Understanding the evolution and function of the mycobacterial Type VII ESX secretion systems (T7SSs) and their substrates

By

Mae Newton-Foot

Dissertation presented for the degree of Doctor of Philosophy in Medical Sciences (Molecular Biology) in the Faculty of Medicine and Health Sciences at Stellenbosch University

Supervisor: Prof. Nicolaas Claudius Gey van Pittius

Co-supervisor: Prof. Robin Mark Warren

March 2013

Declaration

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: March 2013

Summary

Mycobacterium tuberculosis, the causative agent of tuberculosis disease, contains five copies of the ESAT-6 gene cluster, each encoding a dedicated ESX protein secretion system which has been defined as a novel Type-VII secretion system. The ESX have been implicated in virulence and survival of *M. tuberculosis*, and as such present a promising target for novel treatment interventions. This study has investigated the evolution, regulation, functions and substrates of the ESX secretion systems.

The evolutionary history of the ESX secretion systems was established using *in silico* and phylogenetic analyses of the sequenced mycobacteria, closely related actinomycetes and WXG-FtsK clusters from other bacteria. The ESX-4 gene cluster appears to have evolved with the start of the evolution of the mycomembrane, followed by the duplication of ESX-3, which marks the evolution of the genus *Mycobacterium.* The ESX-1 duplication occurred next, followed by ESX-2 and ESX-5 which occur only in the slow growing mycobacteria. Five additional ESX gene clusters were newly identified and named ESX-P1 to -P5. These additional ESX clusters occur, or are predicted to occur, on plasmid DNA, and appear to be progenitors of the genomic ESX-1 to -5 gene clusters, possibly indicating a plasmid-mediated mechanism of ESX duplication and evolution.

The promoters expressing the *M. tuberculosis* ESX-1 to ESX-5 secretion systems were investigated using a promoter probe assay, and characterised using *in silico* analyses. Promoters were identified for ESX-1, -2, -3 and -5.

The functions of the mycobacterial ESX secretion systems were investigated using whole proteomic, secretomic and metabolomic analyses of the fast growing, non-pathogenic *M. smegmatis*, which contains three of the ESX secretion systems, ESX-1, 3, and 4. ESX knockout strains of *M. smegmatis* were generated and used in comparative analyses with wild-type *M. smegmatis*. ESX-1 was highly expressed in wild-type *M. smegmatis*, however no specific pathways showed considerable variation when ESX-1 was deleted. Deletion of ESX-3 resulted in substantial variation to multiple cellular pathways, including amino acid, carbohydrate and fatty acid metabolism and oxidative stress. These and other differences indicate possible perturbed polyamine metabolism in the absence of ESX-3. Although no ESX-4 protein components were detected in wild type *M. smegmatis*, the ESX-4 knockout displayed substantial proteomic variation. Reduced levels of ESX-3 component proteins in the ESX-4 knockout suggest that ESX-4 influences ESX-3 expression. Other variation linked ESX-4 to cell division and molybdenum metabolism.

Secretomic analyses of wild-type and ESX knockout *M. smegmatis* strains were used to search for novel substrates of the *M. smegmatis* ESX secretion systems. No prototype ESX substrates were identified in the culture filtrates, however 10 possible substrates of the ESX-1, -3 and -4 secretion systems, containing the general ESX secretion signal, YxxxD/E, were identified. The functions of some of these proteins correlate with the ESX functions identified in the proteomic and metabolomic analyses.

This study sets the groundwork for future work in understanding the functional roles and expression patterns of these ESX secretion systems and in using global proteomic and metabolomic analyses to understand cellular changes in response to specific signals or genomic changes.

Opsomming

Mycobacterium tuberculosis, die veroorsakende agent van tuberkulose, bevat vyf kopieë van die ESAT-6 geengroep, wat elk 'n toegewyde ESX proteïen sekresiesisteem, omskryf as 'n nuwe Tipe-VII sekresiestelsel, kodeer. Die ESX sekresiesisteme is betrokke by virulensie en oorlewing van *M. tuberculosis*, en is dus belowende teikens vir nuwe behandelings. Hierdie studie het die evolusie, regulasie, funksies en substrate van die ESX sekresiesisteme ondersoek.

Die evolusionêre geskiedenis van die ESX sekresiesisteme is bepaal met behulp van *in silico* en filogenetiese analises van die volgordebepaalde mikobakterieë, nouverwante actinomisete en WXG-FtsK groepe van ander bakterieë. Die ESX-4 geengroep het saam met die evolusie van die mikomembraan ontwikkel, gevolg deur die duplisering van ESX-3, wat die evolusie van die genus *Mycobacterium* merk. Die ESX-1 duplisering het volgende plaasgevind, gevolg deur ESX-2 en ESX-5, wat slegs in die stadig-groeiende mikobakterieë voorkom. Vyf addisionele ESX geengroepe is nuut geïdentifiseer in hierdie studie en is ESX-P1 tot -P5 genoem. Hierdie addisionale ESX groepe is op, of word voorspel om op, plasmied DNS voor te kom, en mag voorlopers van die genomiese ESX-1 tot -5 geengroepe wees, wat moontlik dui op 'n plasmied-gemedieërde meganisme van ESX duplisering en evolusie.

Die promoters wat verantwoordelik is vir die uitdrukking van die *M. tuberculosis* ESX-1 tot ESX-5 sekresiesisteme is ondersoek deur middel van 'n promoter aktiwiteitstoets, en gekarakteriseer deur *in silico* analises. Promoters is geidentifiseer vir ESX-1, -2, -3 en -5.

Die funksies van die mikobakteriële ESX sekresiesisteme is ondersoek deur proteomiese, sekretomiese en metabolomiese analises van die vinnig-groeiende, nie-patogeniese mikobakterium *M. smegmatis*, wat ESX-1, -3 en -4 sekresiesisteme besit. ESX uitslaanmutante van *M. smegmatis* is gegenereer en gebruik in die vergelykende analises met die wilde-tipe *M. smegmatis*. ESX-1 is hoogs uitgedruk in wilde-tipe *M. smegmatis*, maar geen spesifieke metabolise weë het aansienlike variasie getoon wanneer ESX-1 verwyder is. Delesie van ESX-3 het gelei tot aansienlike variasie in verskeie sellulêre weë, insluitend aminosuur-, koolhidraat- en vetsuur-metabolisme en oksidatiewe stres. Hierdie en ander verskille dui op moontlike versteurde poli-amien metabolisme in die afwesigheid van ESX-3. Hoewel geen ESX-4 proteïenkomponente opgespoor is in wilde-tipe *M. smegmatis* nie, vertoon die ESX-4 uitslaanmutant aansienlike proteomiese variasie. Laer vlakke van ESX-3 proteïne dui daarop dat ESX-4 die uitdrukking van ESX-3 beinvloed. Baie van die proteomiese variasie kan geassosieer word met verlaagde ESX-3 uitdrukking, maar ander variasie mag ESX-4 koppel met seldeling en molibdeen metabolisme.

Sekretomiese analises van wilde-tipe en ESX uitslaanmutant *M. smegmatis* stamme is gebruik om nuwe substrate van die *M. smegmatis* ESX sekresiesisteme te identifiseer. Geen prototipe ESX substrate is geïdentifiseer in die kultuurfiltraat, maar 10 moontlike substrate van die ESX-1, -3 en -4 sekresiesisteme, met die algemene ESX sekresiesein, YxxxD/E, is geïdentifiseer. Die funksies van sommige van hierdie proteïene korreleer met die funksies geïdentifiseer in die proteomiese en metabolomiese analises.

Hierdie studie stel die grondslag vir toekomstige werk in die begrip van die funksionele rol en uitdrukkingspatrone van die ESX sekresiesisteme en in die gebruik van globale proteomiese en metabolomiese analises om sellulêre veranderinge in reaksie op spesifieke seine of genomiese veranderinge te verstaan.

Presentations and Publications

Publications

Newton-Foot M, Gey van Pittius NC. The complex architecture of mycobacterial promoters, *Tuberculosis* (2012), http://dx.doi.org/10.1016/j.tube.2012.08.003

Loots, DT, Meissner-Roloff RJ, Newton-Foot M, Gey van Pittius NC. A Metabolomics Approach to Exploring the Function of the ESX-3 Type VII Secretion System of *M. smegmatis.* Submitted to *Metabolomics.*

Poster Presentations

Newton-Foot, M., Warren, R.M., van Helden, P.D., Gey van Pittius, N.C. Investigating divalent metal cation regulation of the mycobacterial ESX-3 secretion system. **Stellenbosch University Faculty of Health Sciences Academic Yearday (2011).**

Newton-Foot, M., Smit, M. Steyn, A., Sampson, S., Warren, R.M., van Helden, P.D., Gey van Pittius, N.C Modelling the *M. tuberculosis* ESX secretion system based on protein-protein interactomes of ESX-3 and ESX-4. **Keystone Symposia J4 Mycobacteria: Physiology, Metabolism and Pathogenesis – Back to basics. Vancouver, British Columbia, Canada (2011).**

Acknowledgements

I would like to thank the following people,

My family, for your love, support, encouragement and patience and for always believing in me!

My friends, for your encouragement and support!

My colleagues and friends, for your support through the successes and frustrations; and for all the advice, laughter and tears we've shared through the years!

Prof Nico Gey van Pittius, Prof Rob Warren and Prof Paul van Helden, for creating such a stimulating scientific environment; for your continuous support, encouragement and stimulation; and for the opportunities you have afforded me to develop scientifically and personally.

Every member of the Division of Molecular Biology and Human Genetics, for your advice, assistance, encouragement, and everything I have learnt from each one of you.

Dr Salome Smit, at the MS Proteomics unit of CAF, Stellenbosch University, who performed the LC-MS analysis. Your experience, patience, encouragement, input and hard work is much appreciated. We are incredibly lucky to have you on our team!

Prof. Du Toit Loots and Reinart J. Meissner-Roloff at the Centre for Human Metabonomics, North-West University, South Africa; with whom we have, and continue to collaborate on the metabolomic analyses of the ESX secretion systems.

Prof NC. Gey van Pittius, RG. van der Merwe and J. Botha for ESX promoter probe constructs.

N. Steyn and M. Smit for *M. smegmatis* ESX knockout strains and constructs.

E. Machowski and J. Rauzier for plasmids.

The National Research Foundation and the German Academic Exchange Service (DAAD), The Harry Crossley Foundation, Stellenbosch University and the Faculty of Medicine and Health Sciences and The Ernst and Ethel Erikson Trust, for funding.

Table of Contents

Detailed Table of Contents

List of Abbreviations

ABC	ATP binding cassette
ABC	ammonium bicoarbonate
ACN	acetonitrile
ADC	albumin, dextrose, catalase supplement
amp	ampicillin
amp^R	ampicillin resistance
APS	ammonium persulphate
ATP	adenosine triphosphate
ATPase	adenosine triphosphatase
B.	*Bacillus*
BCG	Bacille Calmette et Guérin
bp	base pair
BDGP	Berkeley Drosophila Genome Project
BPROM	Prediction of bacterial promoters
BSA	bovine serum albumin
C.	*Corynebacterium*
C-	carboxy-
CAF	Central Analytical Facility
cam	Chloramphenicol
CF	culture filtrate
CFP-10	10kDa culture filtrate protein
cm	centimeter
CoA	coenzyme A
DNA	dioxyribonucleic acid
Dnase	dioxyribonuclease
DOE	Department of Energy
ecc	esx conserved component protein
Ecc	esx conserved component gene
ECL	enhanced chemiluminescence
E. coli	*Escherichia coli*
EDTA	ethylenediaminetetraacetic acid
ERA-NET	European Research Area Network
ESAT-6	early secretory antigenic target of 6 kDa
esp	ESX secretion-associated gene
Esp	ESX secretion-associated protein
EssC	ESAT-6 secretion system C
esx	ESAT-6 family gene

Esx	ESAT-6 family protein
ESX	ESAT-6 secretion system
ESX-1	ESAT-6 secretion system 1
ESX-2	ESAT-6 secretion system 2
ESX-3	ESAT-6 secretion system 3
ESX-4	ESAT-6 secretion system 4
ESX-5	ESAT-6 secretion system 5
$ESX\text{-}2_{AN}$	ancestral ESAT-6 secretion system 2
$ESX\text{-}3_{AN}$	ancestral ESAT-6 secretion system 3
ESX-P1	plasmid ESAT-6 secretion system 1
ESX-P2	plasmid ESAT-6 secretion system 2
ESX-P3	plasmid ESAT-6 secretion system 3
ESX-P4	plasmid ESAT-6 secretion system 4
ESX-P5	plasmid ESAT-6 secretion system 5
FA	formic acid
Fur	ferric uptake regulator
g	gram
GC	guanine and cytosine
GC-MS	gas chromatography–mass spectrometry
GTP	guanine triphosphate
hyg	hygromycin
HRP	horse radish peroxidase
Hyg^R	hygromycin resistance
IDT	Integrated DNA Technologies
IdeR	iron dependant repressor
IPTG	isopropyl-β-D-thiogalactopyranoside
JGI	Joint Genome Institute
kan	kanamycin
kan^R	kanamycin resistance
KAUST	King Abdullah University of Science and Technology
kb	kilobases
KCl	potassium chloride
kDa	kilodalton
KEGG	Kyoto Encyclopedia of Genes and Genomes
kV	kiloVolt
L.	*Listeria*
L	liter
lacZ	β-galactosidase gene
LB	Luria Bertani broth
LC-MS	Liquid chromatography–mass spectrometry
m	milli

M	Molar
M.	*Mycobacterium*
$MgCl_2$	magnesium chloride
min	minute
ml	milliliter
mm	millimeter
mM	milliMolar
ms	millisecond
$m.s^{-1}$	meters per second
MPTR	major polymorphic tandem repeat
MS	mass spectrometry
MSTFA	N-Methyl-N-(trimethylsilyl) trifluoroacetamide
m/z	mass per charge ratio
MycP	mycosin protease
n	nano
N-	amino-
NA	not annotated
NCBI	National Centre for Biotechnology Information
ng	nanogram
NIH	National Institutes of Health
NISC	NIH Intramural Sequencing Center
nl	nanoliter
nm	nanometer
OD	optical density
ORF	open reading frame
oriE	*E. coli* origin of replication
oriM	mycobacterial origin of replication
P.	*Plasmodium*
PBS	phosphate buffered saline
PCA	principle component analysis
PCR	polymerase chain reaction
pe	PE gene
PE	proline-glutamic acid (mycobacterial protein family)
PGF	Production Genomics Facility
PGRS	polymorphic GC-rich sequence (mycobacterial protein family)
PLS-DA	Partial least squares Discriminant Analysis
ppe	PPE gene
PPE	proline-proline-glutamic acid (mycobacterial protein family)
ppm	parts per million
PS	pseudogene
R.	*Rhodococcus*

RD	region of difference
RT	room temperature
s	second
S.	*Saccaromyces*
sacB	levansucrase gene
SAM	S-adenosylmethionine
SAP	shrimp alkaline phosphatase
SDS	sodium dodecyl sulphate
Sec	general secretion machinery
SOB	super optimal broth
SOC	super optimal catabolite repression
TAE	tris-acetic acid-EDTA buffer
Tat	twin arginine translocation
TBS	tris buffered saline
TBS-T	tris buffered saline with Tween-20
TEMED	N,N,N',N'-tetramethylethylenediamine
tet	tetracycline
TIGR	The Institute for Genomic Research
Tm	annealing temperature
TMCS	trimethylchlorosilane
tris	tris(hydroxymethyl)aminomethane
Tween-20	polyoxyethylene sorbitan monolaurate
Tween-80	polyoxyethylene sorbitan monooleate
T7SS	type-VII secretion system
U	unit
URL	uniform resource locator
US DOE	United States Department of Energy
UV	ultraviolet
V	volt
VIP	variable of importance
WCL	whole cell lysate
WT	wild-type
WT_{ms}	wild-type *M. smegmatis* strain
WXG	tryptophan-X-glycine (mycobacterial protein family)
X	variable amino acid
X-gal	5-bromo-4-chloro-3-indolyl-β-galactoside
ZN	Ziehl-Neelsen
Zur	zinc uptake regulator
7H9	Middlebrook 7H9 Broth
7H11	BBL™ Seven H11 Agar Base
µ	micro

µg	microgram
µl	microliter
µm	micron
β	beta
°C	degrees Celcius
Ω	Ohm
%	percentage
xg	times gravitational force
Δ	delta
ΔESX-1_{ms}	*M. smegmatis* Δ*ESX-1* knockout strain
ΔESX-3_{ms}	*M. smegmatis* Δ*ESX-3* knockout strain
ΔESX-4_{ms}	*M. smegmatis* Δ*ESX-4* knockout strain
ΔESX-1ΔESX-3_{ms}	*M. smegmatis* Δ*ESX-1*, Δ*ESX-3* knockout strain
ΔESX-1ΔESX-4_{ms}	*M. smegmatis* Δ*ESX-1*, Δ*ESX-4* knockout strain
ΔESX-3ΔESX-4_{ms}	*M. smegmatis* Δ*ESX-3*, Δ*ESX-4* knockout strain
ΔESX-1ΔESX-3ΔESX-4_{ms}	*M. smegmatis* Δ*ESX-1* Δ*ESX-3*, Δ*ESX-4* knockout strain

List of company origins and abbreviations

Company	Abbreviation	Origin
AEC Amersham		South Africa
Applied Biosystems		United States of America
Becton Dickinson	BD	United States of America
Biosolve		Netherlands
Bio-Rad		United States of America
Burdick and Jackson		United States of America
Eppendorf		Germany
Fermentas		Canada
Integrated DNA Technologies	IDT	United States of America
Kimix		South Africa
Merck		Germany
Millipore		United States of America
New England Biolabs	NEB	United States of America
Promega		United States of America
Qiagen		Germany
Retsch		Germany
Roche		Switzerland
Santa Cruz Biotechnology		United States of America
Sigma		United States of America
Stratagene		United States of America
SUPELCO Analytical		United States of America
Thermo Scientific		Germany

INTRODUCTION

The mycobacterial Type VII ESX secretion systems

1. Introduction

Bacteria rely on the secretion of proteins for survival. In gram negative bacteria, which are typically enveloped by two membranes, this may involve the incorporation of proteins into the inner plasma membrane, secretion into the periplasm, insertion into the outer membrane or export through both membranes into the extracellular environment.[1] Protein secretion in gram positive bacteria is generally considered to be simpler, as these bacteria have only a single membrane layer.[2] Secreted proteins are involved in numerous essential functions, including organelle biogenesis, nutrient acquisition, virulence and efflux of drugs and toxins.

2. Bacterial secretion mechanisms

There are two secretion pathways present in both gram positive and gram negative bacteria, the general secretion system, or Sec system, and the Twin-arginine transporter, or Tat pathway. The Sec secretion system is found in all bacteria and is responsible for the export of unfolded proteins with specific N-terminal signal sequences across the plasma membrane.[3-5] The Tat pathway transports folded proteins, recognised by an N-terminal signal sequence and a double arginine motif, across the plasma membrane.[6] Additional gram positive secretion mechanisms include the polypeptide translocating ABC transporters and the type IV prepilin-like pathway which export specific proteins.[7] Six additional secretion systems (Type I to VI) exist in gram negative bacteria to transport proteins across the inner and outer membranes via unique one- and two-step mechanisms, which may utilise the Sec and Tat secretion machineries for transport across the inner membrane.[8-14]

3. Protein secretion in mycobacteria

The genus *Mycobacterium,* consisting of over 156 species and 13 subspecies,[15] contains both non-pathogenic saprophytic and pathogenic organisms, the most renowned being *M. tuberculosis*, which infects and kills millions of people annually.[16] *Mycobacteria* belong to the class Actinobacteria, within which it is placed in a distinct group, the Mycolata. The Mycolata are gram positive bacteria which possess an inner and outer membrane, the latter of which is composed of large hydroxylated branched-chain fatty acids, called mycolic acids.[17-19] The mycobacterial cell envelope (Figure 1) is composed of the inner plasma membrane consisting of a phospholipid bilayer, the outer mycomembrane, and a complex matrix of peptidoglycan and arabinogalactan, the cell wall, which

spans the periplasmic space and is covalently linked to the mycomembrane. The mycomembrane also contains various free lipid components, and forms a hydrophobic barrier between the cell and its extracellular milieu. A polysaccharide capsule surrounds the cell.[20, 21]

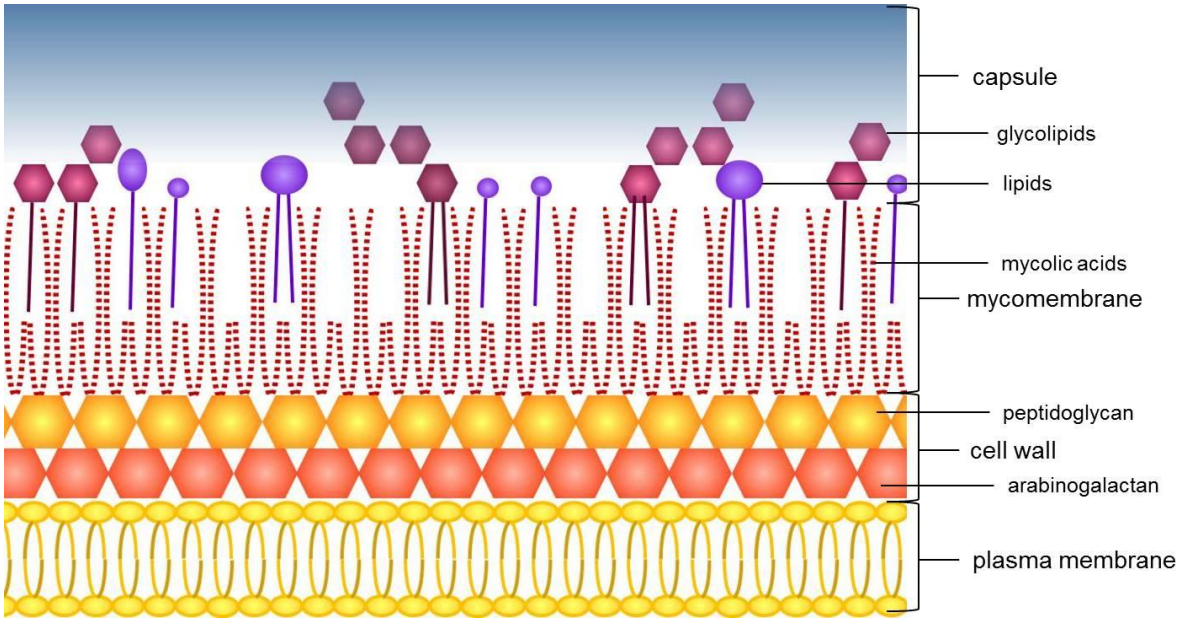


Figure 1. The Mycobacterial cell envelope structure. The mycobacterial cell is surrounded by four layers, the phospholipid bilayer plasma membrane, the cell wall of arabinogalactan and peptidoglycan which is covalently linked to the mycolic acids in the mycomembrane, which also incorporates various lipids and glycolipids, and a polysaccharide capsule.[20, 21]

The complex mycobacterial envelope forms a valuable defence mechanism against toxic compounds, drugs and the host immune system. It is, however, also an obstruction to the secretion and uptake of proteins and other molecules. Mycobacteria possess two Sec secretion machineries (SecA1[22] and SecA2[23]) and a Tat secretion pathway.[24, 25] However, these protein secretion systems only transport proteins across the inner plasma membrane, and not across the mycomembrane. Furthermore, the secretion systems responsible for protein transport across the outer membrane of most gram-negative bacteria (Type-I to -VI) are notably absent in mycobacteria, which are generally considered to be gram positive bacteria. A novel Type-VII secretion system has, however, been identified, and is proposed to be responsible for the secretion of substrate proteins which lack traditional secretion signals across the inner and outer membranes. This secretion system is encoded by the ESAT-6 (or ESX) gene cluster and is named the ESX secretion system.[26-32]

4. The ESAT-6 gene cluster

The ESAT-6 gene cluster was identified during comparative genomic analyses between the attenuated vaccine strain *M. bovis* BCG and virulent *M. tuberculosis* and *M. bovis*. During this analysis a single region, the region of difference 1 (RD1), was found to be absent from all strains of *M. bovis* BCG.[33-35] This region encompasses nine genes annotated Rv3871 to Rv3879c, including the T-cell antigens ESAT-6 (6kDa early-secreted antigenic target) and CFP-10 (10kDa culture filtrate protein) which are secreted, but lack traditional Sec and Tat secretion signals.[27, 36] Sequencing of the *M. tuberculosis* genome identified a cluster of genes across the RD1 of which there are 5 copies in the genome.[26, 27] These clusters include, in addition to genes encoding the ESAT-6- and CFP-10-like (Esx) proteins, PE and PPE family proteins, cytoplasmic ATPases (EccA), membrane ATPases (EccC), transmembrane (EccB and EccE) and integral membrane proteins (EccD), subtilisin-like serine proteases (mycosins, MycP) and other proteins (including some ESX secretion-associated proteins, Esp), and have been named the ESAT-6 or ESX gene clusters 1 to 5 (Figure 2, Table 1). These ESX gene clusters have been identified in species throughout the genus *Mycobacterium*, and have been proposed to have evolved in the order ESX-4 (Rv3450c - Rv3444c), ESX-1 (Rv3868 - Rv3883c), ESX-3 (Rv0282 - Rv0292), ESX-2 (Rv3895c - Rv3884c) and then ESX-5 (Rv1782 - Rv1798), with ESX-4 occurring in various Mycolata outside of the genus *Mycobacterium*.[27]

Several components of these ESX gene clusters have been shown to be required for the secretion of their related Esx, PE and PPE proteins, and it has been proposed that the ESX gene cluster components form dedicated secretion systems responsible for the secretion of these, and other proteins.

5. The mycobacterial ESX secretion systems

The secretion machineries encoded by the five *M. tuberculosis* ESX gene clusters have been named the ESX-1 to -5 secretion systems, and have recently been proposed to be a novel Type VII secretion system.[30, 31] This machinery is hypothesised to be responsible for the transport of proteins across the plasma membrane and uniquely, across the complex mycomembrane of the Mycolata. Although the precise structure and secretion mechanism of these secretion systems have not been clearly defined, the roles of specific proteins have been shown or predicted based on experimental and *in silico* studies and used to propose a model for ESX secretion (Figure 3).

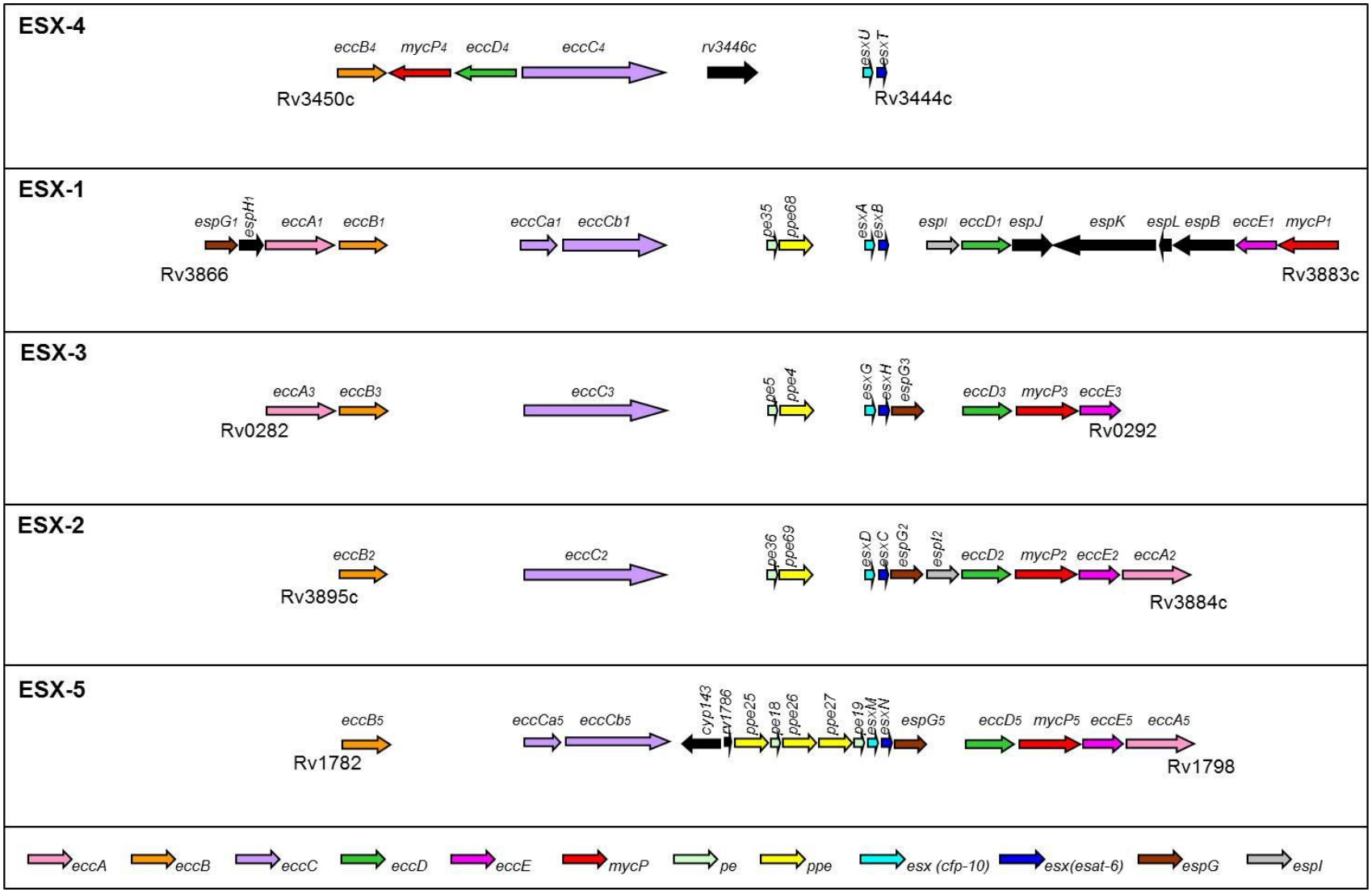


Figure 2. The ESX gene clusters of *M. tuberculosis.* The *M. tuberculosis* ESX gene clusters contain the ESX conserved component genes (*eccA-E*), mycosin genes (mycP), ESX secretion associated genes (*esp*) as well as the *pe, ppe* and *esx* genes.

Table 1. The components of the five ESAT-6 gene clusters of *M. tuberculosis.*

ESX component	Description	Presence of genes in the ESAT-6 gene cluster regions				
		1	2	3	4	5
EspG	ABC transporter family signature	Rv3866 (*espG$_1$*)	Rv3889c (*espG$_2$*)	Rv0289 (*espG$_3$*)		Rv1794
EccA	AAA+ class ATPases, CBXX/CFQX family, SpoVK, 1x ATP/GTP-binding site	Rv3868 (*eccA$_1$*)	Rv3884c (*eccA$_2$*)	Rv0282 (*eccA$_3$*)		Rv1798 (*eccA$_5$*)
EccB	Amino terminal transmembrane protein, possible ATP/GTP-binding motif	Rv3869 (*eccB$_1$*)	Rv3895c (*eccB$_2$*)	Rv0283 (*eccB$_3$*)	Rv3450c (*eccB$_4$*)	Rv1782 (*eccB$_5$*)
EccC	DNA segregation ATPase, ftsK chromosome partitioning protein, SpoIIIE, YukA, 3x ATP/GTP-binding sites, 2x amino-terminal transmembrane protein	Rv3870 (*eccCa$_1$*) -Rv3871 (*eccCb$_1$*)	Rv3894c (*eccC$_2$*)	Rv0284 (*eccC$_3$*)	Rv3447c (*eccC$_4$*)	Rv1783 (*eccCa$_5$*) -Rv1784 (*eccCb$_5$*)
PE	PE	Rv3872 (*pe35*)	Rv3893c (*pe36*)	Rv0285 (*pe5*)		Rv1788 (*pe18*) Rv1791 (*pe19*)
PPE	PPE	Rv3873 (*ppe68*)	Rv3892c (*ppe69*)	Rv0286 (*ppe4*)		Rv1787 (*ppe25*) Rv1789 (*ppe26*) Rv1790 (*ppe27*)
Esx	CFP-10, Esx family protein	Rv3874 (*esxB*)	Rv3891c (*esxD*)	Rv0287 (*esxG*)	Rv3445c (*esxU*)	Rv1792 (*esxM*)
Esx	ESAT-6, Esx family protein	Rv3875 (*esxA*)	Rv3890c (*esxC*)	Rv0288 (*esxH*)	Rv3444c (*esxT*)	Rv1793 (*esxN*)
EspI	ATPases involved in chromosome partitioning, 1x ATP/GTP-binding motif	Rv3876 (*espI*)	Rv3888c			
EccD	Integral inner membrane protein, binding-protein-dependent transport systems inner membrane component signature, putative transporter protein	Rv3877 (*eccD$_1$*)	Rv3887c (*eccD$_2$*)	Rv0290 (*eccD$_3$*)	Rv3448 (*eccD$_4$*)	Rv1795 (*eccD$_5$*)
MycP	Mycosin, subtilisin-like cell wall-associated serine protease	Rv3883c (*mycP$_1$*)	Rv3886c (*mycP$_2$*)	Rv0291 (*mycP$_3$*)	Rv3449 (*mycP$_4$*)	Rv1796 (*mycP$_5$*)
EccE	2x amino-terminal transmembrane protein	Rv3882c (*eccE$_1$*)	Rv3885c (*eccE$_2$*)	Rv0292 (*eccE$_3$*)		Rv1797 (*eccE$_5$*)

Ecc: esx conserved component; Esp: ESX secretion-associated protein; MycP: mycosin protease.

The prototype substrate of the ESX secretion system is the pair of Esx proteins encoded in each gene cluster. There are eleven pairs of *esx* genes (*esxA-esxW*) in the *M. tuberculosis* genome, six of which occur outside of the ESX gene clusters, but which are duplicated from an ESX gene cluster.[26, 37, 38] The Esx proteins are small proteins of approximately 100 amino acids, which although not highly conserved in sequence, each contain a WXG amino acid motif and have a helix-turn-helix structure.[37, 39-42] Each Esx protein pair interacts to form heterodimers or heterotetramers,[40-42] which are secreted as a complex after which they may dissociate to perform their function.[43] The Esx pairs encoded in, and duplicated from, each ESX gene cluster are secreted by the associated secretion system.

