unwinding these nucleic acid complexes. RNA helicases are essential for RNA metabolism. RNA helicases are large families of proteins that share up to 12 motifs involved in nucleotide-triphosphate binding, interaction with RNA, and intramolecular contacts [37]. According to differences in their sequence motifs, they can be attributed to various families [38]. RNA helicases exist in both prokaryotes and in eukaryotes and can be found in many RNA viruses. RNA helicases belonging to the SF3 and SF6 super families form hexameric ring complexes [57]. Superfamillies 3-6 contain DNA helicases, translocases and AAA+ proteins [128]. An exception to this is the PhoH RNA helicase which loosely fits in the SF6 superfamily based on similar homology with the AAA+ family of the SF6 group and their hexameric ring-like structure. The PhoH helicase also contains motifs from other helicase families, which is discussed below.

1.6.3 PhoH domains

The PhoH family is believed to have evolved in bacteria by the loss of a C terminal α/β domain present on the last common ancestor of Super family 1 and 2 helicases. As a result of a gene duplication event within the bacterial lineage, there are two orthologous groups of PhoH-like ATPases: PhoH and YlaK, [40]. 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, [41]. Homologs of PhoH are present in many organisms, and they can be classified into three groups. The first group can be found in most bacteria and is proposed to be functionally linked to phospholipid metabolism as well as RNA modification [75]. Proteins belonging to the second group are members of fatty acid beta-oxidation regulons, which are present in aerobes. The third group is unique to enterobacteria [76, 77].

Kim *et al.* (1993) first described PhoH as a phosphate starvation-induced protein, in *E. coli*. They found that expression of *phoH* was controlled from two promoters; P1 and P2. P1 promoter was found to be induced by phosphate limitation and dependent on PhoB, while induction from P2 was dependent on the concentration of phosphate in the medium. DNA sequencing results confirmed that the *phoH* gene sequence had hypothetical nucleotide-binding motifs [31]. This study also found that the phoH sequence contained similar sequences as the junction sequences of psiH45::lacZ, psiHl4::lacZ and psiH16::LacZ. PsiH was induced in the presence of phosphate, carbon and nitrogen starvation and by UV irradiation, [39].

PhoH-like proteins are defined as ATPases and have been placed into clan CL0023 in PFam, this clan is part of the P-loop NTPase superfamily [13]. PhoH from *E. coli* has the ability to bind ATP, implying that these proteins have the capacity to turnover ATP [31]. Anarntharaman (2002) [40] suggested that PhoH proteins might be involved in RNA metabolism either as helicases or through the involvement in ATP dependent processes in ribonucleoprotein (RNP) complexes in bacteria [40]. Based on the structure of PDB 3B85, it is likely that PhoH-like ATPases adopt a hexameric macromolecular conformation (Figure 1.6). Based on these observations, it is predicted that PhoH proteins will possess helicase activity,

[10, 11]. PyMol was used to generate the electrostatic potential on the molecular surface of the PhoH domain (PDB 3B85) from PhoH2 in *Corynebacterium glutamicum*. As RNA is negatively charged it is possible that it may make contact with the blue positively charged motifs at the entrance to the PhoH hexameric tunnel (Figure 1.6). Andrews and Arcus, 2015, reported that when PhoH is fused to PIN domain (PhoH2) it possesses ATPase and RNA helicase activity [4, 30]. This RNA helicase activity will be important to give the PIN domain access to cellular RNA now exposed by the PhoH domain. The presence of PhoH proteins in various organisms, suggests that their ATPase activity may be recruited to provide energy in various processes, [75, 11].

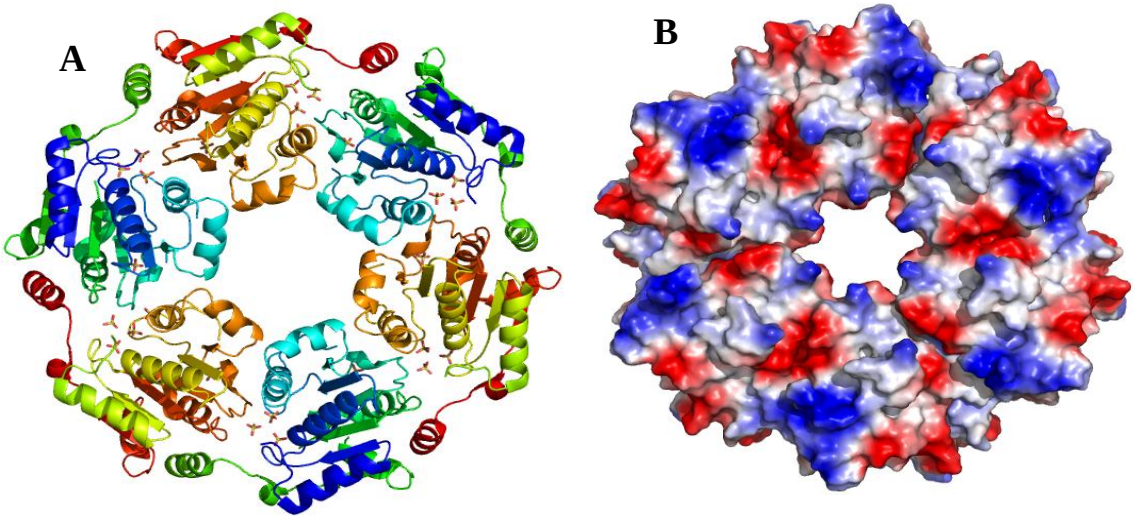

Figure 1. 6: Structural representation of the ring-like hexameric PhoH domain from PhoH2 in *C. glutamicum*. The PhoH domain forms part of a PhoH2 protein, from C. glutamicum. In figure A: sulfate ions are present between monomers (shown as sticks) and show the position where the γ and β phosphate of ATP would bind in the active site. In figure B: the PhoH domain is represented as a surface, showing electrostatic charges. Red areas represent acidic regions (negatively charged), while blue areas represent basic regions (positively charged). Image was made in PyMol V1.6.0.0 using PDB 3B85.

Examination of a PhoH structure (PDB 3B85), from cg2513, of *C. glutamicum* (Figure 1.7), and the PhoH protein HMM sequence (Figure 1.8), show that 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, which include an arginine residue (Box 7 in HMM Figure 1.8 and position 3 in Figure 1.7, A) and conserved sequence RGRTL located between beta sheets β3 and β4, [44]. Structurally PhoH proteins show a P-loop NTPase α/β arrangement with parallel β sheets surrounded by α helices and a β strand topology similar to the SF1/SF2 helicase family

(Figure 1.7 B) [38, 44]. They display walker A and B motifs, as do all other helicase families. PhoH proteins also share a similar motif to the III motif of SF1 [42]. They also have the defining features of AAA+ proteins, the Sensor I and II (SI, SII) in similar positions at the top of β4 and at the base of α7. And the second region of homology (SRH-arginine finger) positioned in a similar arrangement to AAA proteins between β4 and β5 [43]. Furthermore, PhoH has a Q motif characteristic of DEAD box RNA helicases and many members of SF1/2. In structural arrangements, this Q motif as well as Walker A/B, RNA binding and binding RRB1, motif III, SI, SIII interact with the RNA recognition and binding RRB2 and SRH motifs, positioned in the neighbouring monomer of the PhoH domains.

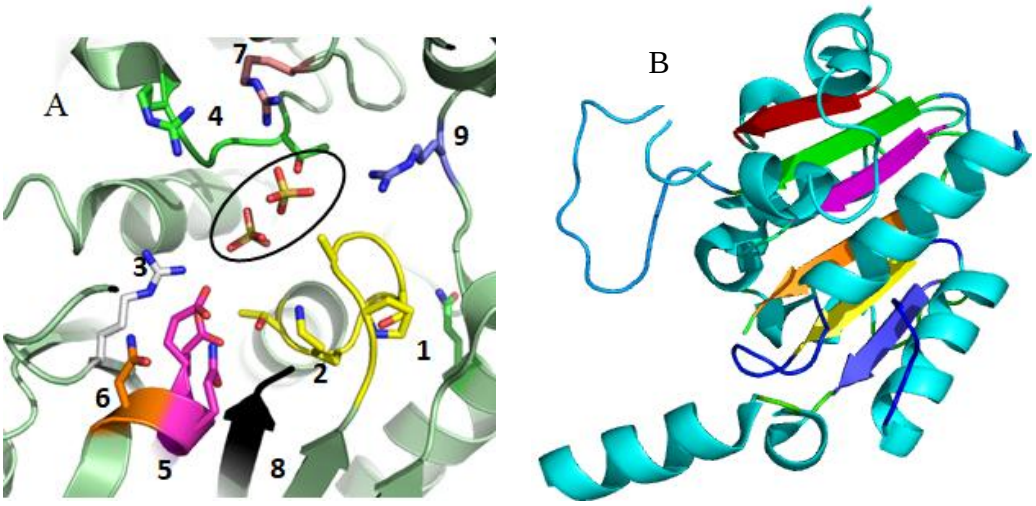

Figure 1. 7: PhoH monomer (PDB 3B85) from PhoH2 protein from *C. glutamicum*. Figure A shows a magnified view of the active site in the PhoH domain. Conserved residues are numbered accordingly to the HMM figure below (Figure 1.8). Sulfate ions are circled. Image was taken with permission from [4]. Figure B depicts the phoH domain showing β strand topology, typical of SF1/SF2 family domains: β1 (yellow), β2 (green), β3 (red), β4 (pink), β5 (orange) and β6 (blue). Both images were generated using PyMol V1.6.0.

The Hidden Markov Model logo (HMM) for the PhoH family from pFam (http://pfam.xfam.org/) as well as the PhoH structure (PDB 3B85) distinguished additional regions of conservation located within the PhoH protein sequence [4]. These two motifs are unique to PhoH proteins (RGRTL and R residue) [10, 11] (3 and 4 in Figure 1.7 and Figure 1.8). Mutation of arginine 56 and arginine 114 respectively resulted in loss of RNA unwinding activity and therefore cleavage

activity [4]. These mutations did not affect the turnover of ATP suggesting involvement of these motifs in nucleic acid recognition and binding.

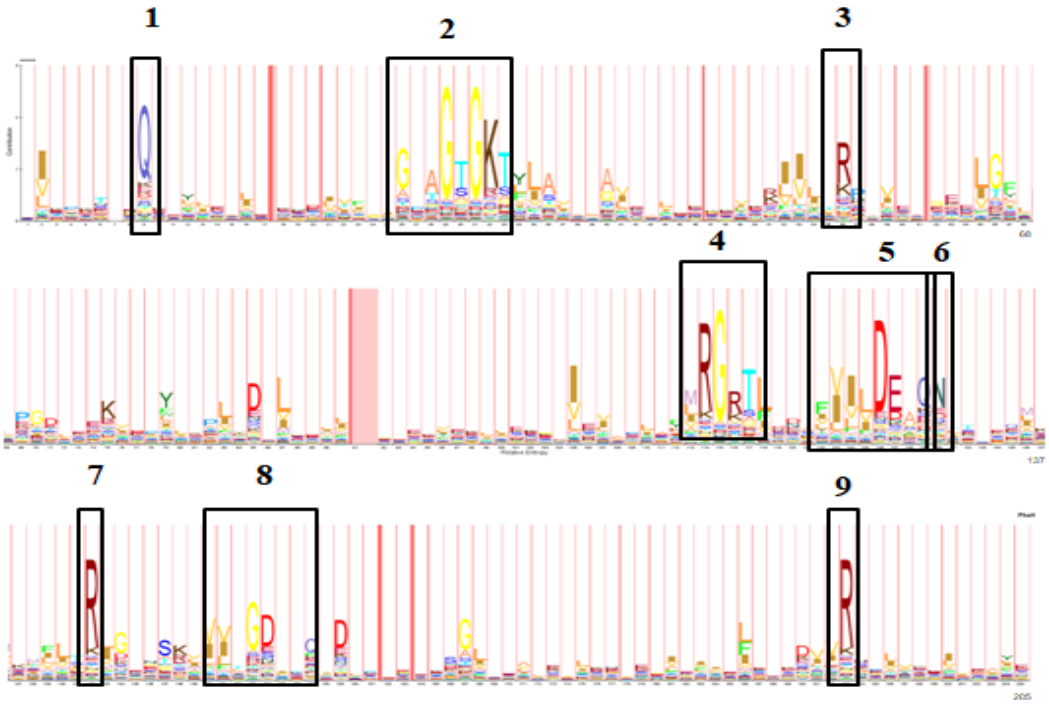

Figure 1. 8: Hidden Markov model (HMM), based on multiple sequence alignment showing conserved motifs located in PhoH proteins. Conserved motifs are described as in [11, 4]. The width of each letter defines the importance of that position within the PhoH family. While the height of each letter describes the importance of the particular amino acid at its position within the PhoH family. The light and dark pink areas show where insertions have occurred. The amino acids contained within boxes display sites of great conservation that are related to that of other helicase families: Q motif SF1/2 (1), Walker A (2), Walker B (5), N residue SI AAA+ (6), R residue SRH/arginine finger AAA (7), motif III Sf1 (8) and R residue SII AAA+ (9) and the two motifs R residue RNA recognition and binding (3), RGRTL, RNA recognition and binding (4) that are only found in PhoH proteins and play roles in nucleic acid interactions. Generated using PFam [104].

1.7 Biochemical characterisation of PhoH2 proteins

In 2015 Andrews and Arcus [4] biochemically characterised PhoH2 from *M. tuberculosis*, *M. smegmatis* and *Thermobispora bispora*, *in vitro*, using traditional RNA helicase substrates, in a series of activity assays. These short RNA oligonucleotides were composed of a duplex region with single stranded overhangs. Using different sequence combinations at the 5' terminal of the RNA oligonucleotide, they were able to deduce a specific sequence of RNA bases, AC [A/U] [A/U] [G/C], that were required for activity in the presence of ATP. Using

site directed mutagenesis, a single amino acid mutation was made at two highly conserved arginines at positions 280 and 339 in *M. smegmatis* and 284 and 343 in *T. bispora* independently located within the PhoH protein sequence. Mutation to alanine at these positions resulted in loss of unwinding and degradation activity, but retention of ATPase activity. This suggested that these motifs within the PhoH protein sequence are involved with RNA recognition and binding and that the role of the PIN domain was to cleave the unwound RNA product [4].

Further, in *M. tuberculosis*, PhoH2 was identified, along with 40 other proteins, as posing as a potential PknH substrate [117]. This study proposes that PknH Ser/Thr protein kinases of *M. tuberculosis* control the expression of a variety of cell wall related enzymes and regulate the *in vivo* growth in mice. Based on these observations it is speculated that PknH kinases phosphorylate several substrates, controlling different metabolic and physiological pathways. Experimental data suggests that the PknH kinase may play a role in *M. tuberculosis* virulence. This study used [117] bioinformatics analysis to identify potential substrates that are phosphorylated by PknH kinases. These results suggest that the PknH kinase has the potential to phosphorylate a wide range of mycobacterial proteins including PhoH2 [117].

A study by Wimmerova *et al.* in 2003 [116] used fold recognition and exploratory data to identify possible glycosyltransferases in the genome of *M. tuberculosis.* The search was conducted against a library composed of all known crystal structures of glycosyltranferases and some related proteins. The top 25 scores were obtained by fold recognition calculations from the *M. tuberculosis* translated genome over the three known structure of glycosyltransferases with BGT folds. Sequences were ranked according to the average pair/ surface score. PhoH2 was ranked number 10 out of 25 proteins. These results suggested that PhoH2 may constitute a putative glycosyltransferase, due to the presence of a Rossaman fold that is found within glycosyltransferases [116]. These folds are common to proteins that belong to the SF6 hexameric helicases as well as other motor proteins [128] that PhoH family proteins roughly falls within, thus it is more likely that PhoH2 is an RNA helicase and RNAse than a glycosyltransferase.

Festa *et al.* (2010) [118] identified proteins of *M. tuberculosis* that are pupylated (degraded) by the prokayrotic ubiquitin-like protein (Pup). Pup in *M. tuberculosis*

is the first known post-translational small protein modifier in prokaryotes. Its role is to target several proteins for degradation (pupylation) by a bacterial proteasome in a similar way to ubiquitin (Ub) mediated proteolysis in eukaryotes. PhoH2 was identified along with three other new Pup substrates as targets for degradation by the proteasome, due to their accumulation in mpa (proteosomal ATPase) and pafA (Pup ligase) mutants. This study predicted that as proteasome function is essential for *M. tuberculosis* pathogenesis, the identification of new proteasomal degradation substrates may provide answers to link proteasome dependent proteolysis and virulence. The inability to turn over potentially hundreds of proteins could greatly compromise bacterial survival when adapting to a new environment. It is currently unknown what effect, if any, the accumulation of metabolic enzymes, like PhoH2, has during an infection. Their data showed that several virulence-associated proteins are regulated by proteasomal degradation [118].

Outside of mycobacteria, a study by Fedida and Lindell in 2017 [119] found that a PIN-PhoH homolog is involved in the infection process of *Synechococcus* by Syn9 phages. *Synechococcus* is an abundant marine cyanobacterium. Transcriptional data of *Synechococcus* strains that were infected with Syn9 phages revealed that while the transcript levels of most host genes declined, the transcript levels for the gene SYNW1946 (PIN-PhoH) increased after infection. Upon inactivation of SYNW1946 there was no impact on host growth or on the length of the Syn9 lytic cycle however there was significantly higher Syn9 genomic DNA replication and progeny production, suggesting that this PIN-PhoH2 protein plays a role in restraining the infection process. The PIN-PhoH2 protein in *Synechococcus* may limit phage progeny production through unwinding and cleavage of phage RNA during infection, the target of its cleavage activity may be mRNA required for replication proteins [119].

In *C. glutamicum.* Larisch *et al.* [113] used genome wide transcription profiles of a wild-type and *ΔsigB* strain of *C. glutamicum* to identify genes whose mRNA levels would be affected by either type of the strain. The data indicated that the mRNA levels of 111 genes were significantly changed in the wild-type strain during the transition phase, while the expression profile of the Δ*sigB* strain showed only minor changes (26 genes). The *phoH2* gene, annotated as *cg2513* in *C. glutamicum,* was identified as having decreased mRNA expression levels at the

transition phase in a *sigB*-proficient *C. glutamicum* strain (RES167) compared to a Δ*sigB C. glutamicum* strain (CL1) [113].

In 2003 Ishige *et al*. [126] used whole genome DNA microarrays to identify genes, from *C. glutamicum,* whose expression was affected by a shift from phosphate sufficient to limiting growth conditions. The *phoH2* gene, annotated as *phoH1* in this study, was shown to have increased gene expression in phosphate limiting conditions [126]. Target DNA binding sites were found for each of the major Pho regulon genes including *phoH1* suggesting that in *C. glutamicum*, *phoH1* is under the control of PhoR [127]. These two studies are the first of their kind to suggest the involvement of a PhoH2 (YlaK homologue) protein inside Actinobacteria taking part in the phosphate starvation response under the regulation of PhoP.

1.8 RNAseq analysis proposes targets for PhoH2 proteins

An experiment by Dr Emma Andrews focused on determining the biological role of PhoH2 proteins in *M. smegmatis* [personal communication]. A growth experiment was carried out with *M. smegmatis* strains mc^{2}155 (wild-type) and mc^{2}155 *ΔphoH2* (*phoH2* deletion strain), under standard growth conditions (LB media with 0.05% tyloxapol) and samples were taken every 24, 48 and 72 hours. RNA was then extracted from samples of the three time points across the growth curve (Figure 1.9) and sent to the Beijing Genomic Institute (BGI) for transcriptomic analyses that included Bioinformatics.

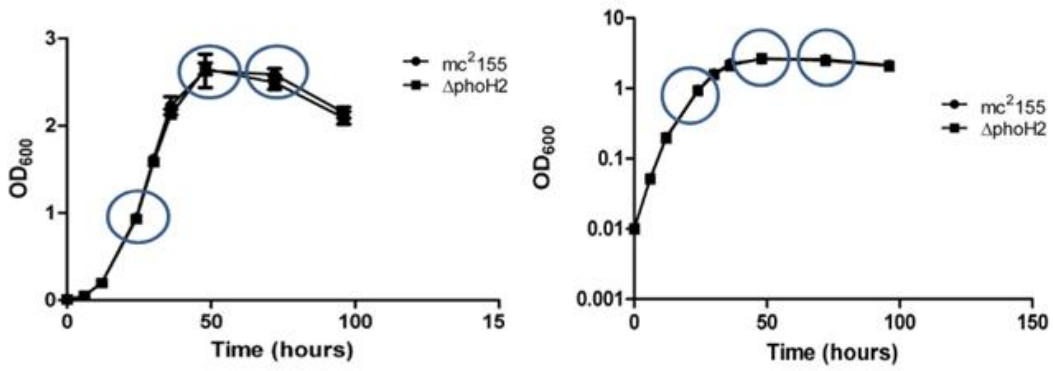

Figure 1. 9: Growth curve of *M. smegmatis* mc^{2}155 and *M. smegmatis* mc^{2}155 *ΔphoH2* in LB tyloxapol media. Growth was measured by optical density (OD) measurements. Growth curves are plotted on linear and log scale. No significant difference in growth was observed between mc²155 and mc²155 ΔphoH2 cultures. Time points selected for RNA extractions for transcriptome analysis are indicated by blue circles (24, 48 and 72 hours).

The 48 hour transcriptome data revealed an up regulation of genes belonging to the SigF regulon, in the *M. smegmatis* mc^2155 *ΔphoH2* strain. This is in comparison to the wild-type *M. smegmatis* mc^2155 strain. It was then hypothesised that the protein PhoH2 may regulate the SigF regulon in *M. smegmatis*, through helicase and RNase activity. [4].

1.9 Sigma factors and SigF

1.9.1 Sigma factors

Sigma factors are bacterial transcription initiation factors that reversibly associate with RNA polymerase [74] (RNAP) and help position the core polymerase to a particular promoter, allowing the expression of a specific set of genes [47, 52, 65]. Changes in environmental factors lead to the replacement of sigma factors in the holoenzyme and the transcriptional regulation of a different set of genes. This in turn, expresses proteins that are required for an organism to survive under stress conditions [53]. Sigma factors generally recognise distinct sets of promoter sequences. Therefore, variation in active sigma factor populations may represent a powerful way to modulate transcription profiles of an organism in accordance with its physiological requirements [52].

There is a specific family of sigma factors that are found within mycobacteria known as the sigma 70 family, [48]. This family of sigma factors can be further broken into four different phylogenetic groups based on function and gene structure, [51]. Group 1, which includes SigA, contains the main and essential sigma factors, found in mycobacteria. Group 2 sigma factors such as SigB are closely related to the essential sigma factors but are not required for bacterial cell growth. SigF is part of group 3, and is known as an alternative sigma factor, [68]. The sigma factors in group 3 are less related to the sigma 70 family and play a role in the activation of regulons in response to a specific signal such as a heat shock. Sigma factors within group 3 can be further classified into different groups of functionally related proteins that take part in sporulation, flagella biosynthesis, or the heat shock response [67]. The final and largest group of sigma factors, group 4 have widely diverged extracytoplasmic functions. This group are also described as alternative sigma factors which provide adaption to a large range of

internal and external stimuli [69]. There are many different forms and numbers of alternative sigma factors that are found between species, whose role is to provide a stress response indicative to the stress stimuli that the species is facing [47, 51].

1.9.2 SigF

SigF from group 3 of the sigma factor families is an alternative sigma factor that is often associated with stress. Alternative sigma factors vary in the roles they play in each species depending upon the stress response they respond to [50, 53]. SigF was first identified by DeMaio *et al.*, 1996, as a stationary-phase stress response sigma factor in *M. bovis,* [45]. SigF is thought not to be an essential sigma factor, as deletion mutants (Δ*sigF)* in *M. tuberculosis* do not have decreased growth in broth culture or human macrophages, *in vitro*. However, the deletion mutant, in *M. tuberculosis*, showed reduced virulence in mouse and guinea pig infection models [111]. SigF is also involved in stress response and sporulation sigma factors within other bacteria [45]. SigF is highly conserved in mycobacteria [70] and in *M. smegmatis* SigF mediates a general stress response as loss of SigF was shown to increase susceptibility to oxidative stress, acidic pH and heat shock [72]. While in *M. tuberculosis* it has been associated with the regulation of virulence [4, 81]. It is speculated that due to the difference between *M. tuberculosis* and *M. smegmatis,* SigF will play a different role in the regulatory circuits of each of these bacteria, due to the varying natures of their environments [73].

1.9.2.1 SigF in *M. smegmatis*

M. smegmatis is a fast growing saprophytic soil bacterium. Due to the wide range of habitats where this organism can be found, it will encounter more diverse conditions than its pathogenic counterparts, such as *M. tuberculosis* [74]. There are 28 sigma factor genes in *M. smegmatis* in comparison to the 13 reported in *M. tuberculosis* [79]. The genome of *M. smegmatis* encodes one sigma factor from group I, II and III represented by SigA, SigB and SigF and 25 sigma factors from group IV [74].

In *M. smegmatis*, SigF is expressed at basal levels during exponential growth and expression increases upon entering stationary phase or during times of stress. SigF directs the activity of RNA polymerase to regulate expression of genes and

proteins that help the bacterium to adapt to stationary phase, oxidative stress and heat/cold shock [72]. Singh *et al*., 2015 [73] used a *sigF* mutant of *M. smegmatis* mc²155 to show that SigF is not essential for growth of the bacterium; and by deleting *sigF* there is a loss of carotenoid pigmentation which resulted in increased susceptibility to H_2O_2 induced oxidative stress in *M. smegmatis* [112]. From these experiments they concluded that SigF modulates the cell surface architecture and lipid biosynthesis expanding the range of SigF functions in this species [74].

The regulon of SigF in *M. smegmatis* includes a variety of genes expressed during exponential and stationary phases of growth that are responsible for oxidative stress, lipid biosynthesis, energy and central intermediary metabolism [74]. In *M. smegmatis sigF* is part of an operon with its antagonist, anti-sigma factor, RsbW encoded upstream and a ChaB family protein encoded upstream of *rsbW* (Figure 1.10). This arrangement is conserved in mycobacteria [46]. It is likely that the regulation of SigF occurs at the posttranslational level, as there is little evidence of transcriptional regulation of the *rsbW-sigF* operon in *M. smegmatis* under standard growth conditions and for most stress conditions applied *in vivo* [47]. Upon over expression of the SigF antagonist, RsbW, in a wild-type strain of *M. smegmatis* a phenotype comparable to *M. smegmatis* mc²155 Δ*sigF* strain was produced [74]. SigF is transcribed from two different promoters, one is a SigF-dependent promoter preceding *chaB* and the other is a SigF-independent promoter preceding *rsbW.* Microarray data provided by Hümpel *et al.* [46], identified 20 genes, immediately preceding the *sigF* locus, that were affected by the deletion of *sigF*. Eleven of these genes contained a *sigF* promoter. The region contains another anti-sigma factor (MSMEG_1787) with similarity to *M. smegmatis* RsbW (43% identity) and three UsfY proteins. MSMEG_1787 has been hypothesised to be a second anti-sigma factor for SigF. In *M. tuberculosis*, the *usfY* gene belongs to the *sigF* locus, but no function has been given to it. It is hypothesised that because these *usfY* genes are close in proximity to the *sigF* gene in both *M. tuberculosis* and *M. smegmatis*, that the proteins may play a role in controlling SigF activity. Hümpel *et al.* [46] used a Simple Modular Architecture Research Tool (SMART) [108] to search for similarities between the five anti-sigma factor antagonists found in *M. tuberculosis* and found that these proteins share a sulphate transporter and anti-sigma factor antagonist domain (STAS).

Four proteins in *M. smegmatis* were identified with such a domain; MSMEG_1786, similar homology to RsfA, MSMEG_6127, similar homology to RsfB, MSMEG_6541 and MSMEG_5551 [47]. Another study by Singh *et al*. also suggests that the regulation of anti-sigma factors is controlled by anti-sigma factor antagonists, RsfA (MSMEG_1786) and RsfB (MSMEG_6127) [74].

A study found 64 genes in exponential phase and 124 genes in stationary phase that are SigF dependent [47]. Several genes that showed SigF-dependent expression are involved in oxidative stress defence e.g a heme-containing catalase and a manganese catalase, several genes are also involved in biosynthesis of carotenoid isorenieratene. This suggests that SigF is a key player for stationary-phase adaptation and stress response in mycobacteria. The data from this study suggests that SigF regulates the biosynthesis of the thermoprotectant trehalose and an uptake system for osmoregulatory compounds, this regulation may explain why the Δ*sigF* strain shows heat susceptibility. Regulatory proteins; SigH3, PhoP, WhiB1 and WhiB may be under the direct control of SigF. Hümpel *et al.* [46] also show 130 potential genes under direct control of SigF, of which more than 50% showed reduced expression in a Δ*sigF* strain [46].

1.9.2.2 SigF in *M. tuberculosis*

As reported above *M. tuberculosis* has 13 sigma factors. SigA, is the primary sigma factor and is continuously expressed and controls the transcription of many housekeeping genes within *M. tuberculosis* [80]. SigB in *M. tuberculosis* encodes a principal sigma factor that is 62% homologous to SigA [131]. SigF the alternative sigma factor shares 32% homology to SigB. The other sigma factor genes (*sigC, sigD, sigE, sigG, sigH, sigI, sigJ, sigK, sigL,* and *sigM*) are defined as extracytoplasmic function sigma factors, which have roles in cell envelope synthesis, secretory functions and periplasmic protein repair and degradation [133]. SigF from *M. tuberculosis* was first discovered within cultured human macrophages, during stationary phase of growth, induced by exposure to cold shock, nutrient starvation and many antibiotics [105]. In *M. tuberculosis* SigF controls the expression of a particular subset of genes by altering RNA polymerase specificity. SigF in *M. tuberculosis* is upregulated by exogenous stress conditions e.g., by the administration of anti-mycobacterial drugs.

In *M. tuberculosis* a Δ*sigF* strain grew to a threefold higher density in stationary phase than the wild-type strain but showed almost similar sensitivity to heat shock, cold shock and hypoxia relative to the parental strain [70]. This same strain showed a decreased virulence of *M. tuberculosis* in a mouse infection model and disease-associated tissue damage in mice and guinea pigs. Loss of *sigF* also leads to an altered cell wall composition due to a lack of virulence-related sulfolipids [106]. Over expression of *sigF* in *M. tuberculosis* resulted in the differential regulation of many cell wall-associated proteins and other genes taking part in the biosynthesis and degradation of surface polysaccharides and lipopolysaccharides, which are thought to take part in host-pathogen interactions [80]. Hartkoorn *et al*. found that the heparin-binding hemagglutinin, HbhA (Rv04750) and PhoY1 were upregulated following SigF over expression and they identified 67 high-affinity SigF binding sites and 16 loci where a SigF promoter directs the expression of a transcript, through genome-wide approaches [139]. These loci include *sigF* itself as well as genes that are part of the lipid and intermediary metabolism and virulence and a transcriptional regulator *Rv2884*. Among these loci are previously published genes, such as *rsbW-sigF, MT1083.2, Rv1284, Rv1823-25, Rv2884 and Rv3301c* [109, 110]. Investigations also found that SigF directed the transcription of the gene for small RNA F6 [80].

The *M. tuberculosis sigF* gene product is structurally related to sigma factors, which are also induced by entry into stationary phase, from *Streptomyces coelicolor* [130], *B subtilis* [129], *Stapylococcus aureus* [131], *Listeria monocytogens* [131] and *M. smegmatis.* SigF from *S. coelicolor* and *B. subtilis* is essential for sporulation in these two species [129, 130]. In 1997 DeMaio *et al.,* described SigF from *M. tuberculosis* as having significant homology to the stress response sigma factor, SigB and the developmental sigma factor, SigF, both from *B. subtilis.* SigB of *B. subtilis* controls responses to environmental stress. *SigB* and *sigF* genes, from *B. subtilis,* are preceded by an open reading frame (*usfX*), like the *sigF* gene from *M. tuberculosis.* The *usfX* encodes a protein that is similar to the *B. subtilis* anti-sigma factors, RsbW and SpoIIAB. Sequence analysis suggests that the *usfX* and *sigF* genes appear to be co-transcribed and translationally coupled. A second open reading frame called *usfY* precedes *usfX*, this *usfX* gene is thought to be similar to the *chaB* gene that precedes the *rsbW* gene in *M. smegmatis,* although no function has been given to

either genes. The *sigF* gene from *M. tuberculosis* was over expressed in *E. coli* to produce polyclonal antibodies and through immunoblotting, SigF was seen to be antigenically closer to SigB than SigF, both from *B. subtilis* [129].

Similar to *M. smegmatis* expression of *sigF* in *M. tuberculosis* is low during the exponential growth phase, a likely reason for this may be because the protein is largely inactive due to the binding of the anti-sigma factor, UsfX, which is encoded by the same operon [80]. The gene encoding SigF is part of a gene cluster comprising *usfY*, *usfX* and *sigF*, where *usfX* and *sigF* appear to be co-transcribed (Figure 1.10**)** [81]. In *M. tuberculosis*, posttranslational regulation of SigF is complex and involves the anti-sigma factor UsfX [47] as well as five anti-sigma factor antagonists; RsfA, RsfB, Rv1364c, Rv1904 and Rv2638 [107]. The latter are not part of the *usfX-sigF* operon and are spread across the genome [74].