The ESX-1, -2, -3 and 5 gene clusters each contain at least one *pe-ppe* gene pair, encoded directly upstream of the *esx* genes.[27] There are also *pe* and *ppe* genes associated with four of the *esx* duplications, as well as several additional duplications throughout the *M. tuberculosis* genome, amounting to 99 *pe* and 69 *ppe* genes, comprising 10% of its coding DNA.[37, 38] The PE proteins contain a conserved 110 amino acid N-terminal domain, with a characteristic proline-glutamic acid (PE) motif at positions 8 and 9,[37, 44, 45] while the PPE proteins have a proline-proline-glutamic acid (PPE) motif from positions 7 to 9 in a unique conserved N-terminal domain of approximately 180 amino acids.[37, 44] Both protein families possess highly variable C-terminal domains which are used to classify them into subfamilies.[45] The functions of the PE and PPE proteins have not been clarified, although some have been shown to influence colony morphology, cellular architecture, cell-cell interactions and antigenic variation, and specific proteins have been shown to be fibronectin-binding and siderophore proteins or involved in phagosome-lysosome fusion, macrophage vacuole acidification, granuloma persistence, replication in macrophages and virulence and essential for *in vitro* or *in vivo* growth.[37, 45-55] Several PE and PPE proteins have been localised to the plasma membrane, outer membrane, cell wall and culture supernatant and shown to be secreted by the ESX secretion systems.[56-65] Various PE-PPE protein pairs have been shown to form 1:1 complexes, with a four-helix structure, and to interact with Esx and other ESX-encoded proteins.[66-70]

In addition to the Esx and PE-PPE proteins, various other ESX secretion-associated proteins (Esp), some of which are secreted by the ESX secretion systems, have been identified. The EspC and EspA proteins and the EspB and EspK proteins are secreted and are also predicted to interact to form four-helix bundles, similar to the Esx and PE/PPE complexes, suggesting that this structure may be a characteristic of ESX secretion substrates.[71] Recently a general mycobacterial Type VII secretion

signal has been identified on the C-terminal of several ESX secreted substrates, including PE25 (PE25/PPE41, ESX-5), LipY (ESX-5), PE35 (PE35/PPE68, ESX-1), CFP-10 (ESAT-6/CFP-10, esxBA, ESX-1) and EspB (EspB/EspK, ESX-1). This YxxxD/E motif is essential for the secretion of these substrates, but does not specify the ESX secretion system responsible for the secretion of the specific substrate.[72]

The ESX gene clusters encode three different ATPases. The *eccC* gene occurs in all five ESX gene clusters and encodes an N-terminal transmembrane ATPase which is thought to provide energy for translocation of the substrate proteins.[26, 27, 39] EccC interacts with the C-terminus of EsxB to translocate the EsxBA complex.[73, 74] EccA is a hexameric protein which undergoes "open-close" movements on ATP binding and hydrolysis, to allow interactions with, and energy transfer to other proteins.[75, 76] EccA interacts with EspC, an additional substrate of ESX-1, to enable secretion of EspF/EspC.[73] The third ATPase is only encoded in ESX-1 and ESX-2, and is annotated EspI. It appears that interactions with multiple ATPases are required for effective ESX secretion.

Several additional membrane proteins are encoded in the ESX gene clusters. The *eccD* gene encodes an integral membrane protein with 11 or 12 transmembrane domains, and is predicted to form the pore through which the proteins are translocated through the plasma membrane.[26, 27, 30] MycP is a subtilisin-like serine protease which is anchored in the plasma membrane and extends into the cell wall.[77] A single substrate of MycP1, EccB, has been identified,[78] but it has been predicted that MycP may be responsible for the cleavage of EccD in order to open the channel and enable secretion.[71] Two additional transmembrane proteins, EccB and EccE are encoded in various ESX gene clusters and are required for ESX secretion.[26, 27, 29] Although their roles have not been clarified, these proteins are predicted to form part of the secretion complex. The protein(s) which are responsible for translocation across the mycomembrane have yet to be identified.

ESX-4 contains only EccB, EccC, EccD and MycP, in addition to the Esx protein pair, suggesting that these may be the core components of the ESX secretion system which are required for secretion of the Esx substrates, while additional components may assist in the translocation of additional substrates including the PE/PPE and Esp secreted proteins, and contribute to the regulation of ESX secretion. The ESX-1 secretion system, its functions, substrates and secretion mechanism have been most comprehensively studied, although more recently ESX-3 and ESX-5 have been the topics of extensive research, while ESX-4 and ESX-2 remain largely unstudied.

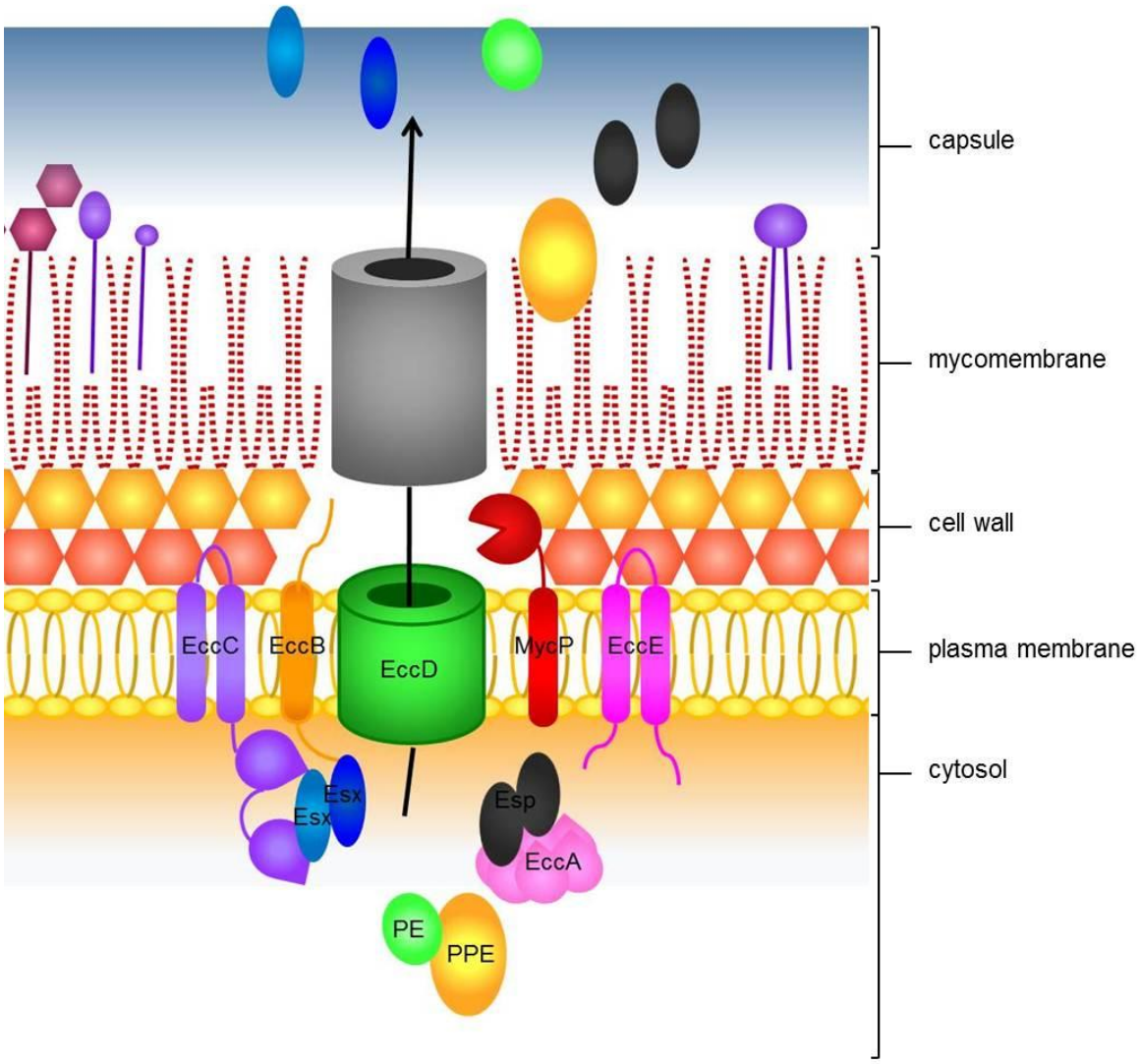


Figure 3. A model of the ESX secretion machinery. The ESX substrates Esx/Esx, PE/PPE and Esp/Esp interact with their partners to form tetra-helical complexes, which interact with the ATPases EccC and EccA via the C-termini of one partner, to provide the energy for translocation through the membrane pore EccD. MycP may assist in translocation by cleaving EccD pore or by interacting with secreted subtrates. The secreted substrates are translocated through an unknown component through the mycomembrane where they may remain associated with the membrane or are released into the extracellular milleu, where they may dissociate from their partners to perform their functions. Interaction and cosecretion of the substrates is required for optimal secretion. The functions of EccB and EccE are unknown, but they are believed to form part of the secretory complex.[30, 39, 40, 43, 57, 63, 67, 70, 71, 73, 74]

5.1. ESX-1

The ESX-1 gene cluster was first identified as traversing the primary attenuating deletion, Region of Difference 1 (RD1), from the *M. bovis* BCG vaccine strain.[33-35] The gene cluster is also absent in other relatively avirulent species, including *M. microti* and the "dassie bacillus", but present in virulent *M. bovis* and *M. africanum*.[79-81] Deletion of ESX-1 from *M. tuberculosis* was furthermore shown to attenuate the organism,[82-84] leading to the deduction that ESX-1 plays an important role in mycobacterial pathogenicity. The *esxA* and *esxB* genes encoded within this region encode two potent T-cell antigens, ESAT-6 (6kDa early secreted antigenic target) and CFP-10 (10kDa culture filtrate protein) which have been identified in the culture supernatant of *M. tuberculosis.*[36] ESAT-6 and CFP-10 do not contain traditional secretion signals and their secretion is dependent on the rest of the ESX-1 components,[29, 32, 82, 84, 85] leading to the proposed function of ESX-1 as a protein secretion system.

Several additional proteins have been shown, or proposed, to be secreted by ESX-1 in *M. tuberculosis,* some of which are retained on the cell surface. These include PE35, PPE68, EspJ, EspK, EspL and EspB, which are encoded within ESX-1, EspE, EspF, EspG and EspH, encoded directly upstream of ESX-1 as well as EspC, EspA and EspD which are encoded in an operon elsewhere in the genome, and the transcriptional regulator EspR.[57, 73, 86-94] These various secreted proteins interact with each other in various combinations, and with different ATPases, including the $EccA_1$ and $EccC_1$ proteins, via their C-termini, to provide the energy for translocation.[70, 73, 74, 87] Various substrates are co-dependent for secretion and full virulence is dependent on expression and secretion of the full contingent of ESX-1 substrates.[57, 73, 87, 88, 93] Other ESX-1 substrates may play a role in stabilising intracellular levels of other proteins[86, 91] or may influence virulence independently of the secretion of other substrates.[93] EspR is involved in transcriptional regulation of the *espACD* operon,[90, 94] thereby regulating ESX-1 dependent virulence, and is also secreted by ESX-1 maintaining a negative feedback loop regulating ESX-1 secretory activity. EspR also binds to 165 other loci in the *M. tuberculosis* genome suggesting that it may coordinate ESX-1 secretion with other cellular functions.[95]

The role of ESX-1 in mycobacterial virulence has been studied extensively. ESX-1 dependent secretion of ESAT-6 has been shown to inhibit T-cell responses and phagosome maturation.[96, 97] Secreted ESAT-6 associates with the membrane of vacuoles in order to destabilise and create pores in the membrane, enabling the escape of the mycobacteria into the macrophage cytosol.[43, 98-100] Perforation of the phagosome membrane also allows the transfer of mycobacterial DNA to the

macrophage cytosol, which activates the cytosolic surveillance pathway and contributes to the increased interferon-γ production observed during tuberculosis infection.[101, 102] ESX-1 dependent phagosome perforation also results in caspase-1 activation and induces the secretion of interleukin-1β and interleukin-18, [103, 104] and activates the inflammasome, exacerbating inflammation and promoting tuberculosis disease.[92]

In the non-pathogenic *M. smegmatis*, ESX-1 has been implicated in conjugal DNA transfer, where it negatively regulates DNA transfer from the donor strain,[105] but is essential for DNA uptake by the recipient strain.[106] ESX-1 is localised at the non-septal pole in *M. smegmatis, M. marinum* and *M. tuberculosis*, with specific conserved proteins implicated in trafficking of the secretion system to the pole.[107, 108] $EccA_1$ has also been linked to mycolic acid synthesis,[109] linking ESX-1 secretion, virulence and mycolic acid synthesis to the non-divisive pole where rapid cell wall synthesis occurs. The *M. tuberculosis* and *M. smegmatis* ESX-1 gene clusters have repeatedly been shown to be functionally equivalent, [105, 107, 110] although conjugal DNA transfer does not occur in *M. tuberculosis* and *M. smegmatis* is avirulent. This suggests that ESX-1 has evolved additional functions in the slow-growing pathogenic mycobacteria, whilst retaining conserved mechanisms. The divergent functions of these ESX-1 secretion systems might be attributed to the additional secretion substrates in *M. tuberculosis* which are not present in *M. smegmatis.*

5.2. ESX-2

ESX-2 is encoded adjacent to ESX-1, but on the complementary strand.[27] This gene cluster is dispensable for the *in vitro* and *in vivo* growth of *M. tuberculosis* and deletions of some components have been found in clinical isolates of *M. tuberculosis.*[54, 111] The ESX-2 secretion system has not been investigated and its functions and substrates remain unknown, but its expression has recently been shown to be induced by WhiB5 [112] and the ESX-1 associated regulator EspR,[95] suggesting that its expression and function may be coordinated with those of other ESX secretion systems.

5.3. ESX-3

ESX-3 is essential for the *in vitro* survival of *M. tuberculosis*, although it is dispensable in the fast-growing saprophyte *M. smegmatis*.[54, 113] Expression of ESX-3 in *M. tuberculosis* is regulated by the iron dependent regulator (IdeR) and the zinc uptake regulator (Zur) in response to iron and zinc concentrations, respectively.[50, 114] In *M. smegmatis*, however, regulation of ESX-3 is only influenced by

iron concentration.[115] High levels of iron and zinc down-regulate expression of this secretion system and ESX-3-mediated protein export is increased during iron-starvation, leading to the hypothesis that ESX-3 is involved in divalent-cation, specifically iron and zinc, homeostasis.[116]

A conditional ESX-3 deletion mutant of *M. tuberculosis* was shown to survive *in vitro* in culture medium containing high concentrations of iron and zinc, suggesting that ESX-3 is somehow involved in the acquisition of these metal cations.[113] Furthermore the ESX-3 mutant could grow in the culture filtrate or boiled supernatant of wild-type *M. tuberculosis*, suggesting that ESX-3 secretes some factor(s) which enable iron acquisition. ESX-3 was shown to be involved in the mycobactin-mediated iron-acquisition pathway, somehow influencing the uptake of iron from exported mycobactins.[113] *M. tuberculosis* has only the mycobactin pathway for iron acquisition, whereas *M. smegmatis* uses an additional pathway, the exochelin system, as its primary iron acquisition mechanism. The involvement of ESX-3 in mycobactin-mediated iron acquisition, and the redundancy of this system in *M. smegmatis,* likely differentiates the essentiality of this region in the two organisms.

ESX-3 has also been associated with immunogenicity. EsxH, previously labelled TB10.4, was found in the culture supernatants of *M. tuberculosis* and found to be highly antigenic.[117] Iron is sequestered by the host during infection, exposing the bacterium to iron starvation, and this may result in increased ESX-3 expression and the associated secretion of EsxH, during infection, resulting in the antigenicity of the protein. However, ESX-3 has also been implicated in the evasion of the innate immune system by *M. smegmatis*, and the *M. tuberculosis* ESX-3 was shown to stimulate bacterial immunity, suggesting that ESX-3 also promotes mycobacterial virulence.[118]

The acquisition of iron and evasion of the immune system are essential for pathogen survival in the host, and the apparent involvement of ESX-3 in both of these pathways highlights it as an important mediator of pathogenicity in *M. tuberculosis.*

5.4. ESX-4

ESX-4 is the most ancestral of the ESX secretion systems, and is present in several species outside of the genus *Mycobacterium*.[27] The smallest of the ESX gene clusters, ESX-4 contains only the $eccB_4$, $eccC_4$, $eccD_4$ and $mycP_4$ genes together with the *esxU-EsxT* gene pair. This may indicate that these are the core ESX machinery components which are required for the secretion of the Esx heterodimer. An additional gene, *Rv3446c,* is also conserved within this gene cluster, although its function has not

been determined and it has not been shown to be secreted. A recent study failed to observe secretion of EsxU in *M. smegmatis* and speculated that ESX-4 is either not transcribed or non-functional in *M. smegmatis*.[119] EsxT and EsxU have never been detected in *M. tuberculosis* culture supernatants, supporting this hypothesis.[120] Although ESX-4 is non-essential for *in vitro* and *in vivo* survival, *M. leprae* is the only mycobacterium which has lost ESX-4, despite deletions of most other ESX gene clusters. The maintenance and conservation of ESX-4 throughout the genus *Mycobacterium* and in other actinomycetes suggests that this region must play a role in the biology of the organism.[27, 30, 54] Expression of $EccC_4$ was shown to be upregulated when FtsI, a protein involved in septum formation was inhibited,[121] possibly linking ESX-4 to cell division. Transcription of *esxU* and *esxT* is regulated by the alternative sigma factor SigM and the global transcriptional regulator WhiB5, which is implicated in maintenance of metabolism during starvation and in reactivation of *M. tuberculosis*.[112, 122, 123] This suggests that the expression and function of ESX-4 may be limited to specific conditions or periods during the growth cycle.

5.5. ESX-5

ESX-5 is the most recently evolved ESX secretion system and has only been identified in the slow-growing mycobacteria.[27] *M. tuberculosis* ESX-5 contains 3 PPE and 2 PE genes, and the PE/PPE gene families have been shown to have expanded from ESX-5 into the rest of the genome.[38] ESX-5 has predominantly been studied in the fish pathogen, *M. marinum*, where it has been shown to be responsible for the export of various PPE, PE, PPE-MPTR and PE-PGRS proteins, as well as LipY, which are cell surface and secreted proteins.[58-60] The secretion and presentation of surface antigens by ESX-5 modulates the host macrophage response, reducing the secretion of pro-inflammatory and toll-like receptor signalling-dependent innate immune cytokines.[59, 124] ESX-5 also induces inflammasome activation and interleukin-1β secretion, and contributes to caspase-independent cell death after translocation of the mycobacteria to the host cytosol.[99, 125]

A recent study in *M. tuberculosis* confirmed the role of ESX-5 in the secretion of the EsxM-EsxN and the other TB9.9 Esx proteins (which were duplicated from ESX-5), as well as several *M. tuberculosis* PE and PPE proteins.[61] This supports the theory that ESX-5 is responsible for the secretion of the majority of PE and PPE proteins, specifically the PE-PGRS and PPE-MPTR proteins, which expanded from the ESX-5 gene cluster.[38] $EccD_5$, $EccA_5$ and the ESX-5 encoded PE and PPE proteins were shown to be required for secretion of EsxN and various PE and PPE proteins.[61] An $EccD_5$ knock-out

strain of *M. tuberculosis* was hypersensitive to detergents and hydrophilic antibiotics, implicating ESX-5 in the formation of a stable cell wall. Furthermore various ESX-5 components were shown to be essential for intracellular growth in macrophages and for virulence in SCID mice.[61]

ESX-5 plays an important role in modulating the human macrophage response during various critical stages during infection, by secreting PE and PPE proteins, thereby altering the immune response to evade the immune system and maintain a persistent infection.

6. Conclusion

The mycobacteria have several features which assist in their survival in general, and in escaping the host immune system, to enable their success as pathogens. It appears that at least the ESX-1 and ESX-5 secretion systems perform important roles in pathogenesis; however their precise biological functions remain unknown. ESX-3 is essential for the survival of *M. tuberculosis* and other pathogenic mycobacteria, and is involved in zinc and iron homeostasis, although the mechanism whereby it functions is unclear. Furthermore the functions of ESX-2 and ESX-4 remain completely unknown. The presence of various of these ESX gene clusters in the fast-growing and non-pathogenic mycobacteria suggests that they perform important biological functions, from which the virulence related functions have evolved. Although *M. tuberculosis* has been studied extensively, little is understood of the mechanisms of pathogenicity of the organism. A better understanding of these secretion systems, which are essential for the virulence and survival of *M. tuberculosis,* may lead to novel developments in the treatment and prevention of tuberculosis disease.

7. Study Design

7.1. Aim

The aim of this study is to investigate the evolution, regulation and functions of the mycobacterial Type VII ESX secretion systems (T7SSs) and their substrates.

7.2. Objectives

1. To investigate the evolutionary history of the Type VII ESX secretion systems (T7SSs) by determining the presence of these systems in different mycobacterial and other actinobacterial species using *in silico* and phylogenetic analyses.

2. To determine the regulation of the ESX secretion systems by identifying promoters responsible for the expression of each of the *M. tuberculosis* ESX gene clusters using a β-galactosidase promoter-probe assay.

3. To investigate the functions of the ESX secretion systems using comparative metabolomics and whole cell lysate and culture supernatant proteomic analyses of *M. smegmatis* wild-type and ESX knock-out strains.

4. To identify novel substrates of the ESX secretion systems using comparative proteomic analyses of the culture supernatants of *M. smegmatis* wild-type and ESX knock-out strains.

MATERIALS AND METHODS

1. Bacterial strains and culture conditions

1.1. Bacterial strains

E. coli JM109, *M. smegmatis* mc^2155[126] and *M. bovis* BCG Pasteur 1173P2 (Table 1) were used in this study.

1.2. Media and culture conditions

Standard culture media are described in Table 2. Antibiotics and supplements are described in Table 3.

E. coli was cultured in liquid Luria-Bertani broth (LB) with shaking, and on LB agar plates, overnight at 37°C. Solid and liquid media were supplemented with antibiotics; ampicillin (50 µg/ml) and kanamycin (50 µg/ml); and solid media additionally with 5-bromo-4-chloro-3-indolyl-β-galactoside (X-gal) and isopropyl-β-D-thiogalactopyranoside (IPTG); as appropriate.

M. smegmatis was routinely grown in Middlebrook 7H9 Broth with shaking, and on BBL™ Seven H11 Agar Base plates, at 37°C for 2-3 days unless otherwise described. Kanamycin (25 µg/ml), sucrose (5%) and X-gal were added as appropriate.

M. bovis BCG was routinely grown in Middlebrook 7H9 Broth with shaking, and on BBL™ Seven H11 Agar Base plates, at 37°C for 2-3 weeks unless otherwise described. Kanamycin (25 µg/ml), was added as appropriate.

Bacterial strains used in this study were stored in 40% glycerol at -80°C.

1.3. Ziehl-Neelsen Staining

All mycobacterial cultures were screened for contamination using Ziehl-Neelsen (ZN) staining according to standard protocol.[127] Briefly, aliquots of mycobacterial cultures were heat-fixed to microscope slides. Slides were flooded with ZN Carbol Fuchsin (BD), heated intermittently with a flame until steaming, and allowed to stand for 5 minutes. Slides were rinsed with water, decolourised with 5% acid-alcohol solution for 2 minutes and rinsed with water. Slides were then counterstained with Methylene Blue (BD) for 1-2 minutes, rinsed with water and allowed to air dry.

Table 1. Bacterial Strains

Strain	Genotype	Features	Reference/Source
E. coli JM109	*recA1, endA1, gyrA96, thi, hsdR17, supE44, relA1, Δ(lac-proAB)/F' [traD36, proAB+, lacIq, lacZΔM15*	Chloramphenicol resistance Methylation: dam+, dcm+ Blue-white selection +	Laboratory strain
M. smegmatis mc²155	*ept-1*	Efficient plasmid transformation mutant of *M. smegmatis* mc²6 Mycobacterial model organism	126
M. bovis BCG Pasteur 1173P2	*ΔRD1*	Attenuated strain of *M. bovis*	Institut Pasteur

Table 2. Culture media

Medium	Composition	Source
Luria-Bertani broth (LB)	1% Tryptone, 0.5% Sodium chloride, 0.5% Yeast extract	
Luria-Bertani agar (LB agar)	1% Tryptone, 0.5% Sodium chloride, 0.5% Yeast extract, 1.2% bacteriological agar	
SOB	2% Tryptone, 0.5% Sodium chloride, 0.5% Yeast extract, 2.5 mM Potassium chloride,	
SOC	SOC supplemented with 20 mM glucose and 10 mM Magnesium chloride after autoclaving	
Middlebrook 7H9 Broth (7H9)	0.47% Middlebrook 7H9 Broth powder, supplemented with 0.2% glycerol, 0.05% Tween-80 and 10% ADC after autoclaving	BD
BBL™ Seven H11 Agar Base (7H11)	1.9% BBL™ Seven H11 Agar Base powder, supplemented with 0.5% glycerol, 0.05% Tween-80 and 10% ADC after autoclaving	BD
Kirchener's broth	0.5% asparagine, 0.3% Na_2HPO_4, 0.4% KH_2PO_4, 0.25% Tri-sodium citrate, 1.07 g.liter-1 $MgSO_4$, 2% glycerol, 0.05% Tween-80	
ADC	5% bovine serum albumin fraction V (Roche), 2% glucose (Kimix), 1000 U/ml catalase (Sigma), filter sterilised	

All media was prepared in deionised water and autoclaved at 121°C for 20 minutes. Concentrations of all dry components are given as % m/v, and liquids as % v/v.

Table 3. Antibiotics and supplements

	Stock concentration	Solvent	Sterilisation	Storage	Supplier	Working concentration	
						E. coli	*M. smegmatis/ M. bovis BCG*
Ampicillin (amp)	50 mg/ml	Deionised water	filtered	-20°C	Roche	50 µg/ml	n/a
Kanamycin* (kan)	50 mg/ml	0.9% NaCl solution	filtered	4°C	Sigma	50 µg/ml	25 µg/ml
Tetracycline (tet)	5 mg/ml	Ethanol		-20°C	Sigma	50 µg/ml	n/a
5-bromo-4-chloro-3-indolyl-β-galactoside (X-gal)	20 mg/ml	N,N'-dimethyl formamide		-20°C, in foil	Roche	100 µl spread on 90 mm plate	
Isopropyl-β-D-thiogalactopyranoside (IPTG)	0.1 M	Deionised water	filtered	-20°C	Biosolve	100 µl spread on 90 mm plate	
Potassium Chloride (KCl)	1 M	Deionised water	autoclaved, 121°C, 20min	RT	Merck		
Magnesium Chloride* ($MgCl_2$)	1 M	Deionised water	filtered	RT	Sigma		
Glucose	1 M	Deionised water	filtered	-20°C	Merck		
Polyoxyethylene-sorbitan monooleate (Tween 80)	20% v/v	Deionised water	filtered	4°C	Sigma		0.2% v/v
Sucrose	50% v/v	Deionised water	filtered	4°C	Merck		5% v/v
ADC		Deionised water	filtered	4°C			10%

*Supplements and antibiotics purchased in solution

Slides were read using a light microscope under the 100X (oil immersion) lens. Uncontaminated mycobacterial cultures contained only acid fast bacilli (pink rods).

1.4. Preparation of electrocompetent cells

1.4.1. Electrocompetent E. coli

E. coli JM109 cells were inoculated into 50 ml LB containing tetracycline (50 µg/ml) and shaken overnight at 37ºC. Overnight cultures were inoculated into fresh antibiotic-containing medium, at a dilution of 1:100, and grown to OD_{600} = 0.7 at 37ºC, with shaking. Cultures were placed on ice and all further steps performed under ice-cold conditions. Cells were harvested by centrifugation at 4000 xg for 10 minutes at 4ºC. The cell pellet was resuspended in an equal culture-volume of ice-cold 10% glycerol and centrifuged at 4000 xg for 10 minutes at 4ºC. The wash step was repeated and the cells pooled in a 50 ml falcon tube and centrifuged at 4000 xg for 10 minutes at 4ºC. The cells were resuspended in ice-cold 10% glycerol using a volume of 2 ml per liter of culture. Aliquots of 100 µl were prepared, frozen in liquid nitrogen and stored at -80ºC.

1.4.2. Electrocompetent M. smegmatis and M. bovis BCG

M. smegmatis mc^2155 and *M. bovis* BCG cells were inoculated into Middlebrook 7H9 medium and grown to OD_{600} = 0.5, at 37ºC with shaking. Cultures were incubated on ice for 1 hour and all further steps performed under ice-cold conditions. Cells were pelleted by centrifugation at 3000 xg for 10 minutes at 4ºC, resuspended in ice-cold 10% glycerol equal to the culture volume and pelleted. The wash step was repeated and the cells resuspended in ice-cold 10% glycerol using a volume of 2 ml per 100 ml of culture. Electrocompetent *M. smegmatis* and *M. bovis* BCG were freshly prepared before use.

2. DNA manipulation

2.1. PCR amplification

Primers were designed using Primer Premier 5.0 (PREMIER Biosoft International) and obtained from Integrated DNA Technologies (IDT). Appropriate restriction endonuclease cutting sites were incorporated as necessary to facilitate cloning procedures. FastStart Taq DNA Polymerase (Roche)

and HotStarTaq (Qiagen) were used for PCR amplification using the GC-rich and Q-buffers respectively, as described by the manufacturers. Elongation times were calculated at 1 minute per 1 kb. Thirty-five cycles were used unless otherwise stated.

2.2. Restriction Digestion

Plasmid DNA was digested using restriction endonucleases as per manufacturer's instructions, using the appropriate buffers. Two to five micrograms of plasmid DNA was digested using 10 U of enzyme in a total volume of 10-20 µl for 2-5 hours at 37°C. All restriction endonucleases were obtained from New England Biolabs (NEB), Roche, Promega or Fermentas.

2.3. Agarose gel electrophoresis

PCR products and restriction digestions were separated by electrophoresis at 80-100 V on 1.0 to 1.5% agarose gels in TAE buffer (40 mM Tris, 20 mM Acetic acid, 1 mM Ethylenediamine tetraacetic acid (EDTA)) with ethidium bromide and visualized using UV light. The GeneRuler™ 100 bp DNA Ladder Plus, ready-to-use (Fermentas) was used to determine band size. When necessary the resultant bands were excised from the gel and purified.

2.4. DNA clean-up

The Wizard® SV Gel and PCR Clean-up System (Promega) was used to purify DNA embedded in agarose gel slices and directly from a PCR or restriction endonuclease reaction.

2.5. DNA quantification

DNA was quantified using the NanoDrop® ND1000 Spectrophotometer.

2.6. DNA sequencing

Automated sequencing was done by the Central Analytical Facility (CAF) at Stellenbosch University, using the ABI 3130XL Genetic Analyzer (Applied Biosystems) in combination with the corresponding DNA primers.

2.7. Cloning vectors

All inserts were cloned into the commercial T-vector, pGem-T-Easy (Promega), and subcloned into the relevant vectors described in Table 4.

Table 4. Vectors

Vector	Description	Reference/Source
pGem-T Easy	*E. coli* cloning T-vector, amp^R, *lac*Z, *ori*E	Promega
pBluescript II KS	Cloning vector, amp^R, *lac*Z, *ori*E	Stratagene
pJem15	Promoter probe vector used for creating transcriptional fusions with *lac*Z, kan^R, *ori*E, *ori*M	A kind gift from J. Rauzier[128]
p2NIL	*E. coli* cloning vector, mycobacterial suicide vector, *ori*E, kan^R	A kind gift from E. Machowski[129]
pGOAL17	Plasmid with *lac*Z and *sac*B genes in PacI cassette, *ori*E, amp^R	A kind gift from E. Machowski[129]

amp^R - ampicillin resistance; lacZ - β-galactosidase gene; *ori*E - origin of replication in *E. coli*; kan^R - kanamycin resistance; *ori*M - origin of replication in mycobacteria; *sac*B - levansucrose activity

2.8. Dephosphorylation

Linearized vectors were dephosphorylated using Shrimp Alkaline Phosphatase (SAP, Roche) to prevent self-ligation of the digested plasmids. One unit of SAP was added to 1 µg digested vector DNA with the appropriate buffers, according to the manufacturer's instructions. The reaction was incubated for 10 minutes at 37°C, followed by deactivation for 15 minutes at 65°C.

2.9. Ligations

PCR and restriction digest products were ligated into appropriate vectors using T4 DNA Ligase (Promega), according to manufacturer's instructions. Inserts were ligated into 100 ng of vector at an insert:vector molar ratio of 3:1 using 3 U of T4 DNA Ligase, and incubated at 4°C overnight. Ligation reactions were deactivated at 70°C for 10 minutes prior to electrotransformation.

2.10. Transformations

All transformations of *E. coli, M. smegmatis* and *M. bovis* BCG were by electroporation of electrocompetent cells, using the Gene Pulser™ (Bio-Rad) in 0.2 cm Gene Pulser® Cuvettes (Bio-Rad).

E. coli was transformed at 2.5 kV, 25 µF, 125 µFd, 200 Ω using 2 ul of a 10 µl ligation reaction, or 1 µg of plasmid DNA, in 45 µl of electrocompetent cells. One milliliter of SOC was added to the transformed cells which were recovered for 30-60 minutes at 37°C before plating on appropriate solid media.

One microgram of purified plasmid DNA was transformed into 200 µl of electrocompetent *M. smegmatis* or *M. bovis* BCG at 2.5 kV, 25 µF, 125 µFd, 1000 Ω followed by the addition of 1 ml 7H9 medium and expression for 3 or 4 hours, respectively, at 37°C before plating on appropriate solid medium.

2.11. Colony PCR

All transformants were confirmed by colony PCR using 1 µl of culture in a 12.5 ul PCR reaction as described under PCR amplification.

2.12. Miniprep plasmid isolation

Plasmid DNA was isolated from *E. coli* using the Wizard® *Plus* SV Minipreps DNA Purification System (Promega), according to manufacturer's instructions, and eluted in 50 µl nuclease-free water (Promega).

2.13. Cloning

Inserts were amplified by PCR using specific PCR primers. The PCR products were separated by electrophoresis, purified and ligated into pGem-T-Easy according to the manufacturer's instructions. The ligation mixture was transformed into competent *E. coli* and selected for on appropriate LB agar. Colonies were picked and inoculated into LB media. Colony PCRs were done to confirm the insert and the construct was isolated. The resultant constructs were sequenced using Sp6 and T7 primers to confirm the sequence of the inserts.

The constructs were digested using the primer-specific restriction endonucleases and the inserts separated by electrophoresis, cut from the gel and purified. The purified inserts were ligated into the pre-digested expression or suicide vectors, as appropriate, and transformed into *E. coli.* Transformed colonies were picked, colony PCR done and the constructs isolated and sequenced to verify the sequence of relevant inserts.

3. Protein techniques

3.1. Culture fractionation

Mycobacterial cultures were separated into cell pellet and culture supernatant by centrifugation at 2500 xg for 10 minutes at 4°C. All subsequent steps were performed on ice.

3.1.1. Whole cell lysate (WCL) sample preparation

The cell pellet was washed twice, and resuspended in the appropriate volume of lysis buffer. A volume of 0.1 mm silica beads equal to the volume of the cell pellet was added and the cells were mechanically lysed in a BIO 101/Savant FastPrep FP120, 6 times for 20 seconds at 4.0 $m.s^{-1}$, with 1 minute intervals on ice. Cell debris was pelleted by centrifugation at 13000 xg, and the supernatant retained as the whole cell lysate fraction (WCL). When appropriate, protease inhibitors (Calbiochem Protease Inhibitor Cocktail set III, Merck) and DNase I (NEB) were added prior to disruption and WCL fractions were filtered using a 0.22 µm syringe filter (Millipore) to remove unlysed cells and to sterilise samples. WCL samples were stored at -20°C or -80°C.

3.1.2. Culture filtrate (CF) sample preparation

The culture supernatant was sequentially filtered through 0.45 µm and 0.22 µm syringe filters (Millipore) to remove any remaining cells, to produce the culture filtrate fraction. CF samples were stored at -20°C or -80°C.

3.2. Protein quantification

The protein concentrations of samples were determined spectrophotometrically using the Bio-Rad Protein Assay with either the 1 x Quick Start™ Bradford Dye Reagent or the RC DC™ Protein Assay. The Quick Start™ Bovine Serum Albumin Standard Set (BSA concentration range 0.125 mg/ml – 2 mg/ml) was used to generate a standard curve.

The assay using the 1 x Quick Start™ Bradford Dye Reagent was done according to standard microtiter plate procedure, as described by the manufacturer. Protein samples were measured in undiluted, 1:10 and 1:100 dilutions, in duplicate. Standards were measured in triplicate. Ten microliters of each sample and standard were pipetted into separate microtiter wells. Two hundred microliters of

Dye Reagent was added to each well, and mixed. The samples were incubated at room temperature for 5 minutes. Absorbance was measured at 595 nm (read within 1 hour).

The RC DC™ Protein Assay was done according to manufacturer's instructions; using undiluted and 5x diluted protein samples, as appropriate. Proteins were precipitated with 125 µl of the RC Reagents I and II, redisolved in 127 µl DC Reagent A', and the concentrations read at 595nm after a 15 minute incubation with 1 ml DC Reagent B (read within 1 hour).

Standard curves were plotted from the BSA standard absorbance readings, and sample concentrations determined from the standard curve.

3.3. SDS-PAGE

3.3.1. Sample preparation

Laemmli Buffer was added to each protein sample to a final concentration of 60 mM Tris-Cl pH 6.8, 2% SDS, 10% glycerol, 5% β-mercaptoethanol, 0.01% bromophenol blue. The samples were mixed and incubated at 95°C for 5 minutes.

3.3.2. Precast gradient gel SDS-PAGE

Criterion™ XT Precast Gels (4-12% Bis-Tris, 1.0 mm, 12 well, 45 µl, BIO-RAD) were used for high resolution separation of complex protein samples in XT MOPS running buffer (BIO-RAD). Samples were electrophoresed at 150 V until the front reached the bottom of the gel. PageRuler™ Prestained Protein Ladder (Thermo Scientific) was run alongside the samples as an indication of size and separation.

3.3.3. Protein gel staining

Protein gels were stained with Imperial™ Protein Stain (Thermo Scientific), for 2 hours with gentle shaking. Gels were destained overnight in deionised water, with gently shaking.

3.3.4. Dot Blotting

Protein samples were spotted on Hybond-P PVDF membrane and allowed to dry. The membrane was blocked in blocking buffer (Table 5) for 1 hour at room temperature, with gentle shaking. The blocked

membrane was rinsed twice in Tris-buffered saline with Tweeen-20 (TBS-T, Table 5) and washed 3 times for 10 minutes in TBS-T with shaking. Five milliliters of primary antibody, at the appropriate dilution, was added to the membrane and shaken at room temperature for 1 hour. The membrane was rinsed twice in TBS-T, and washed 3 times for 10 minutes in TBS-T with shaking. The membrane was then probed with 5 ml secondary antibody (at the appropriate dilution) for 1 hour at room temperature, washed as before and placed in a sealable plastic bag. The membrane was incubated in ECL detection fluid (ECL Plus Western Blotting Detection System, AEC Amersham™) according to manufacturers instructions. The detection fluid was removed, the membrane attached to a piece of Whatmann paper and placed in a cassette. Signal was detected for 1 minute to 1 hour with autoradiographic film.

Table 5. Dot blotting buffers

Buffer	Composition	Diluent	Storage
TBS-T	20 mM Tris (pH 7.6) 137 mM NaCl 0.1% Tween-20	Water	RT
Blocking buffer	10% m/v fat free milk powder 1% m/v Bovine serum albumin 10 mM Azide	TBS-T	RT

4. *In silico* and phylogenetic analyses

4.1. DNA and protein sequences

All protein and DNA sequence information was obtained from publicly available finished and unfinished genome sequencing information. The species and their associated URL's, sequencing source and level of completion are given in Table 6.

4.2. Comparative genomic analyses of the ESX gene clusters

M. tuberculosis ESX protein sequences of interest were used as templates to identify orthologous ESX protein and gene sequences. Blast similarity searches, blastn, tblastn and blastp,[130] were done using NCBI Blast and the genome sequence databases listed in Table 6. Where possible adjacent genomic regions were searched for additional ESX genes to determine clustering and arrangement of genes; for unfinished genomes in contig format this was not always possible and gene cluster arrangement was assumed based on sequence similarity and anticipated arrangement. Large intergenic regions were searched for gene insertions using blastx analyses.[131]

Table 6. Mycobacterial and closely related actinomycete genome sequencing information

Species	Website	Sequencing centre
M. abscessus CIP 104536T	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	Genoscope
M. africanum GM041182	http://www.sanger.ac.uk/cgi-bin/blast/submitblast/mycobacterium	Wellcome Trust Sanger Insititute
M. avium 104	http://cmr.jcvi.org/cgi-bin/CMR/GenomePage.cgi?org=gma	TIGR
M. avium paratuberculosis K-10	http://cmr.jcvi.org/cgi-bin/CMR/GenomePage.cgi?org=ntma03	University of Minnesota
M. bovis AF2122/97	http://genolist.pasteur.fr/BoviList/	Sanger Institut/Pasteur Institut
M. bovis BCG Pasteur 1173P2	http://genolist.pasteur.fr/BCGList/	Pasteur Institut
M. canetti CIPT140010059	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	Wellcome Trust Sanger Insititute
M. chubuense	http://blast.ncbi.nlm.nih.gov/Blast.cgi?	DOE Joint Genome Institute
**M. colombiense* CECT3035	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	NCBI
M. gilvum PYR-GCK	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	US DOE Joint Genome Institute
M. intracellulare	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	Department of Microbiology and Immunology, Seoul National University
**M. kansasii* 12479	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	McGill University
M. leprae TN	http://genolist.pasteur.fr/Leproma/	Sanger Institut/Pasteur Institut
M. marinum M	http://genolist.pasteur.fr/MarinoList/	Wellcome Trust Sanger Institute
**M. massiliense* CCUG48898	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	Chungnam National University
**M. microti* 19422	http://www.sanger.ac.uk/cgi-bin/blast/submitblast/mycobacterium	Sanger Institut
**M. parascrofulaceum* BAA-614	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	Baylor College of Medicine
**M. phlei* RIVM601170	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	KAUST
**M. rhodesiae* NBB3	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	JGI
M. smegmatis mc2155	http://cmr.jcvi.org/cgi-bin/CMR/GenomePage.cgi?org=gms	TIGR
M. sp. JDM601	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	Shanghai Jiao Tong University School of Medicine
M. sp. JLS	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	US DOE Joint Genome Institute
M. sp. KMS	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	DOE Joint Genome Institute
M. sp. MCS	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	DOE Joint Genome Institute
M. sp. MOTT36Y	http://blast.ncbi.nlm.nih.gov/Blast.cgi?	Seoul National University College of Medicine
M. sp. Spyr 1	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	US DOE Joint Genome Institute
M. thermoresistibile 19527	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	NCBI
M. tuberculosis H37Rv	http://genolist.pasteur.fr/TubercuList/	Sanger Institut/Pasteur Institut
**M. tusciae* JS617	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	DOE Joint Genome Institute; JGI-PSF
M. ulcerans Agy99	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	Pasteur Institut
M. vanbaalenii PYR-1	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	DOE Joint Genome Institute
**M. xenopi* RIVM700367	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	Center for Biotechnology, Bielefeld University
N. farcinica IFM 10152	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	Department of Bioactive Molecules, National Institute of Infectious Diseases
R. opacus B4	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	National Institute of Technology and Evaluation
R. erythropolis PR4	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	National Institute of Technology and Evaluation
R. equi 103S	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	International Rhodococcus equi Genome Consortium
G. bronchialis DSM43247	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	US DOE Joint Genome Institute (JGI-PGF)
C. diphtheriae NCTC 13129	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	Sanger Institut
C. pseudotuberculosis	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	Center for Biotechnology, Bielefeld University
B. subtilis str. 168	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	NCBI, Baylor College of Medicine, Human Genome Sequencing Center, Baylor College of Medicine
L. monocytogenes L312	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	NCBI, ERA-NET / Spatelis, Institute for Microbiology and Genetics, Georg-August University Goettingen, Justus-Liebig-University Giessen
S. aureus USA300	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi	NISC

*****Sequencing projects are incomplete.

4.3. Phylogenetic analyses of the ESX gene clusters

Annotated protein sequences were obtained from the protein sequence databases. The protein sequences of conserved components of each ESX gene cluster (EccA, EccB, EccC, EccD, EccE, PE(s), PPE(s), Esx (CFP-10-like), ESX (ESAT-6-like), EspG, EspI, MycP, Rv3446c, EspJ, EspK, EspL, EspB, Cyp143 and Ferredoxin) were combined and multiple sequence alignments of all ESX gene cluster combined protein sequences were done with Clustal W 2.0 [132,133] using the Bioedit Sequence Alignment Editor version 7.1.3.0.[134] Similarly, multiple sequence alignments of a single sequence, composed of all of the combined ESX gene cluster protein sequences (ESX-4, -3, -1, -2, -5 and any additional ESX gene clusters), from each species were done. The PAUP 4.0b10 neighbor-joining algorithm was used for phylogenetic analyses using the combined protein multiple sequence alignments.[135] One thousand subsets were generated for bootstrapping resampling of the data. Confidence intervals of greater than 50% for the internal topology of the trees were considered significant. The *C. diphtheria* and *L. monocytogenes* protein sequences were used as outgroups. Consensus trees were calculated using the majority rule formula and phylograms were generated using Treeview 1.6.6.[136]

5. Promoter assays

5.1. Selection of intergenic regions

Intergenic regions of the *M. tuberculosis* ESAT-6 gene cluster regions 1 to 5 were selected *in silico* from the *M. tuberculosis* H37Rv genome sequence based on size. Intergenic regions of less than 40bp were excluded from the study due to a lower probability of containing a promoter sequence.

5.2. DNA manipulation

The pJEM15 ESX intergenic region-containing promoter probe constructs were obtained from previous studies (Table 7). The intergenic regions were amplified using HotStarTaq (Qiagen) and cloned into the mycobacterial-*E.coli* promoter-probe shuttle vector, pJEM15,[128] a gift from J. Rauzier Institut Pasteur, Paris, France, to create transcriptional fusions of the intergenic regions with a promoterless *lacZ* operon. All constructs were confirmed by DNA sequencing.

Table 7. pJEM15 promoter probe constructs, and relevant primers, used in the identification of the ESX gene cluster promoters.

	Construct	Name of primer	Primer sequence (5' - 3')	Primer length	Tm (ºC)	Intergenic region covered	Product length	Construct generated by
ESX-4	pJem34478	Rv34478f	GGATCCACCAGGATGTCCGCAGTCG	25 bp	64.9	From 41bp inside Rv3447c to Rv3448	154 bp	M. Newton-Foot
		Rv34478r	GGTACCTCTCCCCCTTTATTCTCGGAT	27 bp	60.8			
	pJem34487	Rv34487f	GGATCCTCCCGGATCAGACGTAGGC	25 bp	64	From 21bp inside Rv3448 to Rv3447c	160 bp	
		Rv34487r	GGTACCGCTGGGCGACGGTATGCAG	25 bp	67.1			
	pJem34510	Rv34510f	GGATCCATGTCAGCAGGCGGATGG	24 bp	63.9	From 31bp inside Rv3451 to Rv3450c	180 bp	
		Rv34510r	GGTACCTGGCTCTCCCACGGTGGC	24 bp	67.4			
ESX-3	pJem8182	T028182f	GTTCGCCCGCAACACCCT	18 bp	60	From 40 bp inside Rv0281 to 18 bp inside Rv0282	281 bp	N.C. Gey van Pittius
		T028182r	TTCACCTACGCCCGCCAT	18 bp	58			
	pJem8687	T028687f	GACCGCAACCAAAGAACGC	19 bp	60	From 184 bp inside Rv0286 to 40 bp inside Rv0287	273 bp	
		T028687r	GAGGCCACCAACTGTGGGATA	21 bp	66			
	pJem8990	T028990f	TGGGTCTCCACCTTCAGCC	19 bp	62	From 135 bp inside Rv0289 to 68 bp inside Rv0290	249 bp	
		T028990r	GCCATCTCGGTCAACCTGCT	20 bp	64			
ESX-1	pJem 6364	386364f	TCGAAGCACTGATCCGTCAGATCAA	25 bp	74	From 28bp inside Rv3863 to Rv3864	272 bp	J. Botha
		386364r	TTTATGTGTTTCCTTACGCTCGCCG	25 bp	74			
	pJem6465	386465f	GACAAGAAGGAAGACGAGGAAGGCG	25 bp	78	From 101 bp inside Rv3864 to Rv3865	190 bp	
		386465r	TGGGTTGAACCCTCCTCTACGTTTA	25 bp	74			
	pJem7172	387172f	GTTAAGATTATTTCATTGCCGG	22 bp	60	From 4bp inside Rv3871 to 2bp inside Rv3872	149 bp	
		387172r	ATTTGCCTACTTCTTCCCTCC	21 bp	62			
	pJem7374	T387374f	GCAGGAGCGTGAAGAAGAC	19 bp	60	From 55 bp inside Rv3873 to 94 bp inside Rv3874	241 bp	N.C. Gey van Pittius
		T387374r	CCTGGTCGATCTGGGTTTT	19 bp	58			
	pJerm7576	387576f	TCACTGGGATGTTCGCATAGGGC	23 bp	72	From 19bp inside Rv3875 to Rv3876	134 bp	J. Botha
		387576r	TGACAACCTCTCAGAGTGCGCTCAA	25 bp	76			
	pJem8483	388483f	GTTTGGTAGCAGCCGCATAT	20 bp	60	From 8bp inside Rv3884c to Rv3883c	231 bp	
		388483r	CTCCCAGAACACTCCATTCGTT	22 bp	66			

Table 7. *continued*

	Construct	Name of primer	Primer sequence (5' - 3')	Primer length	Tm (ºC)	Intergenic region covered	Product length	Construct generated by
ESX-2	pJem38854	Rv38854f	GGATCCCCGCTATGTCAGCCTTGA	24 bp	62.6	From 46bp inside Rv3885c to 22bp inside Rv3884c	137 bp	M. Newton-Foot
		Rv38854r	GGTACCTCGTGTCCACCATTCTCG	24 bp	60.9			
	pJem38898	Rv38898f	GGATCCCCGGCGTGTGGTGTGAAT	24 bp	65.5	From 16bp inside Rv3889c to 22bp inside Rv3888c	131 bp	
		Rv38898r	GGTACCGGTCATTCCACGGGTTCG	24 bp	63.5			
	pJem389089	Rv389089f	GGATCCCCATCGCGGGCTTGTTCT	24 bp	65.4	From 20bp inside Rv3890c to 22bp inside Rv3889c	137 bp	
		Rv389089r	GGTACCCATCGACCGTCGTCGTCA	24 bp	64.1			
	pJemP_{AN}	Panf	CGCGACGGTGCGTGTTGT	18 bp	60	From 16bp inside Rv3892c to 34bp inside Rv3891c	178 bp	J. Botha
		Panr	CATCTGCGGTGTTACCTGAA	18 bp	60			
	pJem38932	Rv38932f	GGATCCTTGCCGAACACGCCAGTC	24 bp	64.1	From 41bp inside Rv3893c to Rv3892c	119 bp	M. Newton-Foot
		Rv38932r	GGTACCGCCTACTTACCCTTGCAGC	25 bp	63.1			
	pJem39843	Rv38943f	GGATCCAGGGGATTTTGGTGGGGTAT	26 bp	62.3	From 44bp inside Rv3894c to Rv3893c	310 bp	
		Rv38943r	GGTACCGTTCGCCCCTCCGATTGCTT	26 bp	66.3			
	pJem38976	Rv38976f	GGATCCAGGTTAAGGAATAGTCGCCCTG	28 bp	61.8	From 7bp inside Rv3897c to 23bp inside Rv3896c	186 bp	
		Rv38976r	GGTACCCCAATCCTGTGCCATGTGC	25 bp	64			
ESX-5	pJem178182	Rv1781-82f	GGATCCTCGGGCAAGTTCGACCAG	24 bp	58	From 100bp inside Rv1781 to 100bp inside Rv1782	265 bp	R.G. van der Merwe
		Rv1781-82r	GGTACCGCCACGCTAGATCACCT	23 bp	54			
	pJem178384	Rv1783-84f	GGATCCCAGCGGGTTGCCCACATT	24 bp	58	From 100bp inside Rv1783 to 100bp inside Rv1784	71 bp	
		Rv1783-84r	GGTACCATGGTGTCGGTCCGGCTA	24 bp	58			
	pJem178687	Rv1786-87f	GGATCCGAAATGGCGCTCATCCT	23 bp	52	From 100bp inside Rv1786 to 100bp inside Rv1787	270 bp	
		Rv1786-87r	GGTACCCAACCAGTCCTCCCTCTC	24 bp	58			
	pJem178788	Rv1787-88f	GGATCCCGCTAGGGTCGCGGATC	23 bp	58	From 100bp inside Rv1787 to 100bp inside Rv1788	79 bp	
		Rv1787-88r	GGTACCCATGCGATCTCCTGCTTA	24 bp	54			
	pJem178990	Rv1789-90f	GGATCCATGGCTCGCCCACCCTTC	24 bp	60	From 100bp inside Rv1789 to 100bp inside Rv1790	454 bp	
		Rv1789-90r	GGTACCGTCCATTCCGAACCCTTT	24 bp	54			
	pJem179091	Rv1790-91f	GGATCCGTCTCCGTCGGCGGGATA	24 bp	54	From 100bp inside Rv1790 to 100bp inside Rv1791	427 bp	
		Rv1790-91r	GGTACCGACATGCTGTCTTCCTCT	24 bp	54			
	pJem179192	Rv1791-92f	GGATCCGAGGCGGCCAACGCAG	22 bp	56	From 100bp inside Rv1791 to 100bp inside Rv1792	144 bp	
		Rv1791-92r	GGTACCCATGTTGCCTGTCTCCTTA	25 bp	58			
	pJem179394	Rv1793-94f	GGATCCACAGCGCCGTCGGCTCCA	24 bp	62	From 100bp inside Rv1793 to 100bp inside Rv1794	88 bp	
		Rv1793-94r	GGTACCGATCCATCGCTACCTCAGC	25 bp	60			
	pJem179495	Rv1794-95f	GGATCCAGTGGAAAACACACAGCA	24 bp	52	From 100bp inside Rv1794 to 100bp inside Rv1795	272 bp	
		Rv1794-95r	GGTACCCATCGTCCTGTACCCCCT	24 bp	58			

The amplicons of the ESX intergenic regions were cloned into pJEM15 via one of two methods. The primers, intergenic regions cloned and construct details are given in Table 7. Intergenic regions from ESX-3, ESX-1 and the P_{AN} promoter from ESX-2 were cloned via a 3-step cloning method. Briefly, the selected intergenic regions were amplified and cloned into pGEM-T Easy (Promega). The inserts were digested out of pGEM-T Easy at the vector specific *Eco*RI restriction sites and subcloned into the *Eco*RI site in pBluescript II KS (Stratagene). The *Bam*HI / *Kpn*I fragments were excised from pBluescript II KS and cloned into the corresponding sites in the mycobacterial-*E.coli* promoter-probe shuttle vector pJEM15. The intergenic regions from ESX-4, ESX-2 (excluding the P_{AN} promoter) and ESX-5 were cloned via a 2-step protocol which omitted pBluescript II KS by introducing a *Bam*HI and *Kpn*I into the specific PCR primers. Inserts were excised from pGEM-T Easy using *Bam*HI and *Kpn*I and subcloned directly into compatible sites in pJEM15.

5.3. β-galactosidase activity

M. smegmatis and *M. bovis* BCG were electrotransformed with the intergenic region-containing pJEM15 constructs, and selected on Difco Seven H11 Agar base containing kanamycin (25 μg/ml). Single colonies were picked and grown in Middlebrook 7H9 medium with kanamycin, washed and inoculated into 10 ml Kirchener's broth and grown to mid log phase (OD_{600} = 0.6-0.8). The cells were pelleted and lysed in 500 μl PBS, as described in section 3.1.1. Protein concentrations were determined using the 1 x Quick Start™ Bradford Dye Reagent (section 3.2). β-galactosidase activity was quantitatively assayed in liquid medium according to the method described by Pardee *et. al.* (1959)[137] and Timm *et. al.* (1994)[128], adapted for microtitre plates. Equal quantities of protein (1 μg) were diluted to 50 ul in PBS and incubated with *o*-nitrophenyl-β-D-galactoside (ONPG) in 2X Assay Buffer (200 mM Sodium phosphate, pH7.3; 2 mM $MgCl_2$, 100 mM β-mercaptoethanol, 1.33 mg/ml ONPG) at 37°C for 30 minutes. The reaction was stopped by adding 150 ul of 1 M sodium bicarbonate. β-galactosidase activity was measured spectrophotometrically at 420nm. One unit of β-galactosidase is defined as producing 1 μMole o-nitrophenol per minute from o-nitrophenyl-β-D-galactoside at 28°C, pH 7.0. (1 μMole/ml o-nitrophenol has an optical density at 420 nm of 0.0075). β-galactosidase units were calculated as 200 x OD_{420} /mg of protein/minute, where OD_{420} was corrected by subtracting 1.65 x OD_{550}. The assays were performed triplicate and duplicate in *M. smegmatis* and *M. bovis* BCG, respectively.

5.4. Identification of promoter sequence motifs

The Berkeley Drosophila Genome Project (BDGP) Neural Network Promoter Prediction Search (http://www.fruitfly.org/seq_tools/promoter.html[138]) and the Softberry BPROM Promoter Predictor (http://linux1.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb) were used to screen the DNA sequences of the intergenic regions which showed promoter activity for possible promoter elements. Multiple sequence alignments of promoter-containing intergenic regions were done with the program ClustalW 2.0 on the Clustal W WWW server at the European Bioinformatics Institute website, http://www.ebi.ac.uk/Tools/msa/clustalw2/,[133] to identify sequence motifs which are conserved across different mycobacterial species.

5.5. Divalent metal cation regulation of the ESX-3 promoter

For the investigation of ESX-3 promoter regulation by divalent metal cations, the *M. smegmatis* ESX-3 promoter region was amplified from genomic DNA (f primer: GGATCCGACGCTGAACGAGTG and r primer: GGTACCCTGACCACATCGCGG; Tm = 62) and cloned into the *BamH*I and *Kpn*I restriction sites in pJem15 to produce pJemESX-3_{sm}, using the two-step process described above. The *M. tuberculosis* and *M. smegmatis* ESX-3 pJem promoter constructs, pJem8182 and pJemESX-3_{sm}, were transformed into *M. smegmatis* and pJem8182 additionally into *M. bovis* BCG. Cultures were grown in Kirchener's medium which was treated with Chelex-100 (Biorad) overnight,[50] to chelate any divalent metal cations, filtered and supplemented with $MgS0_4$ (1g/L). The culture medium was supplemented with individual metal cations: $ZnSO_4$ (100µM), Ferric ammonium sulphate (100µM), $ZnSO_4$ and Ferric ammonium sulphate (100µM ea), $CoCl_2$ (7.5µM), $NiCl_2$ (35µM), $CuCl_2$ (50uM), $Pb(CH_3COO)_2$ (10µM) and $MnSO_4$ (0.75µM).[139] Cultures were grown to OD_{600} = 0.6, whole cell lysates were prepared and β-galactosidase activity assayed as described above.

6. Knock-out construction

An ESX knock-out library was created in *M. smegmatis* using homologous recombination[140] to delete each ESX gene cluster present in this mycobacterium, either individually or in combination, as described in Table 8.

6.1. Culture conditions

For knock-out construction, *M. smegmatis* was grown in liquid LB with shaking, and on LB agar plates, for 4-5 days at 37°C (LB was used instead of Middlebrook media, as the sugars in Middlebrook 7H9 medium are toxic to single crossover knock-out cultures containing the counterselectable marker *sacB* gene). Liquid LB medium was supplemented with 0.1% Tween-80, and kanamycin (25 µg/ml, Sigma) when required. Solid LB medium was supplemented with kanamycin (25 µg/ml, Sigma), X-gal (Roche) and 5% sucrose, as required.

6.2. DNA manipulation

Briefly, 500-1000 bp of DNA directly upstream of (Upsmeg) and downstream (Downsmeg) of each ESX gene cluster was amplified from *M. smegmatis* mc^2155 genomic DNA using primers containing specific restriction sites (Table 9). The Upsmeg and Downsmeg fragment pairs were simultaneously cloned into the appropriate restriction sites of the p2NIL suicide vector via three-way cloning, to create p2NIL_UpDownSmeg. The *lacZ-sacB* marker gene cassette was digested out of pGOAL17 using *Pac*I and inserted into the *PacI* site of p2NIL_UpDownSmeg to create the suicide vectors p2NIL_ESX-4 KO, p2NIL_ESX-3 KO, p2NIL_ESX-1 KO.

6.3. Generation of ESX knock-out strains

The knock-out constructs were transformed into *M. smegmatis* mc^2155. Single recombinants were selected on kanamycin and X-gal, and passaged to allow a second recombination event to occur, double recombinants were counterselected on sucrose. Wild type and ESX knock-out *M. smegmatis* were differentiated by PCR using primers specific for either the wild type or knock-out strain (Table 10). The *M. smegmatis* ESX knockouts were confirmed by sequencing across the deleted region. This process was repeated using the single knock-out strains as parental strains to create double and triple *M. smegmatis* ESX knock-out strains.

Table 8. *M. smegmatis* ESX knock-outs

Parent strain	Deletion(s)	Specific region(s) deleted	Name used	Knockout construct made by	Knockout generated by
M. smegmatis	wild-type		WT_{ms}		
M. smegmatis	*ΔESX-4*	MSMEG_1534-MSMEG_1539	$\Delta ESX\text{-}4_{ms}$	M. Smit (MSc)	M. Smit (MSc)
M. smegmatis	*ΔESX-3*	MSMEG_0615-MSMEG_0626	$\Delta ESX\text{-}3_{ms}$	M. Newton-Foot (MSc)	M. Newton-Foot (MSc)
M. smegmatis	*ΔESX-1*	MSMEG_0057-MSMEG_0083	$\Delta ESX\text{-}1_{ms}$	N. Steyn (Hons)	N. Steyn (Hons)
M. smegmatis	*ΔESX-3; ΔESX-4*	MSMEG_0615-MSMEG_0626 MSMEG_1534-MSMEG_1539	$\Delta ESX\text{-}3\Delta ESX\text{-}4_{ms}$		This study
M. smegmatis	*ΔESX-1; ΔESX-4*	MSMEG_0056-MSMEG_0083 MSMEG_1534-MSMEG_1539	$\Delta ESX\text{-}1\Delta ESX\text{-}4_{ms}$		This study
M. smegmatis	*ΔESX-1; ΔESX-3*	MSMEG_0056-MSMEG_0083 MSMEG_0615-MSMEG_0626	$\Delta ESX\text{-}1\Delta ESX\text{-}3_{ms}$		This study
M. smegmatis	*ΔESX-1; ΔESX-3; ΔESX-4*	MSMEG_0056-MSMEG_0083 MSMEG_0615-MSMEG_0626 MSMEG_1534-MSMEG_1539	$\Delta ESX\text{-}1\Delta ESX\text{-}3\Delta ESX\text{-}4_{ms}$		This study

Table 9. Primers used to generate the *M. smegmatis* knock-out constructs

Knock-out construct	Primer name	Primer sequence	Length (bp)	Tm (°C)	Restriction site	WT Product	size (bp)
p2NIL_ESX-4 KO	UpstreamR4 f	GGGGGTCGACGACGAGGAAGAGGCG	25	62	*Sal*I	57bp upstream of MSMEG_1534	778
	UpstreamR4 r	GGGGTTCGAACCGTGCTGTGAACGAAACCC	30	60	*Sfu*I		
	DownstreamR4 f	GGGGTTCGAACGGACTGACTCGTTGGAGCG	30	64	*Sfu*I	100bp downstream of MSMEG_1539	747
	DownstreamR4 r	GGGGAAGCTTCCGCGCAGTCGCCCGT	26	62	*Hind*III		
p2NIL_ESX-3 KO	R3 Upsmeg f	GGGGTACCGGAGCATCCGCTGCAGACC	27	64	*Kpn*I	28bp upstream of MSMEG_0615	828
	R3 Upsmeg r	GGGGAGATCTCTCTCCCTTATGTATGCC	28	54	*Bgl*II		
	R3 Downsmeg f	GGGGAGATCTCGATCCCAGTGCTCCCACA	29	62	*Bgl*II	58bp downstream of MSMEG_0626	830
	R3 Downsmeg r	GGGGAAGCTTCCCGAGCGATCCTTTCC	27	56	*Hind*III		
p2NIL_ESX-1 KO	UpReg1smeg f	GGTACCAGTCGAACCTGGGCAAGCG	25	62	*Kpn*I	219 bp upstream of MSMEG_0056	824
	UpReg1smeg r	AGATCTTCGGCGTATTCGCAGCACC	25	62	*Bgl*II		
	DownReg1smeg f	AGATCTCCCCAACAAGAATTGATCCG	26	60	*Bgl*II	29bp upstream of MSMEG_0083	845
	DownReg1smeg r	AAGCTTTTGTTCAGCGAGCCCGAG	24	58	*Hind*III		

Restriction enzyme sites are underlined

Table 10. *M. smegmatis* ESX knock-out verification primers.

ESX deletion	Validation PCR	Primer name	Primer sequence	Length (bp)	Tm (°C)	Product size (bp)	Product description
ESX-4	R4 F1	R4 F1 f	GAGATCGCGATGGCCATCGCC	21	65	694	Amplifies from within Upsmeg R4 into MSMEG_1534 (only amplifies in WT and SCO)
		R4 F1 r	CGCGTAACCGGTTGCCGTGCC	21			
	R4 F2	R4 F2 f	CCAACGCCTGGGGCGTCTGAT	21	65	647	Amplifies from within Downsmeg R4 into MSMEG_1539 (only amplifies in WT and SCO)
		R4 F2 r	CTGCACGGCAACCTGTCGCAG	21			
	R4 F1F2	R4 F1 f	GAGATCGCGATGGCCATCGCC	21	65	1159	Amplifies across Upsmeg R4 and Downsmeg R4(only amplifies in ΔESX-4_{ms} and SCO)
		R4 F2 r	CTGCACGGCAACCTGTCGCAG	21			
ESX-3	R3 F1	R3 F1 f	GCAGTGGTTCTCCGAGCGTGG	21	60	798	Amplifies from within Upsmeg R3 into MSMEG_0615 (only amplifies in WT and SCO)
		R3 F1 r	ACGACGTCCGACCAGCGTTGG	21			
	R3 F2	R3 F2 f	CGCTGAACCCGTTGTCGACCG	21	60	550	Amplifies from within Downsmeg R4 into MSMEG_0626 (only amplifies in WT and SCO)
		R3 F2 r	GTTCGGTGATGGACCATCTGG	21			
	R3 KO	R3 KO f	TCCTTCTTTGCGCTGGTCTT	20	60	822	Amplifies across Upsmeg R3 and Downsmeg R3 (only amplifies in ΔESX-3_{ms} and SCO)
		R3 KO r	TGTCGCTGCCGTGGTTCT	18			
ESX-1	R1 F1	R1 F1 f	GAGGTGCTCGACAGCGAGGC	20	60	880	Amplifies from within Upsmeg R1 into MSMEG_0056 (only amplifies in WT and SCO)
		R1 F1 r	TCGCCAGCGGAATCAGTGCC	20			
	R1 F2	R1 F2 f	GTTCGGATCGTTGGGATCCTGA	22	60	920	Amplifies from within Downsmeg R1 into MSMEG_0083 (only amplifies in WT and SCO)
		R1 F2 r	GGTGTACCCGTATCCCTTTCCA	22			
	R4 F1F2	R1 F1 f	GAGGTGCTCGACAGCGAGGC	22	60	876	Amplifies across Upsmeg R1 and Downsmeg R1 (only amplifies in ΔESX-1_{ms} and SCO)
		R1 F2 r	GGTGTACCCGTATCCCTTTCCA	22			

7. Metabolomics

7.1. Bacterial Strains

The *M. smegmatis* mc^2155 wild-type (WT_{ms}), ΔESX-4_{ms} and ΔESX-3_{ms} strains were used for metabolomic analyses.

7.2. Culture conditions

WT_{ms}, ΔESX-4_{ms} and ΔESX-3_{ms} were cultured in 100 ml complete Middlebrook 7H9 medium with shaking, at 37 ºC to mid log (OD_{600} = 0.6-0.8). Samples were prepared for extraction as previously described by Jaki *et al.,* (2006).[141] In short, the cells were washed twice with phosphate buffered saline (PBS), collected via centrifugation, snap frozen in liquid nitrogen, lyophilized and stored at -80ºC until extraction and GC/MS metabolomic analysis. Twelve individually cultured samples were prepared for each strain for the ESX-3 metabolome study and six for the ESX-4 metabolome study.

7.3. GC-MS whole metabolome analysis

GC-MS whole metabolome analysis was done in collaboration with Prof. Loots at the Centre for Human Metabonomics, North-West University, South Africa. The metabolite extraction, GC-MS analysis, data processing and statistical analyses described in sections 7.3.1 to 7.3.4 were performed by Mr Reinart J. Meissner-Roloff in Prof. Loots' group.

7.3.1.Extraction method

Five milligrams of each lyophilized *M. smegmatis* sample was weighed into a microcentrifuge tube and extracted as previously described.[142] All reagents are described in Table 11. Briefly, 50 µl of phenyl butyric acid (0.104 µg/ml) was added as an internal standard, followed by chloroform, methanol and water, in the ratio 2.5:7.5:2.5. The sample was briefly vortexed and extracted using a vibration mill with a carbide tungsten bead (Retsch) for 5 minutes at 30 Hz.s-1. The sample was centrifuged for 5 min at 12 000 rpm and the supernatant transferred to a GC-MS sample vial and dried under a stream of nitrogen at room temperature. Samples were derivitised using 50 µl methoxyamine hydrochloride (containing 15 mg/ml pyridine) at 50 ºC for 90 minutes, followed by 50 µl MSTFA with 1% TMCS at 50

ºC for 60 minutes.[143] The samples were transferred to a 0.1 ml insert in a GC-MS sample vial and capped prior to GC-MS analysis.

Table 11. Reagents used for metabolomics analyses

Reagent	Abbreviation	Source
methoxamine hydrochloride		Sigma Aldrich
n-methyl-n-(trimethylsilyl) trifluoroacetamide	MSTFA	Sigma Aldrich
trimethylchlorosilane	TMCS	Sigma Aldrich
pyridine		Sigma Aldrich
3-phenylbutyric acid		Burdick and Jackson
chloroform		Burdick and Jackson
methanol		Burdick and Jackson
hexane		Burdick and Jackson

7.3.2.GC-MS analysis

An Agilent 7890A gas chromatograph coupled to an Agilent 5975B mass selective detector, equipped with a 7683B injector and 7683 autosampler, was used to analyze the extracted samples. A VF1-MS capillary column (30m x 250µm i.d., 0.25 µm film thickness) was used to achieve separation in the GC. The extracts were analyzed by injecting 1 µl of sample in splitless mode. The injector temperature was held at a constant 250 ºC for the total run time and the GC oven temperature was initially held at 50 ºC for 1 min after injection, followed by an increase in temperature of 10 ºC/min, until a temperature of 240 ºC was reached. This was followed by an increase of 20 ºC/min to a final temperature of 300 ºC, at which it was maintained for 7 min. The carrier gas used was helium and the flow rate was kept at a constant 1.2 ml/min. The MS detection was set to scan mode from m/z 50-559.

7.3.3.GC-MS data processing

GC-MS chromatograms were deconvoluted using the freely available AMDIS32-V2.65 (Automated Mass Spectral Deconvolution System) software, where automatic curve resolution was applied to the chromatograms. Raw GC-MS data was analyzed at a threshold of 0.01% of the total signal. Compound peaks were aligned across all samples analysed as previously described.[144] A data matrix was compiled containing the relative concentrations of all the aligned compound peaks present in each individual sample of each group (in order to normalise the data using the internal standard). This data matrix was used in the multi- and univariate statistical data analyses in order to identify the markers characterising the groups.