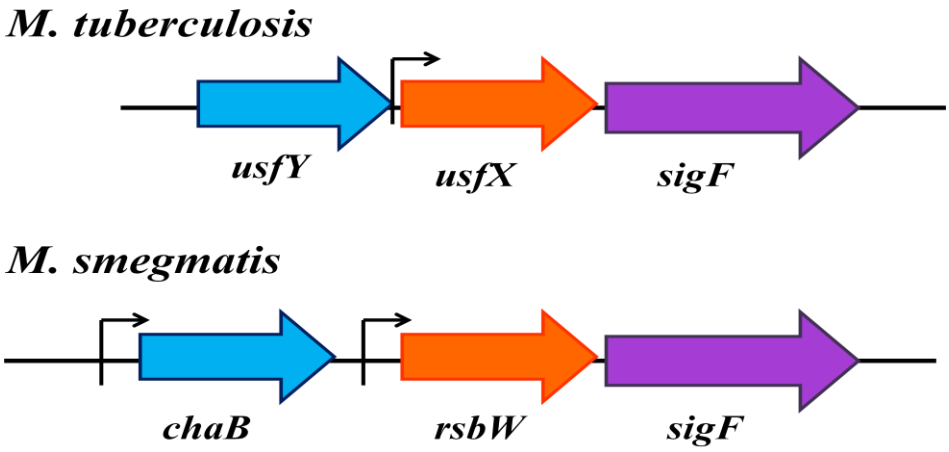

Figure 1. 10: Schematic of the *sigF* gene arrangement in the genomes of; *M. tuberculosis* and *M. smegmatis*. Arrows indicate known promoter regions in the *sigF* operon.

1.10 Aims and Objectives

Recent transcriptome analyses by Dr Andrews showed an upregulation of SigF regulon genes after 48 hours in *ΔphoH2 M. smegmatis* mc²155 (knock-out) when compared to *M. smegmatis* mc²155 (wild-type). From these results it is hypothesised that PhoH2 is acting as a regulator of the alternative sigma factor; SigF and regulating expression via the proteins unwinding and degradation activity.

The objective of this thesis is to understand more behind the activity of *M. smegmatis* PhoH2 proteins against RNA and determine if PhoH2 is a regulator of SigF; either through direct RNase activity on *sigF* mRNA or through the indirect regulation of expression, by targeting *sigF* antagonists, such as RsbW.

Aim One: Carry out unwinding and degradation assays with purified PhoH2 protein from *M. smegmatis* against mRNA substrates, designed to incorporate *sigF*, and its operon genes; *rsbW* and *chaB*, to determine if *sigF* mRNA is a direct target of PhoH2 proteins.

Aim Two: Validate RNA-seq gene expression results from *M. smegmatis* strains; *ΔphoH2* mc^2155 against mc^2155, using qPCR experiments.

2 Chapter two: Methods and Materials

2.1 General methods

2.1.1 Growth media

All dehydrated media was supplied by Difco.
All media was prepared using standard recipes. Liquid media was either prepared in 500ml, 1L glass bottles or in 2L glass flasks. LB (Luria broth) media 1% (*w/v*) bactotryptone, 0.5% (*w/v*) yeast extract, 1% (*w/v*) NaCl pH 8.0 made up to the required volume with MQ H_2O. LBT (Luria broth + tyloxapol) as above with 0.05% (*v/v*) 0.2 µM filtered tyloxapol. Solid media was prepared in 500ml or 1L glass bottles and stored in molten form at 50°C until needed. LB-agar 1% (*w/v*) bactotryptone, 0.5% (*w/v*) yeast extract, 1% (*w/v*) NaCl, 15g.L-1 agar pH 8.0 and LBT-agar as above with 0.05% (*v/v*) tyloxapol All media was autoclaved as per standard sterilisation protocol (121°C for 20 minutes) and liquid media was stored at room temperature (~ 22°C).

2.1.2 Glycerol stocks

Glycerol stocks for long term storage of transformed bacteria were made by the addition of 0.5ml of overnight culture (LB for *E. coli* or LBT for *M. smegmatis* + appropriate antibiotic) to 0.5ml sterile 50% (v/v) glycerol. Glycerol stocks were stored at -80°C. All glycerol stocks used in this study were provided by Dr Emma Andrew's (Proteins & Microbes Lab, University of Waikato, Hamilton).

2.1.3 DNA manipulations

2.1.3.1 *M. smegmatis* mc²155 genomic DNA extraction:

DNA was required as a template for many PCR reactions and as a positive control for PCR's in chapter four (section 4.2.3 and 4.2.4).

M. smegmatis was grown in LBT broth, in 500ml flasks, at 37^{o}C, 200 rmp, for 48 hours. After the 48 hours of growth cells were immediately added to 50ml falcon tubes, 10ml of cells in each tube. One tenth of the cell volume of fresh 5M guanidium isothiocynate (GITC) pH7 was added to each falcon tube and cells were pelleted by centrifugation. The supernatant was removed and the pellet was resuspended in 0.5ml 5M GITC and added to a 2ml screw cape tube that contained around 0.3g of 0.1mm and 2.5mm zirconia beads. Tubes were placed inside a FastPrep cell disrupter (FP120 Thermo savant) and cells were ruptured with three rounds of bead beating. The first round was 25 seconds at setting 4.5, the second round was 30 seconds at setting 6 and the last round was 35 seconds at setting 6. Between rounds the samples were left to cool for 1 minute. Tubes were cooled on ice and then briefly centrifuged at 4^{o}C and 13,000 rpm to remove any foam. Fifty µl of 2M sodium acetate pH4 was added to the samples along with 25µl of basic phenol and chloroform. The samples were mixed for 1 minute. To remove any protein from the samples, they were rotated on a rotor wheel for 10 minutes at room temperature. Samples were then centrifuged at 13,000 rpm at room temperature for 1 minute to separate the phases. The top layer was removed and added to a 1.5ml sterile eppendorf tube with an equal volume of isopropanol (IPA). The samples were then incubated for 20 minutes at room temperature. Samples were centrifuged at 13 000 rpm at room temperature for 20 minutes. The supernatant was removed and discarded and the pellet was washed with 1ml of 70% ethanol. The side of the eppendorf tube was scraped with the pipette tip to disturb any pellet that had collected on the side of tube. The samples were then centrifuged at 13000 rpm at room temperature for 5 minutes. The ethanol was then removed and discarded with the tube been re-centrifuged to ensure all traces of ethanol were removed. The pellet was then air dried at room temperature for 5 minutes. Once the pellet was dry it was resuspended in 50µl of DNase free water and the side of the tube was scraped with the pipette tip to ensure all traces of pellet were resuspended.

2.1.3.2 Agarose Gel Electrophoresis

DNA fragments were separated via agarose gel electrophoresis for visualisation. The percent of agarose in the gel depended on the size of DNA; as a general rule samples < 200bp were run on a 2% (w/v) gel, 200-400bp 1.5% (w/v), 400-1000bp

1% (w/v) and >1000bp 0.8% (w/v) agarose in TAE buffer. All samples were separated on a 1% agarose gel in 1x TAE buffer unless specified otherwise. Samples were mixed with 10x DNA loading dye prior to loading onto the gel. Agarose gels contained 1x SYBR Safe™ DNA gel stain (Invitrogen, USA). Gels were visualised by a blue light box (Invitrogen, USA) and images captured. Band sizes were determined by visual comparison against a 1kb-Plus DNA ladder (Invitrogen, USA).

2.1.3.3 Primer design

All primers for each gene were designed based on the published *M. smegmatis* mc²155 sequence (GenBank accession number NC_008596) with the aid of Geneious Pro (Version 10.2.3) (BioMatters Ltd, NZ). Where possible, primers were designed to have similar properties so that PCR's for the amplification of different genes could be performed in the same run. Primers for qPCR were designed to amplify a product greater than 100bp but smaller than 200bp. Primer dimers and secondary structures were avoided where possible. All primers were run through the Basic Local Alignment Search Tool (BLAST, NCBI) which is a search engine that finds regions of local similarity between sequences to identify possible primers that might produce non-specific banding through similarities with other regions in the *M. smegmatis* genome. All primers were obtained from IDT (USA) and they were supplied purified with standard desalting techniques and were reconstituted in MQ H_2O to a concentration of 100µM with working stock concentrations prepared at 10µM in MQ H_2O.

2.1.3.4 DNA Quantification

DNA was quantified using a Nanodrop ND-1000 Spectrophotometer (Nanodrop Technologies, USA). This measures absorbance of DNA at 260nm, quantifying the amount of DNA. DNA purity and concentration was estimated by measuring A260/A230nm and A260/A280nm ratios using a NanoDrop®ND-1000 Spectrophotometer (NanoDrop Technologies, USA).

2.1.3.5 PCR

All PCR reactions were performed on the DNA Engine, Peltier Thermal Cycler (Bio-Rad, USA). PCR annealing temperature and magnesium concentration were

optimised for each primer set using a temperature gradient designed between 5°C above and below the calculated Tm and a selection of $MgCl_2$ concentrations between 0.5mM and 1.5mM. The KapaHiFi (Roche, USA), HotFirePol® and HotFirePol® Blend (Solis BioDyne, USA) DNA polymerase hot start systems were used to amplify PCR products from DNA templates. PCR reactions were carried out in 20-25µL volumes. For reaction set up, reagents were thawed on ice. When using the KapaHiFi PCR Kit all reactions were set up on ice, since the high proofreading activity of the enzyme will result in rapid primer degradation at room temperature. While when using HOT FIREPol® and HOTFIREPol® Blend DNA polymerase Kits all reactions were set up at room temperature. The HOTFIREPol® Blend PCR kit is a premixed ready to use solution containing all reagents required for PCR (except template, primers and water). Once all reagents were properly thawed and mixed a PCR master mix was set up. The PCR master mix contained the appropriate volume of all reaction components in regards to the number of reactions to be preformed. The required volumes for each component were calculated based on the following table (Table 2.1).

Table 2. 1: PCR kits used for various PCR reactions

Hot FirePol® DNA Polymerase	**KAPA HiFi™ HotStart**	**HOT FIREPol® Blend Master Mix**
1x Buffer B1	1x KAPA 5x HiFi GC Buffer	1x HOT FIREPOL® Blend Master Mix Ready to Load
0.05 U HOT FIRE Pol	1 U KAPA HiFi Hotstart DNA polymerase	10 µM Forward Primer
2.5 mM $MgCl_2$	10 mM KAPA dNTP Mix	10 µM Reverse Primer
200µM dNTP Mix	10 µM Forward Primer	Up to 20µl of nuclease free water.
10 µM Forward Primer	10 µM Reverse Primer	1-10 ng DNA template
10 µM Reverse Primer	1-10 ng DNA template	
1-10 ng DNA template	Up to 20µl of nuclease free water.	
Up to 25µl of nuclease free water.		

All components were added to the master mix except for primers and template. The master mix was then gently mixed and centrifuged. The appropriate volumes of PCR master mix, template and primers were transferred to individual PCR

tubes. One PCR reaction that did not contain the template; was used as the negative control. All reactions were caped, mixed and centrifuged briefly. All PCR reactions were performed using cycling conditions in Table 2.2.

Table 2. 2: PCR cycling conditions, performed on the BioRad PCR cycler

Step	Temperature	Duration	Cycles
Initial denaturation	95^{o}C	3 or 12 minutes[1]	1
Denaturation	95^{o}C	20 seconds	35
Annealing	- [2]	15 seconds	
Extension	72 ^{o}C	15-60 seconds/ kb	
Final extension	72 ^{o}C	1 minute/ kb	1

[1]Intial denaturation for KAPAHiFi PCR kits was 3 minutes, and 12 minutes for HOT FIREPol ® and HOTFIREPol® Blend PCR kits.

[2]Annealing temperature depended on Tm of primers.

PCR's were purified as described below in section 2.1.3.6 and resulting DNA products were quantified as described in section 2.1.3.4.

2.1.3.6 Purification of PCR products from solution or Agarose Gel

A commercial Ultra Clean PCR Clean-Up Kit (Qiagen, Netherlands) was used to purify PCR reaction products. PCR products were purified from solution if there was no obvious contamination or no non-specific banding occurring in the reaction. This was judged by running 2µl of the PCR reaction on a 1% agarose gel, with 1x SYBR Safe™ DNA gel stain and observing band formations under a blue light box (Invitrogen, USA) gel imager. PCR products that contained non-specific banding, surrounding the correct bands, were run out on an agarose gel and viewed under a UV light box and the correct PCR bands were cut out using a sterile scalpel blade cleaned with 70% (v/v) ethanol and the gel pieces were placed inside a sterile 1.5ml eppendorf tube. If the gel image showed no contamination, primer dimmers or non-specific banding, the PCR reactions were pooled and purified using an Ultra Clean PCR Clean-Up Kit (Qiagen, Netherlands). Both solution and agarose PCR product purification could be obtained using the same Qiagen kit, with instructions for each procedure followed according to manufacturer. DNA was eluted in 30µl of Nuclease free water.

2.2 Methods relating to Chapter Three: Biochemical characterisation of PhoH2 proteins

2.2.1 Large Scale Protein Expression

2.2.1.1 Streaking plates

Glycerol stocks of *E. coli BL21* strains containing the plasmid *pET28b-PtsI* PhoH2 were streaked out into LB agar plates supplemented with Kanamycin (50µl/ml). Plates were incubated at 37^{o}C overnight.

2.2.1.2 Seeder cultures

Expression of PhoH2 protein was achieved using LB liquid media supplemented with kanamycin. A sterile loop was used to pick one colony of the desired *E. coli* strain. The colony was then used to inoculate 10ml of LB liquid media with 50µg/ml final kanamycin, which was incubated at 37ºC, 200 rpm overnight and positioned horizontally to achieve maximum aeration.

2.2.1.3 Induction of Expression Cultures

Ten ml of seeder culture was used to inoculate 1L (1:1000) of sterile LB media supplemented with Kanamycin (50µg/ml) and incubated in a shaking incubator at 37ºC, 200 rpm until an OD_{600} was between 0.4-0.6. Cells were induced for expression with 1ml of 1M Isopropyl β-D-thiogalactopyranoside (IPTG), to induce PhoH2 protein expression. The culture was grown in a baffled flask at 37ºC, 200 rpm for 24 hours.

2.2.1.4 Harvesting Cells from Expression Cultures

Cells from large scale expression cultures were pelleted by centrifugation at 4600 rpm, 4ºC for 20 minutes. The supernatant was discarded and pellets were stored at -80ºC.

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

SDS-PAGE gels were prepared in a Hoefer® multiple, mini gel casting system SE275 (Hoefer® Inc, USA). SDS-PAGE gels (pH 8) consisted of; 8.8, 10%, 12%, 15%, 16.5% or 18% acrylamide resolving gel (depending on the protein size) overlaid with a 5% acrylamide stacking gel (pH 6.8) (Table 2.3 and Table 2.4). PhoH2 protein is 48.7kDa and was resolved on 12% gels. All SDS-PAGE gels were made up with 30% acrylamide/bis solution with a ratio of 37.5:1 (Bio-Rad Laboratories, USA), and included 0.1% (w/v) filtered SDS, 1.5M Tris (pH 8.8) in the resolving gel and 1M Tris (pH 6.8) in the stacking gel and were polymerised by the addition of 0.05% (w/v) ammonium persulphate and 0.05% (v/v) TEMED.

Table 2. 3: Reagents required for Resolving Gels at a different %

Volumes in ml	**10%**	**12%**	**15%**	**16.5%**	**18%**
dd H2O	12.05	10.05	7.05	5.55	4.05
30% acrylamide	10.00	12.00	15.00	16.50	18.00
Resolving Buffer(1.5M Tris pH 8.8)	7.50	7.50	7.50	7.50	7.50
10% SDS w/v	0.30	0.30	0.30	0.30	0.30
10% AOS w/v	0.15	0.15	0.15	0.15	0.15
TEMED	0.015	0.015	0.015	0.015	0.015
Total	30.015	30.015	30.015	30.015	30.015

Table 2. 4: Reagents required for 5% Stacking Gel

Volumes in ml	**5%**
dd H2O	8.5
30% acrylamide	2.125
Stacking Buffer (1.0M Tris pH 6.8)	1.6
10% SDS w/v	0.125
10% APS w/v	0.063
TEMED	0.0063
Total	~12.5

Samples for SDS-PAGE analysis 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 prior to loading onto gel. Gels were run in 1x SDS-PAGE running buffer (25mM Tris, 250mM glycine, 0.1 % (w/v) SDS) at constant 70V until the dye front reached the resolving gel and then at 150V until the dye front reached the end of the gel. Ten µl of a Precision Plus Protein™ Unstained Standard (Bio-Rad Laboratories, USA) was run alongside each gel and used as a molecular weight marker.

2.2.1.6 Coomassie Blue Staining for Protein Gel Electrophoresis

Protein electrophoresis gels were stained by colloidal coomassie staining using a modified version of the quick stain method by Wong *et al*. in 2000 [83], with fairbanks staining solutions A and D. Gels were placed in a microwaveable box with 50ml of Fairbanks staining solution A (0.05 % (w/v) coomassie blue R-250, 25% (v/v) isopropanol, 10% (v/v) acetic acid), and then microwaved for 30 seconds on full power, then cooled to room temperature with gentle shaking. The stain was then removed and gels were rinsed gently with tap water and 50ml of Fairbanks staining solution D-10% (v/v) acetic acid, solution D being the destaining solution, was added and the gel was microwaved for 30 seconds on full power and cooled to room temperature with gentle shaking. A rolled up tissue was placed in the warmed Fairbanks D solution to assist destaining. Protein sizes were estimated by comparison with the protein sizes of the Precision Plus Protein™ Unstained Standard (Bio-Rad Laboratories, USA).

2.2.1.7 Purification of His-tagged proteins using IMAC/Nickel chromatography

A 5ml HiTrap™ Chelating HP column (GE Healthcare, UK) connected to either an NGC™ Chromatography System (Bio-Rad, USA) or an ÄKTA Basic™ FPLC system (GE Healthcare, UK) was used to purify proteins by immobilised metal affinity chromatography (IMAC). Elution of proteins was monitored by absorbance at 280nm. Prior to purification columns were primed with 5ml 100mM $NiCl_2$ and washed with 20ml water, followed by 20ml of lysis buffer that

was used in the purification. Columns were stripped of Ni^{2+} ions using 10ml 100mM EDTA pH 8.0 and re-primed between purifications. Cell pellets from large scale expression cultures (1L), that contained the desired protein were removed from the -80 °C freezer and placed on ice to thaw. Once the pellets had thawed they were resuspended in 25ml lysis buffer (50mM Tris pH 8, 200mM NaCl, 20mM imidazole, 1mM EDTA and 1mM DTT) with the addition of an EDTA free protease inhibitor tablet (Roche Applied Science, Switzerland). The cells were sonicated on ice using a large probe for 30s bursts at setting 4.5-5 with 1 min rests in between, on a Misonix Sonicator (USA) until cells were lysed (3min). The lysate was spun (13,000 rpm, 20 min, 4°C) and the supernantent was filtered through 1.2, 0.45 and 0.2µm Minisart filters (Sartorius AG, Germany). If the column was to be connected to an AKTA basic™ FPLC system (GE Sciences), the filtered supernatant would be manually loaded onto a 5ml HiTrap™ Chelating HP column, with a syringe, following pre-equilibration with the appropriate lysis buffer in a 'drop-to-drop' method. The column was attached to an ÄKTA Basic™ FPLC system. The column was washed with 15 - 25ml of lysis buffer, at a flow rate of 1ml/min to remove any excess unbound proteins. Bound proteins were eluted by running a gradient of 0-50% elution buffer (50mM Tris pH 8, 200mM NaCl, 1M imidazole, 1mM EDTA and 1mM DTT) over 50ml at a flow rate of 1ml/min. When IMAC was performed using a NGC™ Chromatography system the filtered supernatant is automatically loaded onto the 5ml HiTrap™ Chelating HP column before attachment to the machine. The sample pump line pumps the supernatant into a 5ml loop where it is then loaded onto the column at 1ml per minute. The remaining purification steps follows the same process as the AKTA Basic™ FPLC system. Fractions containing the protein of interest were purified further by size exclusion chromatography (SEC) using a Superdex column (GE Healthcare, UK). Aliquots of fractions (15µl) containing the protein of interest from both IMAC were added to 5µl of 4 x SDS loading buffer for analysis with 12% SDS-PAGE gel (section 2.2.1.5)

2.2.1.8 Analytical and preparative size exclusion chromatography (SEC) fast protein liquid chromatography (FPLC)

Following IMAC chromatography, Size exclusion chromatography was used to further purify the desired protein. Size exclusion chromatography was performed

using either a HiLoad™ 16/60 Superdex™ 200 column or a Superdex™ 200 10/300 GL column (depending on the amount of protein to be purified). All columns were supplied by GE Healthcare, Sweden. Columns were either attached to the NGC™ Chromatography System (Bio-Rad, USA) or the AKTA Basic™ FPLC system (GE Sciences). Both purification systems were washed with MQ H_2O prior to column attachment. The HiLoad™ 16/60 Superdex™ SEC column was prepared by washing with 25ml of 0.2 μm filtered MQ H_2O and equilibrating with 25ml of SEC buffer (50mM Tris pH 8.0, 200mM NaCl, 1mM EDTA and 1mM DTT), at a flow rate of 1ml/min. The Superdex™ 200 10/300 GL column SEC column was prepared by washing with 130ml of 0.2μm filtered MQ H_2O and equilibrating with 130ml of SEC buffer (50mM Tris pH 8.0, 200mM NaCl, 1mM EDTA and 1mM DTT) at a flow rate of 0.3ml/min, as an overnight method. IMAC fractions containing the desired protein were pooled and concentrated to between 1 and 5ml and filtered with a Nanosep® MF 0.2μm filter (Pall, USA) before loading onto a pre-equilibrated loop. PhoH2 protein elution peak retention volume had been predetermined by Dr Emma Andrews from the Proteins laboratory, at the University of Waikato. Following elution of the protein by size, fractions were collected in 0.5ml volumes, and subsequently run on 12% SDS-PAGE gels (section 2.2.1.5) to identify desired protein and their corresponding chromatograph peaks. After protein separation, SEC columns were washed with filtered MQ H_2O and rinsed with 20% ethanol with the appropriate volumes and flow rates as mentioned above.

2.2.1.9 Anion exchange chromatography

Anion exchange chromatography (AEX) was performed with a HiTrap™ Q 5ml column attached to an ÄKTA™ Basic FPLC system. Fractions that contained the desired protein, from a SEC run, were pooled together and diluted 2x the amount of sample, with binding buffer QA 50mM Tris pH8, 50mM NaCl, 1mM EDTA and 1mM DTT. The dilute sample was then loaded onto a HiTrap Q 5ml column, which was pre equilibrated with MQ water and buffer QA. The column was attached to an ÄKTA™ basic, which was pre equilibrated with MQ water and QA buffer. Bound proteins were eluted by running a gradient of 0-50% elution buffer 50mM Tris pH 8, 1M NaCl, 1mM EDTA and 1mM DTT. Fraction peaks were analysed using 12% PAGE SDS gels following method described in section 2.2.1.5.

2.2.2 Extraction of *E.coli* host nucleic acid from purified PhoH2 proteins

Precipitated PhoH2 protein from an IMAC or a SEC purification was mixed with 10ml of lysis buffer (1x TE buffer, 0.5% (w/v) SDS and fresh proteinase K (1mg/ml)) and incubated for 1 hour at 37^{o}C. * An equal volume of acid phenol/chloroform was then added and solution was mixed. The solution was then centrifuged at 1300 rmp for 5 minutes at RT. The upper aqueous phase was transferred to a new tube and the process from * was repeated twice. An equal volume of chloroform was added to the aqueous layer, to remove phenol. This sample was then centrifuged at 1300 rmp for 5 minutes at RT. Aqueous layer was removed to a new tube and 3 times the volume of 4^{o}C ethanol was added and mixed gently. The sample was incubated at -20^{o}C for 30 minutes. The sample was then spun at 1300 rmp for 15 minutes at 4^{o}C. The supernatant was then discarded and the pellet was resuspended with 1ml of 70% ethanol. The sample was then spun at 1300 rmp for 2 minutes at 4^{o}C and the supernatant was discarded and the pellet was air dried on ice for 5 minutes. The pellet was then resuspened with 15µl of nuclease free water. Nucleic acid concentration was examined by Nanodrop as described in section 2.1.3.4 and the sample was run out on a 1% agarose gel as described in section 2.1.3.2.

2.2.3 Robotic Crystallisation Screens

Crystallisation trials for PhoH2 were set up using the high throughput mosquito® Crystal robotic system (TTP Labtech Ltd, UK), using PEGRx HT, Crystal Screen HT, Index HT and SaltRx HT crystallisation screens (Hampton Research, USA). Aliquots of each screen condition (100µl) were manually pipetted into either a 96 well Intelli-plate (Hampton Research, USA) or triple drop 96-well plate (TTP Labtech Ltd, UK). Sitting drops with a 1:1 ratio of PhoH2 (5mg/ml) and reservoir solutions were robotically set up in the crystallisation well. All plates were sealed with ClearSealTM Film (Hampton Research, USA) and stored in a temperature controlled crystallisation room at 18°C.

2.2.4 Protein concentrating and quantification

Protein samples were concentrated using 20ml, 2ml or 500µL Vivaspin concentrators (GE Healthcare, Sweden) with molecular weight cut offs of 10 kDa. Concentrators were pre-equilibrated with 0.05% (v/v) Tween-80 and protein purification buffer prior to adding protein to help prevent protein adherence. Protein samples were added to the upper reservoir of a pre-equilibrated concentrator and spun at 3000g (20ml), 3600g (2ml) or 9000g (500µL) at 4°C until the desired volume or concentration was achieved. Protein concentration was quantified using a Nanodrop ND-1000 spectrophotometer (Nanodrop Technologies, USA). The Nanodrop measures the absorbance of protein at 280nm and the accompanying software calculates protein concentration in correlation with absorbance using the Beer's Law equation, $A = \varepsilon\ c\ \ell$, where A is the absorbance at 280nm, ε is the theoretical molar extinction coefficient of the protein ($M^{-1}\ cm^{-1}$), c is the concentration (M), and ℓ is the pathlength (cm). The theoretical molar extinction coefficients were calculatedusing the online tool ProtParam (http://web.expasy.org/protparam), by providing the amino acid sequence.

2.2.5 2M NaCl treatment of IMAC purified PhoH2 Protein

Fractions from an IMAC purification containing the desired protein were pooled together and added to an equal amount of SEC buffer (50mM Tris pH 8.0, 1mM EDTA and 1mM DTT, except with 4M NaCl in the SEC buffer. Protein was stored at -4^{o}C overnight, on a plate shaker. Protein was collected the next morning and spun briefly in a centrifuge to collect precipitated pellet to the bottom. Non-precipitated protein was filtered through a Nanosep® MF 0.2µm filter before being loaded into a 5ml loop connected to a NGC™ Chromatography System, with a HiLoad™ 16/60 Superdex™ 200 column attached. SEC purification would proceed using the same method as in section 2.2.1.8, but in a high salt (2M) SEC buffer, to dissociate the co-purified *E. coli* nucleic acid.

2.2.6 RNase treatment of IMAC purified PhoH2 Protein

RNase treatment protocol has been adapted from [140], with concentrations adjusted for different proteins. Fractions from an IMAC purification containing the desired protein were pooled together and concentrated as in section 2.2.1.12. Protein was concentrated to 2.5ml with a Nanodrop reading of 1.4mg/ml. 1.25ml of protein was placed in a sterile eppendorf tube. 225µl of 20mg/ml RNase A solution was added to the protein. The solution was mixed and incubated at 37^{o}C for 30 minutes at 200 rmp, to activate the RNase. The solution was briefly centrifuged to collect any precipitated protein to the bottom of the tube. Non-precipitated protein was filtered through a Nanosep® MF 0.2µm filter before being loaded into a 5ml loop connected to a NGC™ Chromatography System, with a HiLoad™ 16/60 Superdex™ 200 column attached. SEC purification would proceed using the same method as in section 2.2.1.8. After SEC the fractions containing the desired protein were purified by anion exchange to ensure removal of any contamination from RNase A and *E. coli* nucleic acids.

2.2.7 Heat treatment of IMAC purified PhoH2 Protein

Fractions from an IMAC purification containing the desired protein were pooled together and concentrated as in section 2.2.1.12. Protein was concentrated to 2.5ml with a Nanodrop reading of 1.4mg/ml. 1.25ml of protein was placed in a sterile eppendorf tube. Protein was then incubated at 37^{o}C, for 1 hour at 400 rmp. After 30 minutes the tube was removed and the solution was briefly centrifuged to collect any precipitated protein to the bottom of the tube. Non-precipitated protein was filtered through a Nanosep® MF 0.2µm filter before being loaded into a 5ml loop connected to a NGC™ Chromatography System, with a HiLoad™ 16/60 Superdex™ 200 column attached. SEC purification would proceed using the same method as in section 2.2.1.8.

2.2.8 Protein activity assays

2.2.8.1 Generation of positive control oligonucleotide substrates

Self-annealing RNA oligonucleotides were designed and ordered from IDT (Custom Science, USA) by Dr Emma Andrews (at the University of Waikato, proteins and microbes laboratory).

Oligonucleotide name	Oligonucleotide sequence (5' to 3')
5'5	ACAUCUACAUCAGAGUGCGCACUC
5'3	ACUUGUACAUCAGAGUGCGCACUC

The RNA-oligonucleotides were made up to a final working concentration (50µM) in diethylpyrocarbonate (DEPC) or nuclease free water. Lyophilised RNA oligonucleotide stocks were resuspended in nuclease free water to 100uM. This was vortexed and centrifuged then diluted 1:1 in filtered hybridisation buffer (20mM Tris-HCl pH 7.5, 500mM NaCl, 1mM EDTA). The reactions were incubated in a PCR machine (BioRad) at 95^{o}C for 10 minutes then left to cool to room temperature overnight. RNA oligonucleotides were stored at -80°C.

2.2.8.2 Generation of mRNA substrates

Primers (Table 2.5 and Table 2.6) for mRNA substrates were designed as described in section 2.1.3.3.

Table 2. 5: Primers for round 1 PCR

Gene name	Primer name	Primer sequence (5' to 3')	Product size (bp)
SigF	SigF Fwd	GTG ACG TCG GAA TAC GCA GA	754
	SigF Rev	GCT ACT GCA GCT GGT CGC GC	
SigF + RsbW	RsbW + SigF Fwd	TGG CGG AAA CAC CCG CTC GG	1167
	SigF Rev	GCT ACT GCA GCT GGT CGC GC	
Upstream + RsbW + SigF	Up + RsbW +SigF Fwd	ACC GCC ACC GCG GGC CGT GT	1270
	SigF Rev	GCT ACT GCA GCT GGT CGC GC	
ChaB + RsbW + SigF	ChaB + RsbW + SigF Fwd	GTG CCG AAG ACG ACC CGT GA	1690
	SigF Rev	GCT ACT GCA GCT GGT CGC GC	
Upstream + ChaB + RsbW + SigF	Up + ChaB +RsbW + SigF + Fwd	CTT CTC ATC TCA CTG GCG CT	1890
	SigF Rev	CTA CTG CAG CTG GTC GC	
SigF fragment 1	SigF Fwd	GTG ACG TCG GAA TAC GCA	280
	SigF Rev	CTA CTG CAG CTG GTC GC	
SigF fragment 2	SigF 488 Fwd	TAC AAC ACG CTG TCC ATC GA	280
	SigF Rev	CTA CTG CAG CTG GTC GC	
SigF fragment 3	SigF 257 Fwd	TTC GTG TCG TTC GCC GTG CC	254
	SigF Rev	CTA CTG CAG CTG GTC GC	
Upstream + ChaB + RsbW	Up + ChaB +RsbW + SigF + Fwd	CTT CTC ATC TCA CTG GCG CT	1270
	RsbW Rev	TCA CCG CAG CAG GCT CGC	

Primers for PCR round 1 were used in a PCR reaction with either KapaHiFi (Roche,USA), HotFirePol® or HotFirePol® Blend (Solis BioDyne, USA) DNA polymerase hot start systems. PCR reactions were carried out as described in section 2.1.3.5.

PCR products from the first round of PCR were purified following protocol from section 2.1.3.6. Purified PCR product was run on an agarose gel to check the product was the right size and didn't contain any non-specific banding. Two µl of this PCR product was then used as a template for another round of PCR, which was carried out in the same manner as the above PCR reaction, except that the forward primers contained a T7 sequence, upstream of the original primer sequence (Table 2.6).

Table 2. 6: Primers for PCR round 2

Primer name	Primer sequence (5' to 3')
SigF Fwd +T7	TAA TAC GAC TCA CTA TAG GGG TGA CGT CGG AAT ACG CAGA
RsbW + SigF Fwd + T7	TAA TAC GAC TCA CTA TAG GGT GGC GGA AAC ACC CGC TCGG
Up + RsbW + SigF Fwd + T7	TAA TAC GAC TCA CAC TAT AGG GAC CGC CAC CGC GGG CCG TGT
ChaB + RsbW + SigF Fwd + T7	TAA TAC GAC TCA CTA TAG GGG TGC CGA AGA CGA CCC GTGA
Up + ChaB + RsbW + SigF Fwd + T7	TAA TAC GAC TCA CTA TAG GGC TTC TCA TCT CAC TGG CGCT
SigF 257 Fwd + T7	TAA TAC GAC TCA CTA TAG GGT TCG TGT CGT TCG CC
SigF 488 Fwd + T7	TAA TAC GAC TCA CTA TAG GGT ACA ACA CGC TGT CCA

PCR products from second round of PCR were purified following protocol from section 2.1.3.6. Purified PCR product was run on an agarose gel to check the product was the right size and didn't contain any non-specific banding. PCR product was then quantified using a Nanodrop, following the protocol from section 2.1.3.4.

2.2.8.3 Transcription reaction of PCR product to RNA

Once PCR products had been generated with a T7 promoter sequence upstream of the gene of interest, MEGAscript® transcription reaction kit from Thermo Fisher Scientific was used to generate mRNA substrates as below (Table 2.7).