7.3.4. Statistical data analysis

Due to the complexity of the GC-MS data, data pretreatment using a non-parametric transformation[145] was necessary prior to statistical analysis to provide a balanced representation of all the metabolites present. The pretreated data was then processed using multivariate statistical analyses, in the form of a principle component analysis (PCA) and a partial least squares discriminate analysis (PLS-DA). PCA is considered an unsupervised learning method and is often used to summarise the data as principle components (PCs) and determine if natural separation between the sample groups exists, by plotting the principal component scores. The modeling power of each variable in the PCA was also determined and used to rank the variables in order of relevance. PLS-DA is a supervised pattern recognition method. The PLS-DA model can be used to predict the group membership of an unseen sample based on its metabolic profile. The VIP value of each variable (i.e., those metabolites that are important in the PLS-DA projection) was calculated and used to rank the metabolites. For each selected metabolite, a non-parametric Mann-Whitney U test was performed to access the univariate importance of the metabolite and a value of < 0.05 was considered significant. Effect sizes were also calculated for all selected metabolites.[146] Effect sizes provide information concerning the practical relevance of a statistically relevant result. An effect size of >0.5 is considered to be highly relevant, >0.3 potentially relevant and <0.1 irrelevant. All statistical analyses were performed using SPSS version 18, and the R statistical package version 2.13.

8. Proteomics

8.1. Bacterial Strains

The *M. smegmatis* mc^2155 WT_{ms}, $\Delta ESX\text{-}1_{ms}$, $\Delta ESX\text{-}3_{ms}$ and $\Delta ESX\text{-}4_{ms}$ strains were used for all proteomic analyses.

8.2. Sample preparation

8.2.1.Whole cell lysate (WCL) sample preparation

For whole cell lysate proteomic analyses, *M. smegmatis* strains were grown, in duplicate, in 50 ml Middlebrook 7H9 medium, in the absence of bovine serum albumin (bovine serum albumin was omitted from the ADC), with shaking, at 37°C to mid log (OD_{600}= 0.6-0.8). Cells were disrupted, as

described in section 3.1.1, in 600 µl Tris lysis buffer (5 mM Tris, pH 7.6, 0.005% Tween-80) with protease inhibitors and 10 µl DNase, filtered and stored at -80°C. Protein concentrations were determined using the Biorad RC DC™ Protein Assay.

8.2.2.Culture filtrate (CF) sample preparation

For culture filtrate proteomic analyses, *M. smegmatis* strains were grown in Middlebrook 7H9 medium, without bovine serum albumin and catalase to mid log (OD_{600} = 0.6-0.8), washed in the same 7H9 medium, without Tween-80, and reinoculated, in duplicate, in 200 ml Middlebrook 7H9 medium, without bovine serum albumin, catalase and Tween-80, to OD_{600} = 0.005. Cultures were grown with shaking, at 37°C to early log phase (OD_{600}= 0.3). Tween-80 was omitted to reduce leakage of intracellular proteins due to the effect of the detergent on the cell membrane structure. Catalase and BSA were omitted as the high concentrations of these proteins would mask the presence of the extracellular proteins. Culture filtrate fractions were prepared as described in section 3.1.2 and concentrated using Amicon Ultra Centrifugal Filters (Ultracel - 3K, Millipore) with a 3kDa molecular weight cut off. Concentrated samples were desalted with PBS (1000x dilution). Protein concentrations were determined using the Biorad RC DC™ Protein Assay and appropriate amounts of protein were lyophilised and resuspended in 30µl Laemmli buffer.

To determine the amount of cell lysis which occurred during culturing and centrifugation, serial dilutions of WCL and culture filtrate sample prepared from the same culture were dot blotted using a mouse IgG2a monoclonal anti-GroEL2 antibody (Colorado State University, 1:50) and an HRP-conjugated goat anti-mouse IgG polyclonal antibody (Santa Cruz Biotechnology, 1:5000). Lysed cells were determined to be between 10^{-3} to 10^{-4}% of the cells in the culture.

8.3. Sample fractionation

All protein samples were fractionated on Criterion™ XT 4-12% Precast Gels. Sixty micrograms of each WCL sample and 50 µg of each CF sample were fractionated in duplicate (each strain was analysed in technical duplicates of in biological duplicates, i.e. 4 replicates). Each lane (sample relicate) was cut into 10 or 6 fractions for WCL and CF, respectively.

8.4. MS sample preparation

All reagents used for MS sample preparation are described in Table 12.

Table 12. MS sample preparation buffers

Abbreviation	Compound	Source	Working concentration	Diluent
ABC	Ammonium bicarbonate	Sigma	50 mM 25 mM	LCMS Cromasolv water (Sigma) LCMS Cromasolv water (Sigma)
ACN	Acetonitrile	Sigma	50% 70% 100%	LCMS Cromasolv water (Sigma) LCMS Cromasolv water (Sigma)
DTT	DL-Dithiothreitol	Sigma	10 mM	25 mM ABC
IAA	Iodoacetamide	Sigma	55 mM	25 mM ABC
FA	Formic acid	Sigma	0.10%	70% ACN
	Trypsin	Promega	20 µg dissolved in 200 ul resuspension buffer (aliquot and store at -80C) dilute 20 ul resuspended trypsin with 180 ul 50mM ABC activate diluted trypsin at 30C for 15 minutes before use	

8.4.1. Drying of gel pieces

Each gel lane was excised and cut into fractions using a clean scalpel. Each gel fraction was cut into small pieces. The gel pieces were washed 3 times with 300 µl of water, for 10 minutes. Thereafter, the gel pieces were washed alternately with 300 µl of 50% ACN and 300 µl of 50 mM ABC, for 10 minutes each, until the blue colour was reduced. Three hundred microliters of 100% ACN was added to each fraction for about 10 minutes, until the gel pieces turned white. The ACN was removed and the gel pieces dried in a Concentrator*plus* (Eppendorf). Dried gel pieces can be stored at -20°C for 2 months.

8.4.2.Reduction, acetylation and trypsin digestion

One hundred and twenty microlitres of 10 mM DTT was added to the dried gel pieces and incubated for 1 hour at 56°C. The gel pieces were washed with 200 µl of 50 mM ABC and 200 µl 50% ACN for 1 minute each, followed by incubation with 120 µl 55 mM IAA for 60 minutes, in the dark. Gel pieces were washed with 300 µl of 50 mM ABC for 10 minutes and 300µl of 50% ACN for 20 minutes, and dried in the Concentrator*plus* (Eppendorf). Eighty microliters of activated trypsin was added to each gel fraction, and incubated at 37°C for 17 hours.

8.4.3.Peptide extraction

Eighty microlitres of 70% ACN, 0.1% FA was added to each fraction, for 30 minutes and the supernatant was removed to a clean microcentrifuge tube. This was repeated and 100% ACN was added to remove any additional peptide-containing water from the gel pieces. The peptides were dried in the Concentrator*plus* (Eppendorf) and stored at -20°C.

8.4.4.Stage tip cleanup

When an additional C18 stage tip cleanup step was included, the combined peptide-containing supernatant was reduced to 30 µl in the Concentrator*plus* (Eppendorf) and 20 µl of 5% formic acid was added to each sample. Stage tips were prepared by punching out disks of Empore^{TM} Octadecyl C18 Extraction disks (SUPELCO Analytical) and wedging them into a P200 pipette tip. The column was activated with 50 µl methanol and washed with 50 µl 5% formic acid, using a 2.5 ml COMBITIP PLUS (Eppendorf). The sample was run through the stage tip twice, and washed with 50 µl 5% formic acid. Sample peptides were eluted from the tip with 50 µl 80% ACN, 5% formic acid and dried in the Concentrator*plus* (Eppendorf).

8.5. LC-MS/MS analysis

LC-MS/MS analysis was done on the LTQ Orbitrap Velos (Thermo Scientific) in collaboration with Dr S. Smit at the MS Proteomics unit of CAF, Stellenbosch University, who performed the LC-MS/MS analysis, initial data acquisition, differential expression and statistical analyses.

Dried peptides were dissolved in 5% acetonitrile in 0.1% formic acid and 10 µl injections were made for nano-LC chromatography. All experiments were performed on a Thermo Scientific EASY-nLC II connected to a LTQ Orbitrap Velos mass spectrometer (Thermo Scientific, Bremen, Germany) equipped with a nano-electrospray source. For liquid chromatography, separation was performed on a EASY-Column (2 cm, ID 100µm, 5 µm, C18) pre-column followed by a EASY-column (10 cm, ID 75 µm, 3 µm, C18) column with a flow rate of 300 nl/min. The gradient used was from 5-17 % B in 5 min, 17-35% B in 90 min, 35-60% B in 10 min, 60-80% B in 5 min and kept at 80% B for 10 min. Solvent A was 100% water in 0.1 % formic acid, and solvent B was 100% acetonitrile in 0.1% formic acid.

The mass spectrometer was operated in data-dependent mode to automatically switch between Orbitrap-MS and LTQ-MS/MS acquisition. Data was acquired using the Xcaliber software package. The precursor ion scan MS spectra (m/z 400 – 2000) were acquired in the Orbitrap with resolution R = 60000 with the number of accumulated ions being 1×10^6. The 20 most intense ions were isolated and fragmented in linear ion trap (number of accumulated ions 1.5×10^4) using collision induced dissociation. The lock mass option (polydimethylcyclosiloxane; m/z 445.120025) enabled accurate mass measurement in both the MS and MS/MS modes. In data-dependent LC-MS/MS experiments, dynamic exclusion was used with 60 s exclusion duration. Mass spectrometry conditions were 1.8 kV, capillary temperature of 250 °C, with no sheath and auxiliary gas flow. The ion selection threshold was 500 counts for MS/MS and an activation Q-value of 0.25 and activation time of 10 ms were also applied for MS/MS.

8.6. Data Analysis

Thermo Proteome Discoverer 1.3 (Thermo Scientific, Bremen, Germany) was used to identify proteins via automated database searching (Mascot, Matrix Science, London, UK, and Sequest) of all tandem mass spectra against the *M. smegmatis* database obtained from TBDB (TB Database).[147] Carbamidomethyl cysteine was set as fixed modification, and oxidized methionine, N-acetylation, deamidation (NQ), Gln to pyro-Glu and Glu to pyro-Glu was used as variable modifications. The precursor mass tolerance was set to 20 ppm, and fragment mass tolerance set to 0.8 Da. Two missed tryptic cleavages were allowed. Proteins were considered positively identified when they were identified with at least 1 tryptic peptide per protein. Proteins were searched and quantified using the Andromeda function from Maxquant 1.2.2.5.[148] Relative protein concentrations were determined and a minimum two fold difference in protein abundance was considered significant. The T-test with a significance level of 0.05 was used to determine the statistical relevance of differential protein abundances. Protein functions and pathways were determined using Smegmalist[149] (http://mycobrowser.epfl.ch/smegmalist.html) and KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.kegg.jp/kegg/).

RESULTS AND DISCUSSION

1. The evolution of the mycobacterial Type VII ESX secretion systems

1.1. Introduction

The ESX gene clusters are found within all members of the genus *Mycobacterium* and have also been identified in the genomes of closely related actinomycetes outside of the genus.[27] Furthermore, clustered WXG (or Esx-like) and FtsK/SpoIIIE proteins have been identified in gram-positive monoderm genera such as *Bacillus*, *Listeria* and *Saccharomyces.*[39] Here we investigate the presence and absence of the ESX gene clusters in the genomes of the sequenced mycobacteria and other closely related actinomycetes, in comparison to identified WXG-FtsK clusters of other species, in order to determine the evolutionary history of the Type-VII ESX secretion systems.

1.2. Results

1.2.1. The ESX gene clusters of sequenced mycobacteria and other actinomycetes

Whole genome sequences of thirty-two mycobacterial species and seven additional actinomycetes were analysed in this study (Materials and Methods, Table 6). The protein sequences of the *M. tuberculosis* H37Rv ESX-1, 2, 3, 4 and 5 secretion system components were used to identify orthologs in the other species. Intergenic regions were searched to identify any additional gene insertions and the arrangement of each ESX gene cluster was determined and is represented in Figure 1.1. Each of the 39 species studied contains between 1 and 7 ESX gene clusters. The ESX gene clusters were named ESX-1 to ESX-5 based on protein sequence similarity and gene arrangement similarity to the clusters in *M. tuberculosis* H37Rv. Clusters which could not be classified as one of the five ESX gene clusters were named based on the plasmid or contig on which they occur. Three additional WXG-FtsK clusters, identified from the literature,[39, 150] from three separate species, *Staphylococcus aureus*, *Listeria monocytogenes, Bacillus subtilis*, are also included in Figure 1.1.

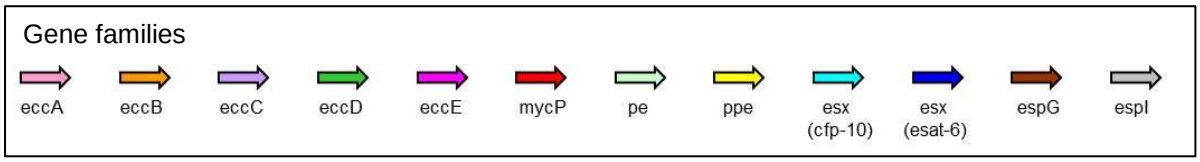
Orthologs of the ESX-4 gene cluster were identified in seven non-mycobacterial actinomycetes of the genera C*orynebacterium, Rhodococcus* and *Nocardia*. An additional ESX gene cluster which includes orthologs of *ecc*A, *ecc*E, *esp*G, *esp*I, *pe* and *ppe* genes, in addition to the ESX-4 components, was identified in *N. farcinica,* and named ESX-3_{AN} (ancestral ESX-3). The arrangement of the genes in this cluster does not resemble that of any of the *M. tuberculosis* ESX gene clusters. *M. abscessus*, the

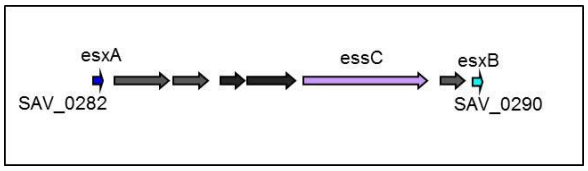
most ancestral mycobacterium studied, and *M. massiliense* contain ESX-4 and ESX-3, while all the other sequenced fast-growing mycobacteria, *M. thermoresistibile*, *M. smegmatis, M. vanbaalenii, M. gilvum, M. sp.* Spyr1, *M. rhodesiae, M. phlei* , *M. sp* JLS*, M. sp.* KMS and *M. sp* MCS, contain copies of ESX-4, ESX-3 and ESX-1, with the exception of *M. chubuense* which appears to have undergone deletions of ESX-3 and most of ESX-1. All of the studied slow-growing mycobacteria contain a copy of ESX-5, while some species appear to have lost their copies of ESX-1 (*M. avium, M. avium paratuberculosis, M. colombiense, M. intracellulare, M. parascrofulaceum, M. ulcerans, M. xenopi, M. sp.* MOTT36Y and *M. sp.* JDM601) and ESX-2 (*M. leprae, M. marinum* and *M. ulcerans*). *M. leprae*, with its reduced genome, has retained only ESX-3, ESX-1 and ESX-5. Partial deletions of ESX-1, previously described as RD1, occur in *M. bovis* BCG and *M. microti.* [33, 34, 79-81]

Additional ESX gene clusters were identified on plasmids within *M. gilvum* (pMFLV01), *M. sp. KMS* (pMKMS01 and pMKMS02), *M. sp. MCS* (Plasmid01) and *M. chubuense* (pMYCCH.01 and pMYCCH.02). pMKMS02 and *M. sp.* MCS Plasmid01 are the same plasmid occurring in both species, while the two ESX gene clusters on the two *M. chubuense* plasmids have the same gene arrangement. The four ESX gene clusters represented on these plasmids were named ESX-P1 (on pMFLV01), ESX-P2 (on pMKMS01), ESX-P3 (on pMKMS02 and *M. sp. MCS* Plasmid01) and ESX-P4 (on pMYCCH.01 and pMYCCH.02), where “P” indicates the plasmid localisation of the ESX. ESX-P1 to -P4 contain all of the core ESX components, including *esp*G and *esp*I. The ESX-P1 also incorporates EspH, while EccA is absent from ESX-P2 (Figure 1.1). The arrangement and composition of these gene clusters most closely resemble ESX-3 and ESX-1.

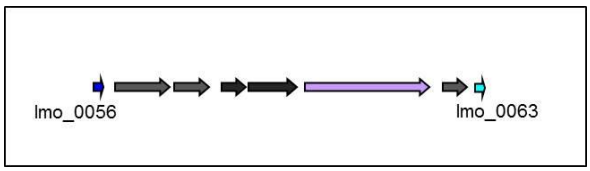
The incomplete genome sequences of *M. parascrofulaceum* and *M. tusciae* each contain 7 copies of the ESX gene cluster, four of which most closely resemble ESX-2 (Figure 1.1). The arrangement and sequences of specific protein components of the clusters vary, supporting their classification as independent gene clusters. Based on gene arrangement and sequence similarity, the *M. tusciae* ESX gene cluster on contig 212 was classified ESX-2_{AN} (ancestral ESX-2), those on contigs 196 and 209 as ESX-P2 and on contig 224 as ESX-P3. The *M. parascrofulaceum* ESX on contig 312/318 was defined as ESX-2, while those on contigs 115, 17 and 109 were classified as ESX-P2, ESX-P4 and ESX-P5 respectively. The *M. parascrofulaceum* and *M. tusciae* sequencing projects are incomplete,

Figure 1.1. The ESX gene clusters of sequenced mycobacteria and related actinomycetes. The WXG-FtsK gene clusters of *S. aureus*, *L. monocytogenes* and *B. subtilis* are as described in the literature. Gene families are represented as coloured arrows as indicated in the legend. ESX gene clusters identified, or predicted to be, on plasmids are named ESX-P1 to -P5. Pseudogenes are indicated as striped arrows. Insertions are indicated. Large insertions or genome rearrangements resulting in genes occurring elsewhere in the genome from the rest of the gene cluster are indicated as transpositions. Genes which are not annotated but for which translated nucleotide results were obtained using tblastn are indicated. The genome sequence of *M. microti* has not been annotated; therefore no gene names are given.