Table 2. 7: Reagents required for transcription reaction

Amount	Component
To 20 µl	Nuclease-free water
2 µl	ATP solution
2µl	CTP solution
2 µl	GTP solution
2 µl	UTP solution
2 µl	10X Reaction Buffer
0.1-2µg	PCR product template
2 µl	Enzyme Mix

The reactions were incubated at 37^{o}C for 2 hours based on the manufacturer's guidelines. After incubation 1µl of TURBO DNase was added and incubated for 15 minutes at 37^{o}C to remove any remaining template PCR product. After treatment with DNase the RNA was cleaned up using sodium acetate ethanol precipitation. Two µl of 3M Na-acetate at pH 5.2 and 60µl of 100% ethanol (RNase free) were added to the RNA and incubated at -20^{o}C for 1 hour. The RNA was then centrifuged at 13,000 rpm for 20 minutes, at 4^{o}C. The supernatant was discarded and the remaining pellet was washed with 1ml of 70% ethanol (RNase free) and centrifuged at 13,000 rpm for 20 minutes at 4^{o}C. All traces of ethanol were removed and the pellet was resuspended in 30µl of nuclease free water. The RNA was stored in a -80^{o}C freezer.

2.2.8.4 RNA Quantification

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

2.2.8.5 Urea Denaturing Polyacrylamide Gel Electrophoresis Protein Analysis (Urea Denaturing-PAGE)

Urea denaturing-PAGE gels were cast in a Hoeffer gel casting system. The gels consisted solely of resolving gel (10, 15 or 20% (v/v) acrylamide depending on size of RNA to be analysed). All urea denaturing-PAGE gels were made up with 30% (w/v) acrylamide with an acrylamide:bisacrylamide ratio of 37.5:1 (Bio-Rad Laboratories, USA) and included 6 M urea and were polymerised by addition of ammonium persulfate (APS) (0.05% (w/v)) and TEMED (0.05% (v/v)). RNA samples mixed in a 1:1 ratio with 2x formamide loading dye were heated to 70°C

for 5 min before loading onto gel(s). A 10x TBE buffer (1M Tris, 1M boric acid, 0.02M EDTA, RNase free H_2O up to 1L) was made. Gels were pre-run in 1x TBE buffer at 50 V for 30 minutes. Wells were then flushed with 1x TBE and samples loaded. Gels were then run at 150 V for one hour or until the xylene cyanol dye front reached the end of the gel.

2.2.8.6 PhoH2 RNA unwinding and cleavage activity assay set up

Helicase and ribonuclease activity assays for PhoH2 were set up as follows:
RNA oligonucleotides were made up to a working concentration of 50 or 25uM. mRNA substrates were also made up to a working concentration of 50 or 25uM with RNase free water. IMAC and SEC purified PhoH2 protein (250uM or 50uM) was incubated with 5'3, 5'5 RNA oligonucleotides or mRNA substrates in a reaction containing 1mM ATP (unless stated otherwise) 6µl of assay buffer (50 mM TRIS pH 7.5, 20mM NaCl, 10mM $MgCl_2$) and were made up to 15ml with RNase free water. Assay reactions were initially at a ratio of 5:1 (protein:RNA) and then changed to 2.5:1 (protein:RNA). Individual assay reactions (Table 2.8) were set up for each time point to reduce the possibility of RNase contamination i.e. four time points, four assay reactions. Test reactions were set up over a time course 5, 15, 30 and 60 min, reactions were then incubated at 37^{o}C for the appropriate time. Negative controls; RNA only (0 hour and 1 hour) assay reactions were set up as above but no PhoH2 was added to the reaction, to ensure there is no RNase contamination of the assay buffer. A protein only control, was introduced to some assays and was set up the same as the reactions above, but without RNA. EDTA and no ATP negative control assay reactions were set up as above with 40mM EDTA in the assay buffer and/or no ATP, to ensure there is no RNase contamination of the protein purification. One reaction was set up for each control and incubated at 37°C, for the assay duration. All reactions were stopped with 10µl of 2x formamide stop solution (80% formamide (v/v), 5 mM EDTA, 0.1% (w/v) bromophenol blue, 0.1% (w/v) xylene cyanol FF) and heating at 70°C for 5 min. Reactions were analysed directly by Urea Denaturing PAGE on a 1x TBE-10% minigel as described in section 2.2.2.5 or on 1% agarose gels, with SYBR safe stain, as described in section 2.1.3.2.

Table 2. 8: Set up of assay reactions

Component	**Controls**			**Test**
	RNA only 2x (RNA from 0 time point, RNA after 60 minutes)	**No ATP x1**	**EDTA x1**	**Time series x4** (consisting of 5, 15, 30 and 60 minute incubations)
RNA substrate (25µM/50 µM)	[1]	[1]	[1]	[1]
PhoH2 protein (250 µM/50µM)	-	[2]	[2]	[2]
ATP (10mM) [3]	1µl	-	1µl	1µl
Assay buffer	6µl	6µl	-	6µl
Assay buffer + EDTA	-	-	6µl	-
RNase free water	To give a final volume of 20µl	To give a final volume of 20µl	To give a final volume of 20µl	To give a final volume of 20µl

[1] Amount of RNA in each reaction tube depended on the amount of RNA required for a final concentration of 25µM or 50µM.

[2] Amount of PhoH2 protein in each reaction tube (minus the RNA only control/s) varied depending on the amount of PhoH2 protein needed to give a final concentration of 250µM or 50µM.

[3] ATP concentrations for standard reactions were at a final concentration of 1mM, unless specified otherwise.

2.2.8.7 Malachite Green ATPase assays

Two different samples of PhoH2 protein were used in the assays. The first sample was heat treated PhoH2 that was purified by SEC (section 2.2.1.8) and the second sample was RNase treated PhoH2 that was purified by SEC (section 2.2.1.8) and anion exchange chromatography (section 2.2.1.10). Assays were set up on master block 0.5ml plates wells. Three reactions were set up for each sample. The first reaction was a control with no protein. The second reaction contained PhoH2 protein and the last reaction used K_2HPO_4 instead of protein. ATP (1mM) and assay buffer (20mM Hepes pH 7.4, 20mM NaCl, 10mM MgCl) were pre-equilibrated in a 96 well clear flat-bottom microplate for 10 minutes, at 37^{o}C. Silicon covers were placed over plates, to avoid evaporation. Protein (62.5µM) was then added to test reactions and K_2HPO_4 (200µM) was added to its control reaction. Plates were then further incubated for 10 minutes at, at 37^{o}C. Then

addition of 30µl of stop solution (1mM malachite green, 8.5mM ammonium molybdate supplemented with 0.1% Triton X-100) was added to the negative control, test and K_2HPO_4 positive control reaction wells, for each protein sample (Table 2.9). Reagents were mixed through pipetting the solution up and down.

Table 2. 9: Set up of Malachite Green ATPase assay in microplate

	Control:	**Test:**	**K_2HPO_4 control:**
Heat treated PhoH2:	90µl Hepes assay buffer and10µl ATP	80 µl Hepes assay buffer, 10µl ATP and 10µl protein.	10µl 1mM ATP, 80µl Hepes assay buffer and 10µl K_2HPO_4
RNase treated PhoH2:	90µl Hepes assay buffer and 10µl ATP	80 µl Hepes assay buffer, 10µl ATP and 10µl protein.	10µl 1mM ATP, 80µl Hepes assay buffer and 10µl K_2HPO_4

2.3 Methods relating to chapter four: Gene expression of mc²155 Δ*phoH2* compared to mc²155 in *M. smegmatis* using qPCR

2.3.1 Growth of *M. smegmatis* mc²155 and *M. smegmatis* mc²155 *ΔphoH2* cultures for DNA or RNA extractions

M. smegmatis mc^{2}155 and *M. smegmatis* mc^{2}155 *ΔphoH2* glycerol stocks were provided by Dr Emma Andrews from the proteins laboratory at the University of Waikato. These glycerol stocks were streaked out onto Luria-Bertani (LB) plates supplemented with 0.05% tyloxapol and incubated at 37^{o}C for 3 days, in a partially sealed container with damp paper towels. Starter cultures were set up in LB media with 0.05% tyloxapol and a single colony from the LB tyloxapol plates. These starter cultures were incubated at 37^{o}C and grown planktonically in a shaker at 200 rpm for 24 hours. After 24 hours the OD_{600} of each sample was measured with a spectrophotometer (Thermofisher scientific Multiskan Go, version 1.00.40), using a single wavelength set at 600nm and the volume calculated to seed 100ml cultures with a starting OD_{600} of 0.01. Cultures were grown in LBT media at 37^{o}C, 200 rmp for 48 hours. After 48 hours of growth cells were immediately added to falcon tubes containing the appropriate amount of fresh 5M guanidium isothiocynate (GITC) pH7 for a DNA extraction or an

RNA extraction. Cells were then pelleted by centrifugation and the supernantent was removed. The remaining pellet was resuspended with the appropriate amount of GITC for RNA or DNA extractions and added to 2ml screw cape tubes that contained around 0.3g of 0.1mm and 2.5mm zirconia beads. The cells were either stored in the -80^{o}C freezer for further use or placed on ice ready to be used for an RNA or DNA extraction.

2.3.2 *M. smegmatis* mc²155 and *M. smegmatis* mc²155 ΔPhoH2 total RNA extraction

Total RNA was extracted from *M. smegmatis* mc^{2}155 and *M. smegmatis* mc^{2}155 *ΔphoH2* cultures, using an extension of the method by Ali Ruthe, Proteins & Microbes Laboratory, University of Waikato developed by Dr Ray Cursons and Dr Emma Andrews.

For the preparation of total RNA from *M. smegmatis*, mc²155 and *ΔphoH2* mc²155 strains were cultured in LBT broth for 48 hours at 37°C. After 48 hours of growth the cells were added to fresh 5M GITC pH 7, with 20ml of GITC being added to every 5ml of cells, and pelleted by centrifugation in 50ml sterile falcon tubes. After centrifugation the supernatant was added to a 2ml screw cap tube that contained around 0.3g of 0.1mm and 2.5mm zirconia beads. 500ul of fresh 5M GITC was added to the tubes. The cells were disrupted by three rounds of bead beating using the Fast Prep cell disrupter (FP120 Thermo Savant, USA). The first setting was 4, to warm up the machine, followed by 6.5 with three rounds of bead beating starting from 20 seconds and increasing by five seconds each time. Between each round the cells were left to cool for one minute. Samples were then centrifuged for 10 seconds to decrease foam.* Tubes were set up on ice and 50ul of 2M sodium acetate pH was added and gently mixed. Samples were then taken to a fume hood, where an equal volume, to the sample volume, of water-saturated acid phenol (pH 4) was added, vortexed and then rotated for 10 minutes at room temperature. Tubes were put back on ice and 100ul of 1-bromo, 3-chloro propanate was added. The samples were then shaken for 1 minute, followed by 5 minute incubation on ice. After 5 minutes the samples were centrifuged at 13000 rpm for 5 minutes to separate out the phases. Addition of water-saturated acid phenol and 1-bromo, 3-chloro propanate results in an upper aqueous phase, while protein partitions into the organic phase (containing the proteins dissolved in

phenol and lipids dissolved in 1-bromo, 3-chloro propanate). The top layer was removed and added to new 1.5ml sterile tubes and the method would be repeated from the * with the volume of acid phenol adjusted accordingly. Once this process was repeated at least once, the top layer was removed and added to a new sterile 1.5ml tube with an equal volume of 100% isopropanol added. Samples were then stored at -40^{o}C overnight. Samples were then spun down in a centrifuge set for 15 minutes, 13000 rpm at -4^{o}C. RNA is insoluble in alcohol, aggregating together and giving a pellet upon centrifugation. After centrifugation all traces of isopropanol were removed and the pellet was washed with 70% RNase free ethanol and spun down for 15 minutes, 1300 rpm at -4^{o}C. All traces of the ethanol were removed and the pellet was resuspended in 100ul of RNase free water and 300ul of viral extraction, with the side of the tube scraped by the pipette tip to ensure all traces of RNA were resuspended. The samples were then placed on a rotor wheel for 5 minutes. 400ul of isopropanol was added and the samples were incubated for 10 minutes at room temperature, with occasional inverting of tubes. The samples were then spun for 15 minutes at 13000 rpm at room temperature. Isporopanol was then removed and the pellet was washed with 1ml of 70% ethanol by resupsending pellet. The samples were then spun for 5 minutes at 13000 rpm at room temperature. All traces of ethanol were removed from the samples and the pellet was dissolved in 200ul of RNase free water and 200ul of 5M $LiCl_2$. Samples were then incubated at -20^{o}C for 1 hour. After an hour the samples were spun at 13000 rpm for 15 minutes at 4^{o}C. Supernatant was removed and the pellet was resupsended in 100ul of RNase free water. 10ul of 3M sodium acetate pH 5.2 and 275ul of ethanol was also added to the samples and then incubated at -20^{o}C for 10 minutes. After 10 minutes the samples were spun down at 13000 rpm for 15 minutes at 4^{o}C. The supernatant was then removed and the pellet was washed with 1ml of 70% ethanol by resuspending the pellet gently through pipetting up and down. The samples were then spun at 13000 rpm for 5 minutes at 4^{o}C and afterwards all traces of ethanol were removed. The pellet was then air dried for 5 minutes on ice and then resuspended in 25ul of RNase free water.

2.3.3 RNA purity and quantity assessment

Once the RNA is extracted it is important to check its quantity and quality. RNA was quantified using a Nanodrop ND-1000 Spectrophotometer (Nanodrop Technologies, USA). This measures absorbance of RNA at 260nm, quantifying the amount of RNA. The quality of RNA was examined by running 2µl of the RNA sample on a 1% agarose elecetrophoresis gel bands are detected by SYBR safe stain.

2.3.4 DNase treatment of total RNA from *M. smegmatis*

M. smegmatis mc²155 total RNA and *M. smegmatis* mc²155 *ΔphoH2* total RNA, 10x reaction buffer, PerfeCTa® DNase I and 10x stop buffer (all from Quanta Biosciences) were thawed on ice. The DNase master mix contained the following: (per sample)

- PerfeCTa® DNase I (1ul)
- 10x Reaction Buffer (2ul)
- DEPC-treated water (6.5ul)

A master mix was made up of the above reagents, based on the number of RNA samples being treated. Four µg of RNA was made up to 11µl with DEPC treated water. Master mix was added to each of the RNA samples, bringing the total volume to 20µl. The reaction mixture was gently vortexed and centrifuged Reactions were incubated for 30 minutes at 37^{o}C. Reactions were stopped with 1µl of 10x Stop Buffer and the solution was gently vortexed and centrifuged. The samples were then incubated for 10 minutes at 65^{o}C. After incubation, the samples were stored on ice for the immediate use to generate cDNA.

2.3.5 cDNA synthesis

qScript™ XLT cDNA SuperMix, from Quanta Biosciences was used to produce cDNA from total RNA. The supplied 5x concentrated master mix provided all necessary components (except RNA template) for first-strand synthesis including: buffer, dNTPs, $MgCl_2$, primers, RNase inhibitor protein, qScript XLT reverse transcriptase and stabilizers. Ten µl of *M. smegmatis* mc²155 and *M. smegmatis*

ΔphoH2 mc²155 DNase treated total RNA, 4µl qScript™ XLT cDNA SuperMix and 6µl of DEPC treated water were gently vortexed and centrifuged. Reactions were incubated in the DNA Engine, Peltier Thermal Cycler (Bio-Rad, USA), following these cycling conditions:

- 5 minute at 25^{O}C
- 60 minutes at 42^{O}C
- 5 minute at 85^{O}C
- Hold at 4^{O}C

The cDNA was stored at -20^{o}C until use.

2.3.6 Primers for qPCR

Primers (Table 2.10) were designed and ordered as described in section 2.1.3.3.

Table 2. 10: Primer sequences used for quantitative PCR

Primer sequence (5' to 3')			
Gene name	**Product size (bp)**	**Forward**	**Reverse**
MSMEG_6384	177	CCG CAA GTA CGA GGA GAT CC	GAG TCG ACG TCG GTC TCT TC
MSMEG_4712	130	GGT TCA CCC TCA CCT GAC AG	GTC GTT GCA GGT TGA CGA AC
MSMEG_0064	164	CTG GAG GCC ACG AAC TTT CT	ATC GGC TTC ATC GGC ATG AT
MSMEG_1804	120	CGG CTC AAG GAA CTC CAC TT	CAC CTC TTC GCA ATC CAT GT
MSMEG_1773	130	TGA TCT CCG AGA TCA AAG GC	CGG TAG TGC AGT TCG TAG CA
MSMEG_0267	143	ATC GCG CGA TTC TCT ACC TG	CAG ACC GTA TTT GGG GAG CA
MSMEG_2415	154	TCG AAG AGG AAC ACA AGG CC	CGG GAG AGC TTG ACG AAC TC
MSMEG_1782	139	CAC CGA GAA GTT CAC CGA CA	GAT GGA GGC GGT GTT GAT GA
MSMEG_1097	158	GAT GTC CAC TTC GTG TCG GT	GTC ATC TCG ACG AGC CAG TC
MSMEG_6213	140	AGA AGA ACT GTC GCA CCT GG	TTA TCT CGT CCT TGC CGT CG
MSMEG_3419	199	AGT TCA TCG ACC GTG GTG C	CGC TCA TCG TAA TCC GGG AT
rrsA&B 16S	152	CAG CTC GTG TCG TGA GAT GT	AGA CCG GCT TTG AAA GGA TT
MSMEG_2758	182	CGT TCC TCA ACC TCA TCC AG	GAT CAC CTG GAC CAT CTG CA

All primers were initially tested with *M. smegmatis* gDNA and then cDNA using PCR. PCR was carried out under similar cycling conditions to that of a qPCR run (section 2.3.8). PCR for primer testing was performed on a DNA Engine, Peltier Thermal Cycler (Bio-Rad, USA) under the following PCR conditions:

1. 95^{o}C for 15 minutes (Initial activation)
2. 95^{o}C for 15 seconds (Denaturation)
3. 54^{o}C to 65^{o}C for 40 seconds (Annealing)
4. 72^{o}C for 1 minute (Elongation)
5. Go to step 2, 30 times
6. 72^{o}C for 8 minutes (Final elongation)
7. End

2.3.7 Quantative PCR Reaction set up

All qPCR reactions were carried out in triplicate 25µl volumes using SYTO 82 (Invitrogen, USA). SYTO 82 was stored as a 0.5mM working stock by adding 10µl of STYO 82 to 90µl of DMSO. The master mix (Table 2.11) was prepared fresh in a PCR cabinet, using nuclease free equipment. The master mix was treated with UV light for 5 minutes before addition of primers.

Table 2. 11: Master mix of reagents and their concentrations for a 1x qPCR reaction

Master mix	**1x (µl)**
Hot fire polymerase	0.1
10x Buffer B1	2.5
25mM MgCl2	1.5
10mM dNTPs	0.5
0.5 mM SYTO 82 dye (In DMSO)	4.8
DEPC water	14.1
Total	25

To x volume of master mix 0.5µl of 10mM forward and reverse primer was added. For each primer pair, three reactions were set up in addition to the three test reactions, two RNA only using *M. smegmatis* mc^{2}155 cDNA or *M. smegmatis* mc^{2}155 *ΔphoH2* cDNA and the third a no template control. All reactions were prepared in a PCR cabinet with nuclease free equipment. Twenty four µl of the reaction mixture was pipetted directly into Mic tubes set out on a loading block. One µl of mc²155 and mc²155 *ΔphoH2* cDNA was used. For the 16S reference

gene cDNA was diluted 1:120 with nuclease free water. One µl of 2mg/ml of *M. smegmatis* mc²155 and *M. smegmatis* mc²155 *ΔphoH2* RNA was used for a RT (RNA template) control, to check for gDNA contamination and 1µl of nuclease free water was used as a NT (no template) control, to check for general contamination. Once all reactions were loaded into the Mic tubes, they were fitted with Mic caps and tubes were placed in a Mic (Magnetic Induction Cycler) rotor with the tube tab in line with the marker located on the rotor label. Any surplus wells were filled with water to optimise temperature uniformity across the rotor. Run was started using Mic software.

2.3.8 Quantative PCR run

Real-time PCR was performed on the Mic qPCR cycler (Bio Molecular Systems, Australia). The reaction conditions were; 1 cycle of 95^{o}C for 15 minutes, followed by 40 cycles of 95^{o}C for 15 seconds, 54^{o}C for 15 seconds, 72^{o}C for 25 seconds and 80^{o}C for 10 seconds with data acquired on the 80^{o}C step in the yellow channel (530-555 nm) for SYTO 82. Immediately following amplification, a melt curve was performed consisting of a temperature range of 72^{o}C to 98^{o}C with the temperature increasing by 0.3^{o}C per second. The identities of the PCR products were verified by the melt curve profile and electrophoresis on a 1% agarose gel, as described in section 2.1.3.2.

2.3.9 Analysis of qPCR runs

Raw data from qPCR runs were first examined using Mic software to determine Cq values. The qPCR curves were displayed with the y-axis in log scale. The second derivative maximum was used as a starting point to locate a Window-of-Linearity (WoL). A cycle threshold was automatically set within the WoL. Cq values were determined by the WOL using the LinReg method. The LinReg PCR algorithm uses the region of exponential growth to calculate Cq values, while minimising variation among the dataset. Using the gradient of the line for each sample in the WoL, the reaction efficiency was also calculated. Once Cq and amplification efficiency values were determined for each reaction, the data was exported to an excel spread sheet. The spread sheet contained formulas for solving gene expression for a gene of interest (GOI) in a wild-type sample compared to a

knock-out sample, normalised against a housekeeping gene (HKG). For relative quantification, the comparative Ct method, also known as the ΔΔCt method [163], where Ct also stands for Cq, was the first method used to determine gene expression values (Equation 1). This method used the average of the triplicates values for each GOI and HKG, for both *M. smegmatis* mc^2155 and *M. smegmatis* mc^2155 *ΔphoH2* samples. The second method is a modified version of the ΔΔCt method, which does not use the average of the triplicates; instead it uses all the Cq values given for each GOI and HKG, but uses the same formula (Equation 1) to determine gene expression values.

Equation 1: Livak (ΔΔCt) formula

1.	***ΔCq*** *wild-type = Cq (HKG wild-type) − Cq (GOI wild-type)*
2.	***ΔCq*** *knock-out = Cq (HK knock-out) − Cq (GOI knock-out)*
3.	***ΔΔCq*** *= ΔCq (wild-type) − ΔCq (knock-out)*
4.	***R*** *= 2 -ΔΔCq.*

R = expression ratio of the gene of interest, normalised to the HKG.

The ΔΔCt assumes that the amplification efficiencies of both the HK gene and the target GOI are approximately equal and are at 100% efficiency.

The last method used to determine gene expression was a variation of Pfaffls method [89], where the values going into the formula (Equation 2) are not averaged, as each value will hold different reaction efficiency. The reaction efficiencies for all values were taken into account when determining gene expression of a wild-type versus a knock-out. The three expression values given at the end of the equation were averaged to give one overall gene expression value. Pfaffls formula is a modification of Wilkening and Bader [164] equation, which takes into account the PCR amplification efficiency for each reaction.

Equation 2: Pfaffl formula

1.	***ΔCq****(GOI) = Cq(GOI) wild-type – Cq(GOI) knock-out*
2.	***ΔCq****(HKG) = Cq(HGK) wild-type – Cq(HKG) knock-out*
3.	***R*** *= E(GOI) ^ ΔCq(GOI) / E(HKG) ^ ΔCq(HKG)*

R = expression ratio of the gene of interest, normalised to the HKG.

E = PCR amplification efficiencies of gene of interest (GOI) or HK gene

^ = denotes 'to the power'

3 Chapter 3: Biochemical characterisation of PhoH2 proteins

3.1 Introduction

PhoH2 is a protein that is made up of an N-terminal PIN domain and a C-terminal PhoH domain. All PIN domains display RNase activity and PhoH domains are helicases with predicted ATPase functionality [4]. Helicases are a class of enzymes that require the energy generated from ATP to separate two annealed nucleic acid strands [11]. Once double stranded nucleic acid strands are separated RNase's are able to access the single stranded RNA products and degrade them through their nuclease activity [40].

SF1, SF2 and AAA+ helicase families show examples of both 5' to 3' and 3' to 5' helicase directionality. SF3 families display only 3' to 5' activity and SF4/5 demonstrate 5' to 3' activity [87]. Structurally PhoH2 proteins contain conserved motifs that are common across all these families and therefore testing both helicase activity and directionality is required. Andrews and Arcus [4] investigated the activity of PhoH2 proteins from *M. tuberculosis* and *M. smegmatis* through strand displacement and degradation assays with RNA oligonucleotides that had either a 5' or a 3' overhang. They found that PhoH2 showed ATP/Mg^{2+} dependent RNA unwinding and cleavage activity, and that 5' tailed duplex RNA substrates with the terminal sequence of A C [A/U] [A/U] [G/C] are specific targets of PhoH2 proteins [4]. Once it was determined that PhoH2 proteins were targeting specific RNA substrates, efforts turned towards finding biological RNA targets of PhoH2 in *M. smegmatis*.

Dr Emma Andrews made a *phoH2* knock-out strain of *M. smegmatis* mc²155, where the gene for *phoH2* was deleted (*ΔphoH2*). Cultures of this strain, along with *M. smegmatis* wild-type strain; mc²155, were grown for 72 hours, under standard growth conditions (LB tyloxapol media) and RNA was extracted from 24, 48 and 72 hour time points of culture growth [personal communication]. Exacted RNA was sent to BGI (Beijing Genomics Institute) for transcriptome analysis

comparing *M. smegmatis* mc²155 *ΔphoH2* (knock-out) RNA and *M. smegmatis* mc²155 (wild-type) RNA. The 48 hour RNA data was examined in detail and it was discovered that there was an up regulation of *sigF* regulon genes in the knock-out RNA sample compared to the wild-type. Dr Emma Andrews hypothesised that PhoH2 may act as a regulator of SigF and therefore its regulon genes, through the PhoH2's RNA helicase and RNase activities [personal communication]. In *M. smegmatis*, *sigF* is part of an operon with its anti-sigma factor, *rsbW*, encoded upstream and a ChaB family protein encoded upstream of *rsbW* [46]. The arrangement of the *sigF* gene, with its anti-sigma factor up stream is highly conserved in mycobacteria [83]. SigF is transcribed from two different promoter regions. The first promoter is found preceding *rsbW* and is *sigF* independent, while the second promoter is found preceding *chaB* and is *sigF* dependent [46].

In order to test if PhoH2 proteins act on *sigF* mRNA nine RNA substrates were designed that spanned the *sigF* mRNA transcript (Figure 3.1). Substrates were designed to include fragments that span the *sigF* gene, to just upstream of the *chaB* gene. Three further shorter substrates (F1, F2 and F3) were designed that were made entirely from fragments of the *sigF* gene, to see if possible secondary structures in the *sigF* gene, would lead to increased PhoH2 activity. A final RNA substrate (upstream + *chaB* + *rsbW*) was designed that excluded the *sigF* gene, to test for activity in the absence of *sigF* to deduce if activity took place within the broader *sigF* mRNA transcript. These RNA substrates were made from PCR products that were amplified using gene specific primers from *M. smegmatis* gDNA. These PCR products were used as template in a second round of PCR that incorporated the T7 promoter at the 5' end. These PCR products were then transcribed into RNA using MegaScript transcription reaction kit and were used as substrates in unwinding and degradation assays with PhoH2 protein, to determine whether *sigF* mRNA is a direct target of PhoH2.

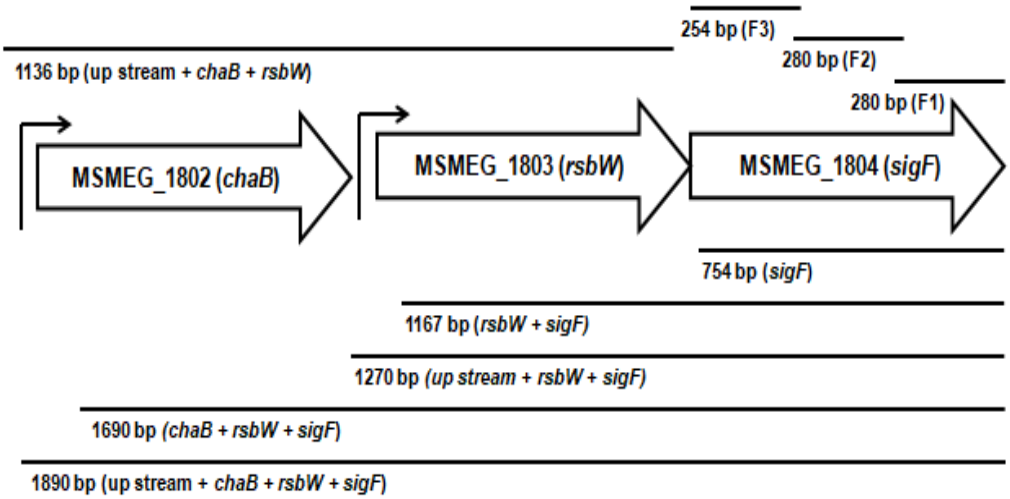

Figure 3. 1: Diagram of the *sigF* operon and surrounding genes. PCR products generated that span the *sigF* mRNA transcript (black lines) were transcribed into RNA and used as substrates for PhoH2 in RNA unwinding and degradation assays. The name of each RNA substrate is indicated under black line along with the size of each substrate. Black arrows indicate promoter regions.

The results from these activity assays will help to confirm whether PhoH2 protein is acting as a new regulator of SigF and its regulon genes and therefore playing a role as a central regulator of the stress response in *M. smegmatis*. If PhoH2 plays this role in *M. smegmatis* it is possible that it may also be playing a similar role in *M. tuberculosis* and this discovery could lead to new drug targets against PhoH2 from *M. tuberculosis*, which is the cause of tuberculosis in humans.

3.2 Results and discussion

3.2.1 IMAC and SEC purification of PhoH2 protein

His-tagged PhoH2 proteins were successfully purified by an IMAC purification as described in section 2.2.1.7. The following chromatogram and corresponding 12% SDS-PAGE gel in Figure 3.2 depicts the resulting peaks and fractions observed from IMAC purifications. With the purification of PhoH2 expressed in *E .coli*, non-specific proteins eluted early off the column at a low concentration of imidazole, whereas PhoH2$_{MSMEG}$ eluted later at an imidazole concentration of ~120-180mM and as a single peak. SDS-PAGE gels show the resulting IMAC fractions (A, B and C). Peak A represents proteins that have not bound to the His-trap column and contains a mixture of different proteins with different molecular weights, seen by the various bands, in the SDS-PAGE gel. Peak B represents a

smaller second peak that contains weak bands around 50kDA and a few bands smaller than 25kDa. Peak C represents the PhoH2 protein peak, which can be visualised in the SDS-PAGE gel as bright bands around 50kDa, the expected molecular weight of $PhoH2_{MSMEG}$ is 48.7kDa. In peak C there is also the presence of some smaller molecular weight bands.

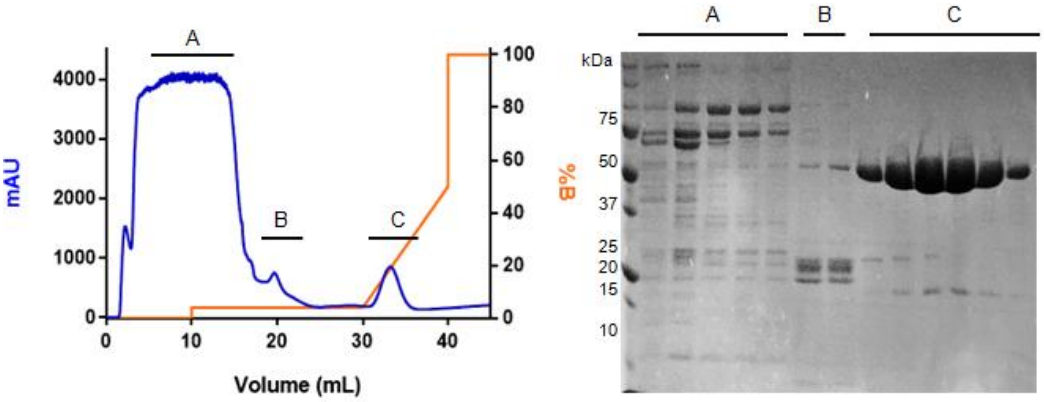

Figure 3. 2: IMAC purification and corresponding 12% SDS-PAGE gels. The chromatogram depicts the UV absorbance (blue) and elution profile of each protein as the concentration of elution buffer increases (orange). The corresponding SDS-PAGE gel depicts the peaks; A, B and C. Fractions and their elution position are shown by the black bar.

Fractions identified by IMAC as containing PhoH2, were further purified by SEC as described in section 2.2.1.8 and typically the elution profile (Figure 3.3) showed a double peak, sometimes with a third peak, eluting late. Peak A, elutes around 8ml, Peak B, elutes around 11ml and Peak C elutes around 25ml. Purified SEC PhoH2 samples were analysed on 12% SDS-PAGE gels (section 2.2.1.5), which show that Peak A contains clean bands around 50kDa, while peak B also contains clean bands around 50kDa, the bands are more intense in line with the increase observed in mAU. Peak C has two weak bands, the first are around 50kDa and the second around 25kDa.

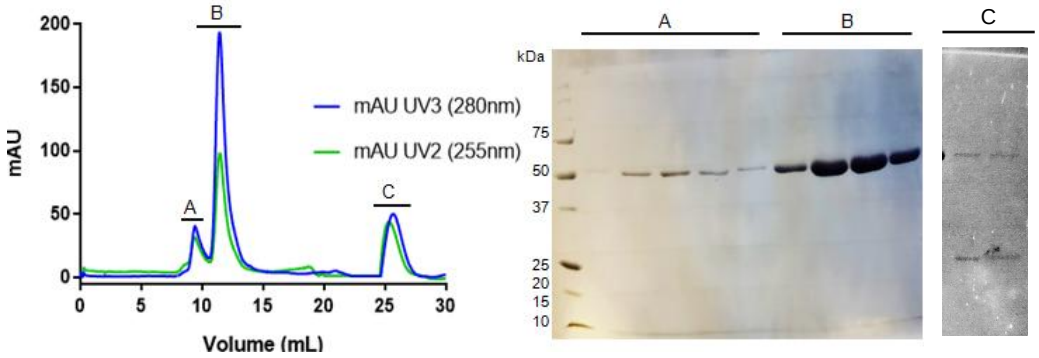

Figure 3. 3: Size exclusion chromatography and corresponding 12% SDS-PAGE gel. The chromatogram depicts the UV absorbance profile. The SDS-PAGE gel depicts the fractions, and their elution position is shown by the black bar and peaks are labelled A, B or C.

Preliminary electron microscopy by Dr Emma Andrews shows that PhoH2 proteins can form ring-like hexameric structures (personal communication), but also other structural conformations, due to the presence of multiple peaks in a SEC purification (Figure 3.3). The first peak to elute off the SEC column is most likely to be PhoH2 bound with *E. coli* host nucleic acid, as further explained in section 3.2.5. The second peak elutes off the column at this volume more consistently and corresponds to the active, hexameric conformation of PhoH2 protein as confirmed by multiangle light scattering and biochemical activity assays (personal communication). This peak may contain less nucleic acid bound to the complex and so elute later off the SEC. The infrequent final peak is likely due to dissociation of the hexameric protein complex, and may represent monomeric forms of the PhoH2 protein.

3.2.2 Optimisation of mRNA substrates

All PCR products that contained the T7 promoter were analysed by 1% agarose gels, with SYBR safe stain (section 2.1.3.2) before being transcribed into RNA. The following figure (Figure 3.4) shows five initial PCR products alongside with their corresponding T7 PCR products. All the initial PCR products match their T7 partners, however, for almost all of the T7 PCR products there is evidence of non-specific amplification, seen by the presence of an extra band above the main product. These extra bands could affect activity assay results and therefore, the desired products that included T7 were further purified (section 2.1.3.6) before

being transcribed into RNA. Originally PCR amplification of assay substrates was achieved by the KAPA HiFi hot start PCR mix (Roche Molecular Systems), but use of this enzyme resulted in continuous non-specific amplification, especially for substrates that contained an up-stream segment, i.e. upstream + *chaB* + *rsbW* + *sigF* and upstream + *rsbW* + *sigF*. The diagram below shows PCR products that were generated using KAPA HiFi hot start PCR mix. To reduce the number of purifications of the PCR products, two new enzyme kits were introduced, to help reduce the amount of non-specific amplification. HotFirePol® and HotFirePol® Blend (Solis BioDyne, USA) DNA polymerase hot start systems. These produced clean, distinct PCR products, with a reduction in the amount of non-specific amplification (data not shown).

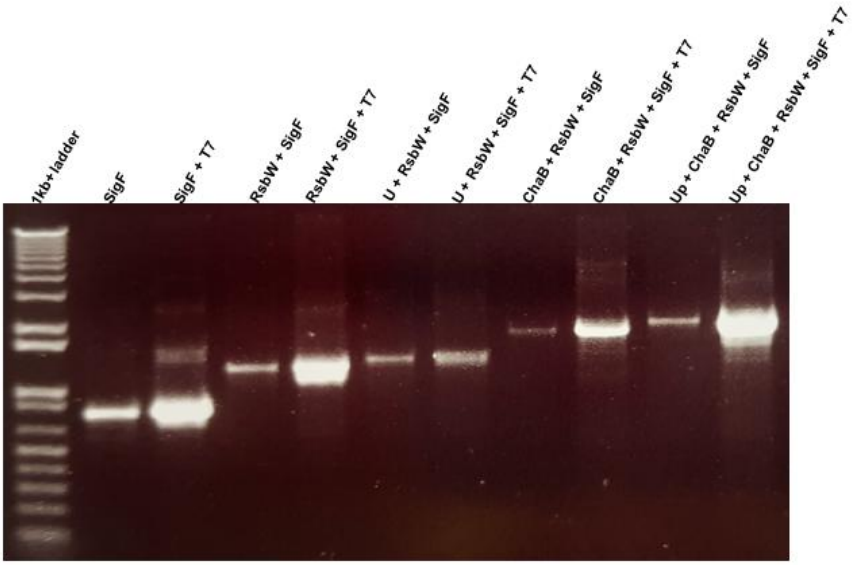

Figure 3. 4: PCR products run out on 1% agarose gels, with SYBR safe stain. Names of each PCR product are displayed at the top of each reaction well, on the gel. Band sizes were determined against a molecular weight 1kb+ ladder.

3.2.3 Reactions used in RNA unwinding and cleavage assays

Most nucleases, including PIN domains, require magnesium ions in order to carry out their nuclease activity [24]. ATPases also require magnesium therefore the addition of EDTA to an assay reaction should chelate any available magnesium ions, and hinder both PIN domain and PhoH domain activity. The addition of 40mM EDTA was included as a control in activity assays in an attempt to inhibit the activity of PhoH2. Helicases are motor proteins that convert the chemical energy of nucleoside triphosphate (often adenosine tripphospahte (ATP))

hydrolysis to the mechanical energy needed to bind and unwind nucleic acid [128]. No ATP controls were set up for activity assays to reduce the activity of the PhoH domain, of PhoH2, from binding to and unwinding the mRNA substrates. All assays contained RNA only controls. These RNA only controls contained all the usual assay components as described in section 2.2.2.6, except for PhoH2 protein. These controls were used as the base to deduce activity by PhoH2 and were included to ensure that the mRNA substrates added to each assay reaction were not degraded before the addition of PhoH2 protein. Some assays contained a protein only control; this will be discussed in the following section (section 3.2.4). Assay reactions had a variation of controls and tests applied, so the following table describes the various labels used to describe the reactions in assay figures. Any further test or controls not mentioned in the table below (Table 3.1) are outlined in the figure caption of that particular assay reaction.

Table 3. 1: Descriptions of the test/control and substrates applied to various assay reactions.

Test/control:	**Details of test/control:**
RNA only T0/T60	RNA only control, with all assay components except for protein. RNA was either incubated for 0 or 60 minutes.
Protein only	This control does not contain any RNA, but all other components including protein.
No ATP	All components without the addition of ATP.
ATP *	Different concentrations of ATP were added to different reaction wells in assays.
RNase	All components kept the same, except for the addition of 1ul 50mg/ml RNase A.
EDTA	Assay buffer containing 40mM EDTA was added
T x min	Test reactions containing all components, incubated for the x number of minutes
mRNA substrates	F1 (280bp) F2 (280bp) F3 (254bp) *sigF* (754bp) R+S (*rsbW* + *sigF*, 1167bp) U+R+S (Upstream + *rsbW* + *sigF*, 1270bp) C+R+S (*chaB* + *rsbW* + *sigF*, 1690bp) U+C+R+S (Upstream + *chaB* + *rsbW* + *sigF*, 1890bp) U+C+R (Upstream + *chaB* + *rsbW*, 1136bp) (-ve) 0457 (MSMEG_0457, 600bp)
Note: * All assays reactions contained 1mM of ATP, except in the case of increasing ATP concentrations trials, where varying amounts of ATP were included, to test how the increase of ATP affected PhoH2 protein function.	

3.2.4 RNA unwinding and cleavage activity assays by PhoH2

IMAC and SEC purified PhoH2 protein was used in RNA unwinding and cleavage assays with nine mRNA test substrates (Table 3.1), a negative control (MSMEG_0457) and a positive oligonucleotide control 5'3. Assays reactions were carried out as described in section 2.2.8.6 and were visualised on agarose gels (section 2.1.3.2), or urea-PAGE gels (section 2.2.8.5). Activity assays are presented in chronological order, to justify why various changes were made to assay reactions over time.

On the 28-08-17 the first six time course assays were set up with 250µM of purified PhoH2 protein and six mRNA substrates; U+R+S, R+S, 0457, *sigF*, U+C+R+S and C+R+S, all at a concentration of 50µM making a ratio of protein to RNA 5:1. These assays were visualised on 10% urea-PAGE gels (section 2.2.2.5), and the resulting reactions were all retarded at the top of the wells (Figure 3.5). Summer *et al.* [134] reports that nucleotide fragments between 2-500bp can be visualised by urea-PAGE denaturing gels, while nucleotide fragments bigger than 500bp will not be able to migrate through the gel [134]. As a majority of these RNA substrates are bigger than 500bp, except for the F1, F2 and F3 mRNA substrates and the 5'-3' positive oligonucleotide, from here on all assay reactions, with mRNA substrates bigger then 500bp were visualised on 1% agarose gels using SYBR safe stain, in 1x TAE buffer (section 2.1.3.2).

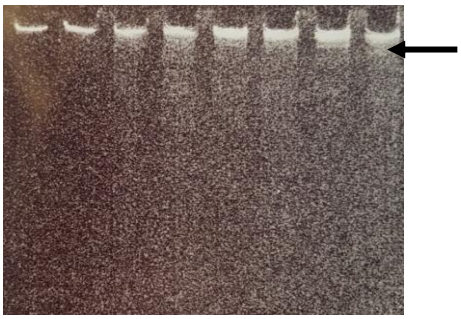

Figure 3. 5: Activity assay reaction results of *sigF* mRNA and PhoH2 protein, run out on a 10% Urea-PAGE gel. Gel shows that the reactions are retarded at the top of the gel, contained within in the wells. Black arrow points to the bands that correspond to the assay reaction.

On the 29-08-17 assay reactions with the substrates U+C+R+S and C+R+S were visualised on 1% agarose gels with SYBR safe stain (Figure 3.6). For U+C+R+S it appears that in the RNA only controls, there was degradation of the mRNA

substrate. With the remaining reactions there is an additional faint, low molecular weight band present. For the C+R+S assay there is very faint banding in the RNA only controls which is also apparent for the other six reactions. In the reactions where protein was added there is a slight band formation towards the bottom of the gel, initially it was thought that these bands were products of RNase activity, by PhoH2 on the C+R+S mRNA substrate, this is discussed later in this section. As these assays show faint or no RNA bands in the RNA only controls, to improve visualisation of the mRNA substrate, from here the ratio of protein to RNA was changed to 2.5:1 (protein (50µM protein:25µM RNA).

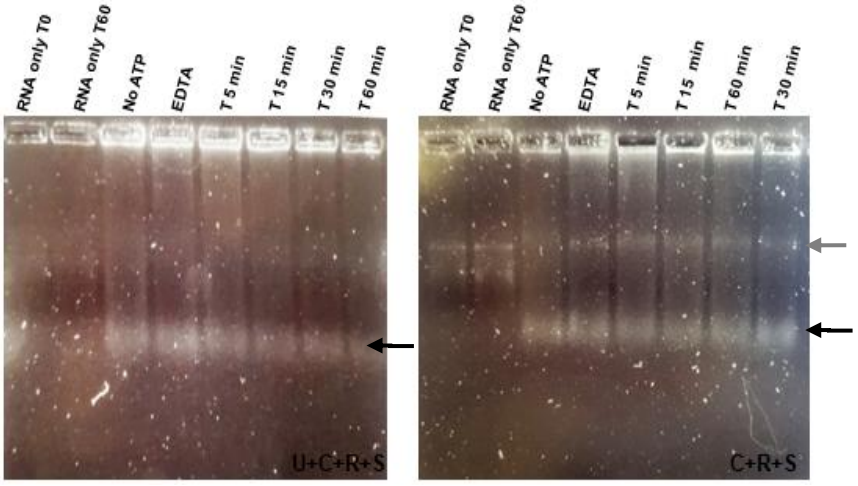

Figure 3. 6: **Activity assay reaction results of U+C+R+S and C+R+S mRNA and PhoH2 protein, run out on a 1% agarose gel.** Labels at the top of the gels show the control or test reaction that each reaction well is displaying. The mRNA substrates used in the assay reactions are displayed in the right bottom corner of each gel. Black arrows in figure point to lower molecular weight nucleic acid band formation in wells. Grey arrow points to higher molecular weight bands.

On the 05-08-17 assays were carried out with *sigF* mRNA and negative control MSMEG_0457 (Figure 3.7). MSMEG_0457 was chosen as a negative control by Dr Andrews as its expression was not detected/unchanged in the *ΔphoH2* mc²155 *M. smegmatis* mutant RNAseq data, suggesting it is not a target of PhoH2 [personal communication]. For the *sigF* assay the RNA only controls show bright bands, indicating the increase in RNA concentration has improved visualisation and that there has been no degradation prior to the assay. For the other reactions the intensity of the RNA decreases over time and by 60 minutes the RNA has almost fully been degraded as indicated by smearing and accumulation of small products at the bottom of the gel. In the reactions that contained protein there is

the presence of bands above the mRNA substrate that are not present in the RNA only controls. In the MSMEG_0457 assay there is little activity of PhoH2 overtime compared with the activity observed on *sigF* mRNA. For the EDTA control there is a faint band present. Of note is the absence of higher molecular weight bands with this assay compared with the *sigF* assay despite the same preparation of protein used.

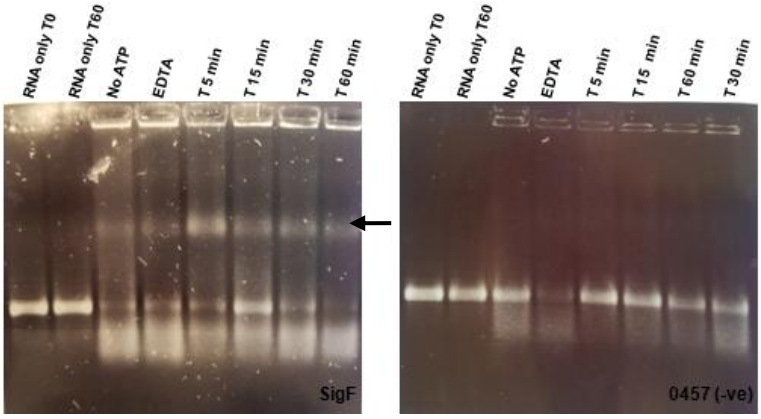

Figure 3. 7: Activity assay reaction results of *sigF* and MSMEG_0457 mRNA with PhoH2 protein, run out on a 1% agarose gel. Labels at the top of the gels show the control or test reaction that each reaction well is displaying. The mRNA substrates used in the assay reactions are displayed in the right bottom corner of each gel. Arrows show high molecular weight products, not present in RNA only controls.

On the 11-09-17 assays (Figure 3.8) were performed with; *sigF* and C+R+S mRNA against newly purified PhoH2 protein. In these assays new controls were included in which RNA was incubated in SEC buffer and IMAC buffer A, to check for RNA nuclease contaminants that could contaminate the protein prep. These controls showed that the buffers that PhoH2 was purified in, do not contain contaminants that cause the degradation of the mRNA substrate. With the remaining reactions, the RNA only controls are intact, however the controls; no ATP and EDTA do not appear to have hindered PhoH2 activity, as the same degree of activity has occurred between these and the test reactions T5 and T60. In these assays, like for MSMEG_0457 assay, there was no apparent higher molecular weight band present.

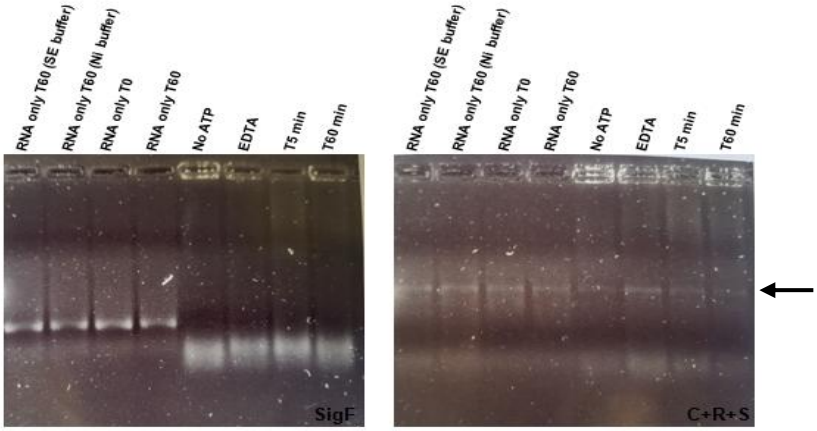

Figure 3. 8: Activity assay reaction results of *sigF* and C+R+S mRNA and PhoH2 protein, run out on a 1% agarose gel. Labels at the top of the gels show the control or test reaction that each reaction well is displaying. The mRNA substrates used in the assay reactions is displayed in the right bottom corner of each gel. SE buffer (Buffer A from SEC) and Ni buffer (Buffer A from IMAC) were used instead of assay buffer with the RNA only controls, to test for contaminants in those buffers. Black arrow indicates faint mRNA banding.

On the 21-09-17 assays were set up with freshly purified PhoH2 protein. Assays were carried out with the following mRNA substrates; oligonucleotide 5'3 (+ve) (Figure 3.9), *sigF*, R+S, C+R+S, MSMEG_0457 (-ve), U+C+R+S and U+R+S (Figure 3.10). PhoH2 has shown efficient activity on the 5'3 oligonucleotide [4]. The 5'3 oligo assay reactions visualised on a 10% urea gel showed that after 5 minutes clear unwinding and degradation activities by PhoH2, despite the poor annealing and therefore multiple conformations of this substrate. There was no activity of PhoH2 in the absence of ATP or presence of EDTA. There is a faint high molecular weight smear with the 30 minute time point and the absence of bands that was not observed with T60.

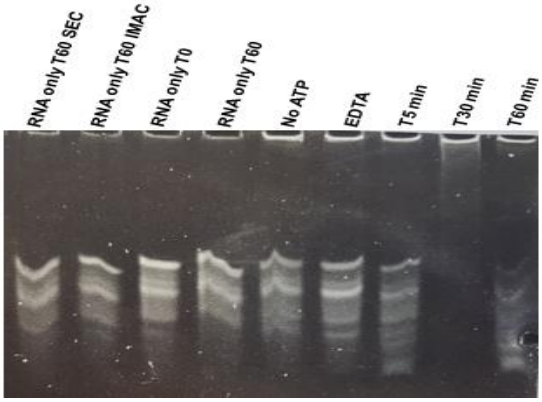

Figure 3. 9: Activity assay reaction results of 5'3 (+ve) oligonucleotide and PhoH2 protein, visualised on a 10% PAGE gel. Labels at the top of the gels show the control or test reaction that each reaction well is displaying. SE buffer (Buffer A from SEC) and Ni buffer (Buffer A from IMAC) were used instead of assay buffer with the RNA only controls, to test for contaminants in those buffers.

For each of the other substrates (Figure 3.10) the RNA only controls showed varying degrees of degradation, despite this, there was clear activity by PhoH2 on each of the RNA substrates within 5 minutes as indicated by the degradation of the RNA substrate and accumulation of smaller products at the bottom of the gel. In these assays, similar to the initial assays, the -ATP control did not hinder the activity of PhoH2. The EDTA control was able to, with each of these substrates, reduce the activity of PhoH2. This can confirm that the activity observed is due to PhoH2 and not the presence of contaminants. Unlike the initial assay performed on MSMEG_0457, in this case, PhoH2 has shown activity on this substrate.

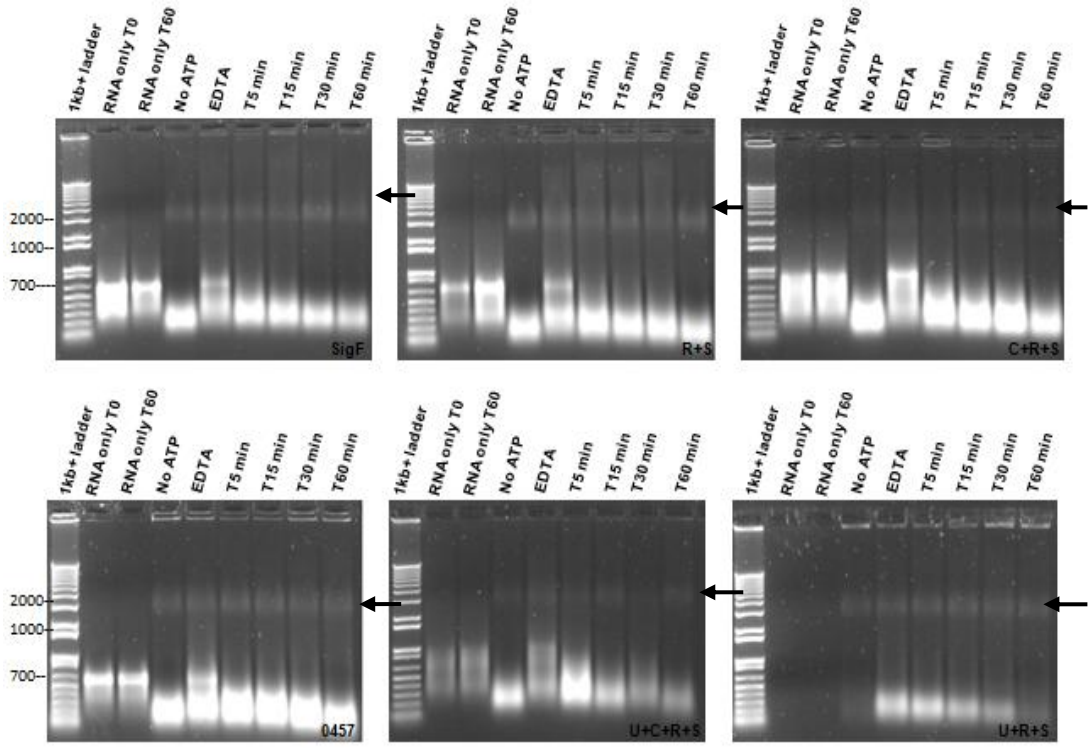

Figure 3. 10: Activity assay reaction results of *sigF*, R+S, C+R+S, 0457, U+C+R+S and U+R+S mRNA with PhoH2 protein, run out on a 1% agarose gel. Labels at the top of the gels show the control or test reaction that each reaction well is displaying. The mRNA substrates used in the assay reactions is displayed in the right bottom corner of each gel. Arrows show high molecular weight products, not present in RNA only controls. Band sizes were determined against a 1kb+ molecular weight ladder.

As observed previously, there is the presence of a higher molecular weight band. This band remains intact over the course of the reaction and does not appear here to affect the visualisation of the activity of PhoH2 on the target RNA substrates.

These next assays (Figure 3.11) were all carried out on the 28-10-17 and contain PhoH2 protein from the previous purification. The mRNA substrates tested in

these assays were the same as above and include F1, F2 and F3. The ratio of PhoH2 protein to mRNA substrate remains 2.5:1. Substrates F1, F2 and F3 were designed from the *sigF* gene (745 bp). F1 corresponds to the sequence from the start of the *sigF* gene to position 280bp. F2 is from 250 to 529 bp and F3 is 500bp to the end of the *sigF* gene 745bp. For substrates R+S, MSMEG_0457, *sigF*, C+R+S, U+R+S and U+C+R in these assays, many of the RNA only controls failed as indicated by the absence of RNA bands on the gel (not shown). There was also the presence of an intense higher molecular weight band at the same size as observed previously (not shown). For F1, F2 and F3, there are multiple bands present with RNA only controls that are similar to what is observed in the test reactions, making deducing activity by PhoH2 unclear and likely not apparent.

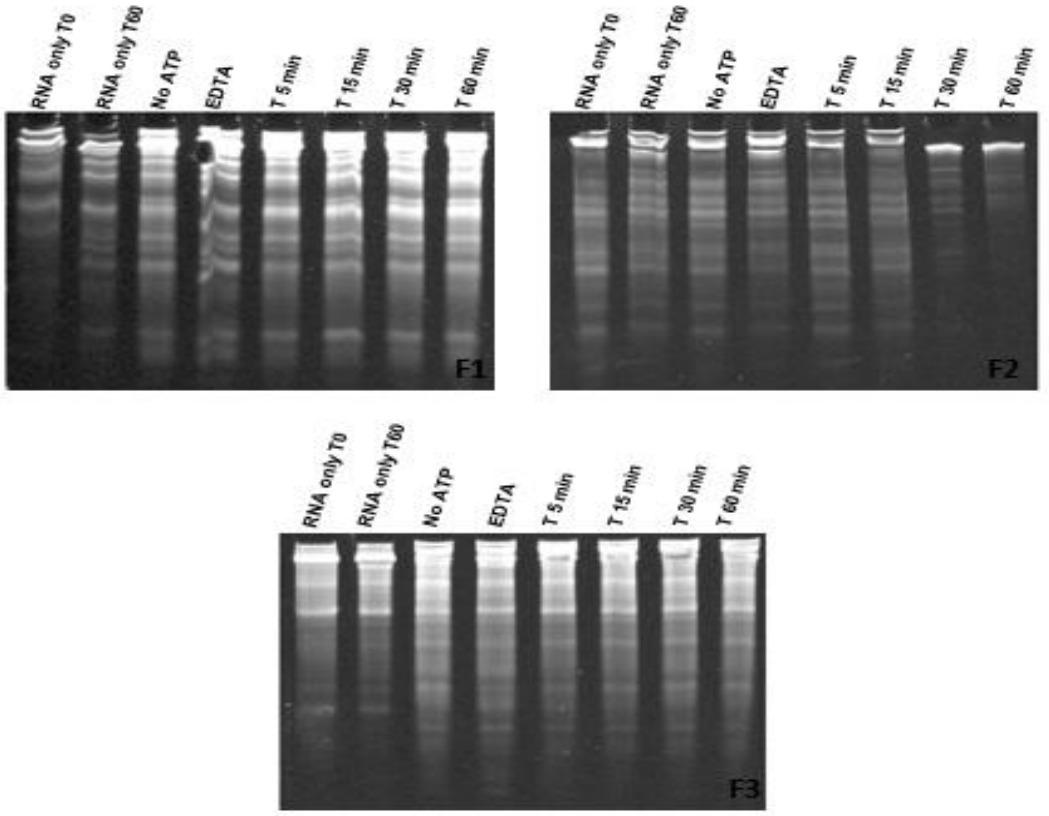

Figure 3. 11: Activity assay reaction results of F1, F2 and F3 mRNA with PhoH2 protein, run out on 8% Urea-PAGE gels. Labels at the top of the gels show the control or test reaction that each reaction well is displaying. The mRNA substrates used in the assay reactions is displayed in the right bottom corner of each gel.

Due to the now consistent presence of a higher molecular weight band that was not observed in the RNA only controls, and sometimes the presence of a smaller molecular weight band (Figure 3.10, U+R+S), a new control, protein only, was included to confirm that this contaminating band was being introduced by the

addition of protein to the reactions. In addition to this protein only control, a higher concentration of ATP (5mM) was tested to determine if this was a limiting factor of these assays and whether increased activity over a shorter time period would be observed with higher concentrations of ATP. These following assays (Figure 3.12) confirmed that the addition of PhoH2 to the reactions was also introducing an additional source of nucleic acid, and that with increased concentrations of ATP, not only was there increased activity on the target substrate, but activity on the contaminating nucleic acid introduced by the protein.

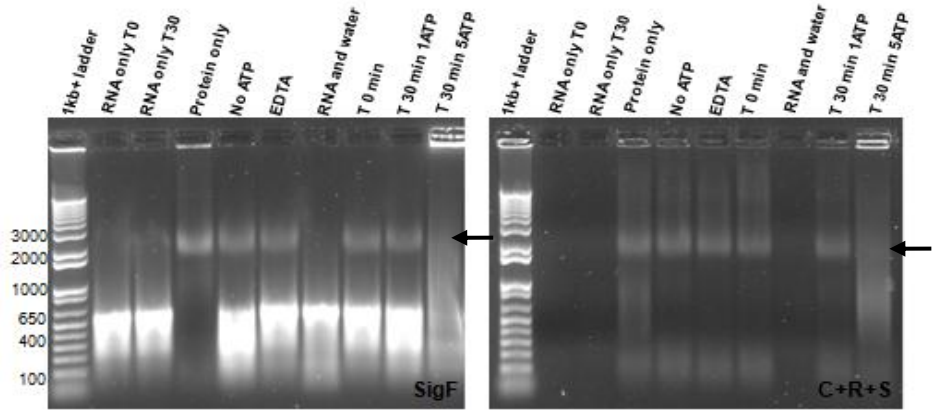

Figure 3. 12: Activity assay reaction results of SigF and C+R+S mRNA with PhoH2 protein, run out on a 1% agarose gel. Labels at the top of the gels show the control or test reaction that each reaction well is displaying. The mRNA substrates used in the assay reactions is displayed in the right bottom corner of each gel. Arrows show high molecular weight products, not present in RNA only controls. Band sizes were determined against a 1kb+ molecular weight ladder.

The next set of assays (Figure 3.13) was performed on *sigF* mRNA, in the presence of increasing concentrations of ATP. An additional EDTA control was also included to determine if increasing the amount of EDTA would reduce PhoH2 activity. The substrates used in these assays were of poor quality, despite this, in the presence of increasing ATP concentrations, there is increased activity by PhoH2 on the contaminating higher molecular weight band. The addition of further EDTA did not decrease the activity of PhoH2 as there is evidence for activity either on the target substrate or contaminating band.

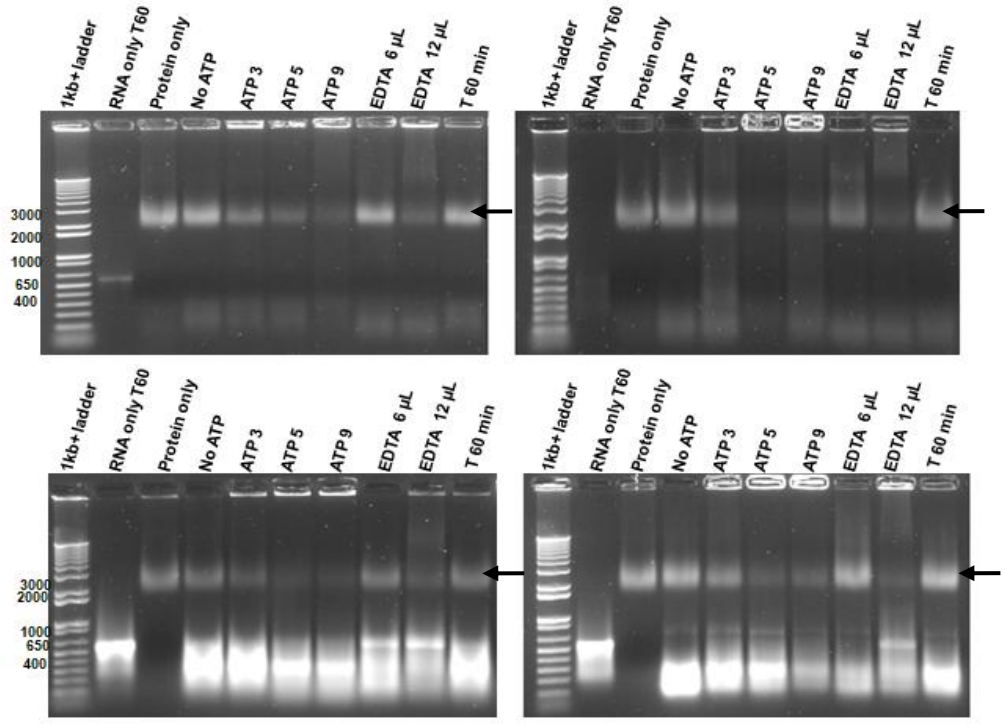

Figure 3. 13: Activity assay reaction results of *sigF* mRNA with PhoH2 protein, run out on a 1% agarose gel. Labels at the top of the gels show the control or test reaction that each reaction well is displaying. Arrows show high molecular weight bands, not present in RNA only controls. Band sizes were determined against a 1kb+ molecular weight ladder.

Due to the increase in activity of PhoH2 observed on both target RNA substrates and the contaminating higher molecular weight band with increasing concentrations of ATP, attempts were made to reduce the contaminant load.

3.2.5 *E. coli* Host Nucleic Acid Contamination of PhoH2 proteins

Contamination of PhoH2 proteins by expression host nucleic acids makes it difficult to determine if the biological activity of the protein is specific for biological mRNA substrates, and/or for the contaminating nucleic acid. Here attempts were made to isolate and determine whether the *E. coli* host nucleic acid was RNA or DNA and several common methods, such as high salt (2M NaCl), heat treatment and RNase treatment were explored as possible protocols for the removal of *E. coli* host nucleic acid bound to purified PhoH2 proteins. It has been noted in the literature that it is common for RNA-binding proteins to retain host RNA during purification. A study by Edinger *et al.*, [175] looked into the purification of hepatitis B virus core particles, from *E. coli*, for drug delivery

applications, and found that there was contamination of the protein by *E. coli* nucleic acid. An RNase treatment protocol was used to remove any RNA contamination from the *E. coli* host. They noted that addition of RNase in the early stage of purification removed contaminating RNA, hence preventing any remnant activity for further studies [175]. Another study [176] also identified *E. coli* host RNA contamination when they purified dsRBEC, an RNA binding protein, from *E. coli*. Contamination of protein by non-specific host nucleic acids can complicate protein purifications [176].

3.2.5.1 Identification of host Nucleic acid contamination in PhoH2 assays

Many of the RNA substrates tested were of poor or degraded quality. Interestingly in these assays there was commonly the presence of higher molecular weight nucleic acid bands in all reactions that contained PhoH2 (Figure 3.14). Sometimes there was also presence of lower molecular weight bands, in similar positions to where mRNA substrate bands appeared (A, B, C, D and F of Figure 3.14). Upon this discovery, an additional control, protein only was included to monitor the extent of the contamination. It was then hypothesised that PhoH2 proteins co-purified with *E. coli* host nucleic acid and that the host nucleic acids were visible in the assay reactions. With increasing concentrations of ATP, activity was observed on both the target RNA and contaminating nucleic acid.