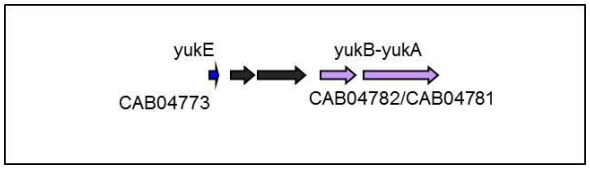




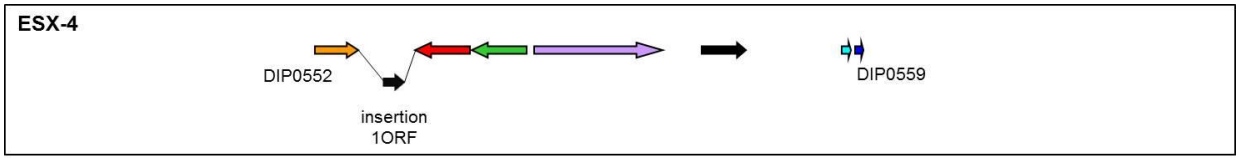
S. aureus



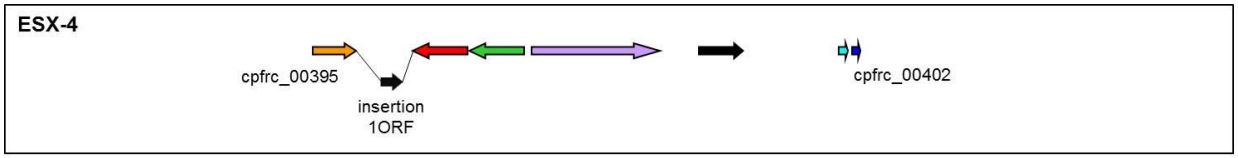
L. monocytogenes



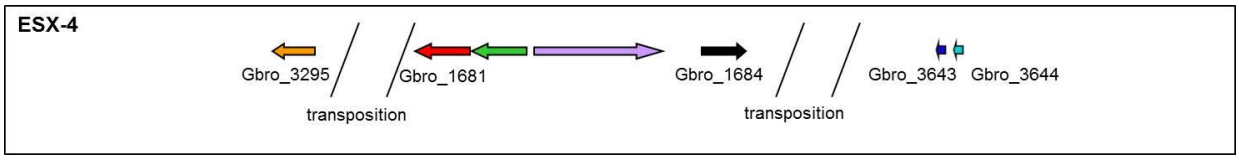
B. subtilis



C. diphtheriae

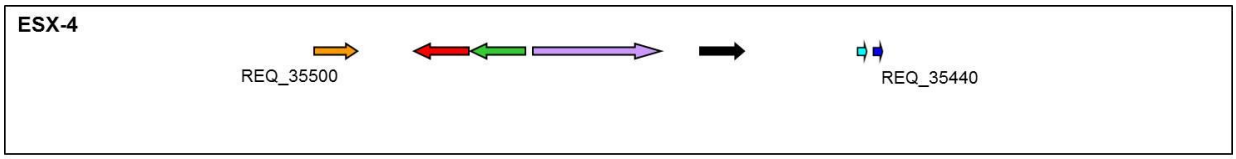


C. pseudotuberculosis

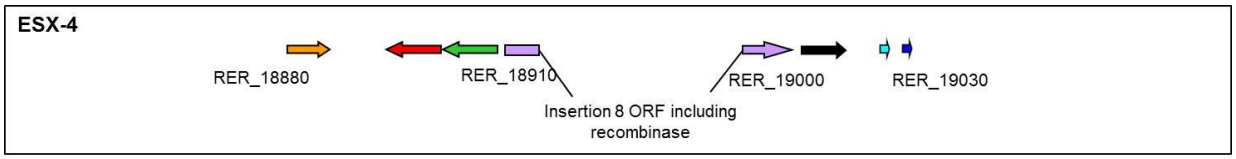


G. bronchialis

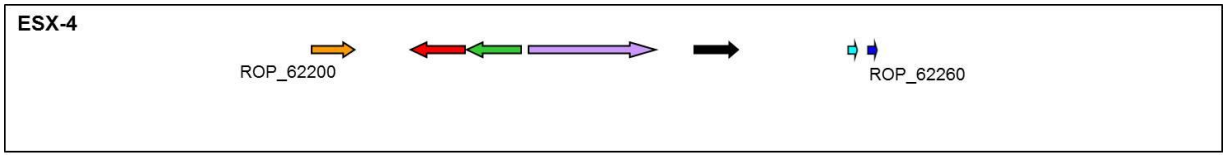
Figure 1.1. *continued*



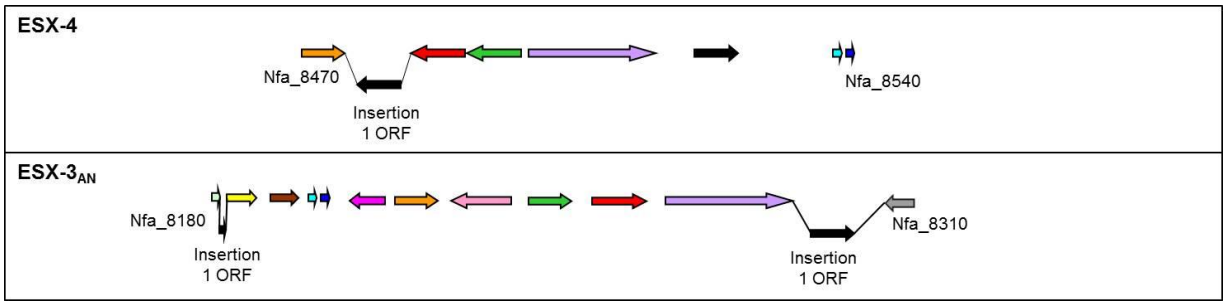
R. equi



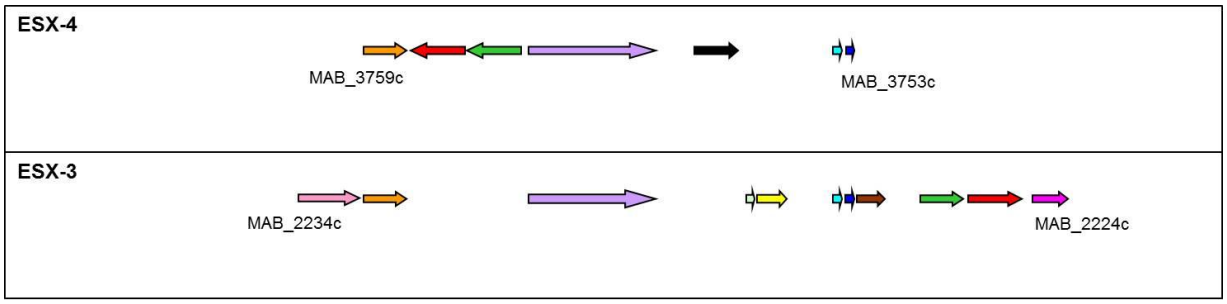
R. erythropolis



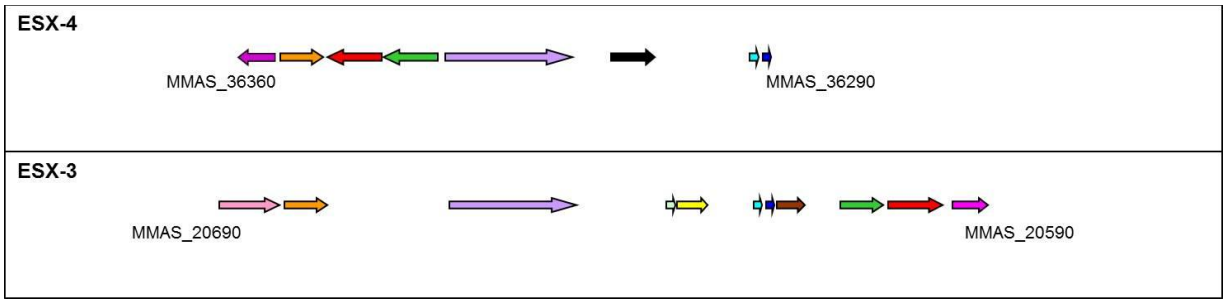
R. opacus



N. farcinica

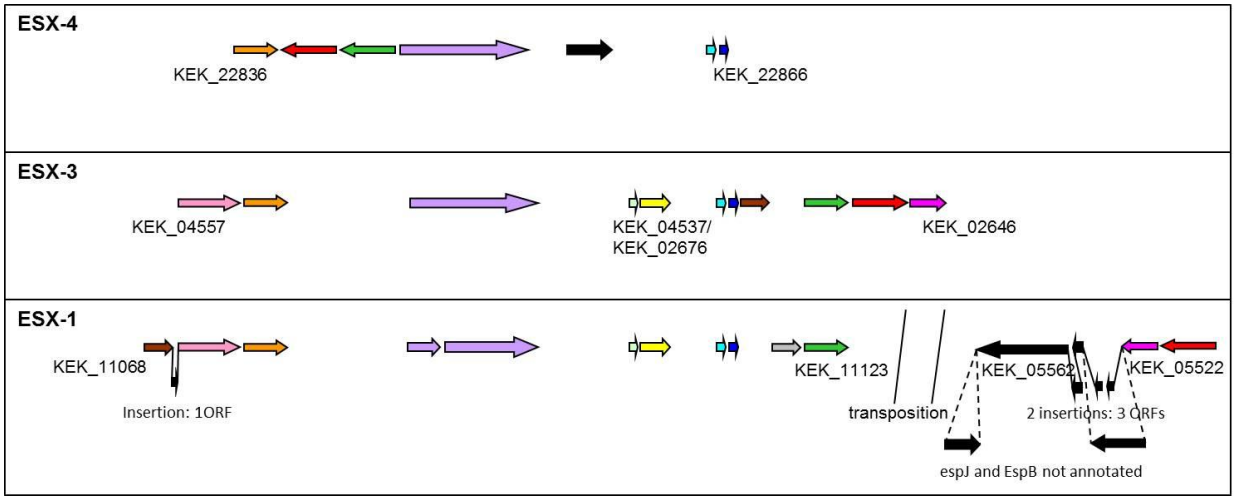


M. abscessus

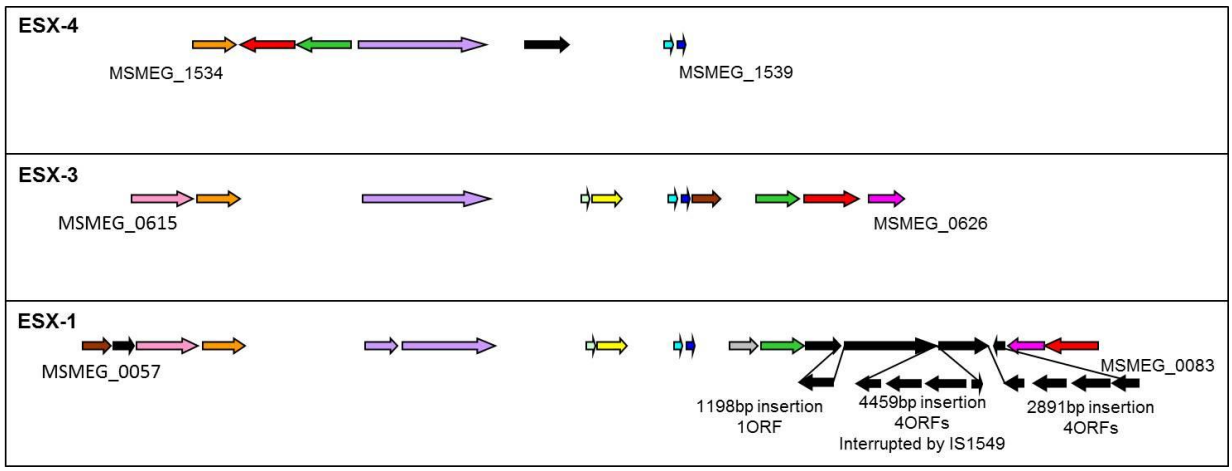


M. massiliense

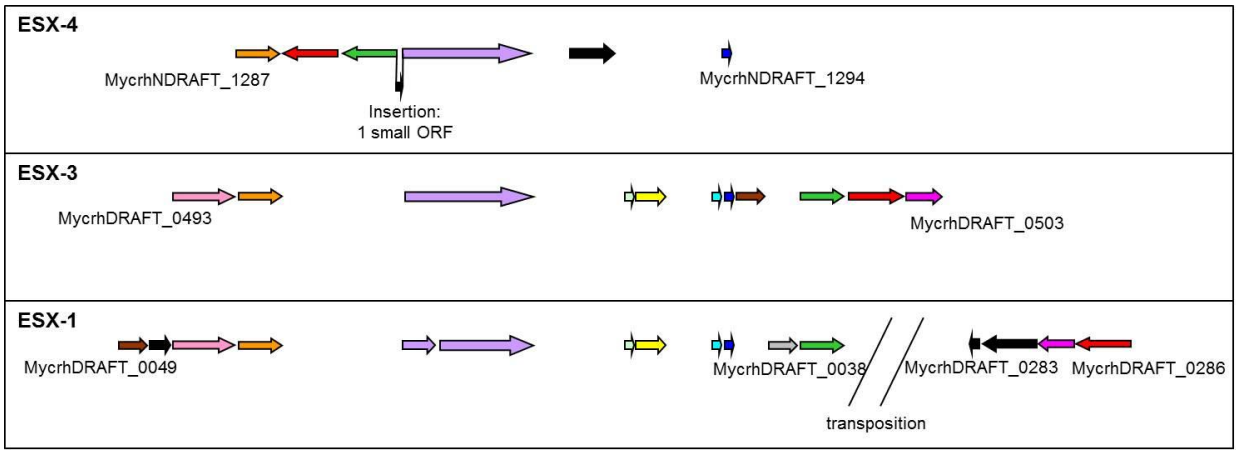
Figure 1.1. *continued*



M. thermoresistibile

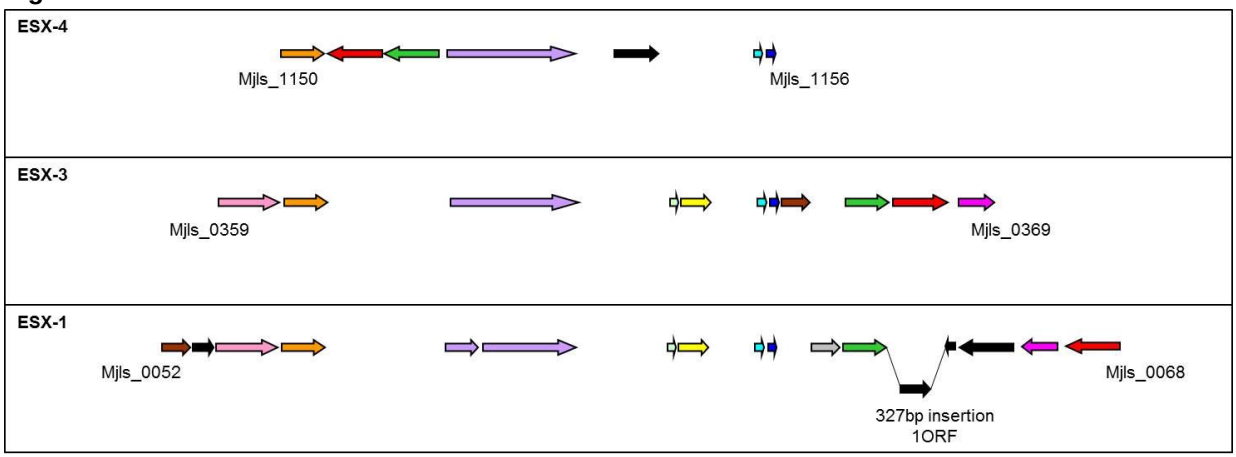


M. smegmatis

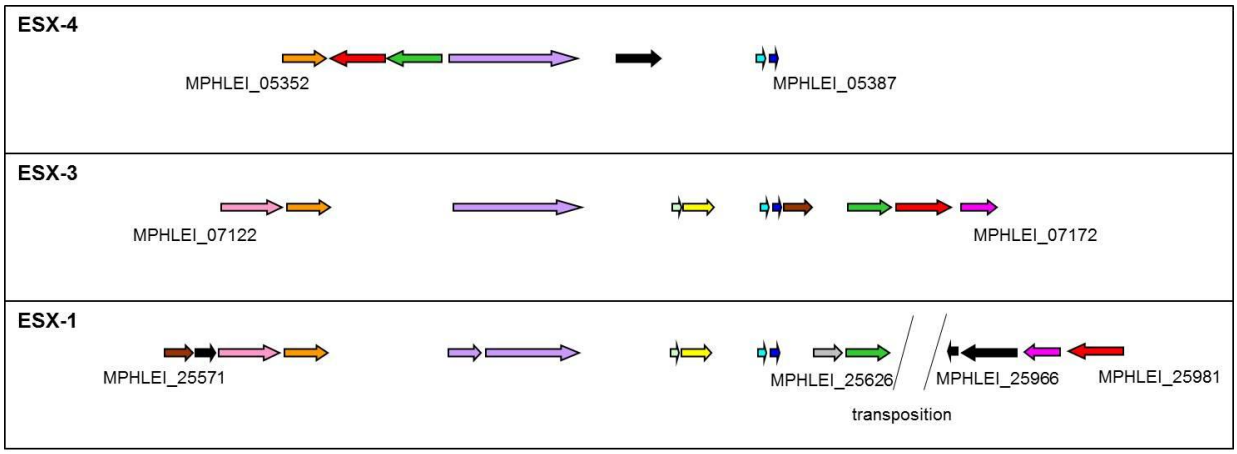


M. rhodesiae

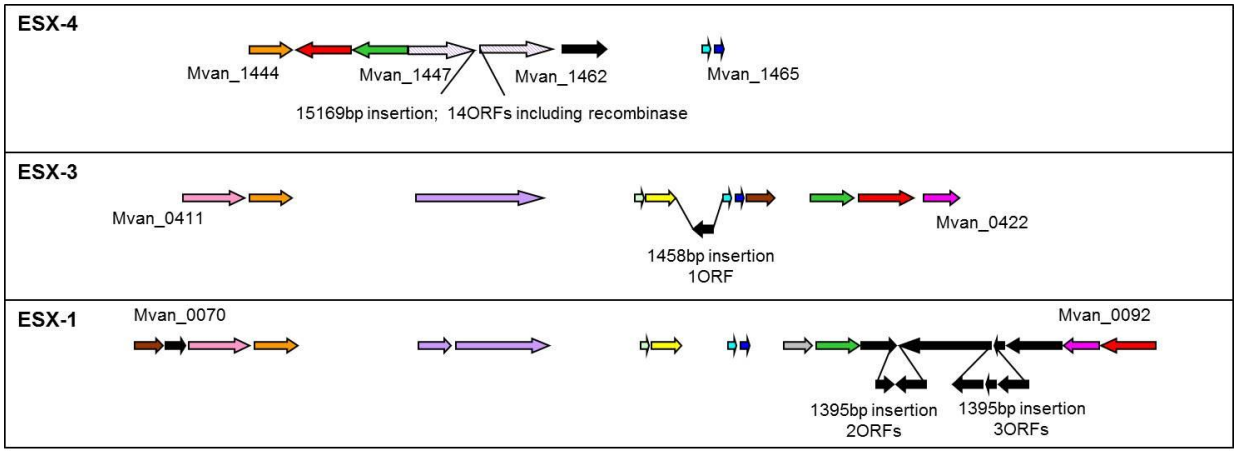
Figure 1.1. *continued*



M. sp. JLS

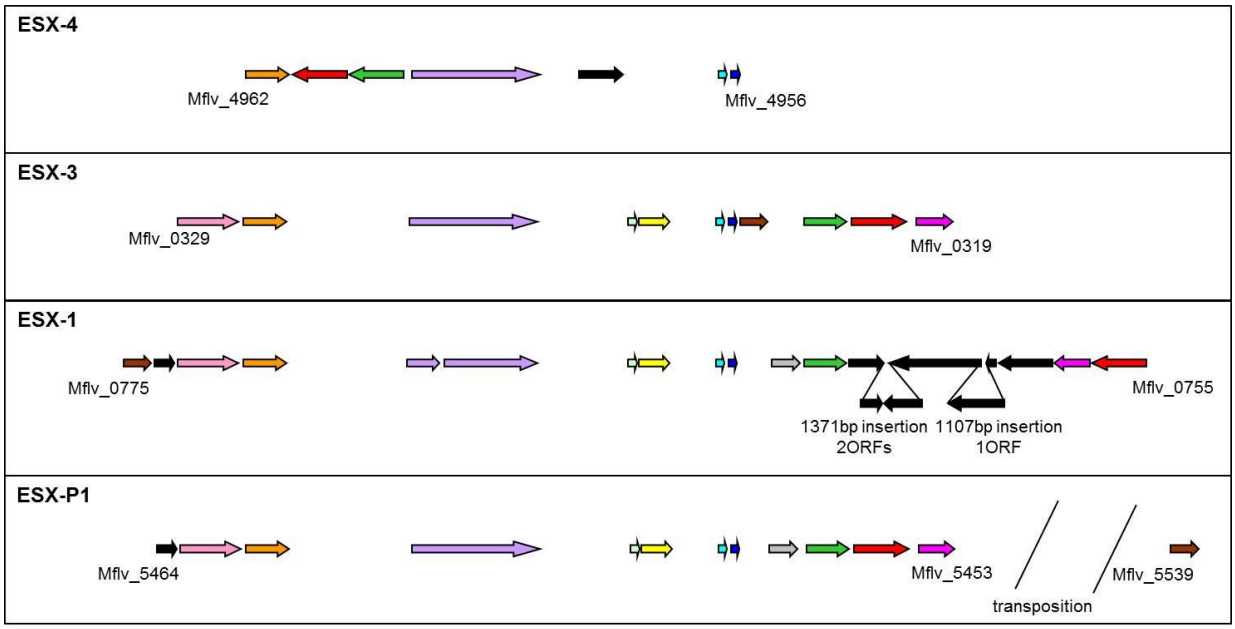


M. phlei

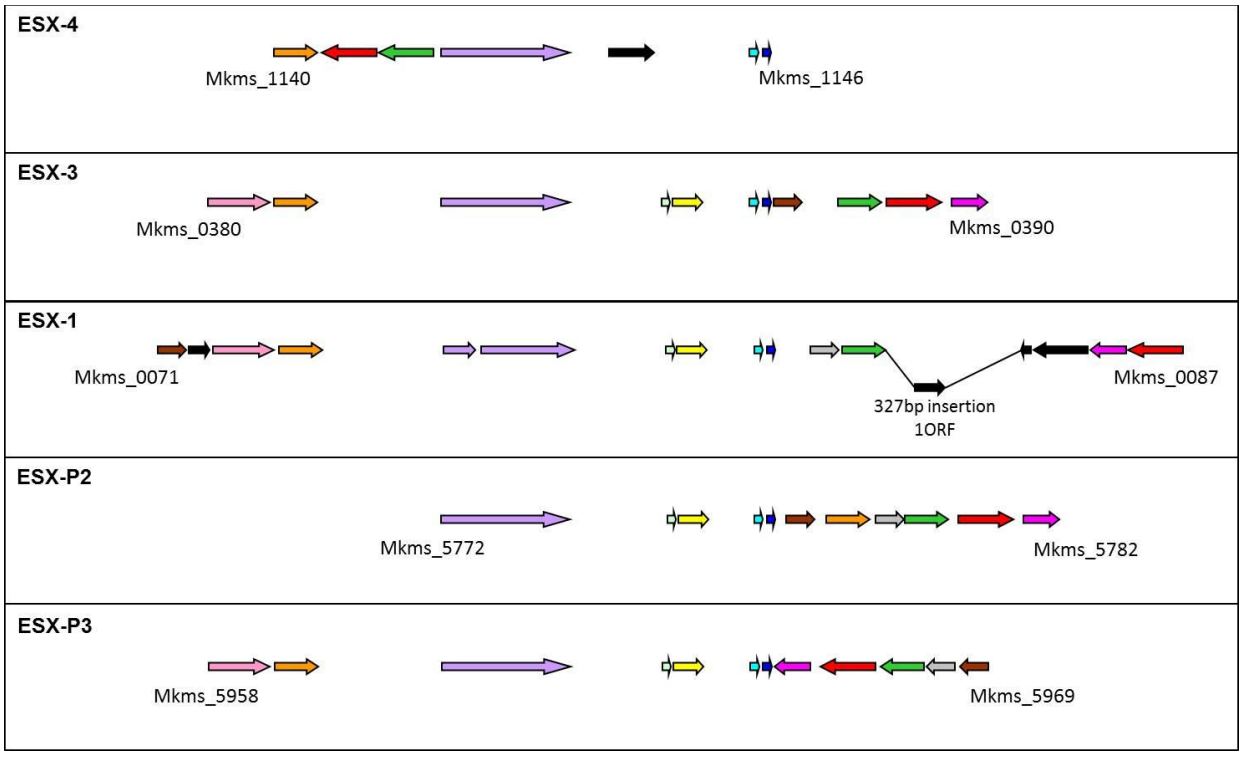


M. vanbaalenii

Figure 1.1. *continued*

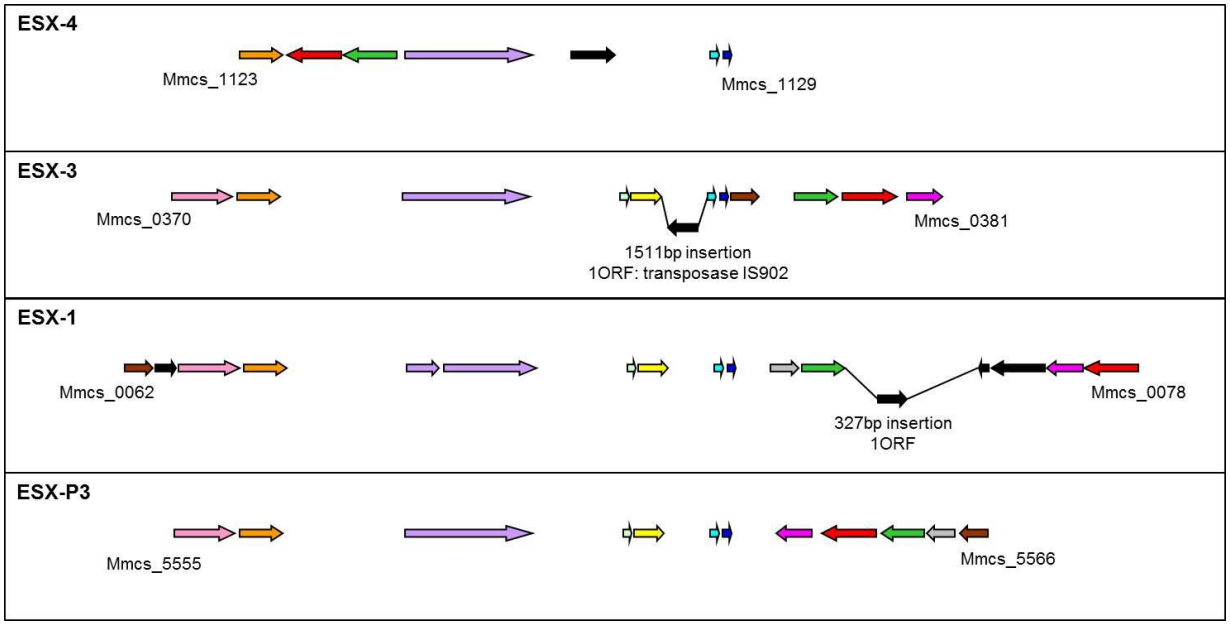


M. gilvum

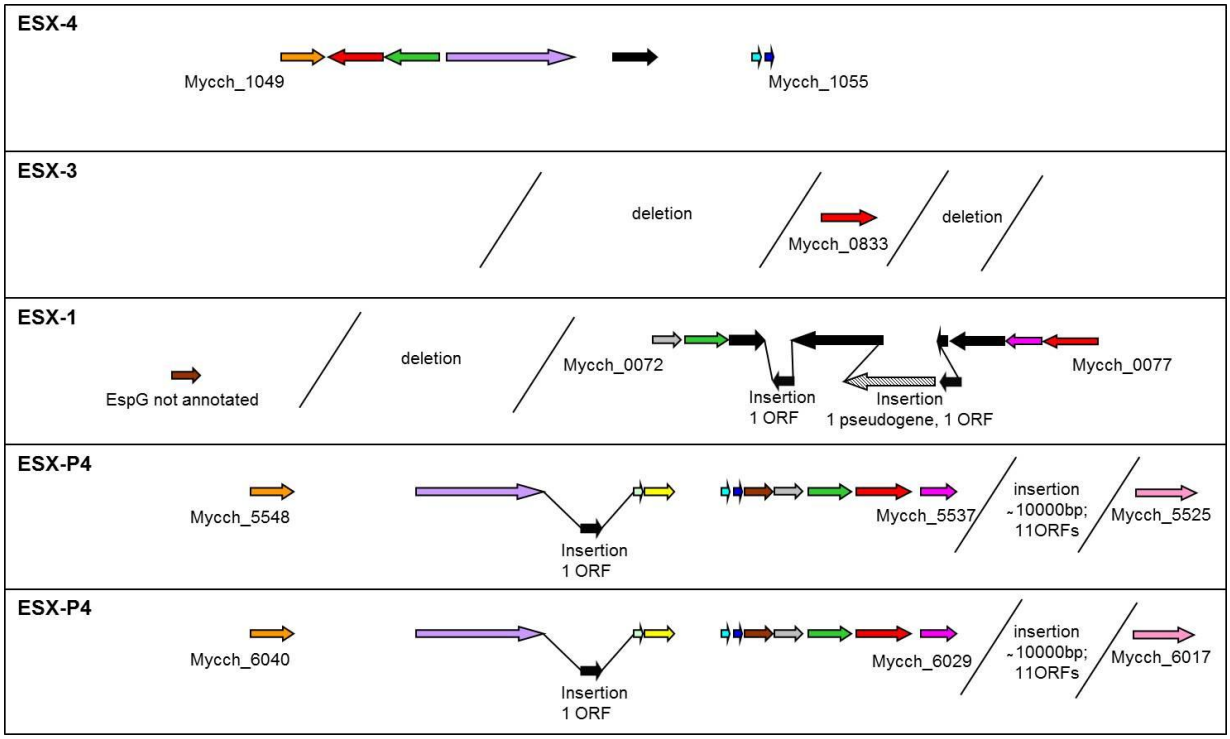


M. sp. KMS

Figure 1.1. *continued*

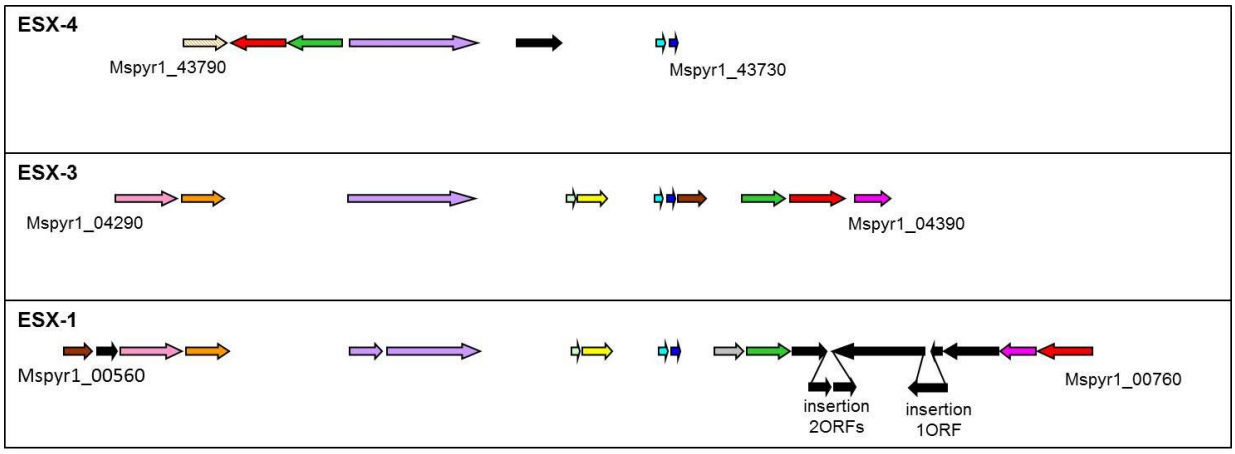


M. sp. MCS

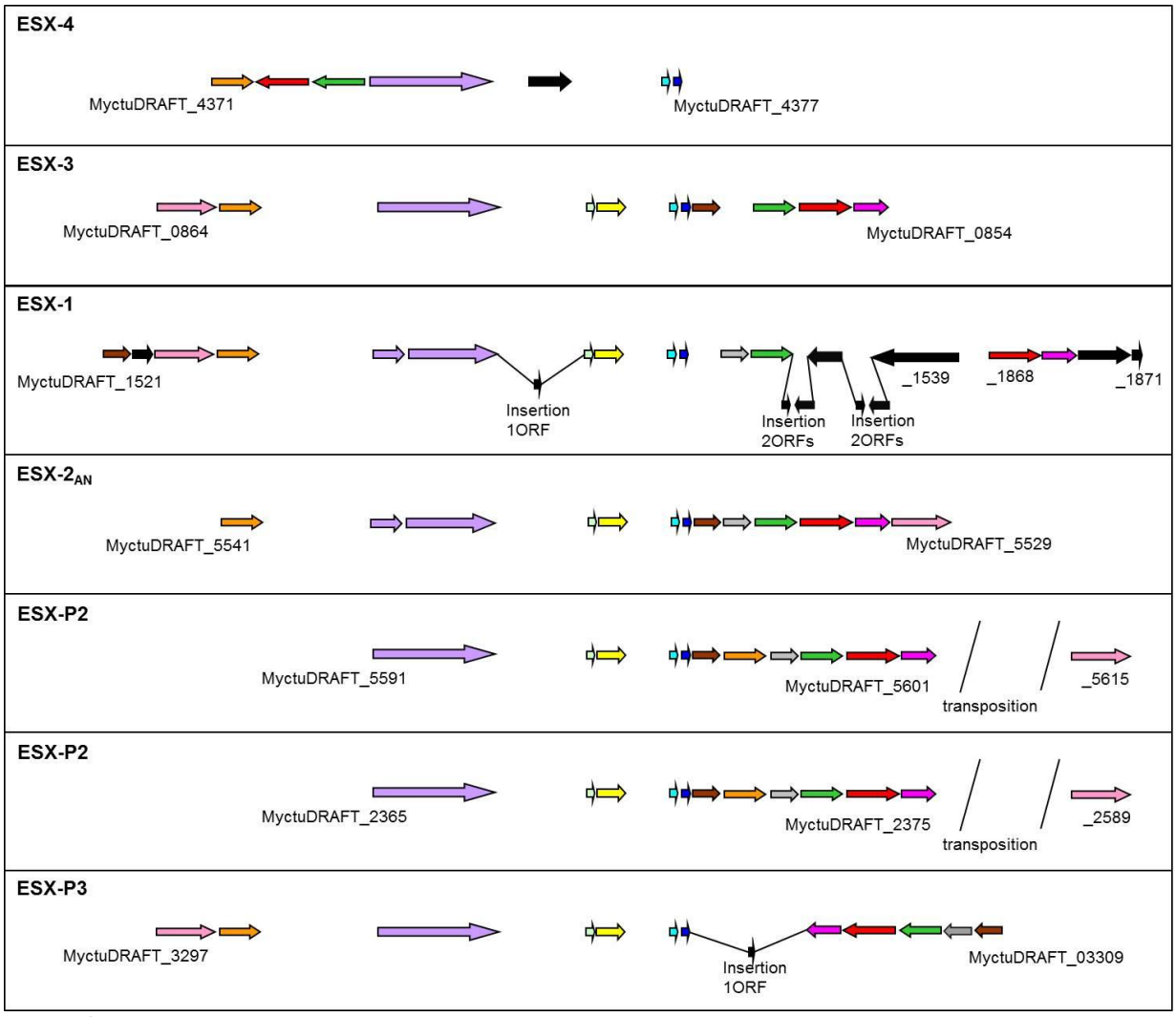


M. chubuense

Figure 1.1. *continued*

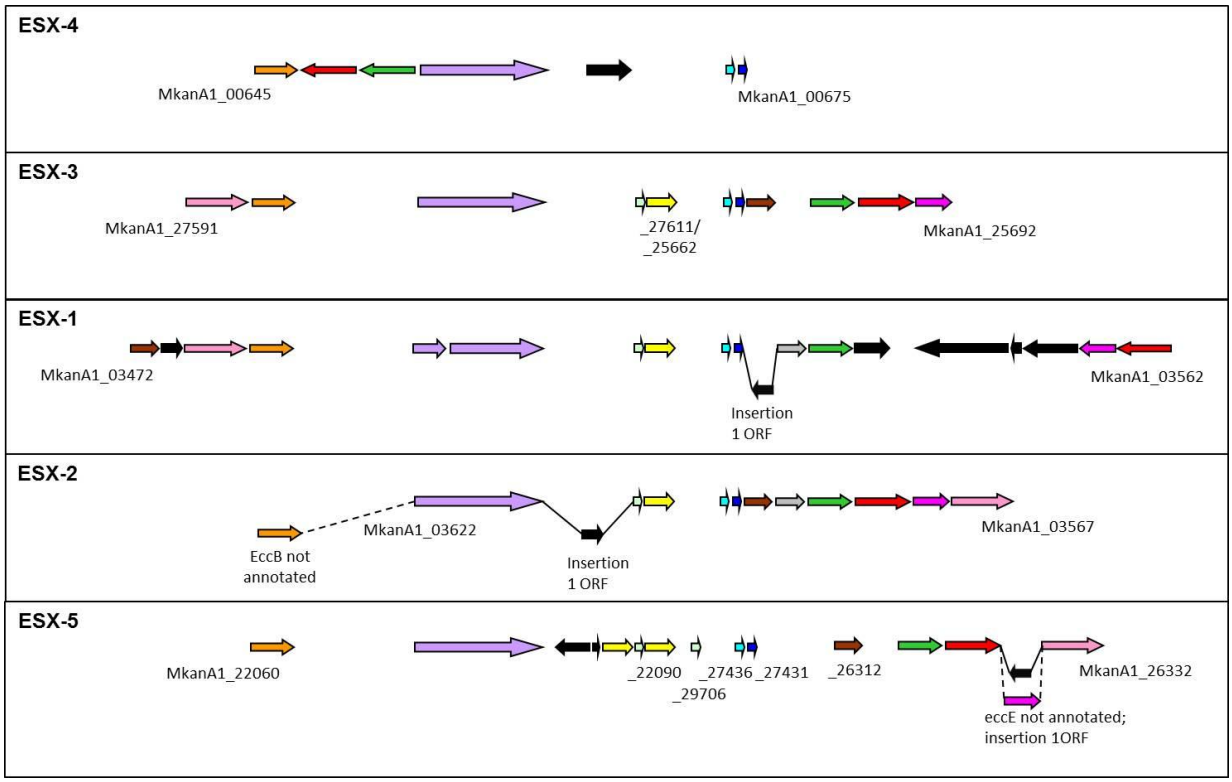


M. sp. Spyr1

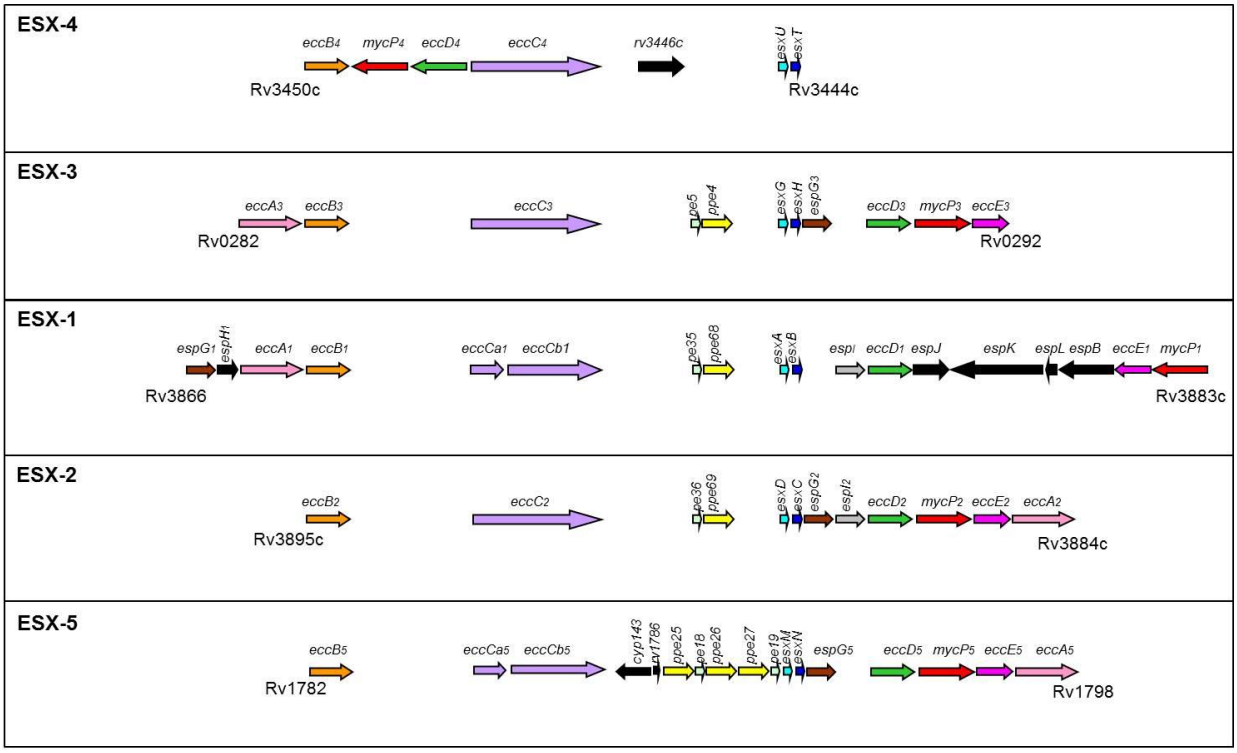


M. tusciae

Figure 1.1. *continued*

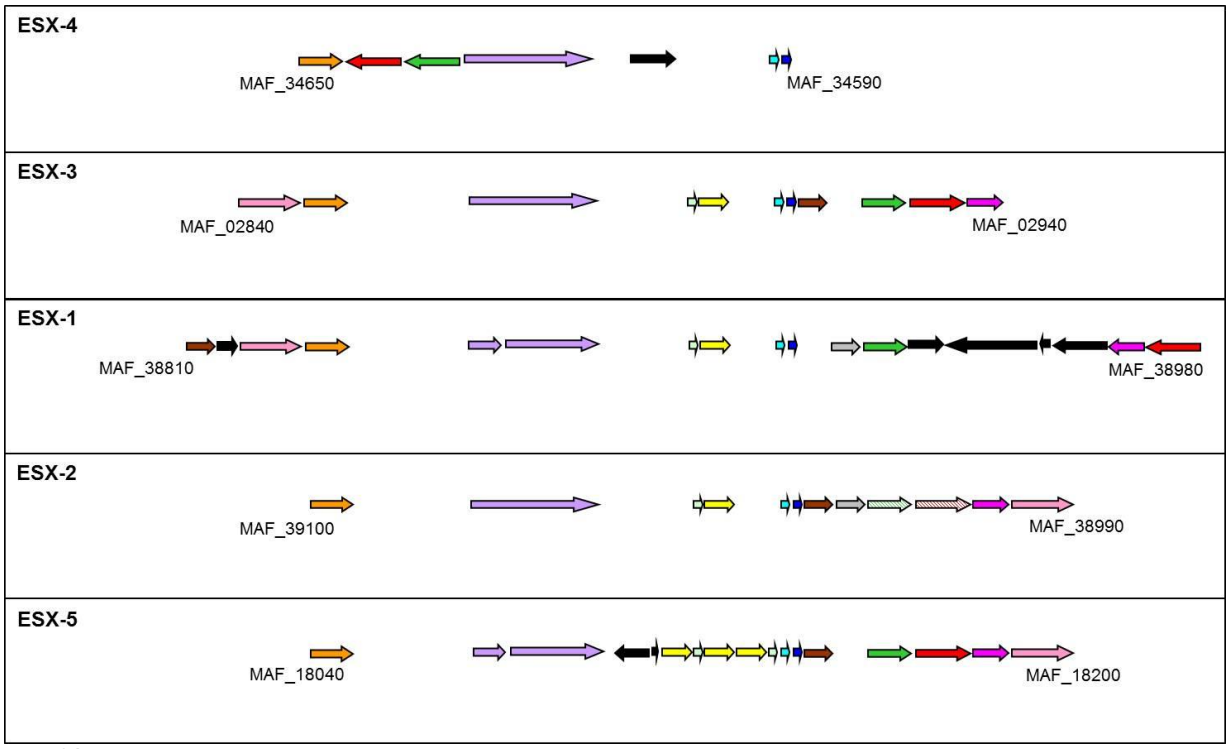


M. kansasii

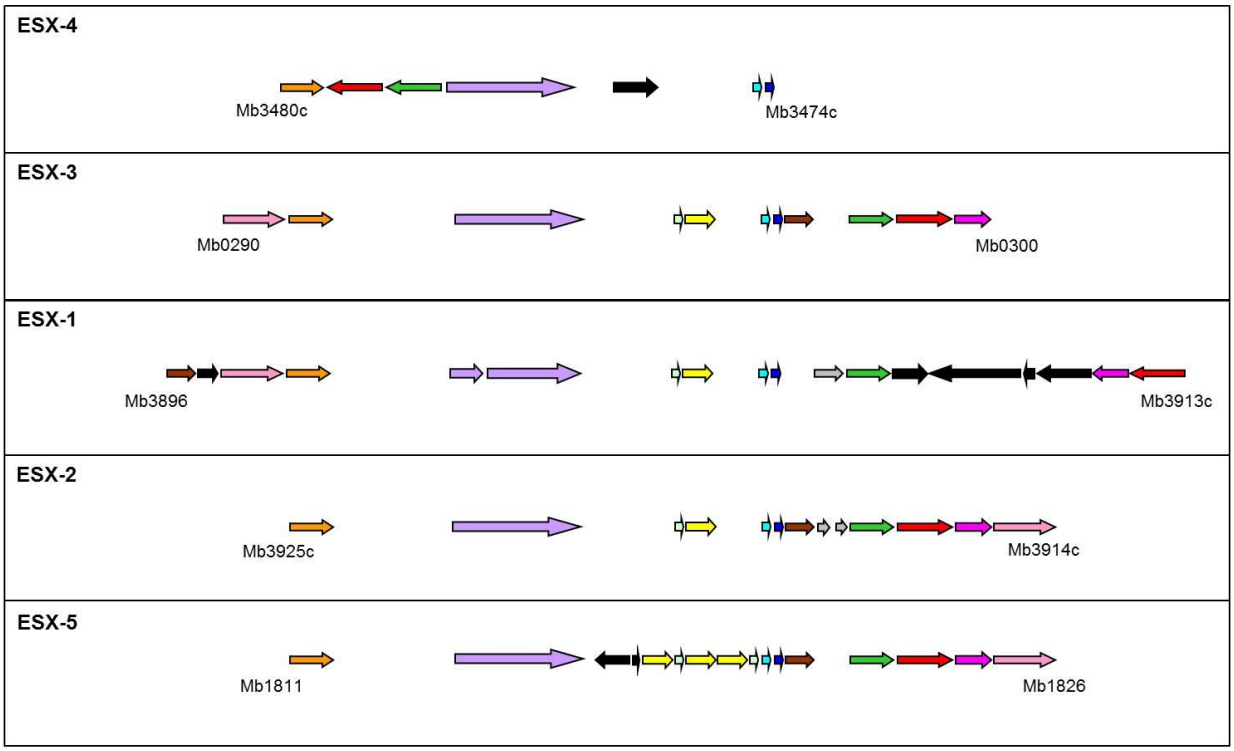


M. tuberculosis

Figure 1.1. *continued*

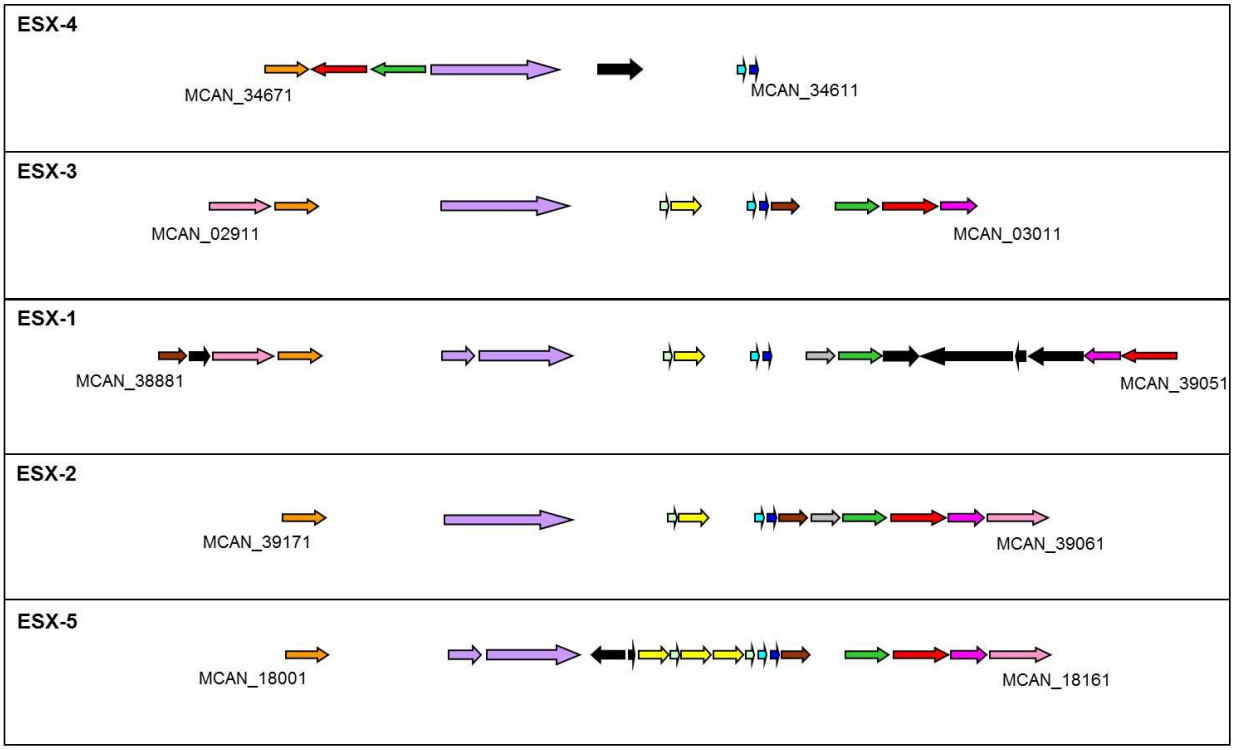


M. africanum

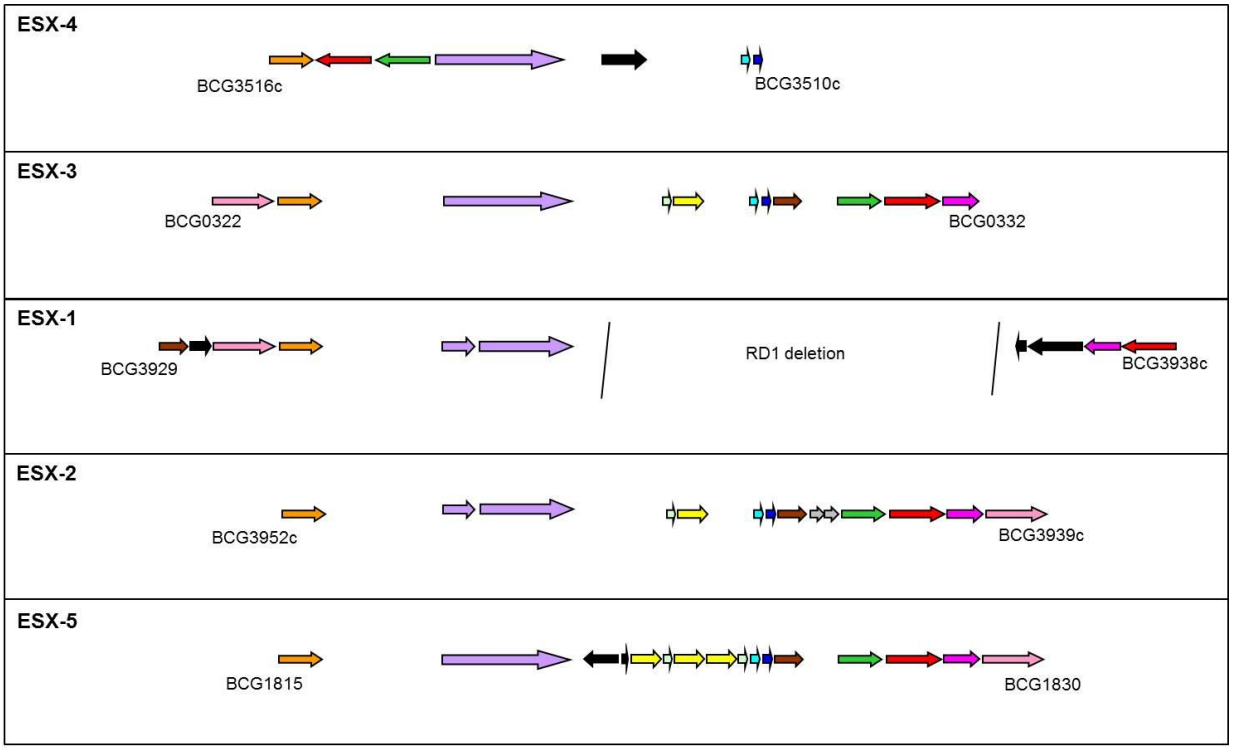


M. bovis

Figure 1.1. *continued*

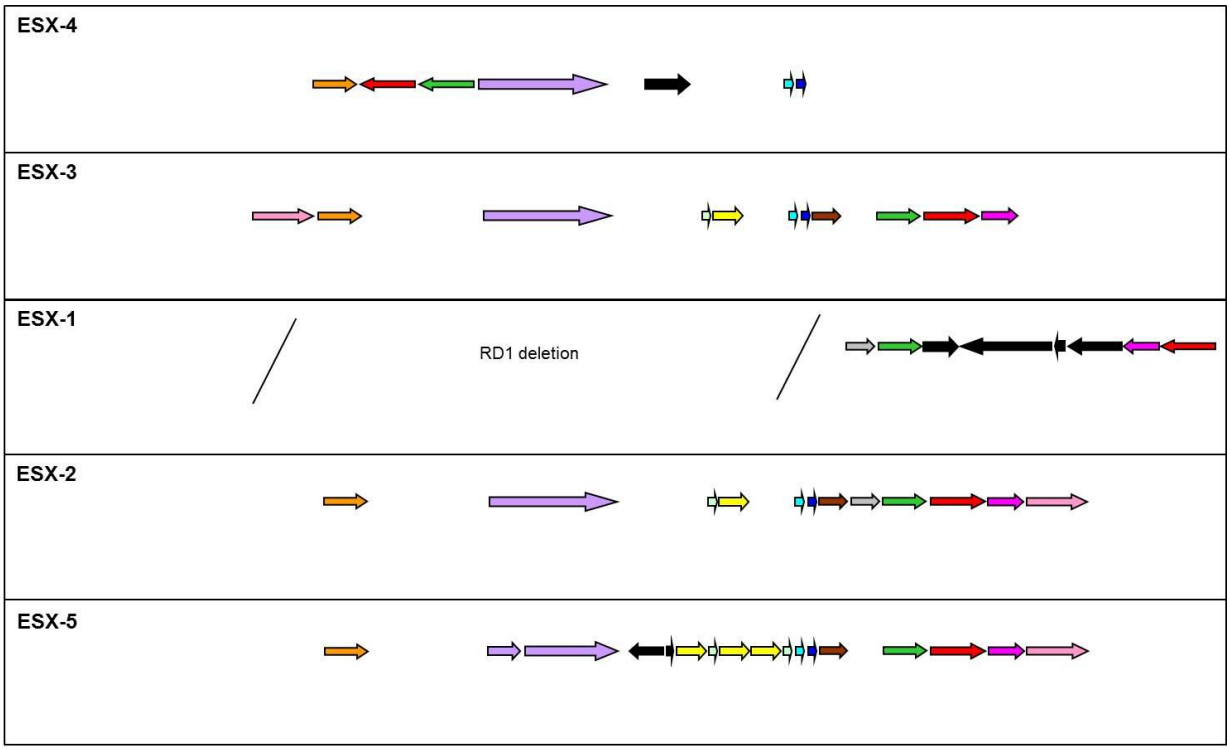


M. canetti

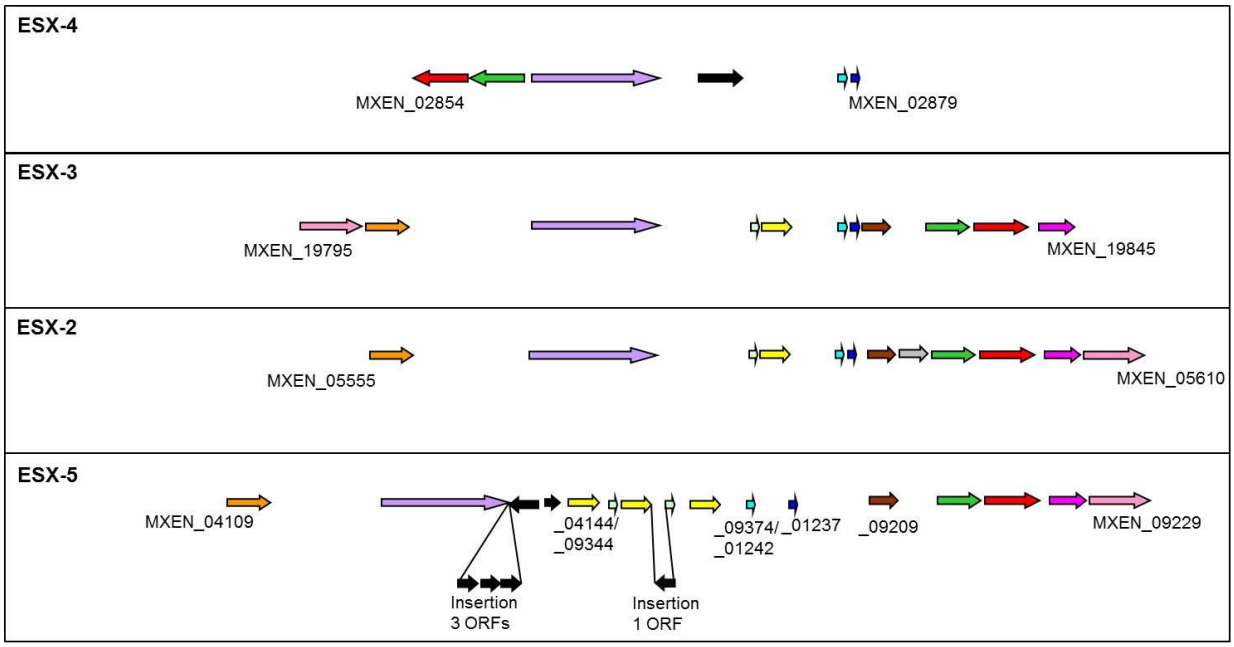


M. bovis BCG

Figure 1.1. *continued*

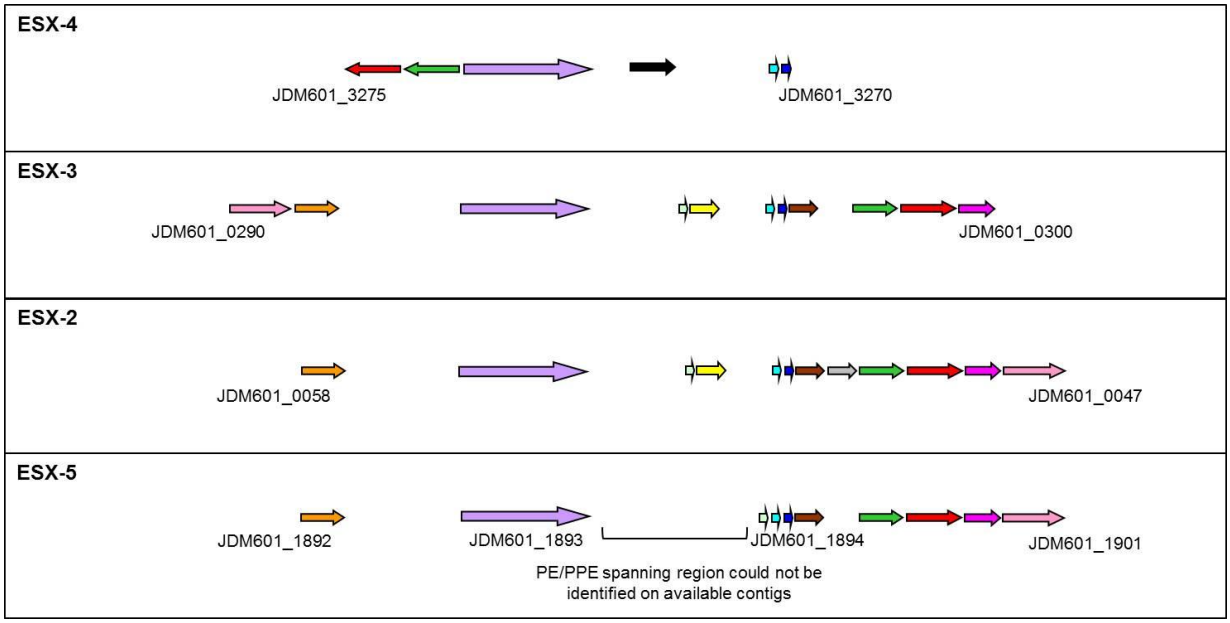


M. microti

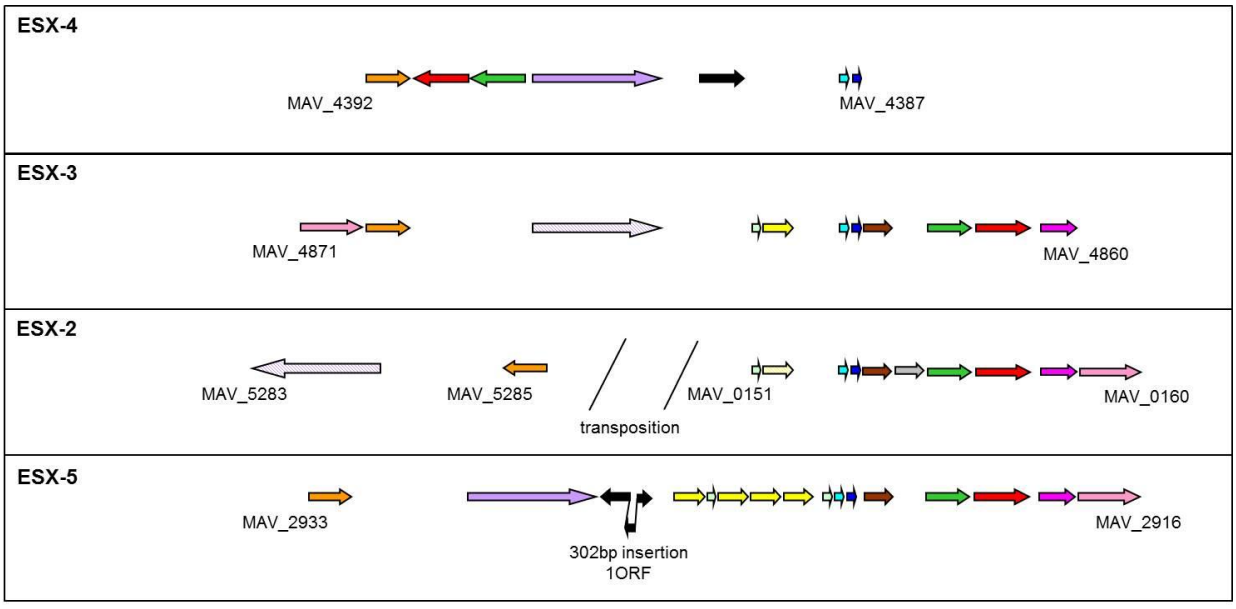


M. xenopi

Figure 1.1. *continued*

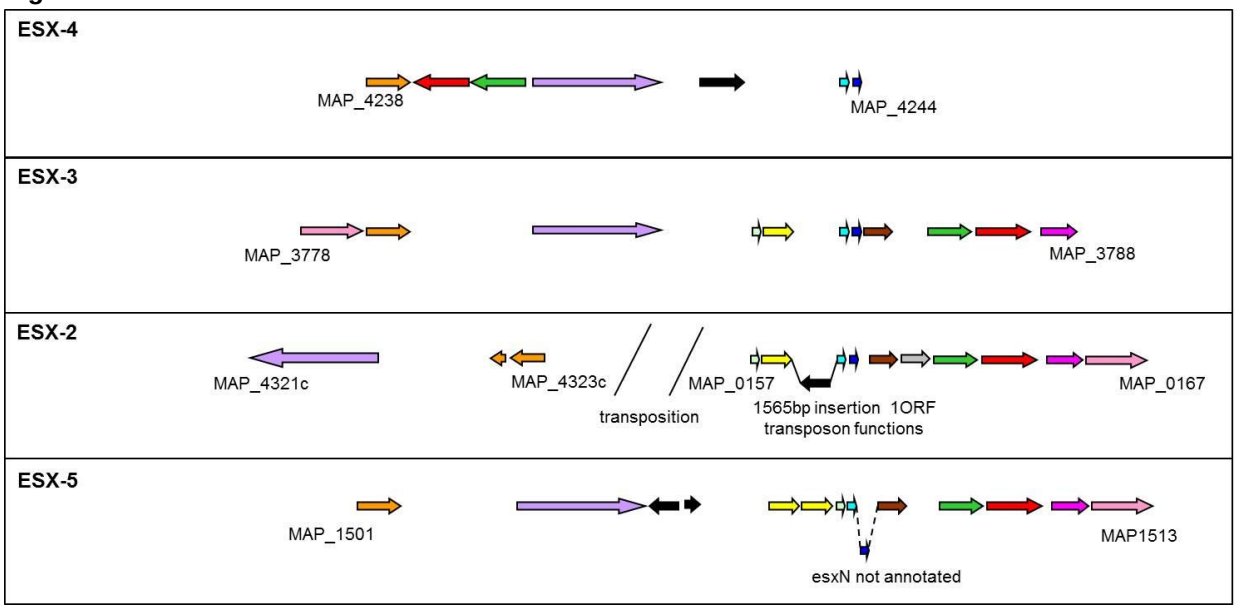


M. sp. JDM601

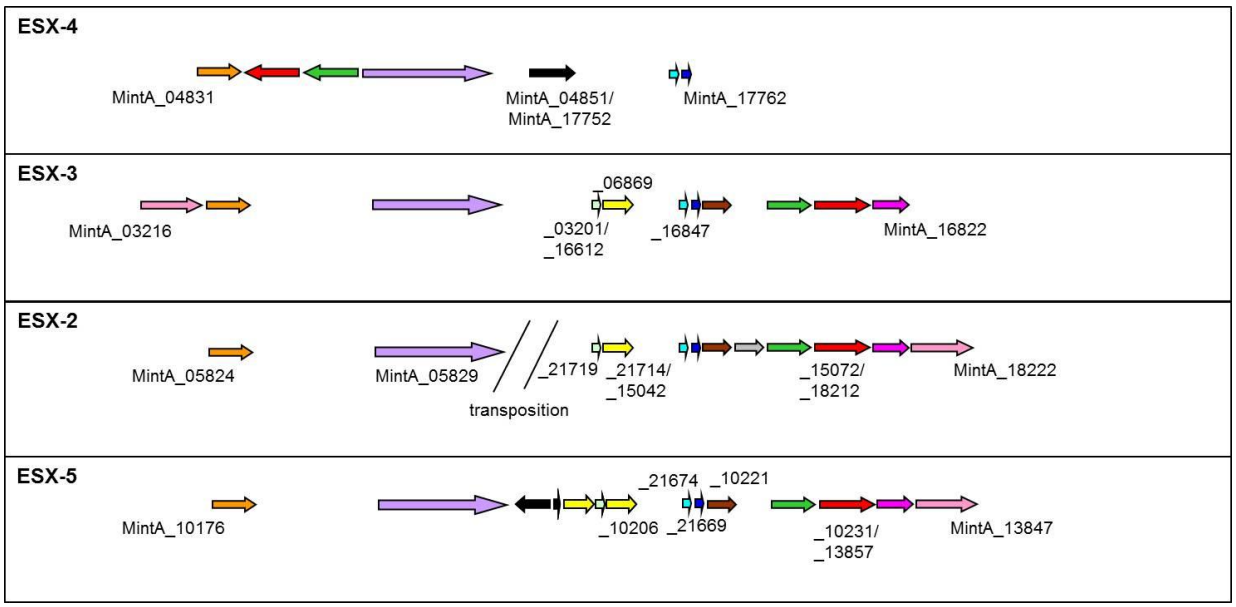


M. avium

Figure 1.1. *continued*

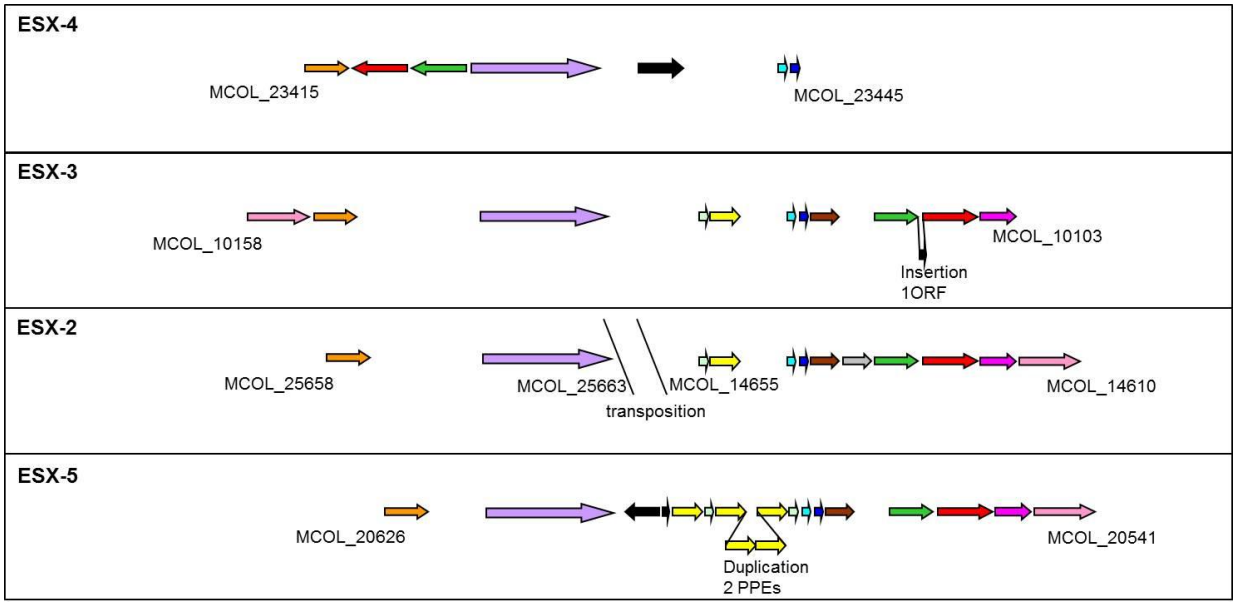


M. avium paratuberculosis

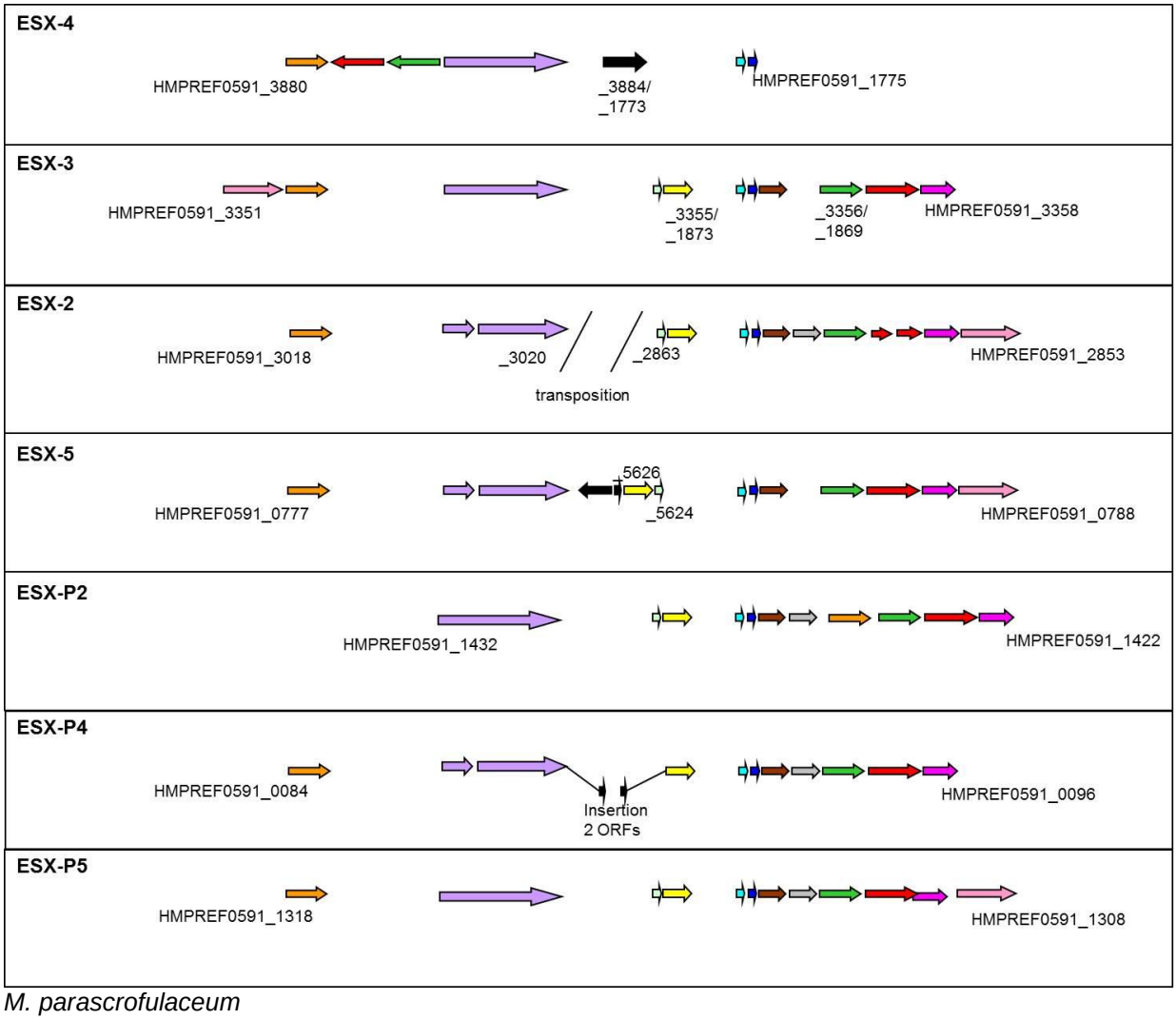


M. intracellulare

Figure 1.1. *continued*

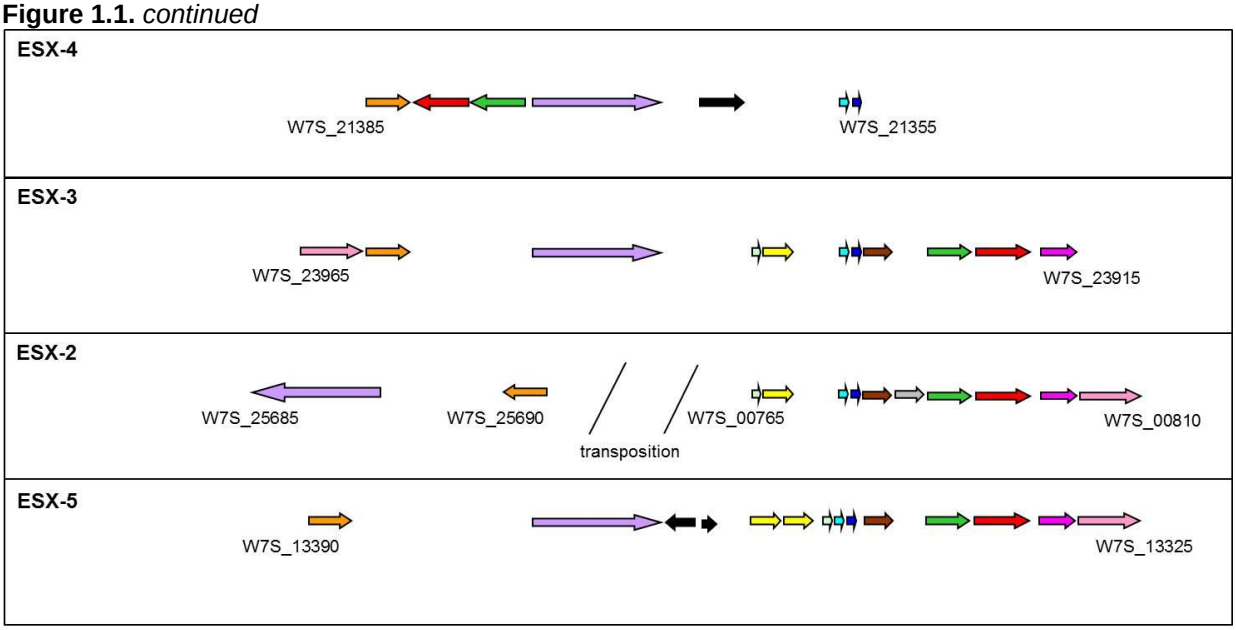


M. colombiense