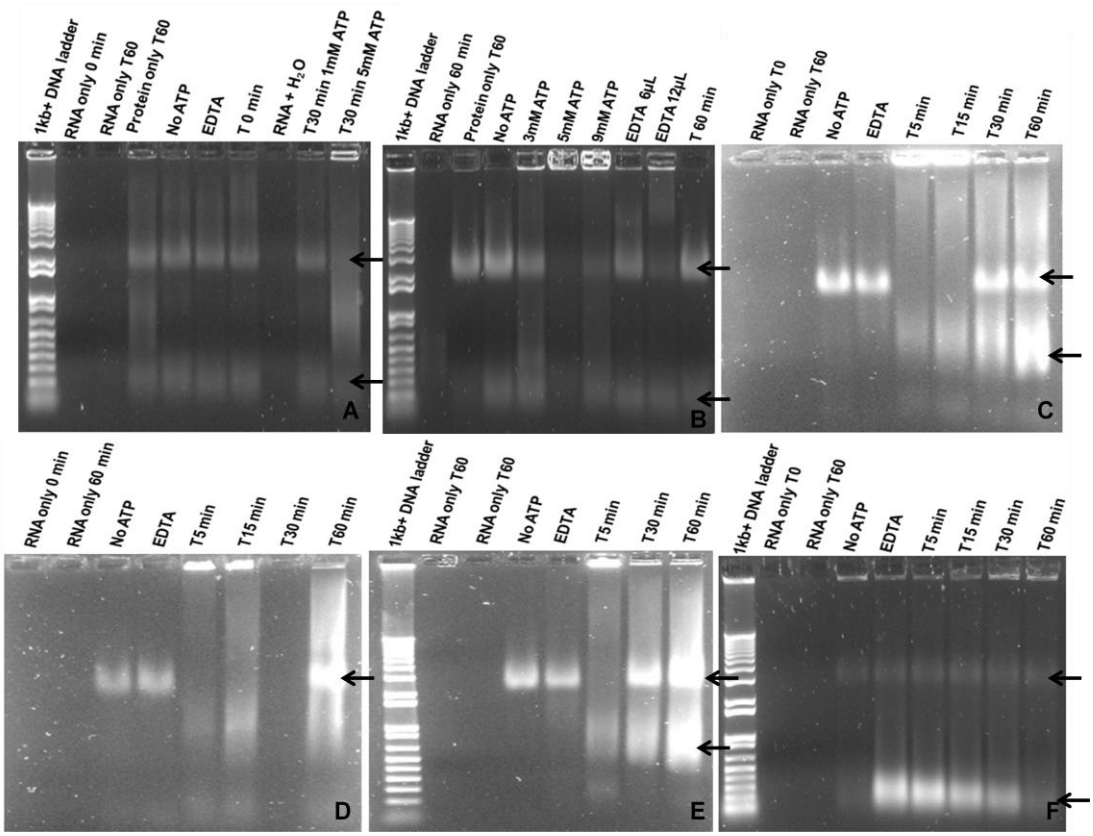

Figure 3. 14: 1% agarose gels showing activity assays with PhoH2 and degraded mRNA substrates. Letters in the right bottom corner of each gel, define different assay reactions with different degraded mRNA substrates. Figure A is C+R+S, figures B and C are *sigF*, figures D and F are U+R+S and figure E is U+R+C.

3.2.5.2 Extracted nucleic acid from PhoH2 protein after purification

Attempts were made to isolate this host *E. coli* nucleic acid from PhoH2 proteins after an IMAC purification (Figure 3.15 B) and after a SEC purification (Figure 3.15 A). The concentration of nucleic acid was measured (Table 3.2), as described in section 2.1.3.4, to examine both concentration and the 260/280 ratio to see if the results would be closer to an expected DNA ratio, around 1.8 or an RNA ratio around 1.9-2. Sample A had a pretty high nucleic acid concentration and had a 260/280 ratio similar to RNA expected ratios. Sample B also had a high nucleic acid concentration and had a 260/280 ratio similar to DNA expected ratios. A reason for these different ratios between samples may be due to sample A extracted after a SE purification and only contained the PhoH2 peak for nucleic acid extraction. Whereas the nucleic acid extracted after IMAC would contain populations of proteins that correspond to peaks one and two on a SEC.

Andrews and Arcus [4] previously observed no activity by PhoH2 on dsRNA or DNA, but there was activity on ssRNA. Therefore it makes sense that the nucleic acid found purified to PhoH2 would be host ssRNA and would give a resulting 260/280 ratio close to 2. Sample B was extracted from the PhoH2 peak after an IMAC purification which often has contaminating proteins carried along with it. As mentioned above, SE purification produces two peaks (Figure 3.3), with PhoH2 having been identified as the second peak. Therefore it is possible that the contaminating proteins after IMAC purification may bind to DNA, dropping the 260/280 ratio down to a DNA ratio of 1.8.

Table 3. 2: Nanodrop readings for nucleic acid extracted from PhoH2 protein

Sample ID :	Nucleic acid concentration:	Unit:	260/280
A	135.9	ng/µl	1.99
B	123.6	ng/µl	1.83

Figure 3.15 shows extracted nucleic acid from precipitated PhoH2 protein after an IMAC and SE purification, run on a 1% agarose gel, as described in section 2.1.3.2. Sample A has a similar high molecular weight band to the band shown in sample B, high molecular weight bands often indicate DNA. There is also a faint smear in both samples, in sample A (SE) it appears around 1500bp, while in sample B (IMAC) it appears just above 2000bp. Both samples also have very low molecular weight faint banding, which it commonly seen for degraded RNA. Sample A appears to have more of this low molecular weight banding compared to sample B.

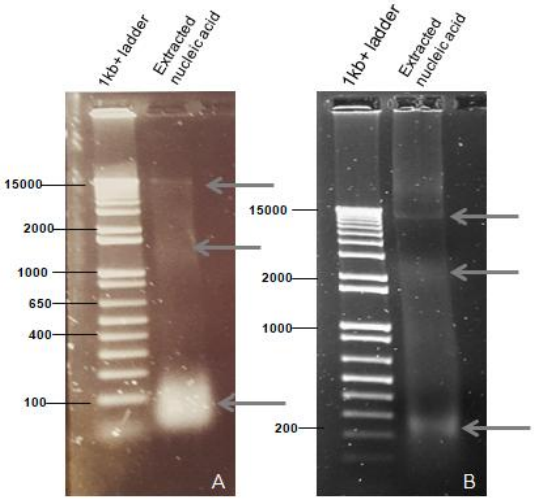

Figure 3. 15: Extracted *E. coli* host nucleic acid from purified PhoH2 protein. Gel A contains extracted nucleic acid from a PhoH2 purified with an IMAC and SEC purification. Gel B contains extracted nucleic acid from a PhoH2 protein purified with an IMAC purification. Arrows indicate the positions of faint nucleic acid bands.

3.2.5.3 DNase and RNase treatment of concentrated PhoH2 nucleic acid

After SEC purification of PhoH2, the three peaks seen in Figure 3.3 were concentrated separately and the concentrated proteins were run on a 1% agarose gel, with SYBR safe stain, following the protocol in section 2.1.3.2. The following figure (Figure 3.16) represents concentrated nucleic acid representing peaks A, B and C from a typical SEC chromomatogram.

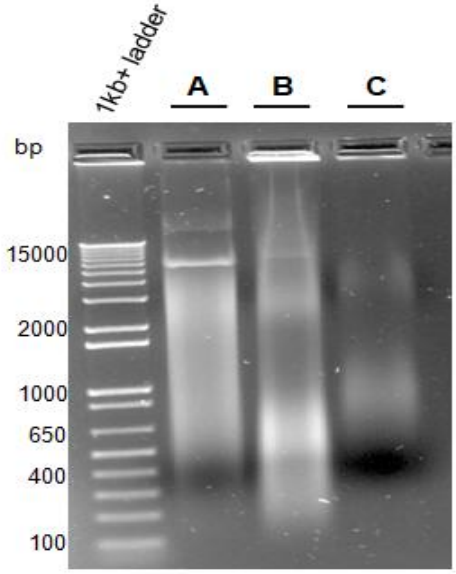

Figure 3. 16: Concentrated IMAC and SEC purified protein peaks run on a 1% agarose gel, with SYBR safe stain. Lane A, representing peak A has two high molecular weight bands, one is above the 15000bp mark and the other is around 7000bp and is similar to the band seen in Figure 3.15. Lane A has a cloud of nucleic acid from around 5000bp to 400bp. In lane B, representing peak B there are several clouds of nucleic acids around 3000bps and a large low resolution band around 650bp. There is also a small high molecular band around 14000bp's, which is sitting higher than the band in lane A. In lane C, which represent peak C, there appears to be less nucleic activity and there are two main clouds of nucleic acid around 4000bp and 850bp.

Two day old, protein samples from peak A and B from a SEC purification were treated with either 5mM of DNase I (Sigma-Aldrich) or 5mM of RNase A (Sigma-Aldrich). Samples were treated for 30 minutes following the manufacturer's recommendations. After 30 minutes the samples were analysed on a 1% agarose gel, with SYBR safe stain, in 1x TBE, following protocol from section 2.1.3.2. The following figure (Figure 3.17) shows peaks A and B treated with DNase in the first gel and RNase in the second gel. The first lanes in each gel labelled A and B are samples that were not treated and are there to compare differences between treated and untreated samples. In the first gel where the last

two lanes were treated with DNase I, lane A has lost its high molecular weight bands, and the cloud of nucleic acid has decreased in intensity and has shifted down the gel. While lane B, appear to be unchanged by the DNase treatment. In the second gel where the last two lanes were treated with RNase A, lane A appears mostly unaffected by the RNase treatment. While in lane B, treatment with RNAase A has cleared the nucleic acid. This suggests that peak A contains DNA and peak B (active form of PhoH2) contains RNA.

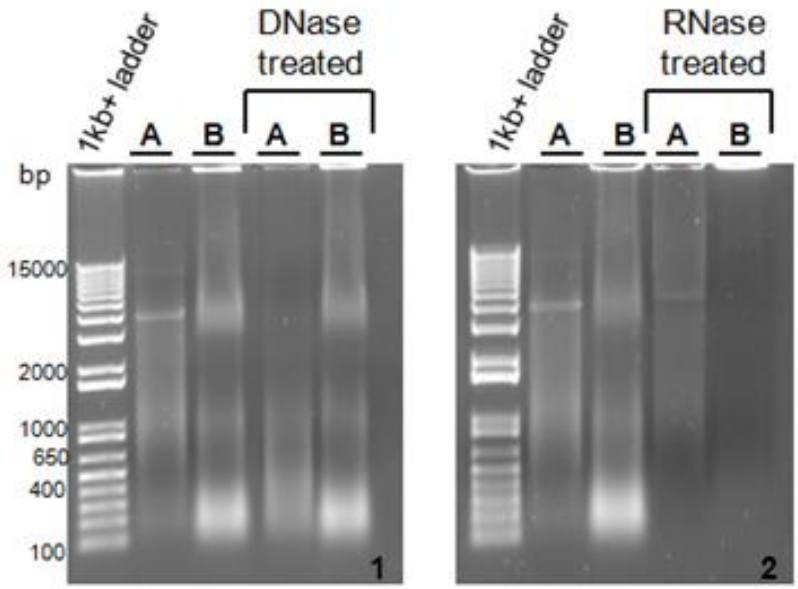

Figure 3. 17: Concentrated IMAC and SEC purified protein peaks run on a 1% agarose gel, with SYBR safe stain, in 1x TBE buffer, treated with DNase I and RNase A. In gel 1, samples in lanes 4 and 5 were DNase treated and in gel 2, samples in lane 4 and 5 were RNase treated. Nucleic acid bands are compared against a molecular weight 1kb+ ladder.

3.2.6 Crystallisation

Robot crystal screen were attempted for PhoH2 protein after a SE purification without little success. There was brown precipitation observed when looking at the crystal screens after 30 minutes, meaning that the protein was at a high enough concentration for crystal development. Unfortunately no crystals developed under any of the four conditions, as described in section 2.2.1.11. It was hypothesised that due to the protein binding to host nucleic acid, there would be a heterogeneous population of PhoH2 protein with different structural formations. Protein crystals are more likely to form if populations of the protein are uniform

and can stack up easily against each other, so this mixed population of PhoH2 proteins may hinder the development of crystal structures. Therefore efforts were turned towards removing this host nucleic acid from purified PhoH2 protein.

3.2.7 Removal of *E. coli* nucleic acid from PhoH2 protein through high salt (NaCl) SEC purification

The use of high salt elution buffers has been noted in the literature as an effective method for the removal of *E. coli* nucleic acids bound to proteins during chromatography purifications. Here 2M NaCl was used in the SEC elution buffer, so as to weaken ionic associations between nucleotides and the protein during their separation on the preparative SEC column. PhoH2 protein was split into two separate samples after an IMAC purification and the first sample was further purified through a standard SEC purification with 200mM NaCl in the buffer. The resulting chromatogram (Figure 3.18) shows the formation of two main elution peaks and a smaller peak around 15ml. This chromatogram represents a standard SEC purification of PhoH2 after IMAC, where the second peak elutes around 11ml and is recognised as the active form of PhoH2 protein. The matching SDS-PAGE gel also seen in Figure 3.18, shows that peaks A and B are of the same molecular weight (50kDa) when run on an SDS-PAGE gel, with peak B showing more distinct banding. The third peak in the chromatogram resulted in no noticeable bands in an SDS-PAGE gel.

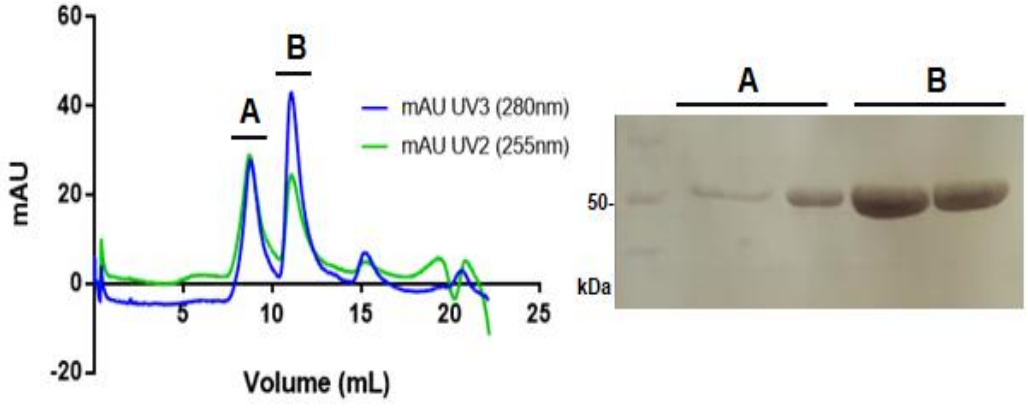

Figure 3. 18: Size exclusion chromatography and corresponding 12% SDS-PAGE gel. The chromatogram depicts the UV absorbance profile. mAU UV3 (280nm) is represented by a blue line. mAU UV2 (255nm) is represented by a green line. The SDS-PAGE gel depicts the fractions, and their elution position is shown by the black bar and peaks are labelled A or B. Sizes of bands in SDS-PAGE are determined against a protein ladder (Bio-Rad).

The second sample of IMAC purified PhoH2 protein, was incubated with an equal volume of 4M NaCl SEC buffer for two hours at 4°C. After two hours the sample was purified through SEC, with a 2M NaCl SEC elution buffer. It was thought that by removing the association of nucleotides to the protein, this would result in a shift of the protein peaks to a later elution volume. Proteins of larger molecular weight elute off SEC first, followed by proteins of smaller molecular weight. The resulting chromatogram (Figure 3.19) shows the formation of multiple peaks, with the first peak seen in normal SEC purifications (Figure 3.3) having decreased significantly, while the second peak appears to be unaffected and is eluting off in the same position as it would in a normal SEC purification corresponding to the active form of PhoH2 (Figure 3.3). The additional peaks present in this chromatogram may be due the presence of smaller subunits or dissociated of PhoH2 proteins, having lost their bound nucleic acid.

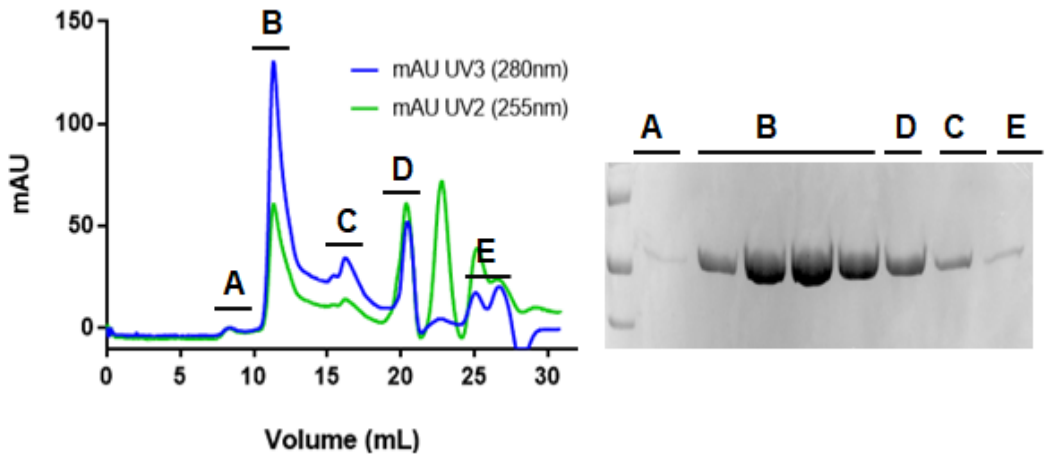

Figure 3. 19: Size exclusion chromatography and corresponding 12% SDS-PAGE gel. The chromatogram depicts the UV absorbance profile. mAU UV3 (280nm) is represented by a blue line. mAU UV2 (255nm) is represented by a green line. The SDS-PAGE gel depicts the fractions, and their elution position is shown by the black bar and peaks are labelled A, B, C, D or E. Sizes of bands in SDS-PAGE are determined against a protein ladder (Bio-Rad).

This high salt treatment of PhoH2 proteins was carried out multiple times, often with varying results and precipitated protein (Figure 3.20, B). Attempts were made to use 2M NaCl treated PhoH2 protein in activity assays with mRNA substrates, but this either resulted in precipitated protein through concentrating

efforts, or there would still be presence on nucleic acids, seen when the resulting assays were run on 1% agarose gels, with SYBR safe stain (section 2.1.3.2)

Diagram A, of figure 3.20, shows activity assays attempted with 2M NaCl treated SEC purified PhoH2 protein and mRNA substrates; U+R+S, C+R+S, R+S and SigF. The protein only control lane in all the following agarose gels contains evidence of nucleic acid and in the *sigF* assay gel, where the mRNA appears to have mostly degraded in the RNA only controls, there is telltale signs of *E. coli* nucleic acid in the other reaction wells. The PhoH2 protein appears to still have activity against the mRNA substrates, seen by the change in appearance of the U+R+S, C+R+S and R+S mRNA bands. Similar to other activity assays, the no ATP controls appears to have no inhibiting affect on PhoH2 activity. Diagram B, in Figure 3.20, shows precipitated protein after going through a 4M salt treatment, before a SEC purification.

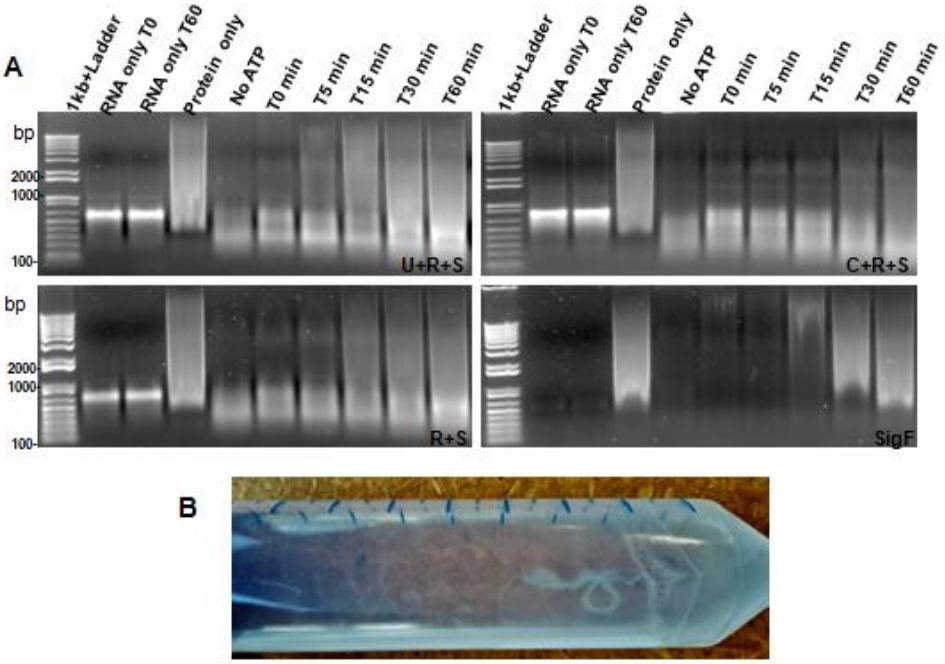

Figure 3. 20: Results of 2M NaCl treated PhoH2 protein. Diagram A shows four activity assays with 2M NaCl treated SEC purified PhoH2 and mRNA substrates. Nucleic acid bands are compared against a molecular weight 1kb+ ladder. Diagram B shows precipitated IMAC PhoH2 protein after a 4M NaCl treatment, at 4^{o}C, before a SEC purification.

Due to the variability in results efforts were turned towards new methods to remove nucleic acids without causing major aggregation of PhoH2 proteins.

3.2.8 Removal of *E. coli* nucleic acid from PhoH2 protein through heat and RNase treatments

PhoH2 protein was purified through IMAC and treated with RNase A, as described in section 2.2.6. RNase treated protein was then purified through SEC (section 2.2.1.8). The resulting chromatogram in Figure 3.21 depicts the resulting peaks and fractions observed from SEC purification. The SDS-PAGE gel also depicted in figure 3.21, shows protein purified from IMAC and SEC purifications and it shows RNase and heat treated PhoH2 protein, after treatment, before a SEC purification. PhoH2 protein from an IMAC purification has produced a large band around 50kDA and two smaller less distinct bands around 35kDa and 20kDa. RNase treated PhoH2 protein has a large band around 50kDa and several smaller bands all under 40kDa. Heat treated PhoH2 protein also has a large band around 50kDa and several smaller bands under 40kDa. Lane A, represents peak A in the chromatogram and only 1 band around 50kDa is observed. Lane B, represent peak B in the chromatogram and has a small band around 50kDa. Lane C, represents peak C in the chromatogram and had no obvious bands present in the SDS-PAGE gel. IMAC, RNase PhoH2, Heat PhoH2, A and B lane in the gel, all share a band around 50kDa. The band of lane B, is less distinct than the bands in the other lanes. This band around 50kDa, most likely represent PhoH2 protein, which is 48.7kDa. Lanes IMAC, RNase PhoH2 and Heat PhoH2 all share a band around 35kDa. The RNase lane has several bands present under the PhoH2 band, with one of these bands being RNase A enzyme (13.7 kDA). In the chromatogram peak A has eluted off in the usual placement of the PhoH2 band, seen in figure 3.3. This indicates that RNase treatment has not affected elution position of the PhoH2 protein. However RNase treatment appears to have affected the position of the first peak that usually elutes off after ~8ml, in a normal SEC purification despite the fact that this peak was thought to contain mostly DNA (Figure 3.3). Peak A here corresponds to PhoH2 that typically elutes after 10 ml. Peaks B or C was thought to be RNAse A however no protein was observed on the corresponding SDS-PAGE gel. The amount of nucleic acid represented by the green peaks in the chromatogram has dropped only slightly from a normal SEC purification.

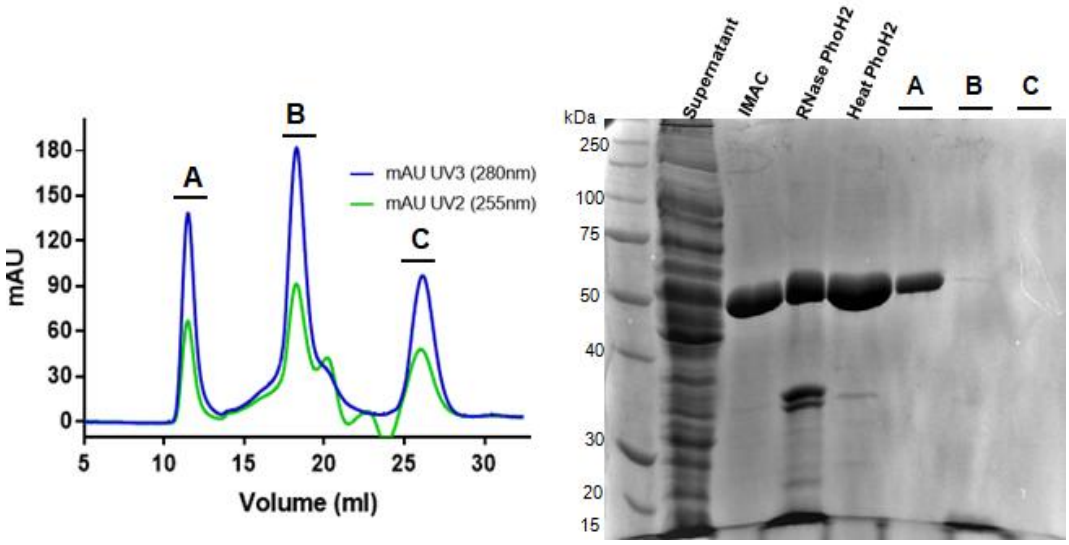

Figure 3. 21: SEC purification and corresponding 12% SDS-PAGE gel. The chromatogram depicts the UV absorbance for UV3 (280nm) in blue and UV2 (255nm) in green. The SDS-PAGE gel depicts the peaks; A, B and C, from the corresponding chromatogram. The gel also shows bands from an IMAC purification of PhoH2 (Supernatant and IMAC lanes) and bands representing RNase and heat treated PhoH2 protein, from an IMAC purification. Fractions and their elution position are shown by the black bar.

Fractions containing RNase treated SEC purified PhoH2 protein (peak B from Figure 3.21) were pooled together and diluted down in low salt AEX (25mM) buffer A, to reduce the salt concentration before being further purified through anion exchange chromatography (AEX) to ensure separation of RNase A from PhoH2. Ion exchange resins contain charged groups; in the case of anion exchange these groups are basic. Samples are loaded under low salt conditions and bound material is eluted by increasing salt concentrations. Proteins bound to the anion exchange resin are bound via non-covalent ionic (salt-bridge) interactions. Proteins will bind to the anion exchange resin if the buffer pH is higher than the pI of the protein [71]. PhoH2 has a theoretical pI of 5.66, and the pH of the buffer was 8. Not only does anion exchange chromatography help remove any unbound nucleic acid in solution with PhoH2 proteins, it ensures that any RNase contamination is not bound to the column due to the high pI of the RNase A enzyme. For the anion exchange purification of PhoH2 a gradient elution was carried out from 50mM NaCl to 1M NaCl. The following chromatogram and corresponding SDS-PAGE gel in Figure 3.22 depict the resulting peaks and fractions observed from an AEX purification. Three peaks are observed in the following chromatogram; A, B and C. Peak A has eluded off the

column before there was an increase in salt concentration, suggesting that this protein was weakly attached to the column and is most likely to be RNaseA. The second peak, B, is of a very small concentration and has eluded off the column around 20% of AEX buffer B. The last peak, C, has the highest concentration out of all the peaks and eluted off the column last, suggesting a very strong interaction with the column. The SDS-PAGE gel shows that all protein peaks from the chromatogram Peak A contains another contaminating band around 40kDa, while peak B, has numerous contaminating bands and peak C, doesn't appear to have any obvious contaminating bands. The chromatogram trace suggests that Peak C should be of a greater concentration than what is reflected in the SDS-PAGE gel for Peak C. Peaks B and C containing PhoH2 were pooled together and used in an activity assay with the mRNA substrate F1.

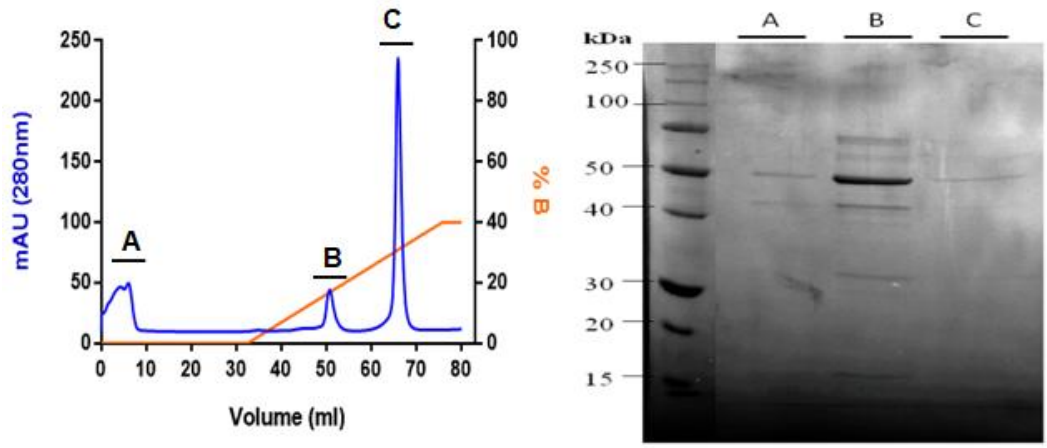

Figure 3. 22: Chromatogram of Anion exchange chromatography and corresponding 12% SDS-PAGE gel. The three peaks on the chromatogram (A, B and C) are depicted on the SDS-PAGE gel and labels represent each peak. Sizes of bands are determined against protein marker (kDa).

In addition to RNase A treatment of PhoH2, PhoH2 protein that was also purified through IMAC was then heat treated for one hour, as described in section 2.2.7. The heat treated protein was then purified through SEC as described in section 2.2. The following chromatogram and corresponding SDS-PAGE gel in Figure 3.23 depict the resulting peaks and fractions observed from an SEC purification. The chromatogram is similar to the profile of a 2M NaCl treated PhoH2 protein (Figure 3.19). The first peak seen in normal SEC purifications (Figure 3.3) is not present in this chromatogram, while peak A, though to be the active form of PhoH2, appears to have been unaffected and is eluting off in the same position as

it would in a normal SEC purification Peak B, can sometimes be present in SEC purifications as seen by Figure 3.3, but not normally to this scale. In the SDS-PAGE gel, peak A shows a clear band around 50kDa and has no obvious contaminating bands, while peak B has no bands present in the gel. The heat treatment has affected the elution profile of the first peak seen in normal SEC purification. It is possible that the heat treatment has disrupted the proteins structure and the protein could be forming monomeric species of a smaller molecular weight. Heat treatment appears to have no affect on the second peak seen in normal SEC purifications, as Peak A is eluting off in the same position.

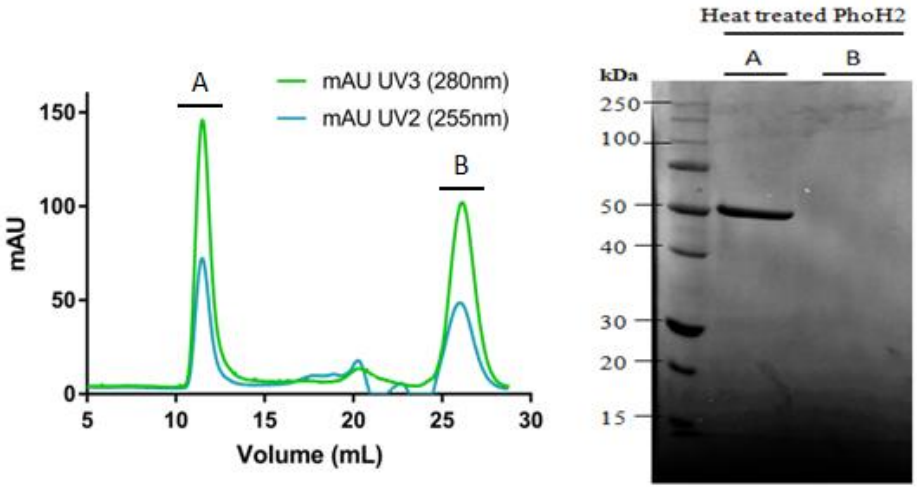

Figure 3. 23: Chromatogram of SEC purification and corresponding 12% SDS-PAGE gel. The two peaks on the chromatogram (A and B) are depicted on the SDS-PAGE gel and A or B labels represent the corresponding peak. Sizes of bands are determined against protein marker (kDa).

3.2.9 Malachite Green ATPase Assays with heat and RNase treated PhoH2

Both treated samples of PhoH2 were tested in simple ATPase assays as described in section 2.2.8. The PhoH domain of PhoH2 is an ATPase that is capable of turning over ATP into ADP and inorganic phosphate (Pi). When malachite green reagent is added to the assay, solutions that contain free phosphate change from orange to dark green/blue colour, while solutions, with no free phosphate, change only slightly to a light green/yellow colour. Hence active protein will turn over ATP and will result in a blue yellow change, while inactive protein will be the same colour as the negative control. Figure 3.24 shows a malachite green ATPase assay with heat treated and RNase treated PhoH2 proteins, after their resulting

purifications as described in section 2.2.7 and section 2.2.6. Heat treated PhoH2 protein shows a blue test, indicating active protein, while the RNase treated protein shows a light green test, the same colour as the negative control, which indicates a non-active protein.