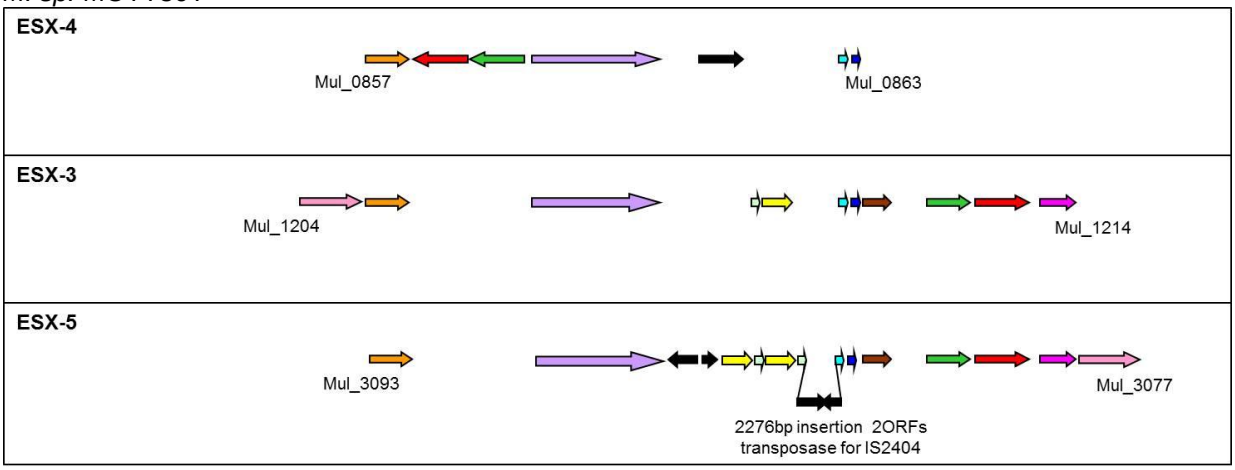


M. parascrofulaceum

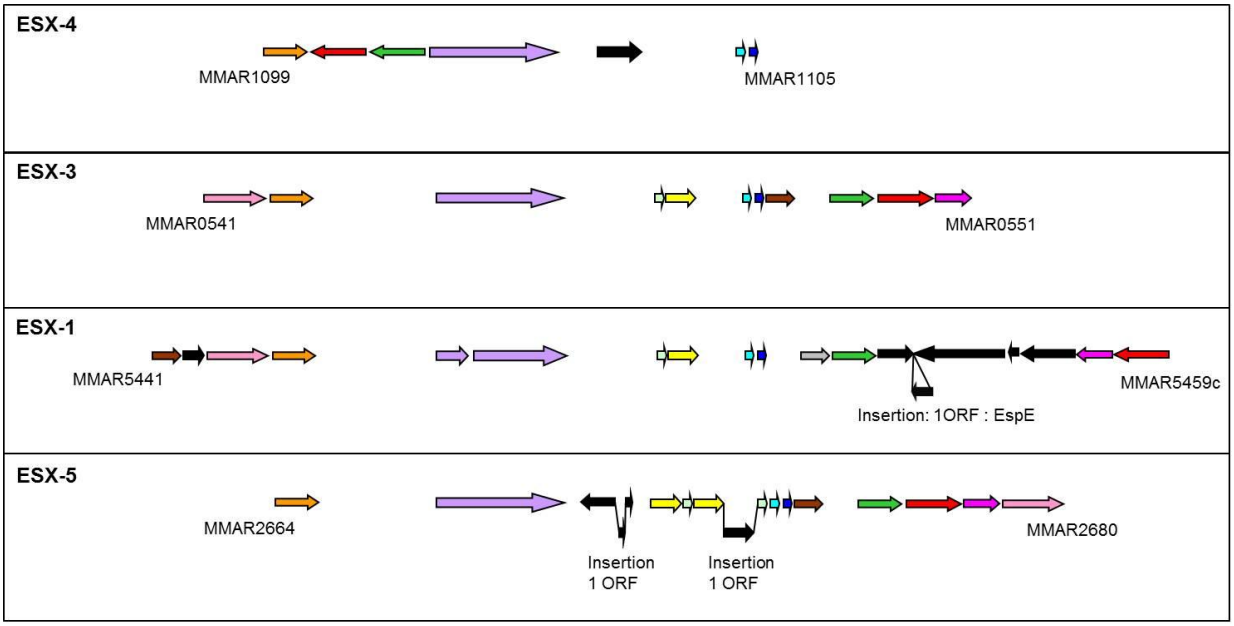
Figure 1.1. *continued*



M. sp. MOTT36Y

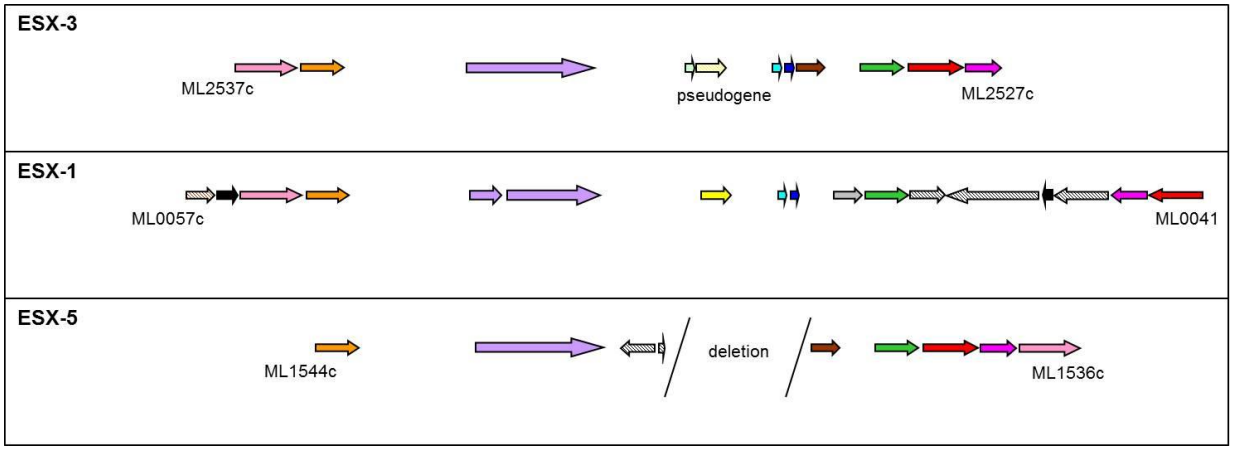


M. ulcerans



M. marinum

Figure 1.1. *continued*



M. leprae

therefore it is not possible to determine whether the ESX gene clusters are plasmid or chromosomally located. Due to the similarity between some of these ESX and the ESX-P clusters, we predict that these ESX are plasmid located, or originate directly from plasmids. In addition to these four ESX-2-like gene clusters, the *M. parascrofulaceum* genome contains ESX-4, 3 and 5, while *M. tusciae* contains ESX-4, 3 and 1.

1.2.2. The phylogeny of the ESX gene clusters

To investigate the phylogeny of the ESX gene cluster, protein sequences of the components of each gene cluster (Supplementary Tables S1.1. to S1.9. EccA, EccB, EccC, EccD, EccE, PE, PPE, Esx (CFP-10), Esx (ESAT-6), EspG, EspH, MycP, Rv3446c, EspI, EspJ, EspK, EspL, EspB, Cyp143 and ferredoxin) were concatenated into a single amino acid sequence and aligned using ClustalW 2.0. A neighbour-joining tree was generated using PAUP and Treeview, using the *C. diphtheria* ESX gene cluster combined protein sequence as the outgroup (data not shown). Incorporation of the WXG-FtsK cluster protein sequences from *S. aureus, L. monocytogenes* and *B. subtilis* using the *S. aureus* sequence as the outgroup generated a similar phylogeny depicting evolution of the ESX gene cluster from a WXG-FtsK cluster progenitor (Figure 1.2).

The 5 ESX gene clusters from the different species cluster together, as anticipated. Interestingly, the *N. farcinica* ESX-3_{AN} gene cluster groups with the ESX-4 gene clusters, most similar to the *G. bronchialis* ESX-4 gene cluster, and not with any of the other mycobacterial ESX gene clusters.

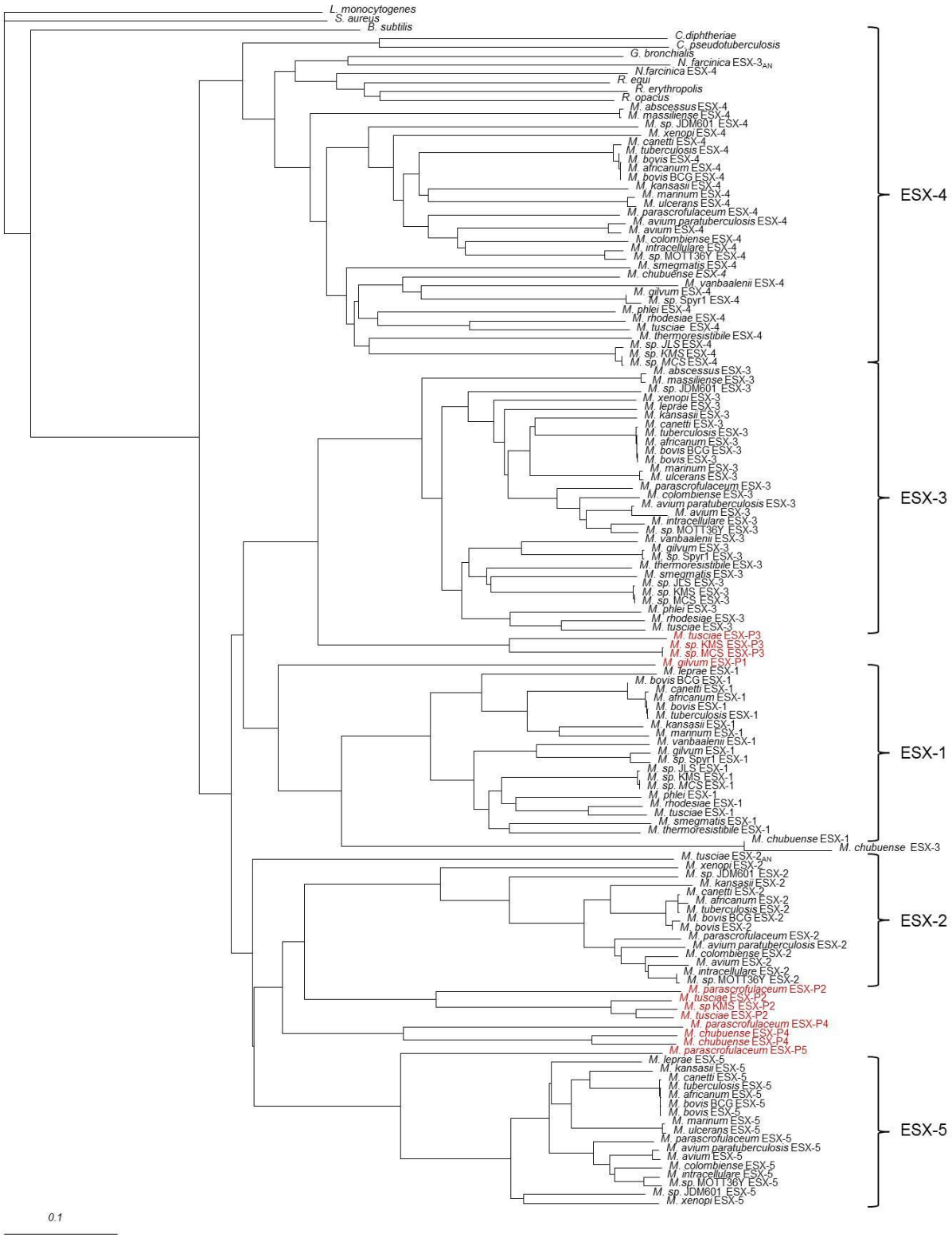


Figure 1.2. The phylogeny of the ESX gene clusters of the mycobacteria and closely related actinomycetes. The tree was drawn using alignments of the combined protein sequences of each ESX gene cluster, using the *S. aureus* WXG-FtsK cluster as the outgroup. The gene clusters have evolved from ESX-4 via duplication events which resulted in ESX-3 and ESX-1, and ESX-2 and ESX-5. The plasmid ESX gene clusters, indicated in red, form outgroups to the ESX-3, -1, -2, and -5 clusters, suggesting that they may have played a role in the evolution of these gene clusters.