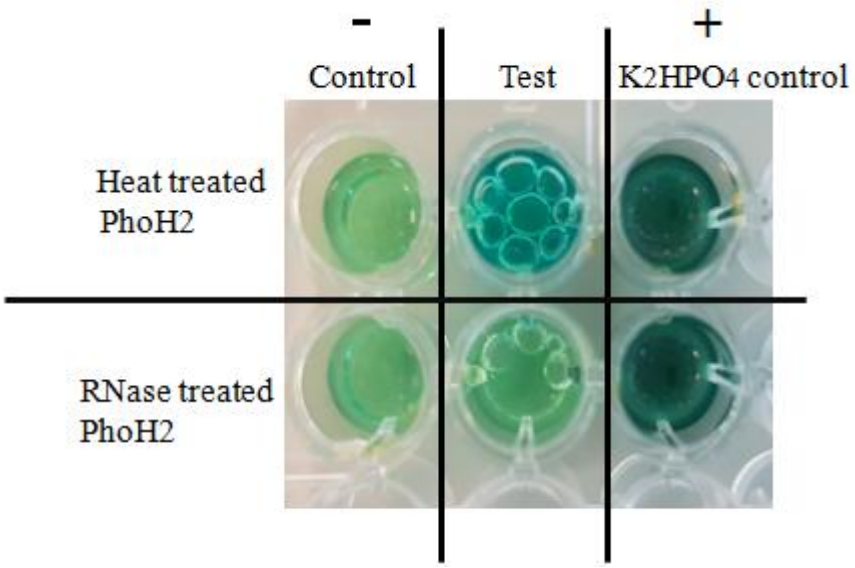

Figure 3. 24: Malachite Green ATPase assay on a microplate. Symbol '–' stands for negative control (assay contains no protein), which is expected to give a yellow or pale green colour. Symbol '+' stands for the positive control, and is expected to turn a dark green/blue colour. Heat treated PhoH2 protein is shown in the top row, while RNase treated PhoH2 protein is shown in the bottom row, underneath the black line. Control, test and K_2HPO_4 columns are labelled at the top of the figure, separated by black lines.

3.2.10 Activity assays with RNase and heat treat PhoH2 protein

The remaining RNase and Heat treated PhoH2 protein was used in an activity assay with F1 mRNA substrate. F1 has previously shown activity with PhoH2 proteins and because it is smaller than 500bp, the resulting assays can be visualised with Urea-PAGE gels, which should show more detail of the reaction than an agarose gel would. Assays were carried out as described in a section 2.2.8.6. Gel A in figure 3.25, shows the assay reaction for heat treated PhoH2 protein, with the mRNA substrate F1. Gel B, in figure 3.25, shows the assay reaction for RNase treated PhoH2 protein, with the mRNA substrate F1. The treatment of PhoH2 with heat has abolished contaminating nucleic acid load as indicated by the absence of any bands in the protein only control. Test and control reactions no ATP and EDTA show activity by PhoH2 on F1 (Figure 3.25, A).

This was not the case with PhoH2 that had been treated with RNase A (Figure 3.25, B).

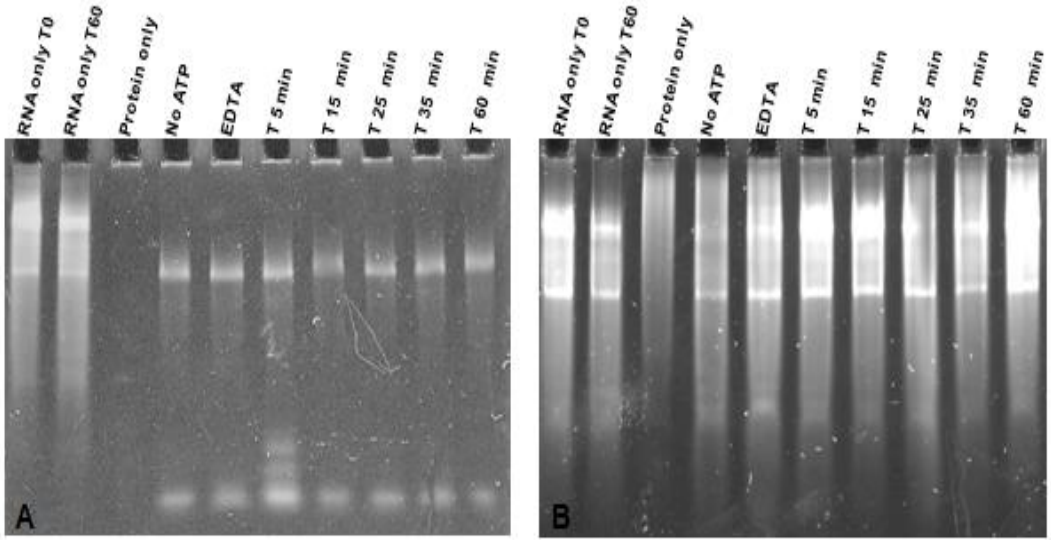

Figure 3. 25: PhoH2 activity assays with F1 mRNA substrate. Assays were run out on 8% urea PAGE gels, and stained with SYBR safe stain. Gel A shows activity of heat treated PhoH2 protein on F1 mRNA substrate. Gel B shows activity of RNase treated PhoH2 protein on F1 mRNA substrate. In both gels; lane 1 and 2 contain RNA only controls, lane 3 contains a protein only control. Lane 4 had no added ATP. Lane 5 had EDTA added to the buffer. Lane 6- 10 show test reactions starting from 5 going to 60 minutes.

The contamination of PhoH2 by *E. coli* nucleic acid was able to be removed from the active form of PhoH2 by heat treatment at 37°C for 30 minutes prior to SEC purification.

3.2.11 Overall trends seen in PhoH2 activity assays

PhoH2 activity was observed across all tested mRNA substrates. U+C+R+S mRNA and U+C+R mRNA weren't able to be used in assay experiments, due to issues with amplification and degradation. An increase in ATP concentration in assay reactions reduces the product band size and in some reaction wells, the band is no longer present and is often replaced by a smear of nucleic acid. The smaller the size of the mRNA substrate the less ATP is required to affect the band, this was observed in assays with F1, F1 and F3 mRNA substrates. Overall most assays are contaminated by *E. coli* nucleic acid bound to PhoH2 proteins after IMAC and SEC purifications. This contaminating nucleic acid is often seen as faint banding

at a high molecular weight on agarose gels and sometimes lower molecular weight around where the mRNA substrate bands appear. This contaminating nucleic acid makes it difficult to truly determine what is occurring in the assay reactions. Although activity on mRNA substrates can still be determined, it is unclear exactly what this activity is. PhoH2 protein is made up of RNase and RNA helicase domains, so depicting the activity of either domain is hard to see, due to the other contaminating bands. In a majority of the assay reactions, the -ATP controls do not appear to hinder PhoH2 activity against mRNA substrates. It is possible that there may be external sources of ATP contamination that are coming through with the purified protein and its bound *E. coli* nucleic acid. Attempts were made to improve this control, by the addition of 6µl of 40mM EDTA, into assay reaction without ATP, but again this appears to have no affect on inhibiting PhoH2 activity. Interestingly it was observed that by increasing the concentration of ATP in an assay reactions, the contaminating bands of *E. coli* nucleic acid decreased in size and where sometimes degraded all together. This suggests that an increase in ATP concentration helps the RNase activity of PhoH2 proteins on not only the added mRNA substrates, but also the *E. coli* nucleic acid that had previously been tightly bound to PhoH2 protein.

3.3 **Conclusions**

PhoH2 protein from *M. smegmatis*, expressed in *E. coli* cells, purifies as a single peak in an IMAC purification and elutes at an imidazole concentration of ~120-180mM. Further purification through SEC sees three resulting peaks coming off the column. The active form of PhoH2 protein was identified as the second peak, which usually elutes off the column in the same position, for numerous purifications. Through activity assays it was identified that PhoH2 proteins purify with *E. coli* host nucleic acids bound to the protein, which are released and acted on when the protein is added to an activity assay with high concentrations of ATP and magnesium present. This contamination of host nucleic acids made identifying activity on the nine mRNA substrates sometimes difficult to understand and made crystallisation virtually impossible due to the heterogenous population of PhoH2 proteins. It was evident that PhoH2 protein does act on *sigF* mRNA, seen by a shift in position of the mRNA band in the gel, to a lower molecular weight and a decrease in the intensity of the mRNA band when PhoH2

was added. PhoH2 also showed activity against the fragments of *sigF*, showing the most activity on fragment F1. In regards to the other mRNA substrates, it appears that PhoH2 does show some activity towards them, but this activity is limited as the size and complexity of the mRNA substrates increased. Unfortunately mRNA substrates Upstream + *chaB* + *rsbW* + *sigF* and Upstream + *chaB* + *rsbW* were difficult to create and therefore it was undetermined whether PhoH2 proteins could act on these mRNA substrates and therefore conclusively deduce that PhoH2 targets SigF. Heat treatment of PhoH2 removed the contaminating RNA from the active form of PhoH2 and PhoH2 remained active post-treatment. Improvements to the negative controls (no ATP and EDTA) are needed to show that reactions without ATP and magnesium abolish PhoH2 activity. If changes cannot be made to improve these negative controls, perhaps a mutant of PhoH2 may be a better choice to show that the activity we are observing in the activity assays is due to the activity of PhoH2.

4 Chapter 4: Gene expression of mc²155 Δ*phoH2* compared to mc²155 in *M. smegmatis* using qPCR

4.1 Introduction

PhoH2 is made up of a C-terminal PhoH domain and an N-terminal PIN domain. The PhoH domain is an RNA helicase, while the PIN domain has RNase activity [4]. PhoH2 is highly conserved across mycobacteria and PhoH2 from *M. tuberculosis* shares 90.7% sequence homology to PhoH2 from its non-pathogenic relative *M. smegmatis* [144].

This chapter will focus on *phoH2* gene expression from *M. smegmatis* as little is known about its biological activity. A recent experiment by Dr Emma Andrews (from the University of Waikato, proteins and microbes laboratory), used RNAseq (RNA-sequencing) to deduce differences between the transcriptomes of *M. smegmatis* strains mc²155 (wild-type) and mc²155 Δ*phoH2* (knock-out) under standard growth conditions (LB tyloxapol media). RNA was extracted at 24, 48 and 72 hour time points and sent to BGI for transcriptome analysis. RNAseq, involves the use of next-generation sequencing technology to study the transcriptome at the nucleotide level [146]. The transcriptomes contain all the messenger RNA molecules and as such reflects genes that are being actively expressed at any given time [145]. Genes within the resulting transcriptomes were shortlisted if they met the following criteria; >150 reads, >100 RPKM (Reads Per Kilobase of transcript, per Million mapped reads) and >2 fold expression change. The shortlisted genes revealed there was upregulation of a suite of genes that comprise the *sigF* regulon in the *phoH2* deletion strain compared to the wild-type at the 48 hour time point. In addition, genes that showed an expression change <2 that fell within the *sigF* regulon were also collated. Literature has described SigF protein as being expressed in all growth stages of *M. smegmatis*, under normal growth conditions [73, 74]. It is hypothesised that during the exponential phase of growth, SigF is only expressed at basal levels, due to its tight regulation by RsbW, its antisigma factor [83]. While in stationary phase, i.e., after 48 hours of growth,

M. smegmatis starts to encounter stressful conditions, like toxin build up, and anti-sigma factor antagonists (RsfA and RsfB) regulate RsbW and stop it from inhibiting SigF, allowing an increase in the expression of SigF protein [74]. Overall RNAseq analysis revealed 129 genes in total as being upregulated in mc²155 Δ*phoH2* compared with mc²155 and 114 of these genes have been revealed to belong to the *sigF* regulon [46, 47 ,74] and accounts for 78% [47]. Following on from these results, Dr Andrews hypothesised that PhoH2 may be acting as a potential regulator of SigF, targeting (a) the *sigF* mRNA transcript directly or (b) via an indirect mechanism, that targets the regulation of anti-sigma factor antagonists by regulating expression through PhoH2's RNA helicase and RNase activity. The regulation of SigF is complex and varies between species and is not fully understood. Current literature describes the *sigF* gene in *M. smegmatis* as being part of an operon with its anti-sigma factor RsbW encoded directly upstream of SigF and a ChaB family protein encoded upstream of *rsbW* [83].

To validate these RNAseq results, RNA was extracted from fresh cultures of *M. smegmatis* mc²155 and mc^{2}155 Δ*phoH2*, after 48 hours of growth and qPCR was used to look at differences in transcript levels of 12 genes; MSMEG_0267, MSMEG_1773, MSMEG_3419, MSMEG_6213, MSMEG_1097, MSMEG_1782, MSMEG_2415, MSMEG_1804, MSMEG_0064, MSMEG_4712, MSMEG_6384 and MSMEG_2758. The first seven genes were chosen as they are part of the *sigF* regulon and had shown significant up regulation in the mc²155 Δ*phoH2* knock-out strain compared to the mc²155 wild-type. The remaining five genes showed down regulation or no change in expression in mc²155 ΔphoH2 compared to mc²155. These genes were chosen to see if the qPCR results would show the same trends as a way to validate RNAseq results.

Quantitative polymerase chain reaction (RT-qPCR) is an accurate and quantitative method for measuring the expression of genes relative to a house-keeping gene, through the creation of complementary DNA (cDNA) transcripts from mRNA and quantifiably measures the amplification of targeted cDNA using a fluorescent dye [147]. qPCR was used to determine if the 12 genes of interest (GOI) are up or down regulated in the mc²155 Δ*phoH2* strain versus the mc²155 from *M. smegmatis* and these expression ratios will be compared to those from RNAseq results. The aim of these qPCR experiments was to not only determine if PhoH2

may regulate *sigF* gene expression, but to also validate Dr Emma Andrews RNAseq results.

4.1.1 Principles behind qPCR

Quantiative PCR (qPCR) was first introduced back in 1993 by Mullis and co workers [153, 154] and is now a widely used molecular technique for the real-time amplification of PCR products. In conventional PCR it is only the end product of the reaction that can be assessed (through use of gel electrophoresis), which can be time consuming and does not accurately show differences in the starting quantity of the template material. qPCR on the other hand, allows for the detection and quantification of the amplicon product, while it is being synthesised.

In order to detect and quantify the accumulation of a specific PCR product, a fluorescent dye, such as SYTO 82, is included into the PCR reaction mixture, that specifically binds to double stranded DNA (dsDNA) producing a significant fluorescent read out. SYTO 82 was chosen for this qPCR experiment as it has a much greater detection limit than SYBR Green, and unlike SYBR Green it does not inhibit PCR reactions and won't conversely bind to GC rich regions [165]. This makes STYO 82 the optimum choice for detecting lowly expressed genes in the GC rich *M. smegmatis* bacterium. At the end of each qPCR cycle an excitation source at a specific wavelength is used to measure the amplitude of SYTO 82 dye bound to dsDNA, which will rise after each cycle, as the amount of product increases. This means that amplicon production can be measured as a function of cycle number. Here the reactions were run for 40 cycles and the threshold (Cq) was selected, which is defined as the cycle at which a significant increase in fluorescence (above the background level) is first detected. Genes that start off with a larger number of mRNA transcripts within a cell, will reach the threshold cycle earlier (i.e. have a lower cycle number) than those genes with a lower number of mRNA transcripts (i.e. have a higher cycle number) [156]. Intercalating dyes, like SYTO 82, are not sequence specific but they allow the generation of melt curves by including a dissociation run at the end of each qPCR program. The double-stranded PCR product is melted during a slow heating step, which results in the release of the intercalculating dye and a reduction of the

fluorescent signal. Amplicon-specific melt peaks are achieved when a decrease in fluoresce is plotted as a function of increasing temperature (-dF/dT) [157].

Overall, qPCR is an extremely sensitive and robust method for detecting the expression levels of (GOI) between samples and can be particularly helpful for detecting lowly expressed mRNA transcripts, or for detecting where there are small changes in expression levels [147]. Hence it is a widely used method for the quantification of mRNA and validation of RNAseq data [155]. qPCR is performed using a two-step approach, where the reverse-transcription reaction is separate from the qPCR reaction. The two-step method produces cDNA that can be used in many experiments with several different genes and with less handling of the RNA, to help reduce RNA degradation [158].

Once a qPCR run is complete, raw data is analysed by one of two main methods for gene quantification: absolute or relative. Absolute quantification uses a standard curve for each reaction run and determines the copy number or concentration of the target transcripts based on their quantification (or threshold) cycle (Cq) values. Relative quantification determines variation in mRNA levels of a target gene relative to a house-keeping (HK) or reference gene, and is expressed as a relative fold change in target expression values [161]. For this qPCR experiment relative gene quantification was chosen as it fitted better with the experimental aim, which was; to determine differences in gene expression between mc²155 and mc²155 Δ*phoH2* samples.

There are many mathematical models available to calculate relative expression, some take into account the amplification efficiency of the PCR reaction and correct for this while others do not, instead assuming an ‘ideal’ efficiency where the PCR product doubles every cycle during the exponential phase [88]. The equation developed by Michael Paffl in 2001 [89] was the first to ensure each individual reaction efficiency was used. This is critical because not all primers used in PCRs have an ideal amplification efficiency, and small efficiency differences between genes can create biased results. An alternative method is the delta-delta Ct (ΔΔCt), published by Livak and Schmittgen in 2001 [163]. This method assumes the amplification efficiency of all primers is 100% which is acceptable as long as it has been validated that the gene of interest and reference

gene have similar efficiencies. Both methods were trialled in determining gene expression values and Pfaffls method was chosen as the most appropriate to analyse the qPCR data, due to varying PCR reaction efficiencies between runs.

4.2 Results and discussion

4.2.1 Growth of *M. smegmatis* mc²155 and mc²155 Δ*phoH2* cultures for RNA extraction

Three starter cultures of *M. smegmatis* strains mc²155 and mc²155 Δ*phoH2* were grown for 24 hours in LB tyloxapol media. The protocol for growth and measurement of *M. smegmatis* cultures is described in section 2.3.1. Cultures that had grown to a similar $OD_{600,}$ to starter cultures from the RNAseq experiment (data not shown), were used to inoculate fresh 100ml *M. smegmatis* cultures with a starting OD_{600} of 0.01 (Table 4.1). These six cultures were grown for a total of 48 hours with measurements taken after 24 and 48 hours. Table 4.1 below shows OD_{600} values for the starter cultures after 24 hours, the volume required to inoculate fresh cultures with a starting OD_{600} of 0.01 and the OD_{600} of the cultures after 24 and 48 hours. There was a lot of variability in OD_{600} values between cultures, throughout the different growth stages. The final 48 hour OD_{600} values show that most of the cultures have a very similar final reading, except for mc²155 1. Cultures that grew to a similar OD600 were chosen for RNA extraction (*M. smegmatis* mc^2155 3, mc^2155 *ΔphoH2* 2 and 3).

Table 4. 1: O.D values for *M. smegmatis* starter cultures and 100ml cultures from strains mc²155 and mc²155 *ΔPhoH2*. Starter cultures with a * symbol next to them were not used for large culture growth.

Cultures:	**Starters OD_{600} after 24hrs:**	**Volume to add for 0.01 OD_{600}:**	**Cultures OD_{600} after 24hrs:**	**Cultures OD_{600} after 48hrs:**
mc^2155 **1***	0.264	3.79 ml	0.8300	2.2479
mc^2155 **2**	0.191	5.24 ml		
mc^2155 **3**	0.228	4.39 ml	0.8749	1.8371
ΔphoH2 mc^2155 **1** *	0.397	2.52 ml		
ΔphoH2 mc^2155 **2**	0.227	4.41 ml	0.8316	1.8496
ΔphoH2 mc^2155 **3**	0.319	3.13 ml	1.6496	1.8938

4.2.2 Extraction of Mycobacterial RNA

RNA extraction is the purification of RNA from biological samples. Intact mycobacterial mRNA is often difficult to extract due to; the short half-life of RNA, the physical strength and mycolic acid content of the cell wall [141], and low amount of RNA compared to DNA in cells [142]. Several methods are used in molecular biology to isolate RNA from samples, with the most common of these using guanidinium isothiocyanate (GITC) phenol-chloroform extraction. GITC is a chaotropic buffer used to help prevent the degradation of RNA, by inhibiting endogenous RNases. This method relies on phase separation of the aqueous sample from the solution containing water-saturated acid phenol, chloroform and GITC, which is a chaotropic denaturing solution and results in an upper aqueous phase and a lower organic phase (mainly phenol). Nucleic acid (RNA) partitions in the aqueous phase, while proteins partition in the organic phase. In the last step, RNA is recovered from the aqueous phase by precipitation with isopropanol or ethanol [143]. RNA quality has a big impact on the results of cDNA synthesis.

The integrity of the RNA was determined by visualisation on agarose gels prior to the generation of cDNA. Samples of *M. smegmatis,* mc²155 3 and mc²155 Δ*phoH2* 2, DNase treated, total RNA were run on a 1% agarose gel, with SYBR safe stain, in 1xTAE buffer (protocol for agarose gels is in section 2.1.3.1). Intact total RNA has distinct 23S and 16S rRNA bands and a less distinct 5S rRNA band. The 23S rRNA band ideally is approximately twice as intense as the 16S band, and both the 23S and 16S bands should be brighter than the 5S band. Degraded RNA will show as heavy smearing at a low molecular weight and will lack the sharp 23S and 16S rRNA bands [148], while partial degradation is indicated by the presence of minor smearing. The RNA samples below (Figure 4.1) show distinct bands present for 23S, 16S and 5S rRNA, as well as some partial degradation. This RNA is of suitable quality for cDNA generation.

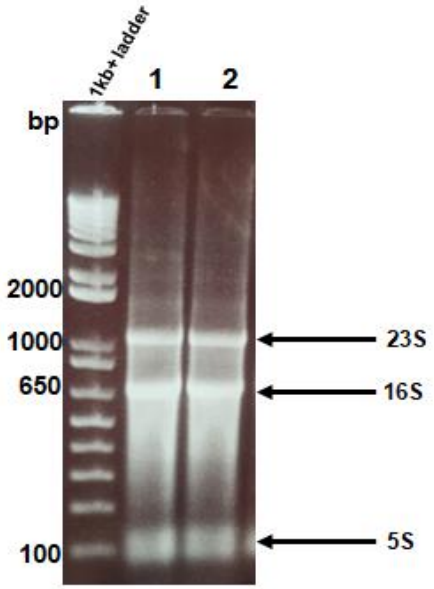

Figure 4. 1: RNA extraction from cultures of *M. smegmatis* strains mc²155 and mc²155 *ΔPhoH2* after 48 hours of growth. Lane 1 is RNA extracted from *M. smegmatis* mc²155 **3** and lane 2 is RNA extracted from *M. smegmatis* mc²155 *ΔphoH2* **2** (Table 4.1). Arrows indicate 23S, 16S and 5S rRNA bands. Bands sizes were determined against a molecular weight (1kb+) ladder.

4.2.2.1 RNA quantification and purity

In addition to knowing the integrity of the RNA, accurate assessment of the yield and its purity is important as well. The accuracy of the yield can be affected by several things; accuracy of the measuring instrument, contamination with gDNA, contamination with salts, and the level of degradation [148]. Here RNA quantity and purity was analysed by spectrophotometric profiling on a NanoDrop™ 1000 spectrophotometer (Thermo Scientific, USA). RNA has a UV absorption maximum at 260nm, allowing the quantity of RNA to be determined. The 280nm absorbance is also measured because this is typically where protein and phenolic compounds have a strong absorbance and this allows the purity of an RNA preparation to be determined using the ratio of absorbance at 260 and 280nm. Ultrapure RNA will give a A260/A280 ratio of 2.1, however anything above 1.8 is considered of sufficient quality to use. It is also very important to observe the 260/230nm reading as it provides information as to the contaminants present which produce peaks in this region (e.g. proteins, chaotropic salts and phenol) [149]. This ratio is used a secondary measure of nucleic acid purity. Expected 260/230 values are commonly in the range of 2.0-2.2, but accepted values are

anything higher than the 260/280 value [148]. The table below [Table 4.2] shows nucleic acid concentration, 260/280 and 260/230 ratios for RNA samples; *M. smegmatis* mc²155 3 and mc²155 Δ*phoH2* 2. Both samples of RNA have high reported yields, although these values are to be treated with caution as these values may not be an accurate representation of the amount of RNA. The 260/280 and 260/230 ratio provides information regarding contamination by DNA and/or other chemicals [150]. Here both samples the ratios are above two, suggesting little contamination within the RNA samples.

Table 4. 2: Nanodrop readings for RNA samples; mc²115 3 and mc²115 Δ*phoH2* 2

Sample ID of RNA:	**Nucleic acid concentration:**	**Unit:**	**260/280**	**260/230**
mc²155 **3**	2602.4	ng/µl	2.01	2.30
mc²155 Δ*phoH2* **2**	1506.6	ng/µl	2.03	2.57

4.2.3 Selection of a Suitable House-keeping (HK) Gene for Relative Quantification

Unfortunately, there is no ideal, universal house-keeping (HK) gene. In eukaryotes, several HK genes have been identified across different organisms, while in prokaryotes there is no consensus on which genes may be good universal HK genes. Quantification of gene expression is affected by several factors, such as RNA quality, primer efficiency, cDNA synthesis and the amount of amplified cDNA template, all of which can influence data analysis. To overcome this issue, expression levels of the target gene should be normalised to an endogenous HK gene whose expression remains unchanged under the experimental conditions. The normalisation of the data against HK genes is an important step in the quantification of gene expression. HK genes are also used to correct for differences between samples, such as variation in the total quantity of RNA, differences within the reverse transcriptions producing the cDNA and variation in qPCR efficiency. The selection of HK genes was initially based on those most commonly used in the literature, such as *sigA.* The gene *sigA* was initially the only HK gene designed for normalisation, however after inconsistencies with amplification more research [152] lead to the inclusion of two more HK genes (*ftsZ* and 16S rRNA) and a different primer set for *sigA* was tried. According to

the Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) guidelines, several HK genes are necessary to obtain accurate data normalisation. These three HK genes, all belonging to a range of different functional and abundance classes, were first trialled with *M. smegmatis* gDNA. A temperature gradient was used to determine the optimum annealing temperature, using a normal PCR, with the same enzyme and conditions used for a qPCR run (protocol in section 2.3.6). Reactions were performed from 54°C to 65°C with the new HK genes (*sigA*1, *sigA*2, *ftsZ* and 16S). Figure 4.2 shows that all the HK genes were successfully amplified with annealing temperatures that ranged from 54°C to 65°C.

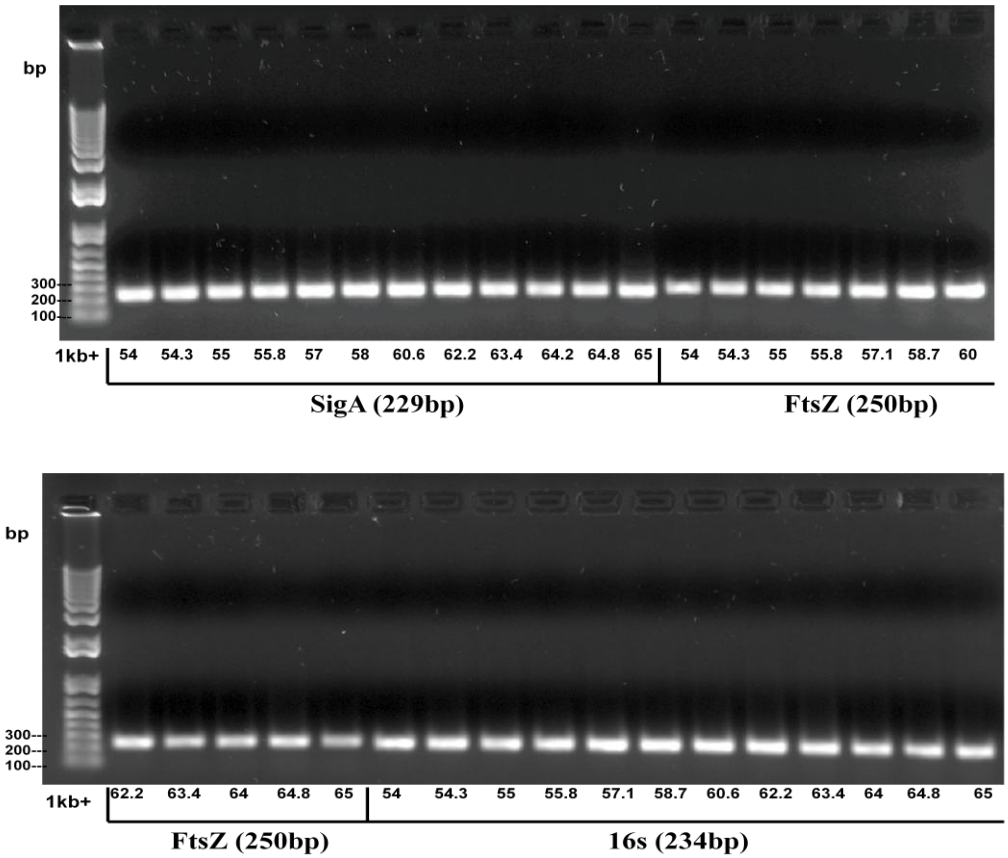

Figure 4. 2: Results of PCR temperature gradient amplification of HK genes with gDNA, run on a 1% agarose gel, with SYBR safe stain. Numbers underneath each band represents the annealing temperatures (°C) that each PCR product was amplified with. Band sizes were determined against a 1kb+ molecular weight ladder.

Once it was confirmed that the HK genes could amplify using gDNA at any temperature between the ranges 54°C to 65°C. A second PCR was carried out with the HK genes, keeping the same PCR conditions, with a single annealing temperature of 54°C. The template for these PCR reactions was the fresh cDNA

that was made from 4µg of DNase treated mc²155 RNA and mc²155 Δ*phoH2* RNA. Figure 4.3 shows that only the 16S HK gene was able to be amplified from gDNA and both of the cDNA samples, made up from mc²155 RNA and mc²155 Δ*phoH2* RNA. The original *sigA* 1 primers were only successful in amplifying SigA from gDNA, whereas, *sigA* 2 primers only successfully amplified SigA from wild-type cDNA. For the *ftsZ* primers amplification was only seen with gDNA and this amplification resulted in multiple bands appearing on an agarose gel. Due to these results only the 16S HK gene could be used in qPCR reactions to normalise the genes of interest. Since there are two copies of the gene, the expression levels for 16S will be higher than the GOI's and so the cDNA for 16S qPCR samples needed to be diluted (1:120) to remedy this issue. It was also decided to include MSMEG_2758 (*sigA*) in qPCR runs, but as a gene of interest, to see if sigA would show a similar gene expression trend to the RNAseq data.

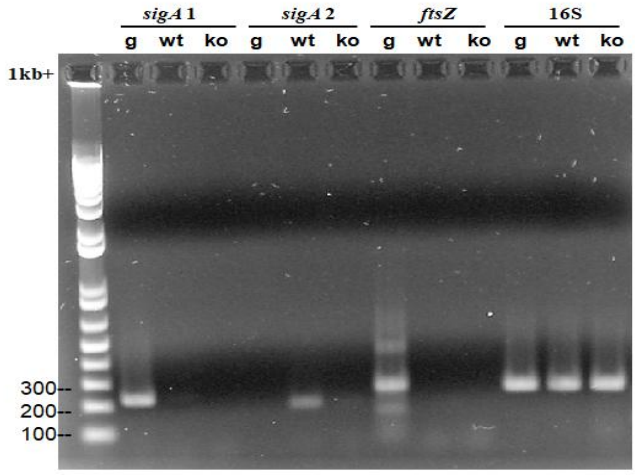

Figure 4. 3: Results of PCR amplification of HK genes with cDNA, run on a 1% agarose gel, with SYBR safe stain. All lanes represented by a 'g' symbol contain HK genes amplified with gDNA, all lanes represented by a 'wt' symbol contain HKG's amplified up by cDNA generated from mc²155 RNA and all lanes represented by a 'ko' symbol contain HK genes amplified up by cDNA generated from mc²155 Δ*phoH2* RNA. Band sizes were determined against a 1kb+ molecular weight ladder.

4.2.4 Genes of interest (GOI)

There was a total of 12 genes of interest (GOI's) used in qPCR experiments. One gene represented *sigF*. Seven of the 12 genes were chosen as they were part of the *sigF* regulon and they had shown a significant (fold change 2>) upregulation in *M. smegmatis* mc²155 *ΔphoH2* strain compared to mc²155 strain. The remaining

genes, outside of the *sigF* regulon, where chosen to see if there expression levels would match those given in RNAseq results. If values were similar between the two techniques, then this would help to validate overall results from the RNAseq data. All GOI's were first tested out using *M. smegmatis* gDNA, to ensure that the primers were able to amplify each respective gene and that there was no non-specific amplification. GOI's were amplified up in a normal PCR reaction, with HotFIREPol enzyme kit and gDNA as a template, as described in section 2.1.3.5. RNA only controls were included to ensure that there was no contaminating gDNA remaining in the RNA prep that was used to make cDNA. No template (NT) controls were also included to insure there was no contamination in the reactions. The following figure (Figure 4.4) shows that all GOI were able to be amplified from gDNA and there is no obvious non-specific amplification. The NT control reactions have no bands present, except for very faint low molecular weight smearing, which are primer dimers. Product band sizes match the expected products for the primer pairs, against a molecular weight 1kb+ DNA ladder.

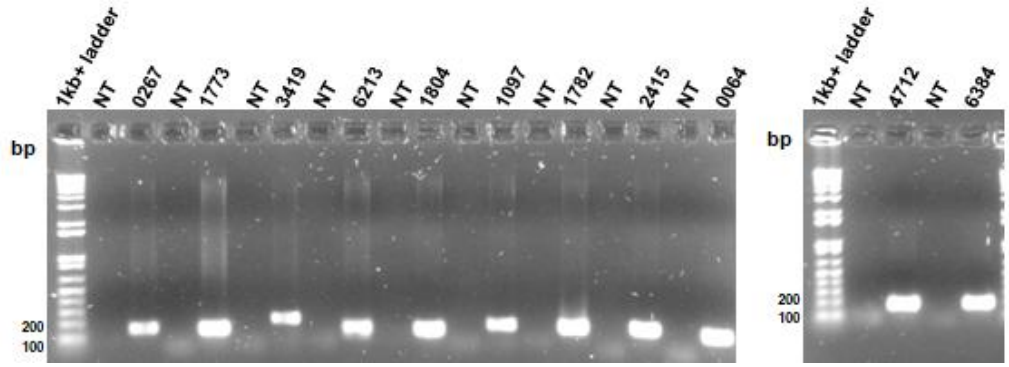

Figure 4. 4: GOI's tested out with *M. smegmatis* gDNA in a normal PCR. PCR's were run on 1% agarose gels, with SYBR safe stain. NT stands for PCR's that do not contain gDNA, with all other reaction components kept the same. Band sizes were determined against a 1kb+ molecular weight ladder.

After it was evident that the primer pairs for all genes were able to amplify using gDNA, one primer set, for the gene MSMEG_1773, was tested with mc²155 and mc²155 Δ*phoH2* cDNA in a normal PCR reaction, using conditions similar to a qPCR run, as described in section 2.3.6, with 54°C as the given annealing temperature. This was done as there was a limited amount of cDNA and the rest would be needed for additional qPCR runs. Figure 4.5 shows the PCR amplification of MSMEG_1773 from mc²155 and mc²155 Δ*phoH2* cDNA. MSMEG_1773 was successfully amplified from both cDNA and gDNA. There

were also no bands present in the NT controls indicating that the reaction mixture did not contain any contaminating DNA. The RNA controls show that there is none/or very little gDNA present in the RNA, this is important as any gDNA present in the RNA, when it is transcribed to cDNA, can create biased results.

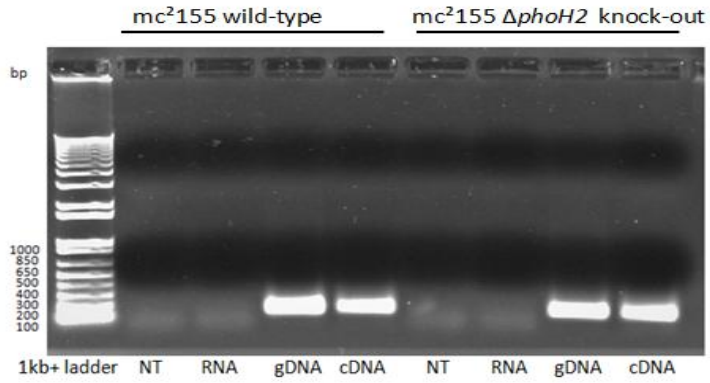

Figure 4. 5: Gel electrophoresis of gene MSMEG_1773 in PCR's with *M. smegmatis* gDNA, RNA and cDNA from mc²155 and mc²155 Δ*phoH2*. NT stands for PCR's that do not contain gDNA, with all other reaction components kept the same. NT, RNA, cDNA and gDNA were all tested as templates for amplification of MSMEG_1773. *M. smegmatis* mc²155 and mc²155 *ΔphoH2* cDNA and RNA were tested separately and corresponding reactions are labelled at the top of the gel. Band sizes were determined against a 1kb+ molecular weight ladder.

4.2.5 Analysis of qPCR runs

All qPCR reactions were performed on a Mic qPCR cycler (bio molecular systems) and analysis of these qPCR runs was done using the Mic software, linked to each qPCR run. Once a qPCR run was complete the analysis option on the Mic software became available. The raw data was displayed for each channel being acquired and was listed in the navigator bar. Underneath the analysis tab was an option for cycling analysis and clicking the + on this tab revealed an option of non-assay yellow. This option comes up as no assay had been assigned to the run and the reactions had been acquired in the yellow channel, due to the use of the SYTO 82 dye. Cycling analysis was used to determine the quantification cycle (C*q*) and reaction efficiency of each sample in the data set. By selecting Cycling Analysis, the software, by default, plotted baseline-corrected curves as fluorescence (*y*-axis), with a maximum fluorescence value of 100 units, against cycle number (*x*-axis), in logarithmic scale, for the target that was chosen. Cycle

data was plotted as cycle number (Cq). There are a number of parameters that can be set for Cycling Analysis, all of which will influence the calculation of C*q* values and efficiency such as; method, ignore cycles before, threshold value, exclusion and fluorescence cut off level. Parameters that are considered when determining Cq values are; method and threshold value. Cq values for each reaction run was determined as described in section 2.3.9. There are four different baseline correction methods available such as; dynamic, fixed length, LinReg and none. The LinReg method, provided by the Mic software, was used to determine the Cq values for all qPCR runs as it had been recommended by many previously published papers [159, 160, 161] and it minimised the generation of errors in quantification due to failures in efficiency calculation caused by either over or underestimation of the baseline values [162]. The LinRegPCR algorithm utilises the region of exponential growth to calculate Cq values.

4.2.6 Quality check of qPCR runs

To ensure accurate gene expression values were being determined a set criteria was created and applied to every run, as a way to reduce error. The criteria was as follows; runs must have an amplification efficiency above 0.89, triplicate values for each sample must be within 0.5 Cq of each other, Cq values for all runs must come up after 10 and before 25 cycles, the melt curve analysis for each run must not show multiple peaks and all end reactions must run as a single band, of the right product size, on gel electrophoresis. Cycling analysis and melt curves were examined for each gene. Reaction runs that did not meet all of the following criteria were not included in the final gene expression calculations. Most of the qPCR runs gave high amplification efficiencies, except for some runs from the genes MSMEG_1097, MSMEG_3419, MSMEG_6213 and MSMEG_0064 (data not shown). Due to the variation of amplification efficiency between runs, it was decided that the pfaffl method, which takes into account amplification efficiencies, was the best method of choice for determining gene expression. The gene MSMEG_1773 continuously showed qPCR reaction runs that obeyed all of the necessary criteria and therefore has been used as an example of how qPCR runs should look when they meet the necessary criteria. In the following figure (Figure 4.6) the gene MSMEG_1773 has been analysed by Mic qPCR software and gel

electrophoresis. Figure A shows a graph, generated by the Mic software, of the reaction runs of MSMEG_1773 amplified up in real-time, using mc²155 and mc²155 Δ*phoH2* cDNA. The amplification curves come up after 10 cycles and triplicate runs all come up in the same place. Figure B shows melt curves that were also generated by the Mic software, for the reaction runs of MSMEG_1773 amplified up in real-time, by mc²155 and mc²155 Δ*phoH2* cDNA. The figure shows that only one peak is present, indicating lack of contamination by primer dimers or non-specific amplification. Figure C shows qPCR reactions, run out on a 1% agarose gel, with SYBR safe stain (section 2.1.3.2). The gel shows there is only one band product present, for each of the runs, and these bands are all of the same size and intensity. Genes that fit within the necessary criteria all had similar cycling and melt curve graphs and had only one band present in their gel electrophoresis (data not shown).

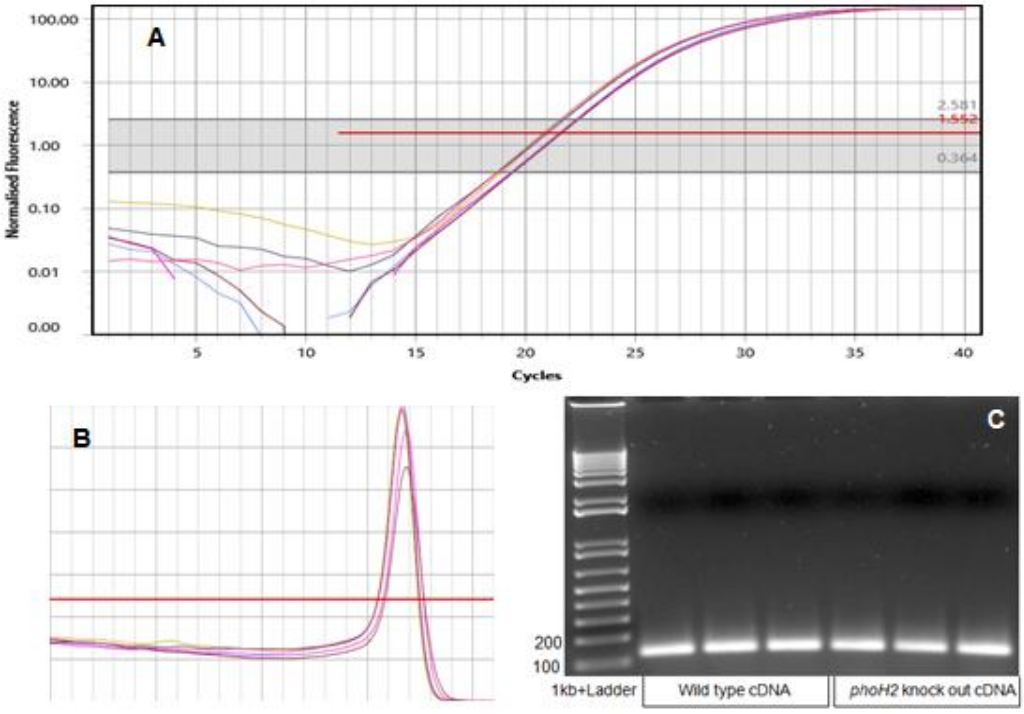

Figure 4. 6: qPCR run analysis of MSMEG_1773 using Mic qPCR software and gel electrophoresis. Figure A was generated from Mic qPCR software and shows normalised fluorescence against cycle number. Y-axis is in log scale. Grey section in the graph shows window of linearity (WOL) where the cq values are determined. All cq values in the graph come up after 10 cycles. In figure B, which was also generated from Mic qPCR software, it is shown that there is only one peak present in the melt curve analysis. Figure C, was generated by gel electrophoresis, using a 1% agarose gel, with SYBR safe stain (section 2.1.3.2). The gel shows one band present for both wild-type and Δ*phoH2* knock-out qPCR reaction runs and the band is of the expected product size (130bp). Band sizes were determined against a molecular weight 1kb+ ladder.

4.2.6.1 Primer dimers and non-specific amplification:

Amplification of primer dimers during qPCR may mask the actual results when using a dsDNA binding dye. Melt curves for qPCR runs were analysed to check for contamination from primer dimers or other non-specific amplification products in the samples. The formation of primer dimers were avoided where possible with good primer design, but were apparent in the melt curves and gel electrophoresis for some of the primer sets used. The following figure (Figure 4.7) shows qPCR runs on gel electrophoresis that display primer dimers and/or non-specific amplification. For the gene MSMEG_1097 there are multiple bands present on the gel, that are around 150bp, which is the expected product size for the primers used and other bands that represent non-specific amplification, which can be seen slightly greater than the 150bp band size of the GOI, and also around 400bp. Non-specific amplification often indicates poor primers. There are also bands below the GOI, under the 100bp mark. These bands are most likely from the formation of primer dimers. On the gel representing the gene MSMEG_2415 there are bands in the gel that are around 150bp, which is the expected product size for the primers used. There are also bands present above the GOI, around 400bp, these bands are evidence of non-specific amplification, which again indicated poor primers. For the gene MSMEG_6384, there are nice distinct bands present around 170bp, which represent the expected product size for the primer set of the GOI. However underneath these bands there is also presence of faint banding underneath the bands for the GOI. Again these bands are most likely due to the formation of primer dimers. For the gene MSMEG_0064 there is a faint band present under one of the gene products, which is also most likely a primer dimer. A big issue is the variability seen between qPCR runs, which is evident with genes like 2415, where only sometimes non-specific amplification occured, seen by multiple peaks in the melt curve analysis (data not shown). Often these variabilities come down to issues with the primers and the best solution for this is to order two sets of primers for each gene and determine through qPCR trials, which set is the most stable and optimal set to use.

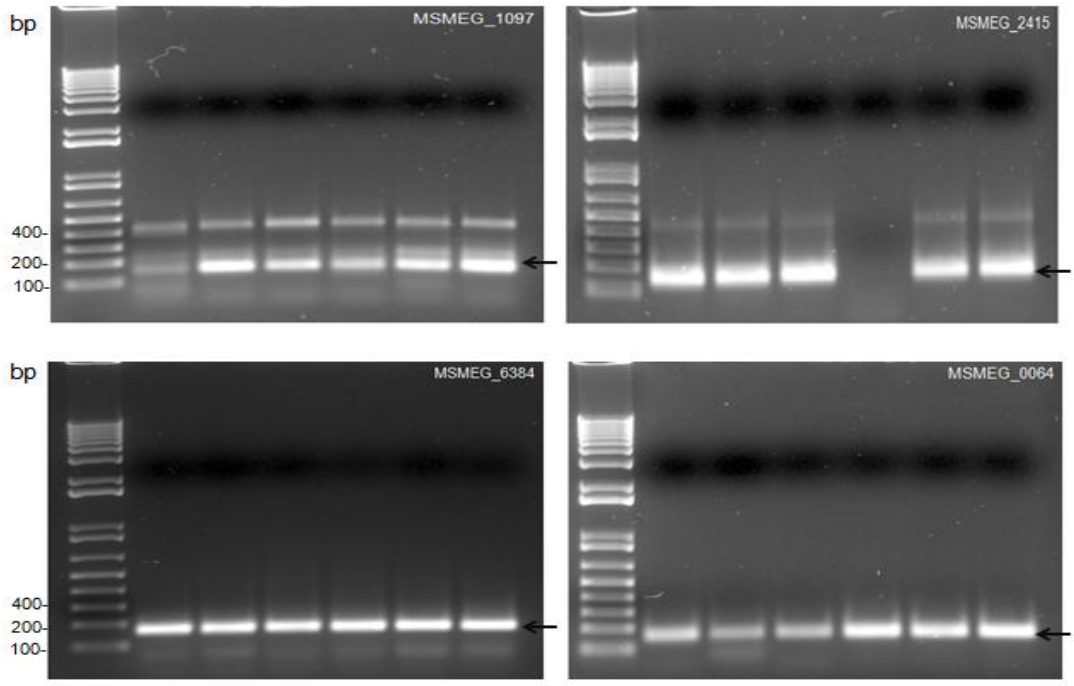

Figure 4. 7: qPCR runs of GOI's on 1% agarose gels, with SYBR safe stain. Gels are shown for qPCR runs for genes, MSMEG_1097, MSMEG_2415, MSMEG_6384 and MSMEG_0064 (names displayed in white, in right top corner of each gel). The six bands in each gel represent three replicates for wild-type and three replicates for knock-out cDNA. Black arrows in each of the gel point at the bands which are of the expected product size. Sizes of bands are determined against a 1kb+ molecular weight ladder run in the first lane of each gel.

4.2.7 Expression levels of GOI's in mc²155 *ΔphoH2* against mc²155 in *M. smegmatis*

Using the Cq values determined by the LinReg method, offered in the Mic qPCR software, an overall picture of expression can be obtained from the cycle numbers where fluorescence levels are detected above background (represented by Cq values). In a typical reaction, a higher starting template concentration will result in an earlier or lower Cq, as generally fewer amplification cycles are needed for the amplification product to rise above the background fluorescence signal. Cq values for wild-type against knock-out were analysed for runs representing the 12 genes of interest. Analysis of cycling runs for MSMEG_0267, MSMEG_1804, MSMEG_0064, and MSMEG_6384, showed that the Cq values for knock-out samples came up first compared to the wild-type, suggesting that in knock-out samples there was a higher starting amount of template of the GOI. While runs for

MSMEG_1773, MSMEG_6213, MSMEG_1097, MSMEG_1782, MSMEG_2415, showed there was no overall difference between the Cq values for knock-out and wild-type samples, this was determined if Cq values between wild-type and knock-out samples were within two cycles of each other. Often it was difficult to determine differences between knock-out and wild-type samples, as some runs contained a variation between triplicate values. But these results might also represent no difference in starting concentration of the GOI, in knock-out samples compared to the wild-type. Two genes; MSMEG_2758 and MSMEG_4712 had runs where the Cq values for the wild-type samples came up earlier compared to the knock-out, suggesting that there was a lower starting amount of the GOI in the knock-out samples compared to the wild-type (data not shown).

These results tie in well with the gene expression values determined, by the Pfaffl method, for each GOI. In the following figure (Figure 4.8) relative fold change, for nine genes, is depicted as a bar graph. Genes MSMEG_1804, MSMEG_0064 and MSMEG_6384 showed up regulation, with fold change values greater than two fold in Δ*phoH2* mc²155 compared to mc²155. MSMEG_1804 is *sigF* in *M. smegmatis* and is hypothesised here to be regulated by PhoH2, hence why there may be an up regulation in the *phoH2* knock-out strain of *M. smegmatis*. MSMEG_0064 encodes for a proline-proline-glutamic acid (PPE) family protein [86], which currently has an unknown function in *M. smegmatis* and is not currently part of the SigF regulon. PE/PPE family proteins, are named after their conserved PE (Pro-Glue) and PPE (Pro-Pro-Glu) domains [78]. The roles of most members of this family are unknown, although selected genes of this family are related to the virulence of *M. tuberculosis* [66]. Several transcriptional regulators involved in the regulation of PE/PPE family proteins have been characterised including; Lsr2 [34], RelA [33] and sigma factors, such as SigF [47,110], SigB [12] and SigD [14]. Genes MSMEG_4712 and MSMEG_2758 gave negative fold change values in mc²155 Δ*phoH2* compared to mc²155, but these values were not significant as they gave fold values less than two. The remaining genes (MSMEG_0267, MSMEG_1773, MSMEG_1783 and MSMEG_2415) gave fold change values less than two, and therefore their gene expression is not significant. Genes part of the sigF regulon contain an * symbol above their bar.

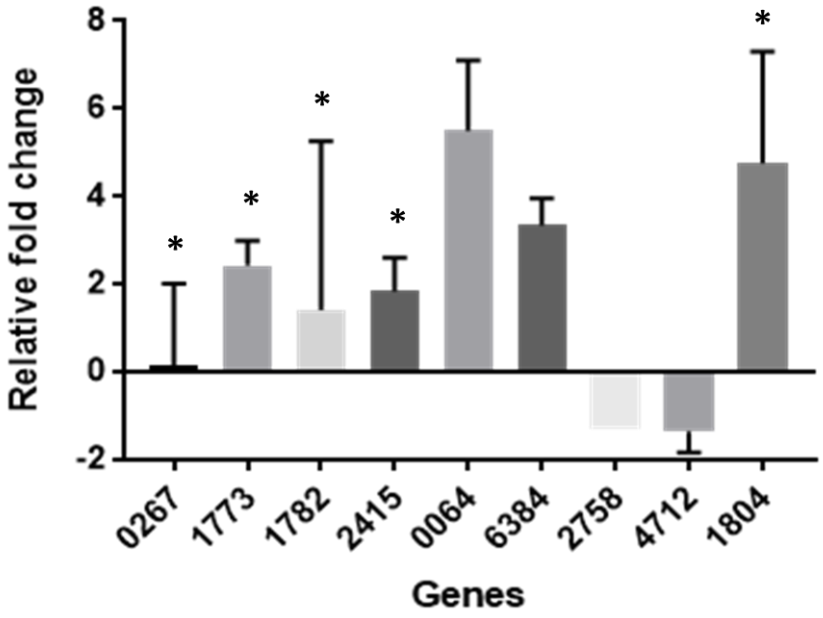

Figure 4. 8: Relative fold change of genes in *ΔphoH2* mc²155 (knock-out) samples versus mc²155 (wild-type), from *M. smegmatis*. Data is presented as a bar graph, with mean and S.D shown by error bars and the name for each gene directly under each bar. Relative gene fold changes (normalised to 16S) which were determined the Pflaffl method (Formula 2.2) and are expressed as fold change. Genes that contain the symbol * above their bar, are part of the sigF regulon. Graph was designed in GraphPad Prism, version 7.04.

4.2.8 Gene expression values for GOI's in qPCR versus RNAseq

The table below (Table 4.3) shows that out of the twelve genes examined using qPCR, only nine genes were used for gene expression analysis. There is quite a lot of variation in gene expression between the qPCR and RNAseq data. Genes MSMEG_1804 and MSMEG_0064 showed the highest expression levels of up regulation, while MSMEG_1782, MSMEG_4712 and MSMEG_2758 showed the lowest levels of down regulation, but these fold change values are not significant. Only two genes; MSMEG_4712 and MSMEG_2758, show similar gene expression levels in both qPCR and RNAseq data. As the aim of these qPCR experiments is to determine if PhoH2 protein is acting as a regulator of SigF and its regulon genes, the focus will now shift towards just *sigF* and genes belonging to the *sigF* regulon, as indicated in table 4.3. Genes within the *sigF* regulon and *sigF* itself, show quite contrasting gene expression values between RNAseq and qPCR data. *sigF* is the only gene from the qPCR data that shows significant up

regulation, while the rest of the genes have a low level of expression. While in the RNAseq data genes belonging to the *sigF* regulon show up regulation, while *sigF* itself, shows very little expression difference between mc²155 and mc²155 Δ*phoH2*.

Table 4. 3: Expression levels of genes in *M. smegmatis* strains; mc²155 against mc²155 Δ*phoH2* for qPCR and RNAseq results. The symbol * stands for genes who are part of the *sigF* regulon. Gene product descriptions are according to the UniProt database. N/A stands for genes whose expression levels were undetermined. Gene whose fold change was <2 fold, were labelled not significant. Fold change values are shown in brackets.

Gene name:	**Gene product:**	**qPCR gene expression:**	**RNAseq gene expression:**
MSMEG_0267 *	Esterase	Not significant (1.46)	Up regulated (6.32)
MSMEG_1773 *	Hypothetical protein	Not significant (1.93)	Up regulated (3.90)
MSMEG_3419 *	Hypothetical protein	N/A	Up regulated (5.44)
MSMEG_6213 *	Manganese containing catalase	N/A	Up regulated (4.86)
MSMEG_1804 *	SigF	Up regulated (4.89)	Not significant (0.5 <2 fold)
MSMEG_1097 *	Glycosyltransferase	N/A	Up regulated (5.22)
MSMEG_1782 *	oxidoreductase	Not significant (1.30)	Up regulated (6.14)
MSMEG_2415 *	Cation binding protein	Not significant (1.84)	Up regulated (5.12)
MSMEG_0064	PPE family protein	Up regulated (5.49)	Down regulated (-2.87)
MSMEG_4712	Pyruvate dehydrogenase	Not significant (-1.36)	Not significant (-1.05)
MSMEG_6384	Catalase / peroxidase	Up regulated (3.34)	Not significant (-1.51)
MSMEG_2758	SigA	Not significant (-1.30)	Not significant (-1.05)

The following figure (Figure 4.9) shows fold change expression values determined from qPCR analysis and RNAseq analysis, for the genes listed in Table 4.3. The expression values are shown as a bar graph to compare the differences between the expression levels of the different genes, with the different techniques.

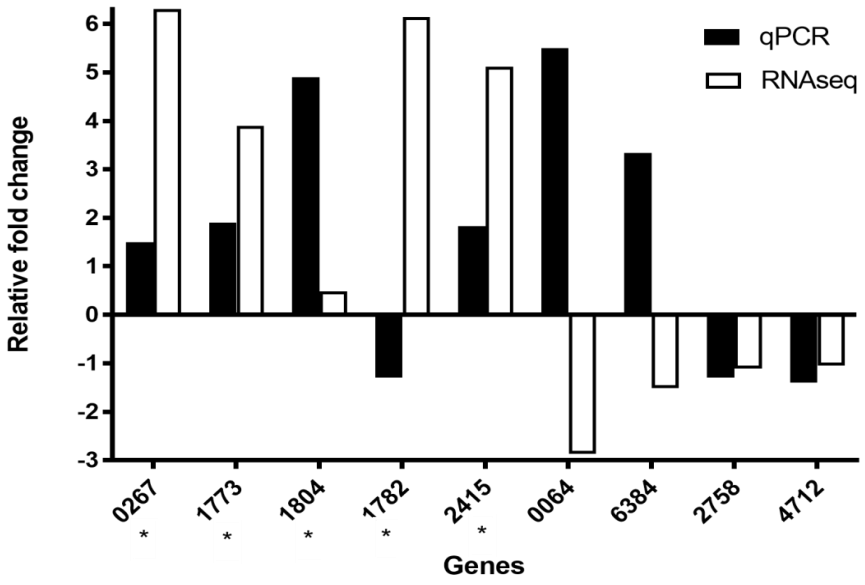

Figure 4. 9: Comparison of relative fold changes in qPCR versus RNAseq. Data is expressed as a bar graph and shows relative fold change of genes in *ΔphoH2* knock-out versus wild-type, from *M. smegmatis*, in qPCR compared with RNAseq. Genes that have an * next to their name are part of the sigF regulon. Graph was designed in GraphPad Prism, version 7.04.

4.3 Conclusions

qPCR is a powerful tool for analysing differences in gene expression between samples, unfortunately due to variability in the set up and analysis of qPCR techniques, qPCR can become inadequately standardised, complex and often gives varying conclusions. However, MIQE have released a set of guidelines detailing how best to proceed with qPCR setup and analysis, to ensure that future publications have accurate and reliable results. This research project presents an optimised RNA isolation and qPCR method, for the analysis of gene expression levels between Δ*phoH2* mc²155 and mc²155, from *M. smegmatis*, which have followed the MIQE guidelines where possible. In qPCR analysis of gene expression differences with Δ*phoH2* mc²155 compared to mc²155, in

M. smegmatis, showed that genes MSMEG_1804 and MSMEG_0064, gave significant levels of up regulation, while all other genes showed no significant difference in regulation between the strains. In comparison with the results from RNAseq analysis only two genes (MSMEG_2758 and MSMEG_4712) had similar expression values, while a majority of the other genes had quite contrasting expression values. In regards to *sigF* and its regulon genes, *sigF* itself in qPCR was significantly upregulated, while in RNAseq there was no significant difference in regulation expression. The *sigF* regulon genes also showed variation between the two techniques, with genes MSMEG_0267, MSMEG_1773, MSMEG_1782 and MSMEG_2415 being up regulated in RNAseq, and having no significant gene expression difference in qPCR. This difference may have come down to variability between techniques or through qPCR complications, such as the use of only one housekeeping gene. Another explanation for these differences seen in *sigF* and it regulon genes, in qPCR and RNAseq, may be due to capturing the genes at different stages of expression between the two different experiments used to harvest RNA, one for RNAseq and a second for qPCR. It is possible that RNA samples taken for RNAseq capture, the consequence of the misregulation of *sigF* by PhoH2 and therefore upregulation of *sigF* regulon genes and the RNA samples used for qPCR capture the initial misregulation (significant misregulation of *sigF*).

5 Chapter five: Conclusion

The protein PhoH2 can be found across many different bacterial species, including mycobacteria. Previous studies on PhoH2 proteins have identified them as a two domain protein, with an N-terminal PIN domain, with RNase activity and a C-terminal PhoH helicase domain. PhoH2 proteins form hexameric ring-like structures that are capable of unwinding and cleaving RNA. PhoH2 has ATP and magnesium dependent RNase and helicase activity towards specific sequences of RNA oligonucleotides with 5' terminal overhangs [4]. Deletion strain; mc²155 Δ*phoH2* in *M. smegmatis*, saw an upregulation of *sigF* regulon genes compared to *M. smegmatis* mc²155 (wild-type), with RNAseq analysis. The intent of this thesis was to understand more behind the biology of PhoH2 proteins and determine if they do play a role in the regulation of the expression of *SigF*, through direct RNase activity on *sigF* mRNA or indirectly, by targeting *sigF* antagonists, such as RsbW.

The first aim of this thesis was to design mRNA substrates from *sigF* and genes of its operon; *rsbW* and *chaB*. These mRNA substrates where then used in RNA unwinding and cleavage activity assays with IMAC and SEC purified PhoH2 protein from *M. smegmatis*, expressed in *E. coli*. These assays revealed that PhoH2 proteins display activity toward *sigF* mRNA that increased with increasing ATP concentration and time saw an increase in activity of the protein against the mRNA substrate. It was undetermined as to whether PhoH2 proteins were also active against *rsbW* and *chaB* mRNA without *sigF*, as no assays with the mRNA substrate Upstream + *chaB* + *rsbW* were successfully attempted, due to degradation of mRNA substrate before addition to assays. The exact details of PhoH2 activity was hard to determine, due to the presence of host *E. coli* nucleic acid contamination, co-purifying with PhoH2 proteins especially with the longer RNA substrates. When activity assays were visualised using Urea-PAGE gels or agarose gels, contaminating nucleic acid was present in these assays and interfered with interpretation of PhoH2 activity against added mRNA substrates. Another issue that hindered this determination of PhoH2 activity on mRNA substrates, was due to the size of a majority of the designed mRNA substrates. For

RNA to be able to pass through the gel matrix of Urea-PAGE gels, it needs to be smaller than 500bp. A majority of the mRNA substrates were bigger than 500bp, so assays reactions with these substrates were run out on 1% agarose gels, with SYBR safe stain. Agarose gels could show the resulting RNA products from activity assays, but in much less detail and clarity compared to assays examined using Urea-PAGE gels, as seen with the F1, F2 and F3 mRNA activity assays. Contaminating *E. coli* nucleic acid was extracted from IMAC and SEC purified PhoH2 protein and visualised on an agarose gel, which showed similar banding patterns to what was observed with samples of SEC PhoH2 protein peaks. With treatment of DNase or RNase, the first peak from a SEC purification lost nucleic acid through both RNase and Dnase treatments, while the second peak lost most of it nucleic acid from the RNase treatment suggesting the peak corresponding to the active form of PhoH2 contained mostly RNA. Various treatments were applied to the PhoH2 protein in an attempt to remove this *E. coli* nucleic acid. High salt (2M NaCl), heating and RNase A treatments on PhoH2 proteins were able to reduce the amount of contamination. Unfortunately the removal of nucleic acid from PhoH2, often resulted in protein precipitation or a reduction in protein activity. Activity of heat treated and RNase treated PhoH2 proteins were tested with ATPase malachite green assays. Heat treated PhoH2 protein changed to a blue colour with ATP, suggesting the protein was still active and able to turn over ATP, while RNase treated PhoH2 did not change colour with the addition of ATP, suggesting inactive protein. Future experiments with this protein could focus on exploring new techniques to remove *E. coli* nucleic acid, from PhoH2, with the maintenance of functioning protein. An example of this may be to use ATP during PhoH2 protein purifications, to encourage activity on the nucleic acid. The nucleic acid could then be removed from solution through ion exchange chromatography. Another future experiment to consider might be to look into sequencing *E. coli* nucleic acid that co-purifies with PhoH2 protein, to understand more about the binding targets of PhoH2 proteins.

The second aim of this thesis was to validate RNAseq results, with qPCR experiments. RNA was successfully extracted from *M. smegmatis* strains; mc²155 (wild-type) and mc²155 Δ*phoH2* (knock-out) and converted into cDNA that was used as a template to amplify up 12 genes of interest, in qPCR reactions. Genes

MSMEG_3419, MSMEG_6213, MSMEG_1097 and MSMEG_0064 were excluded from final expression analysis, as their runs were inconsistent. Gene expression values, for the remaining genes were determined by comparing Cq values from mc²155 runs against Δ*phoH2* mc²155 runs and normalising relative gene expression values against a housekeeping gene (16S). Final fold change values were determined using Pfaffls formula and results showed that genes MSMEG_1808 and MSMEG_0064 showed the greatest up regulation in Δ*phoH2* mc²155 compared to mc²155. MSMEG_1804 is *sigF* in *M. smegmatis* and is thought to be regulated by PhoH2, hence why there may be an up regulation in the mc²155 Δ*phoH2* strain of *M. smegmatis*. MSMEG_0064 encodes for a PPE family protein whose role in *M. smegmatis* is unknown. Gene expression results varied between RNAseq and qPCR, only two genes (MSMEG_4712 and MSMEG_2758) showed similar gene expression values. In regards to *sigF* and its regulon genes, in qPCR *sigF* showed up regulation, while the regulon genes showed no significant difference and in RNAseq, *sigF* showed no regulation difference, while the regulon genes showed up regulation. It is hypothesised that a reason for this difference seen in *sigF* and its regulon genes, in qPCR and RNAseq, may be due to capturing the genes at different points of mis-regulation, between the two different experiments. Future experiments relating to this project should take more samples at varying time points close to the 48 hour growth mark, in *M. smegmatis*. These new samples can be analysed by qPCR to understand more about the expression of *sigF* and it regulon genes, and the role of PhoH2 in this process during *M. smegmatis* stationary phase of growth. For further qPCR experiments it is recommended that more housekeeping genes should be introduced into gene expression calculations, to help overcome the bias that one housekeeping gene may introduce.

References

1.Ang, M., Vasconcelos-Santos, D. V., Sharma, K., Accorinti, M., Sharma, A., Gupta, A., and Chee, S. (2016). Diagnosis of Ocular Tuberculosis. *Ocular Immunology and Inflammation, 26*(2), 208-216. doi:10.1080/09273948.2016.1178304.

2.Global tuberculosis report 2017. (2017, December 06). Retrieved from http://www.who.int/tb/publications/global_report/en/

3.Chetty, S., Ramesh, M., Singh-Pillay, A., & Soliman, M. E. (2017). Recent advancements in the development of anti-tuberculosis drugs. *Bioorganic & Medicinal Chemistry Letters, 27*(3), 370-386. doi:10.1016/j.bmcl.2016.11.084.

4.Andrews, E. S., & Arcus, V. L. (2015). The mycobacterial PhoH2 proteins are type II toxin antitoxins coupled to RNA helicase domains. *Tuberculosis, 95*(4), 385-394.

5.Aguilar-Ayala, D. A., Tilleman, L., Nieuwerburgh, F. V., Deforce, D., Palomino, J. C., Vandamme, P., Martin, A. (2017). The transcriptome of *Mycobacterium tuberculosis* in a lipid-rich dormancy model through RNAseq analysis. *Scientific Reports, 7*(1). doi:10.1038/s41598-017-17751-x.

6.Pagán, A. J., & Ramakrishnan, L. (2014). Immunity and Immunopathology in the Tuberculous Granuloma. *Cold Spring Harbor Perspectives in Medicine, 5*(9). doi:10.1101/cshperspect.a018499.

7.Smith, I. (2003). *Mycobacterium tuberculosis* pathogenesis and molecular determinants of virulence. *Clinical microbiology reviews, 16*(3), 463-496.

8.Peddireddy, V., Doddam, S. N., & Ahmed, N. (2017). Mycobacterial Dormancy Systems and Host Responses in Tuberculosis. *Frontiers in Immunology, 8*. doi:10.3389/fimmu.2017.00084.

9.Ryndak, M. B., Singh, K. K., Peng, Z., & Laal, S. (2015). Transcriptional Profile of *Mycobacterium tuberculosis* Replicating in Type II Alveolar Epithelial Cells. *Plos One, 10*(4). doi:10.1371/journal.pone.0123745.