Figure 1.2 indicates that a duplication of ESX-4 has evolved to produce the other ESX gene clusters via divergent evolution which produced the ESX-3 and ESX-1 gene clusters, and the ESX-2 and ESX-5 gene clusters. The ESX-P1 to ESX-P5 gene clusters form outgroups to all four ESX gene cluster duplications subsequent to ESX-4. ESX-P3 associates with ESX-3 and ESX-P1 with ESX-1; ESX-P2 associates with ESX-2, while ESX-P4 forms an outgroup to the ESX-2/ESX-P2 cluster; and ESX-P5 clusters with ESX-5. ESX-2_{AN} from *M. tusciae* predates the entire ESX-2/ESX-5 cluster, supporting its definition as the ancestor of ESX-2.

1.2.3. The phylogeny of the mycobacteria based on ESX evolution

A phylogenetic analysis of the mycobacteria and related actinomycetes based on their ESX gene clusters was done using the combined protein sequences of all of the ESX gene clusters of each species (Figure 1.3.). *M. chubuense* was omitted from these analyses as the unusual combination of ESX clusters resulting from the deletion of ESX-3 and parts of ESX-1, in the presence of the two ESX-P2 plasmid ESX clusters, distorted the phylogeny. The *Mycolata* have evolved from a single ancestor into two groups, those which contain only ESX-4 (A), with the exception of *N. farcinica* which contains ESX-$3_{AN,}$ which appears to have evolved divergently of the mycobacterial ESX-3 (Figure 1.2); and the *Mycobacteria*, which have additional ESX gene clusters in addition to ESX-4 (B). The duplication event which resulted in ESX-3 marks the evolution of the mycobacteria. All mycobacteria contain at least ESX-4 and ESX-3. *M. abscessus* and *M. massiliense* are the most ancestral mycobacteria, containing only these two ESX gene clusters (C). The duplication which resulted in ESX-1 separates the rest of the mycobacteria (D), which are further divided into 2 groups. The group which contain only ESX-4, -3 and -1 includes the rest of the fast-growers (E) while the duplication of ESX-2 and ESX-5 marks the evolution of the slow-growing mycobacteria (F). The slow-growers can be subgrouped further into those which have subsequently lost ESX-1 (G) and those which did not undergo this ESX-1 deletion event (H). The *M. avium* complex species (I) can be distinguished from the rest of the species which have lost ESX-1 by the transposition of EccB and EccC from ESX-2. Of the species which have retained ESX-1, *M. marinum* and *M. ulcerans* have undergone a deletion of ESX-2 (J), and *M. ulcerans* has further lost ESX-1 (K). *M. leprae,* which has undergone extensive gene reduction, has retained only ESX-3, -1 and -5 (L), while *M. kansasii* and the *M. tuberculosis* complex species have retained all five ESX gene clusters (M). The RD1 deletions have removed parts of ESX-1 from *M. microti* and *M. bovis* BCG (N).

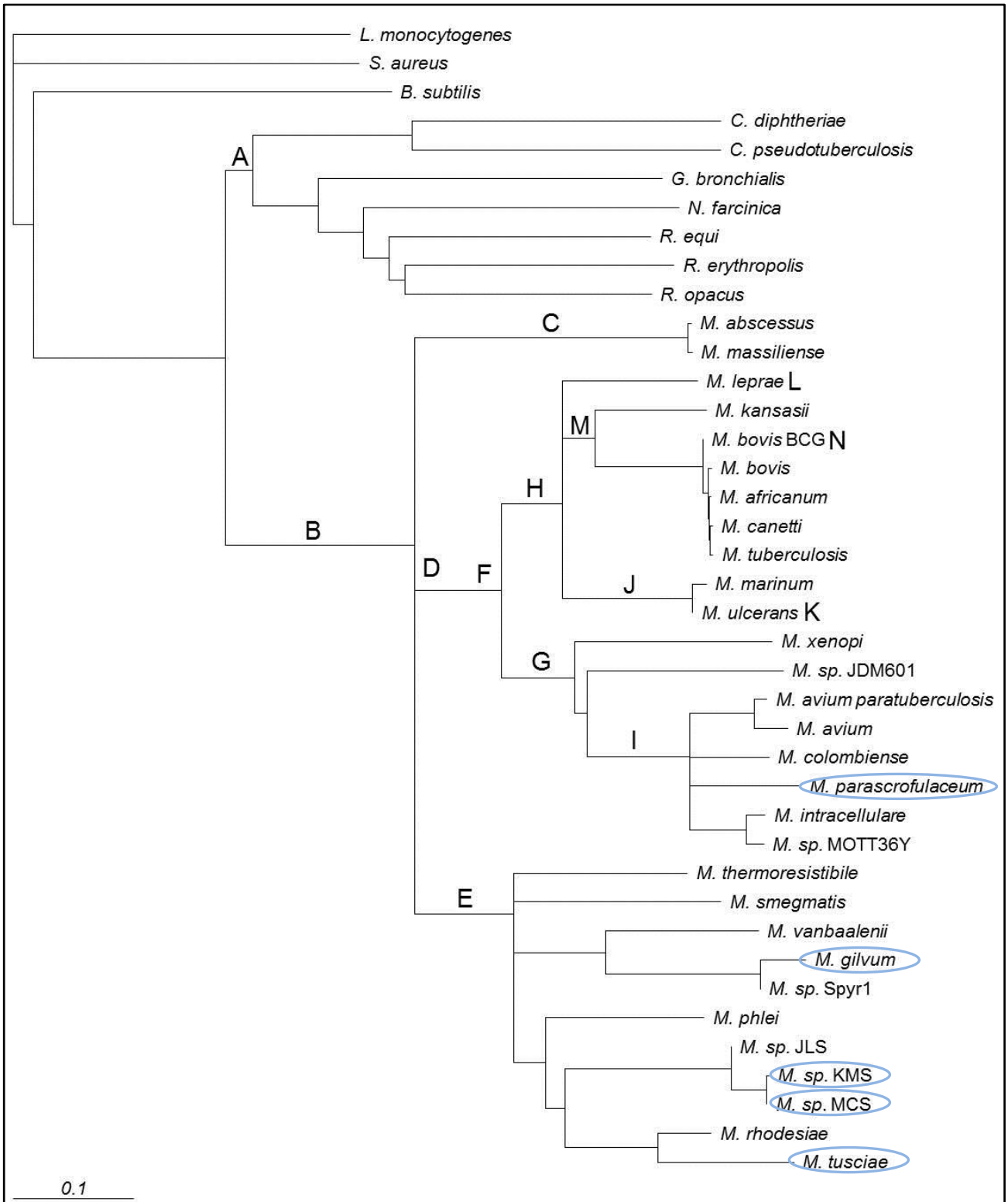


Figure 1.3. The phylogeny of the sequenced mycobacteria and closely related actinomyetes based on the evolution of the ESX gene clusters. The tree was drawn using alignments of the combined protein sequences of all the ESX gene cluster regions in each species, using *S. aureus* as the outgroup. Letters A-N represent ESX duplication and deletion events, as described in the text. The species which contain plasmid located ESX-P gene clusters are circled in blue.

1.3. Discussion

The ESX gene clusters each encode a dedicated protein secretion system, the ESX secretion systems, which are responsible for the secretion of at least the Esx, Esp, PE and PPE proteins. It has been proposed that these ESX secretion systems be classified as a novel Type-VII secretion system, defined by its ability to transport proteins across the outer mycomembrane peptidoglycan-arabinogalactan-mycolic acid matrix.[30] Although secreted protein substrates of the ESX secretion systems have been identified in culture filtrates, the mechanism whereby they are secreted and the components responsible for translocation across the mycomembrane have not been elucidated. This has generated some controversy as some suggest that identification of a mycomembrane translocation mechanism and components are essential for the classification of a novel type of secretion system. Indeed, it has been suggested that ESX secretion systems occur outside of the *Mycolata* (species containing a mycomembrane-like outer membrane, including *Corynebacteria*, *Rhodococci*, *Nocardia* and *Mycobacteria*) and are therefore not typically involved in trans-mycomembrane secretion.[151] Esx-like (WXG100) proteins have been identified in a number of other *Actinobacteria*, *Firmicutes* and *Chloroflexi* which are typically monoderm bacteria. However, only a protein with FtsK/SpoIIIE domains (EccC-like) is frequently associated with the WXG100 proteins throughout the *Firmicutes* and non-mycolata *Actinobacteria.*[151] This suggests that although the FtsK/SpoIIIE-like protein may be involved in translocation of the WXG100 proteins across a monoderm plasma membrane, the additional components of the ESX secretion systems, encoded in the ESX gene clusters, may enable the secretion of additional substrates and facilitate translocation across the mycomembrane as a Type VII secretion system.

In this chapter we have explored the evolution of the mycobacterial Type-VII ESX gene clusters from the WXG-FtsK cluster in *S. aureus*, *L. monocytogenes* and *B. subtilis*, via the ancestral region in *the Corynebacteria, Nocardia* and *Rhodococci* to the 5 ESX gene clusters in *M. tuberculosis,* and the distribution of these gene clusters in the other mycobacteria. We have identified an additional ESX gene cluster in *N. farcinica* (ESX-3_{AN}), which appears to be an ancestral ESX-3 which has evolved divergently of the mycobacterial ESX-3, as it clusters phylogenetically with the ESX-4 gene clusters. The ESX-3, -1, -2 and -5 gene clusters have evolved from a duplication of the ancestral ESX-4 cluster. The presence and absence of the ESX gene clusters in the various mycobacteria defines the order of duplication of the ESX gene clusters as ESX-4, ESX-3, ESX-1 and then ESX-2 and ESX-5.[30]

The presence of ESX-3 marks the evolution of the mycobacteria, with *M. abscessus* and *M. massiliense* possessing only ESX-4 and ESX-3; while the rest of the fast growers also contain ESX-1. The progenitor of the slow-growers contained both ESX-2 and ESX-5, although various deletions of ESX regions have subsequently occurred in some species. One exception to this is *M. tusciae*, which is a slow-growing mycobacterium, which is phylogenetically grouped with the fast-growers. This classification is consistent with 16S rDNA sequencing results which showed that *M. tusciae* is most closely related to the fast-growing mycobacteria *M. farcinogenes, M. komossense* and *M. aichiense.*[152] This species contains several ESX gene clusters which appear to be precursors of the ESX-2 and ESX-5 gene clusters, which are associated with a slow-growing phenotype, and may contribute to its conflicting phenotype.

Interestingly, we identified a number of ESX-containing plasmids in the fast-growing mycobacteria *M. ssp.* KMS and MCS, *M. gilvum* and *M. chubuense,* and seven ESX gene clusters in each of *M. tusciae* and *M. parascrofulaceum*. The plasmid ESX gene clusters as well as 3 and 4 of the *M. parascrofulaceum* and *M. tusciae* ESX gene clusters, respectively, form outgroups to the ESX-3, ESX-1, ESX-2 and ESX-5 gene clusters, and appear to be precursors thereof. The ESX gene clusters on pMKMS02, *M sp.* MCS Plasmid01 and the *M. tusciae* contig 212 appear to be precursors of ESX-3 (ESX-P3) and pMFLV01 of ESX-1 (ESX-P1). The ESX gene clusters on pMKMS01, pMYCCH.01, pMYCCH.02, *M. tusciae* contigs 209 and 196 and *M. parascrofulaceum* contig 115 are precursors of ESX-2 (ESX-P2), while *M. parascrofulaceum* contig 17 ESX is a predecessor of all of these (ESX-P4). *M. parascrofulaceum* contig 109 ESX is a precursor of ESX-5 (ESX-P5), and *M. tusciae* contig 212 ESX predates all ESX-2 and ESX-5 gene clusters and their direct precursors ($ESX\text{-}2_{AN}$).

The *M. parascrofulaceum* and *M. tusciae* genome sequences have not been completed, and as such it remains unclear whether the additional ESX duplications are genomic or plasmid encoded. The close phylogenetic relationships between some *M. tusciae* and *M. parascrofulaceum* ESX gene clusters and the plasmid ESX gene clusters from *M. spp*. KMS and MCS, *M. gilvum and M. chubuense* suggests that at least some of these gene clusters may be, or have been, located on plasmids. If this is the case, it points to plasmid-mediated duplication and evolution of the ESX gene clusters.

We propose an evolutionary pathway for the ESX gene clusters whereby they evolved from the original WXG-FtsK cluster into a complete ESX-4 gene cluster together with the evolution of the

mycomembrane in order to facilitate transport of proteins across it. ESX-4 incorporates genes encoding a mycosin protease (*mycP*), a transmembrane protein (*eccB*) and the predicted transmembrane pore (*eccD*). These proteins may be involved in a more complex secretion mechanism to transport proteins into and across the unique and complex outer membrane. MycP may be involved in the cleavage of proteins or other cell structures to enable the insertion of the EccD pore protein into the mycomembrane, through which proteins can be translocated, or in the release of secreted substrates from the secretion machinery. EccB may be involved in protein-protein interactions or stabilisation of the secretion complex.

We hypothesise that ESX-4 was subsequently incorporated into plasmid DNA and evolved to form ESX-P3, which was inserted into the genome as ESX-3, and marks the evolution of the mycobacteria. ESX-P1 evolved divergently from ESX-P3 and was inserted into the genome subsequent to the evolution of *M. abscessus* and *M. massiliense*. ESX-P3 further diverged to form ESX-2_{AN}, which evolved to form the ESX-2 pre-precursor ESX-P4 and the ESX-5 precursor ESX-P5. The ESX-P4 evolved further to form ESX-P2 which was inserted into the genome and evolved to the current ESX-2. ESX-P5 was also incorporated into the genome and evolved as the current ESX-5. The incorporation of ESX-2 and ESX-5 into the mycobacterial genome marks the evolution of the slow-growing mycobacteria. The evolutionary pathway of ESX evolution is represented in Figure 1.4. Due to the extrachromosomal nature of the plasmids, bacteria are able to retain multiple plasmids containing different ESX clusters, enabling multiple evolutionary paths to occur simultaneously. Additionally, these plasmids may be lost from the bacteria before or after their incorporation into the chromosome; although plasmid DNA which confers a fitness advantage may be preferentially retained or incorporated.

The presence of multiple ESX-2- and ESX-5-like ESX gene clusters, in the absence of ESX-2 and ESX-5, in *M. tusciae,* suggests that this species may be an evolutionary bridge between the fast- and slow growers. The conflicting genotypic and phenotypic classification regarding its growth rate provides further support for this, and may furthermore implicate the duplication of the ESX gene

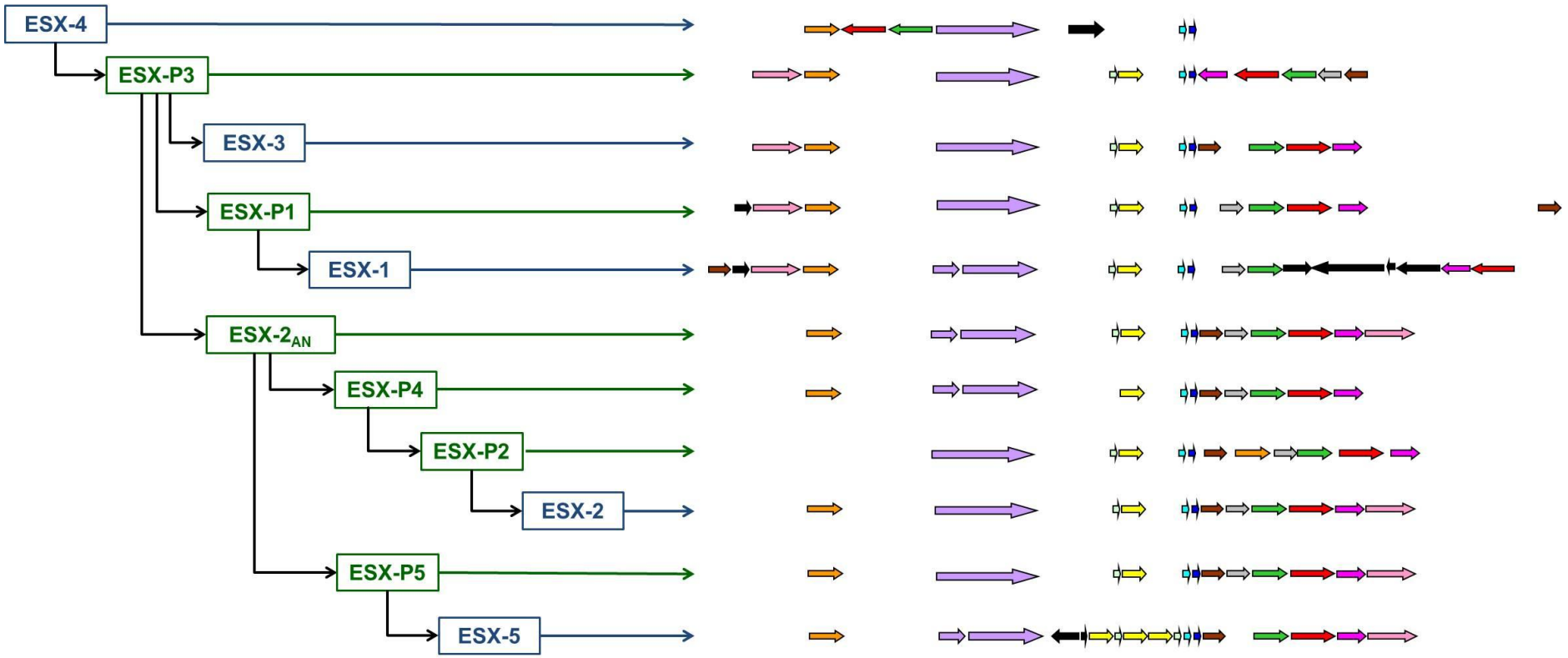


Figure 1.4. The evolutionary pathway of ESX evolution. The five *M. tuberculosis* ESX gene clusters evolved in the order 4, 3, 1, 2 and 5 via several plasmid located intermediaries, as described in the text. Black arrows represent duplication events and/or evolution of the ESX gene cluster. The genomic ESX are indicated in blue; plasmid located ESX are indicated in green.

clusters in the reduced growth rate of the slow-growing mycobacteria. ESX-5 is the only ESX gene cluster which is present in all of the slow-growers but absent in all of the fast-growers, and although deletion of this region does not directly increase the growth rate of *M. marinum* or *M. tuberculosis,* [61, 125] investigation of its potential association with growth rate may provide important understanding of the evolution of the often pathogenic slow-growing mycobacteria.

1.4. Conclusion

The duplication and evolution of the ESX gene clusters and their secretion systems have clearly impacted on the evolution, diversity and success of the mycobacteria. Although the function(s) of ESX-4 have yet to be determined, the presence and maintenance of this gene cluster throughout the mycobacteria and related actinomycetes suggests that it plays an important, though non-essential, role in bacterial metabolism. ESX-3 has co-evolved with the mycobacteria, and although essential for *in vitro* growth of *M. tuberculosis,* it is not essential in the fast-growing *M. smegmatis*,[54] and has been completely lost from the genome of *M. chubuense*. Genetic reduction during evolution of the slow-growers may have eliminated the redundancy of ESX-3 in the fast-growers. ESX-1 has been implicated in virulence, and its deletion in attenuation of the pathogenic mycobacteria.[82-84] However, this gene cluster is present throughout most of the mycobacteria, including non-pathogenic and saphrophytic fast-growing organisms, suggesting that the primary function of this gene cluster is not virulence, and that the virulence-function has evolved more recently in pathogenic organisms. The function(s) of ESX-2 have also not been elucidated, and although its duplication correlates evolutionarily with both the slow-growing and pathogenic phenotype, it has been lost from some of these bacteria. Only ESX-5 has been retained throughout the slow-growing and pathogenic mycobacteria. ESX-5 has been implicated in immune evasion and may be the ESX gene cluster most involved in pathogenicity and the slow growing phenotype.[124] The presence of additional copies of ESX gene clusters in various species and on plasmids has pointed to a novel plasmid-mediated mechanism of genome evolution, via duplication and insertion of ESX gene clusters, in mycobacteria.

Supplementary Tables S1.1. to S1.9. The components of the ESX gene clusters of the sequenced mycobacteria and related actinomycetes. PS - annotated pseudogene; NA - gene is not annotated but orthologous sequence identified using tblastn; *M. microti* was omitted as the genome sequence has not been annotated

Table S1.1. The components of the ESX gene clusters of *S. aureus, L. monocytogenes, B. subtilis, C. diphtheria, C. pseudotuberculosis* and *G. bronchialis*

		S. aureus	*L. monocytogenes*	*B. subtilis*	*C. diphtheriae*	*C. pseudotuberculosis*	*G. bronchialis*
ESX-4	**EccB**				DIP0552	cpfrc_00395	Gbro_3295
	MycP				DIP0554	cpfrc_00397	Gbro_1681
	EccD				DIP0555	cpfrc_00398	Gbro_1682
	EccC	SAV0287/ SAV0288	lmo0061	CAB04782/ CAB04781	DIP0556	cpfrc_00399	Gbro_1683
	Rv3446c				DIP0557	cpfrc_00400	Gbro_1684
	EsxU	SAV0290	lmo0063		DIP0558	cpfrc_00401	Gbro_ 3643
	EsxT	SAV0282	lmo0056	CAB04773	DIP0559	cpfrc_00402	Gbro_3644
		SAV0287-SAV0288 is a single gene			Insertion DIP0553	Insertion cpfrc_00396	2 transposition events
ESX-3		ABSENT	ABSENT	ABSENT	ABSENT	ABSENT	ABSENT
ESX-1		ABSENT	ABSENT	ABSENT	ABSENT	ABSENT	ABSENT
ESX-2		ABSENT	ABSENT	ABSENT	ABSENT	ABSENT	ABSENT
ESX-5		ABSENT	ABSENT	ABSENT	ABSENT	ABSENT	ABSENT

Table S1.2. The components of the ESX gene clusters of *E. equi, E. erythropolis, R. opacus, N. farcinica, M. abscessus* and *M. massiliense*

		R. equi	*R. erythropolis*	*R. opacus*	*N. farcinica*	*M. abscessus*	*M. massiliense*
ESX-4	**EccB**	REQ_35500	RER_18880	ROP_62200	Nfa_8470	MAB_3759c	MMAS_36350
	MycP	REQ_34490	RER_18890	ROP_62210	Nfa_8490	MAB_3758	MMAS_36340
	EccD	REQ_34480	RER_18900	ROP_62220	Nfa_8500	MAB_3757	MMAS_36330
	EccC	REQ_34470	RER_18910/ RER19000	ROP_62230	Nfa_8510	MAB_3756c	MMAS_36320
	Rv3446c	REQ_34460	RER_19010	ROP_62240	Nfa_8520	MAB_3755c	MMAS_36310
	EsxU	REQ_34450	RER_19020	ROP_62250	Nfa_8530	MAB_3754c	MMAS_36300
	EsxT	REQ_35440	RER_19030	ROP_62260	Nfa_8540	MAB_3753c	MMAS_36290
			8 ORF insertion in EccC, includes recombinase		Insertion Nfa_8480		MMAS_36360 is EccE
ESX-3	**EccA**				Nfa_8260	MAB_2234c	MMAS_20690
	EccB				Nfa_8250	MAB_2233c	MMAS_20680
	EccC				Nfa_8290	MAB_2232c	MMAS_20670
	PE5				Nfa_8180	MAB_2231c	MMAS_20660
	PPE4				Nfa_8200	MAB_2230c	MMAS_20650
	EsxG	ABSENT	ABSENT	ABSENT	Nfa_8220	MAB_2229c	MMAS_20640
	EsxH				Nfa_8230	MAB_2228c	MMAS_20630
	EspG				Nfa_8210	MAB_2227c	MMAS_20620
	EccD				Nfa_8270	MAB_2226c	MMAS_20610
	MycP				Nfa_8280	MAB_2225c	MMAS_20600
	EccE				Nfa_8240	MAB_2224c	MMAS_20590
					Nfa_8310 is EspI; 2 Insertions Nfa_8190, Nfa_8300		
ESX-1		ABSENT	ABSENT	ABSENT	ABSENT	ABSENT	ABSENT
ESX-2		ABSENT	ABSENT	ABSENT	ABSENT	ABSENT	ABSENT
ESX-5		ABSENT	ABSENT	ABSENT	ABSENT	ABSENT	ABSENT

Table S1.3. The components of the ESX gene clusters of *M. thermoresistibile, M. smegmatis, M. rhodesiae, M. sp.* JLS, *M. phlei* and *M. vanbaalenii*.

		M. thermoresistibile	*M. smegmatis*	*M. rhodesiae*	*M. sp.* JLS	*M. phlei*	*M. vanbaalenii*
ESX-4	**EccB**	KEK_22836	MSMEG_1534	MycrhNDRAFT_1287	Mjls_1150	MPHLEI_05352	Mvan_1444
	MycP	KEK_22841	MSMEG_1533	MycrhNDRAFT_1288	Mjls_1151	MPHLEI_05357	Mvan_1445
	EccD	KEK_22846	MSMEG_1535	MycrhNDRAFT_1289	Mjls_1152	MPHLEI_05362	Mvan_1446
	EccC	KEK_22851/ KEK_22856	MSMEG_1536	MycrhNDRAFT_1291	Mjls_1153	MPHLEI_05367	Mvan_1447/ Mvan_1462
	Rv3446c	KEK_22861	MSMEG_1537	MycrhNDRAFT_1292	Mjls_1154	MPHLEI_05372	Mvan_1463
	EsxU	KEK_22866	MSMEG_1538	MycrhNDRAFT_1293	Mjls_1155	MPHLEI_05377	Mvan_1464
	EsxT		MSMEG_1539	MycrhNDRAFT_1294	Mjls_1156	MPHLEI_05382	Mvan_1465
				insertion 1ORF MycrhNDRAFT_1290			Insertion in EccC, splits gene in 2; 14 ORFs including recombinase
ESX-3	**EccA**	KEK_04557	MSMEG_0615	MycrhNDRAFT_0493	Mjls_0359	MPHLEI_07122	Mvan_0411
	EccB	KEK_04552	MSMEG_0616	MycrhNDRAFT_0494	Mjls_0360	MPHLEI_07127	Mvan_0412
	EccC	KEK_04547	MSMEG_0617	MycrhNDRAFT_0495	Mjls_0361	MPHLEI_07132	Mvan_0413
	PE5	KEK_04542	MSMEG_0618	MycrhNDRAFT_0496	Mjls_0362	MPHLEI_07137	Mvan_0414
	PPE4	KEK_04537/ KEK_02676	MSMEG_0619	MycrhNDRAFT_0497	Mjls_0363	MPHLEI_07142	Mvan_0415
	EsxG	KEK_02671	MSMEG_0620	MycrhNDRAFT_0498	Mjls_0364	MPHLEI_07147	Mvan_0417
	EsxH	KEK_02666	MSMEG_0621	MycrhNDRAFT_0499	Mjls_0365	MPHLEI_07152	Mvan_0418
	EspG	KEK_02661	MSMEG_0622	MycrhNDRAFT_0500	Mjls_0366	MPHLEI_07157	Mvan_0419
	EccD	KEK_02656	MSMEG_0623	MycrhNDRAFT_0501	Mjls_0367	MPHLEI_07162	Mvan_0420
	MycP	KEK_02651	MSMEG_0624	MycrhNDRAFT_0502	Mjls_0368	MPHLEI_07167	Mvan_0421
	EccE	KEK_02646	MSMEG_0626	MycrhNDRAFT_0503	Mjls_0369	MPHLEI_07172	Mvan_0422
			MSMEG_0625 is outside of ESX-3				insertion 1 ORF Mvan_0416
ESX-1	**EspG**	KEK_11068	MSMEG_0057	MycrhNDRAFT_0049	Mjls_0052	MPHLEI_25571	Mvan_0070
	EspH		MSMEG_0058	MycrhNDRAFT_0048	Mjls_0053	MPHLEI_25576	Mvan_0071
	EccA	KEK_11078	MSMEG_0059	MycrhNDRAFT_0047	Mjls_0054	MPHLEI_25581	Mvan_0072
	EccB	KEK_11083	MSMEG_0060	MycrhNDRAFT_0046	Mjls_0055	MPHLEI_25586	Mvan_0073
	EccC	KEK_11088/ KEK_11093	MSMEG_0061/ MSMEG_0062	MycrhNDRAFT_0045/ MycrhNDRAFT_0044	Mjls_0056/ Mjls_0057	MPHLEI_25591	Mvan_0074/ Mvan_0075
	PE35	KEK_11098	MSMEG_0063	MycrhNDRAFT_0043	Mjls_0058	MPHLEI_25596	Mvan_0076
	PPE68	KEK_11103	MSMEG_0064	MycrhNDRAFT_0042	Mjls_0059	MPHLEI_25601	Mvan_0077
	EsxB	KEK_11108	MSMEG_0065	MycrhNDRAFT_0041	Mjls_0060	MPHLEI_25606	Mvan_0078
	EsxA	KEK_11113	MSMEG_0066	MycrhNDRAFT_0040	Mjls_0061	MPHLEI_25611	Mvan_0079
	EspI	KEK_11118	MSMEG_0067	MycrhNDRAFT_0039	Mjls_0062	MPHLEI_25616	Mvan_0080
	EccD	KEK_11123	MSMEG_0068	MycrhNDRAFT_0038	Mjls_0063	MPHLEI_25621	Mvan_0081
	EspJ	not annotated	MSMEG_0069				Mvan_0082
	EspK	KEK_05562	MSMEG_0071				Mvan_0085
	EspL	KEK_05542	MSMEG_0081	MycrhNDRAFT_0283	Mjls_0065	MPHLEI_25966	Mvan_0089
	EspB	not annotated	MSMEG_0076	MycrhNDRAFT_0284	Mjls_0066	MPHLEI_25971	Mvan_0090
	EccE	KEK_05527	MSMEG_0082	MycrhNDRAFT_0285	Mjls_0067	MPHLEI_25976	Mvan_0091
	MycP	KEK_05522	MSMEG_0083	MycrhNDRAFT_0286	Mjls_0068	MPHLEI_25981	Mvan_0092
		Transposition after EccD; 2 insertions; 3ORFs KEK_05557, KEK_05537. KEK_05532	3 insertions; 9 ORFs MSMEG_0070; MSMEG_0072-75; MSMEG_0077-80; EspB transposed and inverted	transposition after EccD, deletion of EspJ and EspK	insertion 1ORF Mjls_0064		2 insertions; 5ORFs Mvan_0083-84; Mvan_0086-88
ESX-2		ABSENT	ABSENT	ABSENT	ABSENT	ABSENT	ABSENT
ESX-5		ABSENT	ABSENT	ABSENT	ABSENT	ABSENT	ABSENT

Table S1.4. The components of the ESX gene clusters of *M. sp.* MCS, *M. sp.* KMS, *M. gilvum* and *M. chubuense*.

		M. sp. MCS	*M. sp.* KMS	*M. gilvum*	*M. chubuense*
ESX-4	**EccB**	Mmcs_1123	Mkms_1140	Mflv_4962	Mycch_1049
	MycP	Mmcs_1124	Mkms_1141	Mflv_4961	Mycch_1050
	EccD	Mmcs_1125	Mkms_1142	Mflv_4960	Mycch_1051
	EccC	Mmcs_1126	Mkms_1143	Mflv_4959	Mycch_1052
	Rv3446c	Mmcs_1127	Mkms_1144	Mflv_4958	Mycch_1053
	EsxU	Mmcs_1128	Mkms_1145	Mflv_4957	Mycch_1054
	EsxT	Mmcs_1129	Mkms_1146	Mflv_4956	Mycch_1055
ESX-3	**EccA**	Mmcs_0370	Mkms_0380	Mflv_0329	
	EccB	Mmcs_0371	Mkms_0381	Mflv_0328	
	EccC	Mmcs_0372	Mkms_0382	Mflv_0327	
	PE5	Mmcs_0373	Mkms_0383	Mflv_0326	
	PPE4	Mmcs_0374	Mkms_0384	Mflv_0325	Deleted
	EsxG	Mmcs_0376	Mkms_0385	Mflv_0324	
	EsxH	Mmcs_0377	Mkms_0386	Mflv_0323	
	EspG	Mmcs_0378	Mkms_0387	Mflv_0322	
	EccD	Mmcs_0379	Mkms_0388	Mflv_0321	
	MycP	Mmcs_0380	Mkms_0389	Mflv_0320	Mycch_0833
	EccE	Mmcs_0381	Mkms_0390	Mflv_0319	Deleted
		insertion; 1 ORF Mmcs_0375 transposase			
ESX-1	**EspG**	Mmcs_0062	Mkms_0071	Mflv_0775	Not annotated
	EspH	Mmcs_0063	Mkms_0072	Mflv_0774	
	EccA	Mmcs_0064	Mkms_0073	Mflv_0773	
	EccB	Mmcs_0065	Mkms_0074	Mflv_0772	
	EccC	Mmcs_0066/ Mmcs_0067	Mkms_0075/ Mkms_0076	Mflv_0771/ Mflv0770	Deleted
	PE35	Mmcs_0068	Mkms_0077	Mflv_0769	
	PPE68	Mmcs_0069	Mkms_0078	Mflv_0768	
	EsxB	Mmcs_0070	Mkms_0079	Mflv_0767	
	EsxA	Mmcs_0071	Mkms_0080	Mflv_0766	
	EspI	Mmcs_0072	Mkms_0081	Mflv_0765	Mycch_0072
	EccD	Mmcs_0073	Mkms_0082	Mflv_0764	Mycch_0073
	EspJ			Mflv_0763	Mycch_0074
	EspK			Mflv_0760	Mycch_0076
	EspL	Mmcs_0075	Mkms_0084	Mflv_0758	Mycch_0079
	EspB	Mmcs_0076	Mkms_0085	Mflv_0757	Mycch_0080
	EccE	Mmcs_0077	Mkms_0086	Mflv_0756	Mycch_0081
	MycP	Mmcs_0078	Mkms_0087	Mflv_0757	Mycch_0082
		Insertion 1ORF; Mmcs_0074 replaces EspJ and EspK	Insertion 1ORF; Mkms_0083 replaces EspJ and EspK	2 insertions; 3 ORFs Mflv_0762-61, Mflv_0759	2 insertions 2 ORFs, Mycch0075, Mycch_0078 and Mycch_0077 EspK-like pseudogene
ESX-2		ABSENT	ABSENT	ABSENT	ABSENT
ESX-5		ABSENT	ABSENT	ABSENT	ABSENT
Plasmid ESX	**EccA**	Mmcs_5555	Mkms_5958	Mflv_5463	Mycch_5525 Mycch_6017
	EccB	Mmcs_5556	Mkms_5778 Mkms_5959	Mflv_5462	Mycch_5548 Mycch_6040
	EccC	Mmcs_5557	Mkms_5772 Mkms_5960	Mflv_5461	Mycch_5547 Mycch_6039
	EccD	Mmcs_5564	Mkms_5780 Mkms_5967	Mflv_5455	Mycch_5539 Mycch_6031
	EccE	Mmcs_5562	Mkms_5782 Mkms_5965	Mflv_5453	Mycch_5537 Mycch_6029
	pe	Mmcs_5558	Mkms_5773 Mkms_5961	Mflv_5460	Mycch_5545 Mycch_6037
	ppe	Mmcs_5559	Mkms_5774 Mkms_5962	Mflv_5459	Mycch_5544 Mycch_6036
	esx	Mmcs_5560	Mkms_5775 Mkms_5963	Mflv_5458	Mycch_5543 Mycch_6035
	esx	Mmcs_5561	Mkms_5776 Mkms_5964	Mflv_5457	Mycch_5542 Mycch_6034
	EspG	Mmcs_5566	Mkms_5777 Mkms_5969	Mflv_5539	Mycch_5541 Mycch_6033
	EspI	Mmcs_5565	Mkms_5779 Mkms_5968	Mflv_5456	Mycch_5540 Mycch_6032
	MycP	Mmcs_5563	Mkms_5782 Mkms_5966	Mflv_5454	Mycch_5538 Mycch_6030
	EspH			Mflv_5464	
		Plasmid01 **ESX-P3**	pMKMS01 **ESX-P2** pMKMS02 **ESX-P3**	pMFLV01; EspG transposed; Insertion 1 ORF Mycch0046 **ESX-P1**	pMYCCH.01; 2 insertions 1 ORF Mycch_5546 and 11 ORFs **ESX-P4** pMYCCH.02, 2 insertions 1 ORF Mycch_6038 and 11 ORFs **ESX-P4**

Table S1.5. The components of the ESX gene clusters of *M. sp.* Spyr1 and *M. tusciae*.