10.Dillon, B. J. (2017). *Caldalkalibacillus thermarum* PhoH2: Solving the solubility puzzle (Doctoral dissertation, University of Waikato).

11.Andrews, E. S. V. (2013). *The biology and biochemistry of PhoH2 proteins* (Doctoral dissertation, University of Waikato).

12.Fontán, P. A., Voskuil, M. I., Gomez, M., Tan, D., Pardini, M., Manganelli, R., ... & Smith, I. (2009). The *Mycobacterium tuberculosis* sigma factor σB is required for full response to cell envelope stress and

hypoxia *in vitro*, but it is dispensable for *in vivo* growth. *Journal of bacteriology*, *191*(18), 5628-5633.

13.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.

14.Sampson, S. L. (2011). Mycobacterial PE/PPE proteins at the host-pathogen interface. *Clinical and Developmental Immunology*, *2011*.

15.Frampton, R., Aggio, R. B., Villas-Bôas, S. G., , V. L., & Cook, G. M. (20*11). Tox*in-Antitoxin Systems of *Mycobacterium smegmatis* Are Essential for Cell Survival. *Journal of Biological Chemistry*, *287*(8), 5340-5356. doi:10.1074/jbc.m111.286856.

16.Ahidjo, B. A., Kuhnert, D., Mckenzie, J. L., Machowski, E. E., Gordhan, B. G., Arcus, V., 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). doi:10.1371/journal.pone.0021738.

17.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.

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

19.Winther, K., Tree, J. J., Tollervey, D., & Gerdes, K. (2016). VapCs of *Mycobacterium tuberculosis* cleave RNAs essential for translation. *Nucleic acids research*, *44*(20), 9860-9871.

20.Gerdes, K., Christensen, S. K., & Løbner-Olesen, A. (2005). Prokaryotic toxin–antitoxin stress response loci. *Nature Reviews Microbiology*, *3*(5), 371.

21.Maisonneuve, E., Shakespeare, L. J., Jørgensen, M. G., & Gerdes, K. (2011). Bacterial persistence by RNA endonucleases. *Proceedings of the National Academy of Sciences*, *108*(32), 13206-13211.

22.Maisonneuve, E., Castro-Camargo, M., & Gerdes, K. (2013). (p) ppGpp controls bacterial persistence by stochastic induction of toxin-antitoxin activity. *Cell*, *154*(5), 1140-1150.

23.Matelska, D., Steczkiewicz, K., & Ginalski, K. (2017). Comprehensive classification of the PIN domain-like superfamily. *Nucleic acids research*, *45*(12), 6995-7020.

24.Wall, D., & Kaiser, D. (1999). Type IV pili and cell motility. *Molecular microbiology*, *32*(1), 01-10.

25.Miallau, L., Faller, M., Chiang, J., Arbing, M., Guo, F., Cascio, D., & Eisenberg, D. (2009). Structure and proposed activity of a member of the VapBC family of toxin-antitoxin systems VapBC-5 from *Mycobacterium tuberculosis. Journal of Biological Chemistry*, *284*(1), 276-283.

26.Gerdes, K., Christensen, S. K., & Løbner-Olesen, A. (2005). Prokaryotic toxin–antitoxin stress response loci. *Nature Reviews Microbiology*, *3*(5), 371.

27.Arcus, V. L., Rainey, P. B., & Turner, S. J. (2005). The PIN-domain toxin–antitoxin array in mycobacteria. *Trends in microbiology*, *13*(8), 360-365.

28.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.

29.Arcus, V. L., Bäckbro, K., Roos, A., Daniel, E. L., & Baker, E. N. (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.

30.Kim, S. K., 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. doi:10.1128/jb.175.5.1316-1324.1993.

31.Koonin, E. V., & Rudd, K. E. (1996). Two domains of superfamily I helicases may exist as separate proteins. *Protein science*, *5*(1), 178-180.

32.Mohanty, B.K. and Kushner, S.R. (2016) Regulation of mRNA decay in bacteria. Annu. Rev. Microbiol. **70**, 25–44doi:10.1146/annurev-micro-091014-104515.

33.Dahl, J. L., Kraus, C. N., Boshoff, H. I., Doan, B., Foley, K., Avarbock, D., ... & Barry, C. E. (2003). The role of RelMtb-mediated adaptation to stationary phase in long-term persistence of *Mycobacterium tuberculosis* in mice. *Proceedings of the National Academy of Sciences*, *100*(17), 10026-10031.

34.Gordon, B. R., Li, Y., Wang, L., Sintsova, A., Van Bakel, H., Tian, S., ... & Xia, B. J. 468 Liu. 2010. Lsr2 is a nucleoid-associated protein that targets AT-rich sequences and virulence genes in 469 *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA*, *107*, 5154-5159.

35.Aït-Bara, S., Carpousis, A.J. and Quentin, Y. (2015) RNase E in the γ-Proteobacteria: conservation of intrinsically disordered noncatalytic

region and molecular evolution of microdomains. Mol. Genet. Genomics 290,847–862doi:10.1007/s00438-014-0959-5.

36.Cooper, C. R., Daugherty, A. J., Tachdjian, S., Blum, P. H., & Kelly, R. M. (2009). Role of vapBC toxin–antitoxin loci in the thermal stress response of *Sulfolobus solfataricus*.

37.Fairman-Williams, M.E., Guenther, U.-P. and Jankowsky, E. (2010) SF1 and SF2 helicases: family matters. Curr. Opin. Struct. Biol. 20, 313–324doi:10.1016/j.sbi.2010.03.011.

38.Wanner, B. L., and R. McSharry. 1982. Phosphate-controlled gene expression in *Escherichia coli* K12 using Mu dl-directed *lacZ* fusions. J. Mol. Biol. 158:347-363.

39.Anantharaman, V., Koonin, E. V., & Aravind, L. (2002). Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic acids research*, *30*(7), 1427-1464.

40.Kazakov, A. E., Vassieva, O., Gelfand, M. S., Osterman, A., & Overbeek, R. (2003). Bioinformatics classification and functional analysis of PhoH homologs. *In silico biology*, *3*(1, 2), 3-15.

41.Fairman-Williams, M. E., Guenther, U. P., & Jankowsky, E. (2010). SF1 and SF2 helicases: family matters. *Current opinion in structural biology*, *20*(3), 313-324.

42.Lupas, A. N., & Martin, J. (2002). AAA proteins. *Current opinion in structural biology*, *12*(6), 746-753.

43.Leipe, D. D., Wolf, Y. I., Koonin, E. V., & Aravind, L. (2002). Classification and evolution of P-loop GTPases and related ATPases1. *Journal of molecular biology*, *317*(1), 41-72.

44.DeMaio, J., Zhang, Y., Ko, C., & Bishai, W. R. (1997). *Mycobacterium tuberculosis* sigF is part of a gene cluster with similarities to the *Bacillus subtilis sigF* and *sigB* operons. *Tubercle and lung disease*, *78*(1), 3-12.

45.Lauten, E. H., Pulliam, B. L., Derousse, J., Bhatta, D., & Edwards, D. A. (2010). Gene Expression, Bacteria Viability and Survivability Following Spray Drying of *Mycobacterium smegmatis*. *Materials*, *3*(4), 2684-2724. doi:10.3390/ma3042684.

46.Hümpel, A., Gebhard, S., Cook, G. M., & Berney, M. (2010). The SigF Regulon in *Mycobacterium smegmatis* Reveals Roles in Adaptation to Stationary Phase, Heat, and Oxidative Stress. *Journal of Bacteriology*, *192*(10), 2491-2502. doi:10.1128/jb.00035-10.

47.Gruber, T. M., & Gross, C. A. (2003). Multiple sigma subunits and the partitioning of bacterial transcription space. *Annual Reviews in Microbiology*, *57*(1), 441-466.

48.Mattison, K., Wilbur, J. S., So, M., & Brennan, R. G. (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.

49.Rodrigue, S., Provvedi, R., Jacques, P. É., Gaudreau, L., & Manganelli, R. (2006). The σ factors of *Mycobacterium tuberculosis*. *FEMS microbiology reviews*, *30*(6), 926-941.

50.Paget, M. S., & Helmann, J. D. (2003). The sigma70 family of sigma factors. *Genome Biology*,*4*(1), 203. doi:10.1186/gb-2003-4-1-203.

51.Gupta, A. M., & Mandal, S. (2017). Distribution of Sigma factors delineates segregation of virulent and a virulent *Mycobacterium. bioRxiv*, 235986.

52.Manganelli, R., Proveddi, R., Rodrigue, S., Beaucher, J., Gaudreau, L., & Smith, I. (2004). σ factors and global gene regulation in *Mycobacterium tuberculosis*. *Journal of bacteriology*, *186*(4), 895-902.

53.Jankowsky, E., & Fairman, M. E. (2007). RNA helicases—one fold for many functions. *Current opinion in structural biology*, *17*(3), 316-324.

54.Lohman, T. M., & Bjornson, K. P. (1996). Mechanisms of helicase-catalyzed DNA unwinding. *Annual review of biochemistry*, *65*(1), 169-214.

55.Hall, M. C., & Matson, S. W. (1999). Helicase motifs: the engine that powers DNA unwinding. Molecular microbiology, 34(5), 867-877.

56.Enemark, E. J., & Joshua-Tor, L. (2008). On Helicases and other motor proteins. *Current Opinion in Structural Biology*, *18*(2), 243–257. http://doi.org/10.1016/j.sbi.2008.01.007.

57.Sun, B., & Wang, M. D. (2016). Single-molecule perspectives on helicase mechanisms and functions. *Critical reviews in biochemistry and molecular biology*, *51*(1), 15-25.

58.Singleton, M. R., Dillingham, M. S., & Wigley, D. B. (2007). Structure and mechanism of helicases and nucleic acid translocases. *Annu. Rev. Biochem.*, *76*, 23-50.

59.Wessel, R., Schweizer, J., & Stahl, H. (1992). Simian virus 40 T-antigen DNA helicase is a hexamer which forms a binary complex during bidirectional unwinding from the viral origin of DNA replication. *Journal of Virology*, *66*(2), 804-815.

60.San Martin, M. C., Stamford, N. P. J., Dammerova, N., Dixon, N. E., & Carazo, J. M. (1995). A structural model for the *Escherichia coli* DnaB

helicase based on electron microscopy data. *Journal of structural biology*, *114*(3), 167-176.

61.Egelman, E. H., Yu, X., Wild, R., Hingorani, M. M., & Patel, S. S. (1995). Bacteriophage T7 helicase/primase proteins form rings around single-stranded DNA that suggest a general structure for hexameric helicases. *Proceedings of the National Academy of Sciences*, *92*(9), 3869-3873.

62.Fouts, E. T., Yu, X., Egelman, E. H., & Botchan, M. R. (1999). Biochemical and electron microscopic image analysis of the hexameric E1 helicase. *Journal of Biological Chemistry*, *274*(7), 4447-4458.

63.Gorbalenya, A. E., & Koonin, E. V. (1993). Helicases: amino acid sequence comparisons and structure-function relationships. *Current opinion in structural biology*, *3*(3), 419-429.

64.Burgess, R. R., & Anthony, L. (2001). How sigma docks to RNA polymerase and what sigma does. *Current opinion in microbiology*, *4*(2), 126-131.

65.Paget MS. Bacterial Sigma Factors and Anti-Sigma Factors: Structure, Function and Distribution. Wigneshweraraj S, Hinton DM, eds. *Biomolecules*. 2015;5(3):1245-1265. doi:10.3390/biom5031245.

66.Fishbein, S., Van Wyk, N., Warren, R. M., & Sampson, S. L. (2015). Phylogeny to function: PE/PPE protein evolution and impact on *Mycobacterium tuberculosis* pathogenicity. *Molecular microbiology*, *96*(5), 901-916.

67.Lonetto, M., Gribskov, M., & Gross, C. A. (1992). The sigma 70 family: sequence conservation and evolutionary relationships. *Journal of bacteriology*, *174*(12), 3843.

68.Heimann, J. D. (2002). The extracytoplasmic function (ECF) sigma factors.

69.Chen, P., Ruiz, R. E., Li, Q., Silver, R. F., & Bishai, W. R. (2000). Construction and Characterization of a *Mycobacterium tuberculosis* Mutant Lacking the Alternate Sigma Factor Gene, *sigF*. *Infection and immunity*, *68*(10), 5575-5580.

70.Lee, Y. C. (1996). Carbohydrate analyses with high-performance anion-exchange chromatography. *Journal of Chromatography A*, *720*(1-2), 137-149.

71.Gebhard, S., Huempel, A., McLellan, A. D., & Cook, G. M. (2008). The alternative sigma factor SigF of *Mycobacterium smegmatis* is required for survival of heat shock, acidic pH and oxidative stress. *Microbiology*, *154*(9), 2786-2795.

72.Singh, A. K., & Singh, B. N. (2008). Conservation of sigma F in mycobacteria and its expression in *Mycobacterium smegmatis*. *Current microbiology*, *56*(6), 574-580.

73.Singh, A. K., Dutta, D., Singh, V., Srivastava, V., Biswas, R. K., & Singh, B. N. (2015). Characterization of *Mycobacterium smegmatis* sigF mutant and its regulon: overexpression of SigF antagonist (MSMEG_1803) in *M. smegmatis* mimics *sigF* mutant phenotype, loss of pigmentation, and sensitivity to oxidative stress. *MicrobiologyOpen*, *4*(6), 896-916.

74.Pandin, C., Caroff, M., & Condemine, G. (2016). Antimicrobial peptide resistance genes in the plant pathogen *Dickeya dadantii*. *Applied and environmental microbiology*, *82*(21), 6423-6430.

75.Goldsmith, D. B., Crosti, G., Dwivedi, B., McDaniel, L. D., Varsani, A., Suttle, C. A., ... & Breitbart, M. (2011). Development of phoH as a novel signature gene for assessing marine phage diversity. *Applied and environmental microbiology*, *77*(21), 7730-7739.

76.Goldsmith, D. B., Parsons, R. J., Beyene, D., Salamon, P., & Breitbart, M. (2015). Deep sequencing of the viral phoH gene reveals temporal variation, depth-specific composition, and persistent dominance of the same viral phoH genes in the Sargasso Sea. *PeerJ*, *3*, e997.

77.Waagmeester, A., Thompson, J., & Reyrat, J. M. (2005). Identifying sigma factors in *Mycobacterium smegmatis* by comparative genomic analysis. *Trends in microbiology*, *13*(11), 505-509.

78.Deng, W. (2015). PE11 (Rv1169c) selectively alters fatty acid components of *Mycobacterium smegmatis* and host cell interleukin-6 level accompanied with cell death. *Frontiers in microbiology*, *6*, 613.

79.Hartkoorn, R. C., Sala, C., Uplekar, S., Busso, P., Rougemont, J., & Cole, S. T. (2012). Genome-Wide Definition of the SigF Regulon in *Mycobacterium tuberculosis*. *Journal of Bacteriology*, *194*(8), 2001-2009. doi:10.1128/jb.06692-11.

80.DeMaio, J., Zhang, Y., Ko, C., Young, D. B., & Bishai, W. R. (1996). A stationary-phase stress-response sigma factor from *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences*, *93*(7), 2790-2794.

81.Sachdeva, P., Misra, R., Tyagi, A. K., & Singh, Y. (2010). The sigma factors of *Mycobacterium tuberculosis*: regulation of the regulators. *The FEBS journal*, *277*(3), 605-626.

82.Sechi, L. A., Felis, G. E., Ahmed, N., Paccagnini, D., Usai, D., Ortu, S., ... & Zanetti, S. (2007). Genome and transcriptome scale portrait of sigma factors in *Mycobacterium avium* subsp. *paratuberculosis*. *Infection, Genetics and Evolution*, *7*(4), 424-432.

83.Wong, C., Sridhara, S., Bardwell, J. C., & Jakob, U. (2000). Heating greatly speeds Coomassie blue staining and destaining. *Biotechniques*, *28*(3), 426.

84.Girling, D. J., Darbyshire, J. H., Humphries, M. J., & O'Mahoney, G. (1988). Extra-pulmonary tuberculosis. *British medical bulletin*, *44*(3), 738-756.

85.Russell, D. G., VanderVen, B. C., Lee, W., Abramovitch, R. B., Kim, M. J., Homolka, S., ... & Rohde, K. H. (2010). *Mycobacterium tuberculosis* wears what it eats. *Cell host & microbe*, *8*(1), 68-76.

86.Nair, S., Ramaswamy, P. A., Ghosh, S., Joshi, D. C., Pathak, N., Siddiqui, I., ... & Mukhopadhyay, S. (2009). The PPE18 of *Mycobacterium tuberculosis* interacts with TLR2 and activates IL-10 induction in macrophage. *The Journal of Immunology*, jimmunol-0901367.

87.Byrd, A. K., & Raney, K. D. (2012). Superfamily 2 helicases. *Frontiers in bioscience (Landmark edition)*, *17*, 2070.

88.Simon, P. (2003). Q-Gene: processing quantitative real-time RT–PCR data. *Bioinformatics*, *19*(11), 1439-1440.

89.Pfaffl, M. W. (2001). A new mathematical model for relative quantification in real-time RT–PCR. *Nucleic acids research*, *29*(9), e45-e45.

90.Robson, J., McKenzie, J. L., Cursons, R., Cook, G. M., & Arcus, V. L. (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.

91.Miallau, L., Faller, M., Chiang, J., Arbing, M., Guo, F., Cascio, D., & Eisenberg, D. (2008). Structure and proposed activity of a member of the VapBC family of toxin-antitoxin systems: VapBC-5 from *Mycobacterium tuberculosis*. *Journal of Biological Chemistry*.

92.Sharrock, A. V. (2013). *Characterisation of VapBC Toxin-Antitoxins from Mycobacterium tuberculosis* (Doctoral dissertation, University of Waikato).

93.Brzozowska, I., & Zielenkiewicz, U. (2013). Regulation of toxin–antitoxin systems by proteolysis. *Plasmid*, *70*(1), 33-41.

94.Diago-Navarro, E., Hernández-Arriaga, A. M., Kubik, S., Konieczny, I., & Díaz-Orejas, R. (2013). Cleavage of the antitoxin of the parD toxin–antitoxin system is determined by the ClpAP protease and is modulated by the relative ratio of the toxin and the antitoxin. *Plasmid*, *70*(1), 78-85.

95.Van Melderen, L., Thi, M. H. D., Lecchi, P., Gottesman, S., Couturier, M., & Maurizi, M. R. (1996). Atp-dependent degradation of ccda by lon

protease effects of secondary structure and heterologous subunit interactions. *Journal of Biological Chemistry*, *271*(44), 27730-27738.

96.Van Melderen, L. (2010). Toxin–antitoxin systems: why so many, what for?. *Current opinion in microbiology*, *13*(6), 781-785.

97.Jin, G., Pavelka, M. S., & Butler, J. S. (2015). Structure-function analysis of VapB4 antitoxin identifies critical features of a minimal VapC4 toxin-binding module. *Journal of bacteriology*, JB-02508.

98.Ogura, T., & Hiraga, S. (1983). Mini-F plasmid genes that couple host cell division to plasmid proliferation. *Proceedings of the National Academy of Sciences*, *80*(15), 4784-4788.

99.Gerdes, K., & Molin, S. (1986). Partitioning of plasmid R1: structural and functional analysis of the parA locus. *Journal of molecular biology*, *190*(3), 269-279.

100.Gerdes, K. (2000). Toxin-antitoxin modules may regulate synthesis of macromolecules during nutritional stress. *Journal of bacteriology*, *182*(3), 561-572.

101.Hayes, F. (2003). Toxins-antitoxins: plasmid maintenance, programmed cell death, and cell cycle arrest. *Science*, *301*(5639), 1496-1499.

102.Dörr, T., Vulić, M., & Lewis, K. (2010). Ciprofloxacin causes persister formation by inducing the TisB toxin in *Escherichia coli*. *PLoS biology*, *8*(2), e1000317.

103.Cruz, J. W., Sharp, J. D., Hoffer, E. D., Maehigashi, T., Vvedenskaya, I. O., Konkimalla, A., ... & Woychik, N. A. (2015). Growth-regulating *Mycobacterium tuberculosis* VapC-mt4 toxin is an isoacceptor-specific tRNase. *Nature communications*, *6*, 7480.

104.Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., ... & Salazar, G. A. (2015). The Pfam protein families database: towards a more sustainable future. *Nucleic acids research*, *44*(D1), D279-D285.

105. Betts, J. C., Lukey, P. T., Robb, L. C., McAdam, R. A., & Duncan, K. (2002). Evaluation of a nutrient starvation model of *Mycobacterium tuberculosis* persistence by gene and protein expression profiling. *Molecular microbiology*, *43*(3), 717-731.

106.Geiman, D. E., Kaushal, D., Ko, C., Tyagi, S., Manabe, Y. C., Schroeder, B. G., ... & Bishai, W. R. (2004). Attenuation of late-stage disease in mice infected by the *Mycobacterium tuberculosis* mutant lacking the SigF alternate sigma factor and identification of SigF-dependent genes by microarray analysis. *Infection and immunity*, *72*(3), 1733-1745.

107.Beaucher, J., Rodrigue, S., Jacques, P. É., Smith, I., Brzezinski, R., & Gaudreau, L. (2002). Novel *Mycobacterium tuberculosis* anti-σ factor antagonists control σF activity by distinct mechanisms. *Molecular microbiology*, *45*(6), 1527-1540.

108.Letunic, I., Doerks, T., & Bork, P. (2008). SMART 6: recent updates and new developments. *Nucleic acids research*, *37*(suppl_1), D229-D232.

109.Homerova, D., Surdova, K., Mikusova, K., & Kormanec, J. (2007). Identification of promoters recognized by RNA polymerase containing *Mycobacterium tuberculosis* stress-response sigma factor σF. *Archives of microbiology*, *187*(3), 185-197.

110.Williams, E. P., Lee, J. H., Bishai, W. R., Colantuoni, C., & Karakousis, P. C. (2007). *Mycobacterium tuberculosis* SigF regulates genes encoding cell wall-associated proteins and directly regulates the transcriptional regulatory gene phoY1. *Journal of bacteriology*, *189*(11), 4234-4242.

111.Gebhard, S., Huempel, A., McLellan, A. D., & Cook, G. M. (2008). The alternative sigma factor SigF of *Mycobacterium smegmatis* is required for survival of heat shock, acidic pH and oxidative stress. *Microbiology*, *154*(9), 2786-2795.

112.Saviola, B. (2014). Pigments and Pathogenesis. *J Mycobac Dis*, *4*(5).

113.Larisch, C., Nakunst, D., Hüser, A. T., Tauch, A., & Kalinowski, J. (2007). The alternative sigma factor SigB of *Corynebacterium glutamicum* modulates global gene expression during transition from exponential growth to stationary phase. *BMC Genomics*, *8*, 4. http://doi.org/10.1186/1471-2164-8-4.

114.Masuda, H., & Inouye, M. (2017). Toxins of prokaryotic toxin-antitoxin systems with sequence-specific endoribonuclease activity. *Toxins*, *9*(4), 140.

115.Al-Hinai, Mohab & Fast, Alan & T Papoutsakis, Eleftherios. (2012). Novel System for Efficient Isolation of Clostridium Double-Crossover Allelic Exchange Mutants Enabling Markerless Chromosomal Gene Deletions and DNA Integration. Applied and environmental microbiology. 78. 8112-21. 10.1128/AEM.02214-12.

116.Wimmerová, M., Engelsen, S. B., Bettler, E., Breton, C., & Imberty, A. (2003). Combining fold recognition and exploratory data analysis for searching for glycosyltransferases in the genome of *Mycobacterium tuberculosis. Biochimie*, *85*(7), 691-700.

117.Zheng, X., Papavinasasundaram, K. G., & Av-Gay, Y. (2007). Novel substrates of *Mycobacterium tuberculosis* PknH Ser/Thr kinase. Biochemical and biophysical research communications, 355(1), 162-168.

118.Festa, R. A., McAllister, F., Pearce, M. J., Mintseris, J., Burns, K. E., Gygi, S. P., & Darwin, K. H. (2010). Prokayrotic ubiquitin-like protein (Pup) proteome of *Mycobacterium tuberculosis*. *PloS one*, *5*(1), e8589.

119.. Fedida, A., & Lindell, D. (2017). Two synechococcus genes, two different effects on cyanophage infection. Viruses, 9(6), 136.

120.Senissar, M., Manav, M. C., & Brodersen, D. E. (2017). Structural conservation of the PIN domain active site across all domains of life. *Protein Science*, *26*(8), 1474-1492.

121.Clissold PM, Ponting CP (2000) PIN domains in nonsense-mediated mRNA decay and RNAi. Curr Biol 10:R888-90.

122.Iyer, L. M., Leipe, D. D., Koonin, E. V., & Aravind, L. (2004). Evolutionary history and higher order classification of AAA+ ATPases. *Journal of structural biology*, *146*(1-2), 11-31.

123.Unterholzner, S. J., Poppenberger, B., & Rozhon, W. (2013). Toxin–antitoxin systems: biology, identification, and application. *Mobile genetic elements*, *3*(5), e26219.

124.Yamaguchi, Y., & Inouye, M. (2016). Toxin–Antitoxin Systems in Bacteria and Archaea. *Stress and Environmental Regulation of Gene Expression and Adaptation in Bacteria*, 97-107.

125.Demidenok, O. I., Kaprelyants, A. S., & Goncharenko, A. V. (2014). Toxin–antitoxin vapBC locus participates in formation of the dormant state in *Mycobacterium smegmatis*. *FEMS microbiology letters*, *352*(1), 69-77.

126.Ishige, T., Krause, M., Bott, M., Wendisch, V. F., & Sahm, H. (2003). The phosphate starvation stimulon of *Corynebacterium glutamicum* determined by DNA microarray analyses. *Journal of bacteriology*, *185*(15), 4519-4529.

127.Bott, M., & Brocker, M. (2012). Two-component signal transduction in *Corynebacterium glutamicum* and other corynebacteria: on the way towards stimuli and targets. *Applied microbiology and biotechnology*, *94*(5), 1131-1150.

128.Khemici, V., & Linder, P. (2016). RNA helicases in bacteria. *Current opinion in microbiology*, *30*, 58-66.

129.DeMaio, J., Zhang, Y., Ko, C., & Bishai, W. R. (1997). *Mycobacterium tuberculosis sigF* is part of a gene cluster with similarities to the *Bacillus subtilis sigF* and *sigB* operons. *Tubercle and lung disease*, *78*(1), 3-12.

130.Potúčková, L., Kelemen, G. H., Findlay, K. C., Lonetto, M. A., Buttner, M. J., & Kormanec, J. (1995). A new RNA polymerase sigma factor, σF

is required for the late stages of morphological differentiation in *Streptomyces spp*. *Molecular microbiology*, *17*(1), 37-48.

131.Wu, S., de Lencastre, H. E. R. M. I. N. I. A., & Tomasz, A. (1996). Sigma-B, a putative operon encoding alternate sigma factor of *Staphylococcus aureus* RNA polymerase: molecular cloning and DNA sequencing. *Journal of bacteriology*, *178*(20), 6036-6042.

132.Becker, L. A., Cetin, M. S., Hutkins, R. W., & Benson, A. K. (1998). Identification of the Gene Encoding the Alternative Sigma Factor B from *Listeria monocytogenes* and Its Role in Osmotolerance. *Journal of bacteriology*, *180*(17), 4547-4554.

133.Raman, S., Puyang, X., Cheng, T. Y., Young, D. C., Moody, D. B., & Husson, R. N. (2006). *Mycobacterium tuberculosis* SigM positively regulates Esx secreted protein and nonribosomal peptide synthetase genes and down regulates virulence-associated surface lipid synthesis. *Journal of bacteriology*, *188*(24), 8460-8468.

134.Summer, H., Grämer, R., & Dröge, P. (2009). Denaturing urea polyacrylamide gel electrophoresis (Urea PAGE). *Journal of visualized experiments: JoVE*, (32).

135.Ramage, H. R., Connolly, L. E., & Cox, J. S. (2009). Comprehensive functional analysis of *Mycobacterium tuberculosis* toxin-antitoxin systems: implications for pathogenesis, stress responses, and evolution. *PLoS genetics*, *5*(12), e1000767.

136.Hopper, S., Wilbur, J. S., Vasquez, B. L., Larson, J., Clary, S., Mehr, I. J., ... & So, M. (2000). Isolation of *Neisseria gonorrhoeae* Mutants That Show Enhanced Trafficking across Polarized T84 Epithelial Monolayers. *Infection and immunity*, *68*(2), 896-905.

137.Magnuson, R. D. (2007). Hypothetical functions of toxin-antitoxin systems. *Journal of Bacteriology*, *189*(17), 6089-6092.

138.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.

139.Hartkoorn, R. C., Sala, C., Magnet, S. J., Chen, J. M., Pojer, F., & Cole, S. T. (2010). Sigma factor F does not prevent rifampin inhibition of RNA polymerase or cause rifampin tolerance in *Mycobacterium tuberculosis*. *Journal of bacteriology*, *192*(20), 5472-5479.

140.Kang, J., Lee, M. S., & Gorenstein, D. G. (2007). Application of RNase in the purification of RNA-binding proteins. *Analytical biochemistry*, *365*(1), 147.

141.Lewin, B. (1990). The messenger RNA template. *Genes IV*, 171-184.

142.Verma, A., Sampla, A. K., & Tyagi, J. S. (1999). *Mycobacterium tuberculosis* rrn Promoters: Differential Usage and Growth Rate-Dependent Control. *Journal of bacteriology*, *181*(14), 4326-4333.

143.Chomczynski, P., & Sacchi, N. (1987). Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Analytical biochemistry*, *162*(1), 156-159.

144.Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.

145.Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature reviews genetics*, *12*(2), 87.

146.Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, *10*(1), 57.

147.Bustin, S. A. (2000). Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *Journal of molecular endocrinology*, *25*(2), 169-193.

148.Fleige, S., & Pfaffl, M. W. (2006). RNA integrity and the effect on the real-time qRT-PCR performance. *Molecular aspects of medicine*, *27*(2-3), 126-139.

149.Desjardins, P., & Conklin, D. (2010). NanoDrop microvolume quantitation of nucleic acids. *Journal of visualized experiments: JoVE*, (45).

150.Taylor, S., Wakem, M., Dijkman, G., Alsarraj, M., & Nguyen, M. (2010). A practical approach to RT-qPCR—publishing data that conform to the MIQE guidelines. *Methods*, *50*(4), S1-S5.

151.Desroche, N., Beltramo, C., & Guzzo, J. (2005). Determination of an internal control to apply reverse transcription quantitative PCR to study stress response in the lactic acid bacterium *Oenococcus oeni*. *Journal of microbiological methods*, *60*(3), 325-333.

152.Bustin, S. A., Benes, V., Garson, J. A., Hellemans, J., Huggett, J., Kubista, M., ... & Vandesompele, J. (2009). The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clinical chemistry*, *55*(4), 611-622.

153.Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., & Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, *230*(4732), 1350-1354.

154.Mullis, K., Faloona, F., Scharf, S., Saiki, R. K., Horn, G. T., & Erlich, H. (1986, January). Specific enzymatic amplification of DNA *in vitro*: the polymerase chain reaction. In *Cold Spring Harbor symposia on*

quantitative biology (Vol. 51, pp. 263-273). Cold Spring Harbor Laboratory Press.

155.Canales, R. D., Luo, Y., Willey, J. C., Austermiller, B., Barbacioru, C. C., Boysen, C., ... & Ma, Y. (2006). Evaluation of DNA microarray results with quantitative gene expression platforms. *Nature biotechnology*, *24*(9), 1115.

156.Oscorbin, I. P., Belousova, E. A., Zakabunin, A. I., Boyarskikh, U. A., & Filipenko, M. L. (2016). Comparison of fluorescent intercalating dyes for quantitative loop-mediated isothermal amplification (qLAMP). *BioTechniques*, *61*(1), 20-25.

157.Gudnason, H., Dufva, M., Bang, D. D., & Wolff, A. (2007). Comparison of multiple DNA dyes for real-time PCR: effects of dye concentration and sequence composition on DNA amplification and melting temperature. *Nucleic acids research*, *35*(19), e127.

158.Nolan, T., Hands, R. E., & Bustin, S. A. (2006). Quantification of mRNA using real-time RT-PCR. *Nature protocols*, *1*(3), 1559.

159.Schefe, J. H., Lehmann, K. E., Buschmann, I. R., Unger, T., & Funke-Kaiser, H. (2006). Quantitative real-time RT-PCR data analysis: current concepts and the novel "gene expression's C T difference" formula. *Journal of molecular medicine*, *84*(11), 901-910.

160.Godiwala, N. T., Vandewalle, A., Ward, H. D., & Leav, B. A. (2006). Quantification of *in vitro* and *in vivo Cryptosporidium parvum* infection by using real-time PCR. *Applied and environmental microbiology*, *72*(6), 4484-4488.

161.Lee, C., Kim, J., Shin, S. G., & Hwang, S. (2006). Absolute and relative QPCR quantification of plasmid copy number in *Escherichia coli*. *Journal of biotechnology*, *123*(3), 273-280.

162.Ruijter, J. M., Pfaffl, M. W., Zhao, S., Spiess, A. N., Boggy, G., Blom, J., ... & Derveaux, S. (2013). Evaluation of qPCR curve analysis methods for reliable biomarker discovery: bias, resolution, precision, and implications. *Methods*, *59*(1), 32-46.

163.Livak, K. J., & Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *methods*, *25*(4), 402-408.

164.Wilkening, S., & Bader, A. (2004). Quantitative real-time polymerase chain reaction: methodical analysis and mathematical model. *Journal of biomolecular techniques: JBT*, *15*(2), 107.

165.Mao, F., Leung, W. Y., & Xin, X. (2007). Characterization of EvaGreen and the implication of its physicochemical properties for qPCR applications. *BMC biotechnology*, *7*(1), 76.