		M. sp. Spyr1	*M. tusciae*			
ESX-4	**EccB**	Mspyr1_43790	MyctuDRAFT_4371			
	MycP	Mspyr1_43780	MyctuDRAFT_4372			
	EccD	Mspyr1_43770	MyctuDRAFT_4373			
	EccC	Mspyr1_43760	MyctuDRAFT_4374			
	Rv3446c	Mspyr1_43750	MyctuDRAFT_4375			
	EsxU	Mspyr1_43740	MyctuDRAFT_4376			
	EsxT	Mspyr1_43730	MyctuDRAFT_4377			
ESX-3	**EccA**	Mspyr1_04290	MyctuDRAFT_0854			
	EccB	Mspyr1_04300	MyctuDRAFT_0855			
	EccC	Mspyr1_04310	MyctuDRAFT_0856			
	PE5	Mspyr1_04320	MyctuDRAFT_0857			
	PPE4	Mspyr1_04330	MyctuDRAFT_0858			
	EsxG	Mspyr1_04340	MyctuDRAFT_0859			
	EsxH	Mspyr1_04350	MyctuDRAFT_0860			
	EspG	Mspyr1_04360	MyctuDRAFT_0861			
	EccD	Mspyr1_04370	MyctuDRAFT_0862			
	MycP	Mspyr1_04380	MyctuDRAFT_0863			
	EccE	Mspyr1_04390	MyctuDRAFT_0864			
ESX-1	**EspG**	Mspyr1_00560	MyctuDRAFT_1521			
	EspH	Mspyr1_00570	MyctuDRAFT_1522			
	EccA	Mspyr1_00580	MyctuDRAFT_1523			
	EccB	Mspyr1_00590	MyctuDRAFT_1524			
	EccC	Mspyr1_00600/ Mspyr1_00610	MyctuDRAFT_1525/ MyctuDRAFT_1526			
	PE35	Mspyr1_00620	MyctuDRAFT_1528			
	PPE68	Mspyr1_00630	MyctuDRAFT_1529			
	EsxB	Mspyr1_00640	MyctuDRAFT_1530			
	EsxA	Mspyr1_00650	MyctuDRAFT_1531			
	EspI	Mspyr1_00660	MyctuDRAFT_1532			
	EccD	Mspyr1_00670	MyctuDRAFT_1533			
	EspJ	Mspyr1_00680	MyctuDRAFT_1539			
	EspK	Mspyr1_00710	MyctuDRAFT_1536			
	EspL	Mspyr1_00730	MyctuDRAFT_1871			
	EspB	Mspyr1_00740	MyctuDRAFT_1870			
	EccE	Mspyr1_00750	MyctuDRAFT_1869			
	MycP	Mspyr1_00760	MyctuDRAFT_1868			
		2 insertions; 3 ORFs Mspyr1_00690-700; Mspyr1_00720	Insertion 1ORF MtctuDRAFT_1527			
ESX-2	**EccB**		MyctuDRAFT_5541	MyctuDRAFT_2371	MyctuDRAFT_5597	MyctuDRAFT_3298
	EccC		MyctuDRAFT_5540/ MyctuDRAFT_5539	MyctuDRAFT_2365	MyctuDRAFT_5591	MyctuDRAFT_3299
	PE36		MyctuDRAFT_5538	MyctuDRAFT_2366	MyctuDRAFT_5592	MyctuDRAFT_3300
	PPE69		MyctuDRAFT_5537	MyctuDRAFT_2367	MyctuDRAFT_5593	MyctuDRAFT_3301
	EsxD		MyctuDRAFT_5536	MyctuDRAFT_2368	MyctuDRAFT_5594	MyctuDRAFT_3302
	EsxC	ABSENT	MyctuDRAFT_5535	MyctuDRAFT_2369	MyctuDRAFT_5595	MyctuDRAFT_3303
	EspG		MyctuDRAFT_5534	MyctuDRAFT_2370	MyctuDRAFT_5596	MyctuDRAFT_3309
	EspI		MyctuDRAFT_5533	MyctuDRAFT_2372	MyctuDRAFT_5598	MyctuDRAFT_3308
	EccD		MyctuDRAFT_5532	MyctuDRAFT_2373	MyctuDRAFT_5599	MyctuDRAFT_3307
	MycP		MyctuDRAFT_5531	MyctuDRAFT_2374	MyctuDRAFT_5600	MyctuDRAFT_3306
	EccE		MyctuDRAFT_5530	MyctuDRAFT_2375	MyctuDRAFT_5601	MyctuDRAFT_3305
	EccA		MyctuDRAFT_5529	MyctuDRAFT_2589	MyctuDRAFT_55615	MyctuDRAFT_3294
				EccA transposed	EccA transposed	Insertion MyctuDRAFT_3304; EspG-EccE inverted
			CONTIG 212 **ESX-2$_{AN}$**	CONTIG 196 **ESX-P2**	CONTIG 209 **ESX-P2**	CONTIG 224 **ESX-P3**
ESX-5		ABSENT	ABSENT			

Table S1.6. The components of the ESX gene clusters of *M. kansasii*, *M. tuberculosis*, *M. africanum*, *M. bovis*, *M. canetti*, *M. bovis* BCG.

		M. kansasii	*M. tuberculosis*	*M. africanum*	*M. bovis*	*M. canetti*	*M. bovis* BCG
ESX-4	**EccB**	MkanA1_00645	Rv3450c	MAF_34650	Mb3480c	MCAN_34671	BCG3516c
	MycP	MkanA1_00650	Rv3449	MAF_34640	Mb3479	MCAN_34661	BCG3515
	EccD	MkanA1_00655	Rv3448	MAF_34630	Mb3478	MCAN_34651	BCG3514
	EccC	MkanA1_00660	Rv3447c	MAF_34620	Mb3477c	MCAN_34641	BCG3513c
	Rv3446c	MkanA1_00665	Rv3446c	MAF_34610	Mb3476c	MCAN_34631	BCG3512c
	EsxU	MkanA1_00670	Rv3445c	MAF_34600	Mb3475c	MCAN_34621	BCG3511c
	EsxT	MkanA1_00675	Rv3444c	MAF_34590	Mb3474c	MCAN_34611	BCG3510c
ESX-3	**EccA**	MkanA1_27591	Rv0282	MAF_02840	Mb0290	MCAN_02911	BCG0322
	EccB	MkanA1_27596	Rv0283	MAF_02850	Mb0291	MCAN_02921	BCG0323
	EccC	MkanA1_27601	Rv0284	MAF_02860	Mb0292	MCAN_02931	BCG0324
	PE5	MkanA1_27606	Rv0285	MAF_02870	Mb0293	MCAN_02941	BCG0325
	PPE4	MkanA1_27611/ MkanA1_25662	Rv0286	MAF_02880	Mb0294	MCAN_02951	BCG0326
	EsxG	MkanA1_25667	Rv0287	MAF_02890	Mb0295	MCAN_02961	BCG0327
	EsxH	MkanA1_25672	Rv0288	MAF_02900	Mb0296	MCAN_02971	BCG0328
	EspG	MkanA1_25677	Rv0289	MAF_02910	Mb0297	MCAN_02981	BCG0329
	EccD	MkanA1_25682	Rv0290	MAF_02920	Mb0298	MCAN_02991	BCG0330
	MycP	MkanA1_25687	Rv0291	MAF_02930	Mb0299	MCAN_03001	BCG0331
	EccE	MkanA1_25692	Rv0292	MAF_02940	Mb0300	MCAN_03011	BCG0332
ESX-1	**EspG**	MkanA1_03472	Rv3866	MAF_38810	Mb3896	MCAN_38881	BCG3929
	EspH	MkanA1_03477	Rv3867	MAF_38820	Mb3897	MCAN_38891	BCG3930
	EccA	MkanA1_03482	Rv3868	MAF_38830	Mb3898	MCAN_38901	BCG3931
	EccB	MkanA1_03487	Rv3869	MAF_38840	Mb3899	MCAN_38911	BCG3932
	EccC	MkanA1_03492/ MkanA1_03497	Rv3870/ Rv3871	MAF_38850/ MAF_38860	Mb3900/ Mb3901	MCAN_38921/ MCAN_38931	BCG3933/ BCG3934
	PE35	MkanA1_03502	Rv3872	MAF_38870	Mb3902	MCAN_38941	BCG3934
	PPE68	MkanA1_03507	Rv3873	MAF_38880	Mb3903	MCAN_38951	
	EsxB	MkanA1_03512	Rv3874	MAF_38890	Mb3904	MCAN_38961	RD1 deletion
	EsxA	MkanA1_03517	Rv3875	MAF_38900	Mb3905	MCAN_38971	
	EspI	MkanA1_03527	Rv3876	MAF_38910	Mb3906	MCAN_38981	
	EccD	MkanA1_03532	Rv3877	MAF_38920	Mb3907	MCAN_38991	
	EspJ	MkanA1_03537	Rv3878	MAF_38930	Mb3908	MCAN_39001	
	EspK	MkanA1_03542	Rv3879c	MAF_38940	Mb3909c	MCAN_39011	BCG3935c
	EspL	MkanA1_03547	Rv3880c	MAF_38950	Mb3910c	MCAN_39021	BCG3936c
	EspB	MkanA1_03552	Rv3881c	MAF_38960	Mb3911c	MCAN_39031	BCG3936c
	EccE	MkanA1_03557	Rv3882c	MAF_38970	Mb3912c	MCAN_39041	BCG3937c
	MycP	MkanA1_03562	Rv3883c	MAF_38980	Mb3913c	MCAN_39051	BCG3938c
		Insertion MkanA1_03522					
ESX-2	**EccB**	NA	Rv3895c	MAF_39100	Mb3925c	MCAN_39171	BCG3952c
	EccC	MkanA1_03622	Rv3894c	MAF_39090	Mb3924c	MCAN_39161	BCG3951c/BCG3950c
	PE36	MkanA1_03612	Rv3893c	MAF_39080	Mb3923c	MCAN_39151	BCG3949c
	PPE69	MkanA1_03607	Rv3892c	MAF_39070	Mb3922c	MCAN_39141	BCG3948c
	EsxD	MkanA1_03602	Rv3891c	MAF_39060	Mb3921c	MCAN_39131	BCG3947c
	EsxC	MkanA1_03597	Rv3890c	MAF_39050	Mb3920c	MCAN_39121	BCG3946c
	EspG	MkanA1_03592	Rv3889c	MAF_39040	Mb3919c	MCAN_39111	BCG3945c
	EspI	MkanA1_03587	Rv3888c	MAF_39030	Mb3918c	MCAN_39101	BCG3944c/ BCG3943c
	EccD	MkanA1_03582	Rv3887c				
	MycP	MkanA1_03577	Rv3886c	MAF_39020	Mb3917c	MCAN_39091	BCG3942c
	EccE	MkanA1_03572	Rv3885c	MAF_39010	Mb3916c	MCAN_39081	BCG3941c
	EccA	MkanA1_03567	Rv3884c	MAF_39000	Mb3915c	MCAN_39071	BCG3940c
		Insertion MkanA1_03617		MAF_38990	Mb3914c	MCAN_39061	BCG3939c
ESX-5	**EccB**	MkanA1_22060	Rv1782	MAF_18040	Mb1811	MCAN_18001	BCG1815
	EccC	MkanA1_22065	Rv1783/Rv1784	MAF_18050/ MAF_18060	Mb1812	MCAN_18011/ MCAN_18021	BCG1816
	Cyp143	MkanA1_22070	Rv1785c	MAF_18070	Mb1813c	MCAN_18031	BCG1817
	Ferredoxin	MkanA1_22075	Rv1786	MAF_18080	Mb1814	MCAN_18041	BCG1818
	PPE25	MkanA1_22080	Rv1787	MAF_18090	Mb1815	MCAN_18051	BCG1819
	PE18	MkanA1_22085	Rv1788	MAF_18100	Mb1816	MCAN_18061	BCG1820
	PPE26	MkanA1_22090	Rv1789	MAF_18110	Mb1817	MCAN_18071	BCG1821
	PPE27		Rv1790	MAF_18120	Mb1818	MCAN_18081	BCG1822
	PE19	MkanA1_29706	Rv1791	MAF_18130	Mb1819	MCAN_18091	BCG1823
	EsxM	MkanA1_27436	Rv1792	MAF_18140	Mb1820	MCAN_18101	BCG1824
	EsxN	MkanA1_27431	Rv1793	MAF_18150	Mb1821	MCAN_18111	BCG1825
	EspG	MkanA1_26312	Rv1794	MAF_18160	Mb1822	MCAN_18121	BCG1826
	EccD	MkanA1_26317	Rv1795	MAF_18170	Mb1823	MCAN_18131	BCG1827
	MycP	MkanA1_26322	Rv1796	MAF_18180	Mb1824	MCAN_18141	BCG1828
	EccE	not annotated	Rv1797	MAF_18190	Mb1825	MCAN_18151	BCG1829
	EccA	MkanA1_26332	Rv1798	MAF_18200	Mb1826	MCAN_18161	BCG1830
		Insertion MkanA1_26327					

Table S1.7. The components of the ESX gene clusters of *M. xenopi, M. sp.* JDM601, *M. avium, M. avium paratuberculosis*, *M. intracellulare* and *M. colombiense*.

		M. xenopi	*M. sp.* JDM601	*M. avium*	*M. avium paratuberculosis*	*M. intracellulare*	*M. colombiense*
ESX-4	**EccB**			MAV_4392	MAP_4238	MintA_04831	MCOL_23415
	MycP	MXEN_02879	JDM601_3275	MAV_4393	MAP_4239c	MintA_04836	MCOL_23420
	EccD	MXEN_02874	JDM601_3274	MAV_4391	MAP_4240c	MintA_04841	MCOL_23425
	EccC	MXEN_02869	JDM601_3273	MAV_4390	MAP_4241	MintA_04846	MCOL_23430
	Rv3446c	MXEN_02864	JDM601_3272	MAV_4389	MAP_4242	MintA_04851/ MintA_17752	MCOL_23435
	EsxU	MXEN_02859	JDM601_3271	MAV_4388	MAP_4243	MintA_17757	MCOL_23440
	EsxT	MXEN_02854	JDM601_3270	MAV_4387	MAP_4244	MintA_17762	MCOL_23445
ESX-3	**EccA**	MXEN_19795	JDM601_0290	MAV_4871	MAP_3778	MintA_03216	MCOL_10158
	EccB	MXEN_19800	JDM601_0291	MAV_4870	MAP_3779	MintA_03211	MCOL_10153
	EccC	MXEN_19805	JDM601_0292	MAV_4869	MAP_3780	MintA_03206	MCOL_10148
	PE5	MXEN_19810	JDM601_0293	MAV_4868	MAP_3781	MintA_03201	MCOL_10143
	PPE4	MXEN_19815	JDM601_0294	MAV_4867	MAP_3782	MintA_06869/ MintA_16847	MCOL_10138
	EsxG	MXEN_19820	JDM601_0295	MAV_4866	MAP_3783	MintA_16847	MCOL_10133
	EsxH	MXEN_19825	JDM601_0296	MAV_4865	MAP_3784	MintA_16842	MCOL_10128
	EspG	MXEN_19830	JDM601_0297	MAV_4864	MAP_3785	MintA_16837	MCOL_10123
	EccD	MXEN_19835	JDM601_0298	MAV_4863	MAP_3786	MintA_16832	MCOL_10118
	MycP	MXEN_19840	JDM601_0299	MAV_4862	MAP_3787	MintA_16827	MCOL_10108
	EccE	MXEN_19845	JDM601_0300	MAV_4860	MAP_3788	MintA_16822	MCOL_10103
				MAV_4861 falls outside of ESX-3			Insertion MCOL_10113
ESX-1		ABSENT	ABSENT	ABSENT	ABSENT	ABSENT	ABSENT
ESX-2	**EccB**	MXEN_05555	JDM601_0058	MAV_5285	MAP_4321c	MintA_05824	MCOL_25658
	EccC	MXEN_05560	JDM601_0057	MAV_5283	MAP4322c/ MAP4323c	MintA_05829	MCOL_25663
	PE36	MXEN_05565	JDM601_0056	MAV_0151	MAP_0157	MintA_21719	MCOL_14655
	PPE69	MXEN_05570	JDM601_0055	MAV_0152	MAP_0158	MintA_21714/ MintA_15042	MCOL_14650
	EsxD	MXEN_05575	JDM601_0054	MAV_0153	MAP_0160	MintA_15047	MCOL_14645
	EsxC	MXEN_05580	JDM601_0053	MAV_0154	MAP_0161	MintA_15052	MCOL_14640
	EspG	MXEN_05585	JDM601_0052	MAV_0155	MAP_0162	MintA_15057	MCOL_14635
	EspI	MXEN_05590	JDM601_0051	MAV_0156	MAP_0163	MintA_15062	MCOL_14630
	EccD	MXEN_05595	JDM601_0050	MAV_0157	MAP_0164	MintA_15067	MCOL_14625
	MycP	MXEN_05600	JDM601_0049	MAV_0158	MAP_0165 MAP_0165 MAP_0166	MintA_15072/ MintA_18212	MCOL_14620
	EccE	MXEN_05605	JDM601_0048	MAV_0159	MAP_0167	MintA_18217	MCOL_14615
	EccA	MXEN_05610	JDM601_0047	MAV_0160	MAP_0167	MintA_18222	MCOL_14610
				Transposition of EccB and EccC	Transposition of EccB and EccC; EccB split into two genes; Insertion MAP_0159c	Transposition of EccB and EccC	Transposition of EccB and EccC
ESX-5	**EccB**	MXEN_04109	MJDM601_1892	MAV_2933	MAP_1501	MintA_10176	MCOL_20626
	EccC	MXEN_04114	MJDM601_1893	MAV_2932	MAP_1502	MintA_10181	MCOL_20621
	Cyp143	MXEN_04134		MAV_2931	MAP_1503c	MintA_10186	MCOL_20616
	Ferredoxin	MXEN_04139		MAV_2929	MAP_1504	MintA_10191	MCOL_20611
	PPE25	MXEN_04144/ MXEN_09344		MAV_2928		MintA_10196	MCOL_20606
	PE18	MXEN_09349		MAV_2927		MintA_10201	MCOL_20601
	PPE26	MXEN_09354		MAV_2926	MAP_1505	MintA_10206	MCOL_20596
	PPE27	MXEN_09369		MAV_2925	MAP_1506		MCOL_20591
	PE19	MXEN_09364	MJDM601_1894	MAV_2923	MAP_1507		MCOL_20576
	EsxM	MXEN_09374/ MXEN_01242	MJDM601_1895	MAV_2922	MAP_1508	MintA_21674	MCOL_20571
	EsxN	MXEN_01237	MJDM601_1896	MAV_2921	NA	MintA_21669	MCOL_20566
	EspG	MXEN_09204	MJDM601_1897	MAV_2920	MAP_1509	MintA_10221	MCOL_20561
	EccD	MXEN_09209	MJDM601_1898	MAV_2919	MAP_1510	MintA_10226	MCOL_20556
	MycP	MXEN_09214	MJDM601_1899	MAV_2918	MAP_1511	MintA_10231/ MintA_13857	MCOL_20551
	EccE	MXEN_09219	MJDM601_1900	MAV_2917	MAP_1512	MintA_13852	MCOL_20546
	EccA	MXEN_09224	MJDM601_1901	MAV_2916	MAP_1513	MintA_13847	MCOL_20541
		2 insertions 4 ORFs MXEN-04119-04134; MXEN_09359	DNA sequence spanning absent genes could not be identified in available contigs	Insertion MAV_2930; additional PPE MAV_2924			Duplication of 2 PPE genes: MCOL_20586 and MCOL_20581

Table S1.8. The components of the ESX gene clusters of *M. parascrofulaceum* and *M. sp.* MOTT36Y.

		M. parascrofulaceum				*M. sp.* MOTT36Y
ESX-4	**EccB**		HMPREF0591_3880			W7S_21385
	MycP		HMPREF0591_3881			W7S_21380
	EccD		HMPREF0591_3882			W7S_21375
	EccC		HMPREF0591_3883			W7S_21370
	Rv3446c		HMPREF0591_3884/HMPREF0591_1773			W7S_21365
	EsxU		HMPREF0591_1774			W7S_21360
	EsxT		HMPREF0591_1775			W7S_21355
ESX-3	**EccA**		HMPREF0591_3351			W7S_23965
	EccB		HMPREF0591_3352			W7S_23960
	EccC		HMPREF0591_3353			W7S_23955
	PE5		HMPREF0591_3354			W7S_23950
	PPE4		HMPREF0591_3355/HMPREF0591_1873			W7S_23945
	EsxG		HMPREF0591_1872			W7S_23940
	EsxH		HMPREF0591_1871			W7S_23935
	EspG		HMPREF0591_1870			W7S_23930
	EccD		HMPREF0591_1869/HMPREF0591_3356			W7S_23925
	MycP		HMPREF0591_3357			W7S_23920
	EccE		HMPREF0591_3358			W7S_23915
ESX-1			ABSENT			ABSENT
ESX-2	**EccB**	HMPREF0591_3018	HMPREF0591_1319	HMPREF0591_1425	HMPREF0591_0084	W7S_25690
	EccC	HMPREF0591_3019/ HMPREF0591_3020	HMPREF0591_1318	HMPREF0591_1432	HMPREF0591_0085/ HMPREF0591_0086	W7S_25685
	PE36	HMPREF0591_2863	HMPREF0591_1317	HMPREF0591_1431		W7S_00765
	PPE69	HMPREF0591_2862	HMPREF0591_1316	HMPREF0591_1430	HMPREF0591_0089	W7S_00770
	EsxD	HMPREF0591_2861	HMPREF0591_1315	HMPREF0591_1429	HMPREF0591_0090	W7S_00775
	EsxC	HMPREF0591_2860	HMPREF0591_1314	HMPREF0591_1428	HMPREF0591_0091	W7S_00780
	EspG	HMPREF0591_2859	HMPREF0591_1313	HMPREF0591_1427	HMPREF0591_0092	W7S_00785
	EspI	HMPREF0591_2858	HMPREF0591_1312	HMPREF0591_1426	HMPREF0591_0093	W7S_00790
	EccD	HMPREF0591_2857	HMPREF0591_1311	HMPREF0591_1424	HMPREF0591_0094	W7S_00795
	MycP	HMPREF0591_2856/ HMPREF0591_2855	HMPREF0591_1310	HMPREF0591_1423	HMPREF0591_0095	W7S_00800 W7S_00805
	EccE	HMPREF0591_2854	HMPREF0591_1309	HMPREF0591_1422	HMPREF0591_0096	W7S_00810
	EccA	HMPREF0591_2853	HMPREF0591_1308			
		Transposition of EccA and EccB; MycP split into 2 genes; CONTIG 312/318 **ESX-2**	CONTIG 109 **ESX-P5**	EccB inserted between EspI and EccD; CONTIG 115 **ESX-P2**	insertion; 2 ORFs HMPREF0591_0087-88; CONTIG 17 **ESX-P4**	Transposition of EccB and EccC
ESX-5	**EccB**		HMPREF0591_0777			W7S_13390
	EccC		HMPREF0591_0778/HMPREF0591_0779			W7S_13385 W7S_13380
	Cyp143		HMPREF0591_0780			W7S_13375
	Ferredoxin		HMPREF0591_0781			
	PPE25		HMPREF0591_0782			
	PE18		HMPREF0591_5626			
	PPE26					W7S_13370
	PPE27					W7S_13365
	PE19					W7S_13360
	EsxM		HMPREF0591_5624			W7S_13355
	EsxN		HMPREF0591_0783			W7S_13350
	EspG		HMPREF0591_0784			W7S_13345
	EccD		HMPREF0591_0785			W7S_13340
	MycP		HMPREF0591_0786			W7S_13335
	EccE		HMPREF0591_0787			W7S_13330
	EccA		HMPREF0591_0788			W7S_13325
			region extending between PPE25 and EsxN is unclear			

Table S1.9. The components of the ESX gene clusters of *M. ulcerans*, *M. marinum* and *M. leprae*.

		M. ulcerans	*M. marinum*	*M. leprae*
ESX-4	**EccB**	Mul_0857	MMAR1099	
	MycP	Mul_0858	MMAR1100	
	EccD	Mul_0859	MMAR1101	
	EccC	Mul_0860	MMAR1102	
	Rv3446c	Mul_0861	MMAR1103	
	EsxU	Mul_0862	MMAR1104	
	EsxT	Mul_0863	MMAR1105	
ESX-3	**EccA**	Mul_1204	MMAR0541	ML2537c
	EccB	Mul_1205	MMAR0542	ML2536c
	EccC	Mul_1206	MMAR0543	ML2535c
	PE5	Mul_1207	MMAR0544	ML2534c
	PPE4	Mul_1208	MMAR0545	ML2533c (PS)
	EsxG	Mul_1209	MMAR0546	ML2532c
	EsxH	Mul_1210	MMAR0547	ML2531c
	EspG	Mul_1211	MMAR0548	ML2530c
	EccD	Mul_1212	MMAR0549	ML2529c
	MycP	Mul_1213	MMAR0550	ML2528c
	EccE	Mul_1214	MMAR0551	ML2527c
ESX-1	**EspG**		MMAR5441	ML0057c (PS)
	EspH		MMAR5442	ML0056c
	EccA		MMAR5443	ML0055c
	EccB		MMAR5444	ML0054c
	EccC		MMAR5445/MMAR5446	ML0053c
	PE35		MMAR5447	ML0052c
	PPE68		MMAR5448	ML0051c
	EsxB		MMAR5449	ML0050c
	EsxA		MMAR5450	ML0049c
	EspI		MMAR5451	ML0048c
	EccD		MMAR5452	ML0047c
	EspJ		MMAR5453	ML0046c (PS)
	EspK		MMAR5455c	ML0045 (PS)
	EspL		MMAR5456c	ML0044
	EspB		MMAR5457c	ML0043 (PS)
	EccE		MMAR5458c	ML0042
	MycP		MMAR5459c	ML0041
			Insertion MMAR5454c (EspE)	
ESX-2		ABSENT	ABSENT	ABSENT
ESX-5	**EccB**	Mul_3093	MMAR2664	ML1544c
	EccC	Mul_3092	MMAR2665	ML1543c
	Cyp143	Mul_3091	MMAR2666	ML1542 (PS)
	Ferredoxin	Mul_3090	MMAR2668	ML1541c (PS)
	PPE25	Mul_3089	MMAR2669	
	PE18	Mul_3088	MMAR2670	
	PPE26	Mul_3087	MMAR2671	
	PPE27			
	PE19	Mul_3086	MMAR2673	
	EsxM	Mul_3083	MMAR2674	
	EsxN	Mul_3082	MMAR2675	
	EspG	Mul_3081	MMAR2676	ML1540c
	EccD	Mul_3080	MMAR2677	ML1539c
	MycP	Mul_3079	MMAR2678	ML1538c
	EccE	Mul_3078	MMAR2679	ML1537c
	EccA	Mul_3077	MMAR2680	ML1538c
		Insertion 2 ORFs including IS2404 transposase	2 insertions MMAR_2667 and MMAR_2672	

2. The regulation of the mycobacterial Type VII ESX secretion systems

2.1. Introduction

Identification of the promoters which drive expression of the ESX secretion systems may assist in the identification of their regulatory mechanisms and functions. Some promoters have been identified upstream of, and within the ESX gene clusters. They include the P_{ESREG3} promoters upstream of ESX-3,[50, 114, 153] the P_{esat-6} promoter upstream of the esxB-esxA operon (transcribing ESAT-6 and CFP-10) in ESX-1[36] and the P_{AN} promoter upstream of the esxD-esxC operon in ESX-2.[154] However, a comprehensive search for the promoters driving expression of all 5 ESX gene clusters has not been done. In this study we have identified intergenic promoters from the *M. tuberculosis* ESX gene clusters using β-galactosidase promoter probe assays and *in silico* promoter prediction.

2.2. Results

2.2.1.ESAT-6 gene cluster organisation

The ESAT-6 gene clusters of *M. tuberculosis* contain variable numbers of genes, in different arrangements. The positions of the genes in *M. tuberculosis* H37Rv were obtained from TubercuList,[155] and the number of base pairs separating the genes determined. Intergenic regions of 40bp and larger were selected for analysis to determine the presence of promoters driving transcription of these gene cluster regions. Two intergenic regions were selected from ESX-4, one of which separates two genes in opposite orientations, and the sequence was therefore analysed in both directions. Four intergenic regions were selected from ESX-3, including one directly upstream of the gene cluster. Six intergenic regions were selected from ESX-1, seven from ESX-2, and nine from ESX-5. These include intergenic regions, regions directly upstream of ESX-1 and -5 and intergenic regions between additional genes further upstream of the gene clusters. The positions and sizes of the selected intergenic regions are indicated in Figure 2.1.

2.2.2.Promoter identification

To determine which intergenic regions contain promoter elements which allow transcription of the ESX gene clusters, the selected intergenic regions were amplified from *M. tuberculosis* H37Rv genomic DNA and cloned into the *E. coli*-mycobacterial promoter-probe vector pJEM15. The pJEM constructs were transformed into *M. smegmatis*, which is used as a mycobacterial host organism for promoter assays.

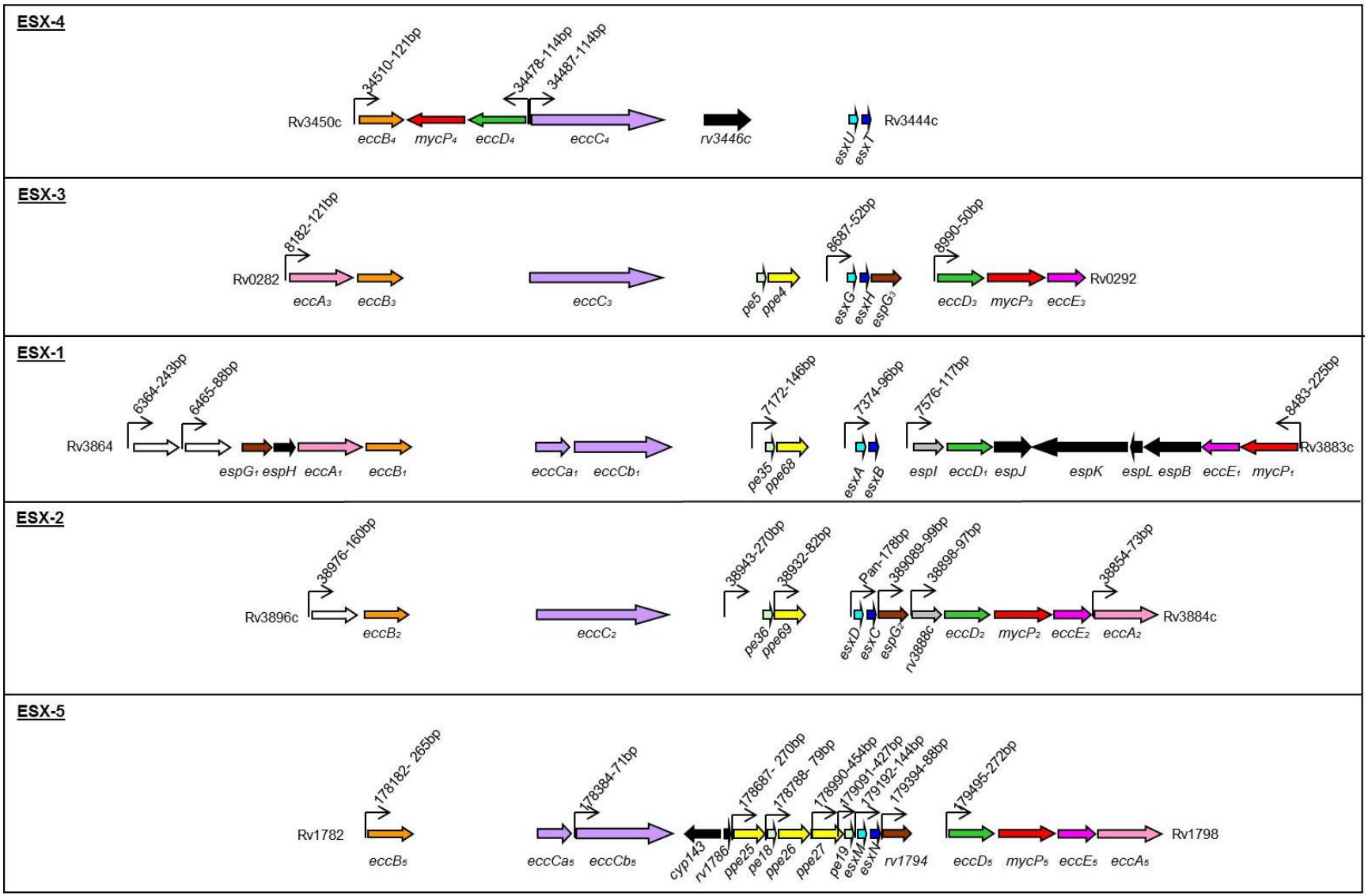
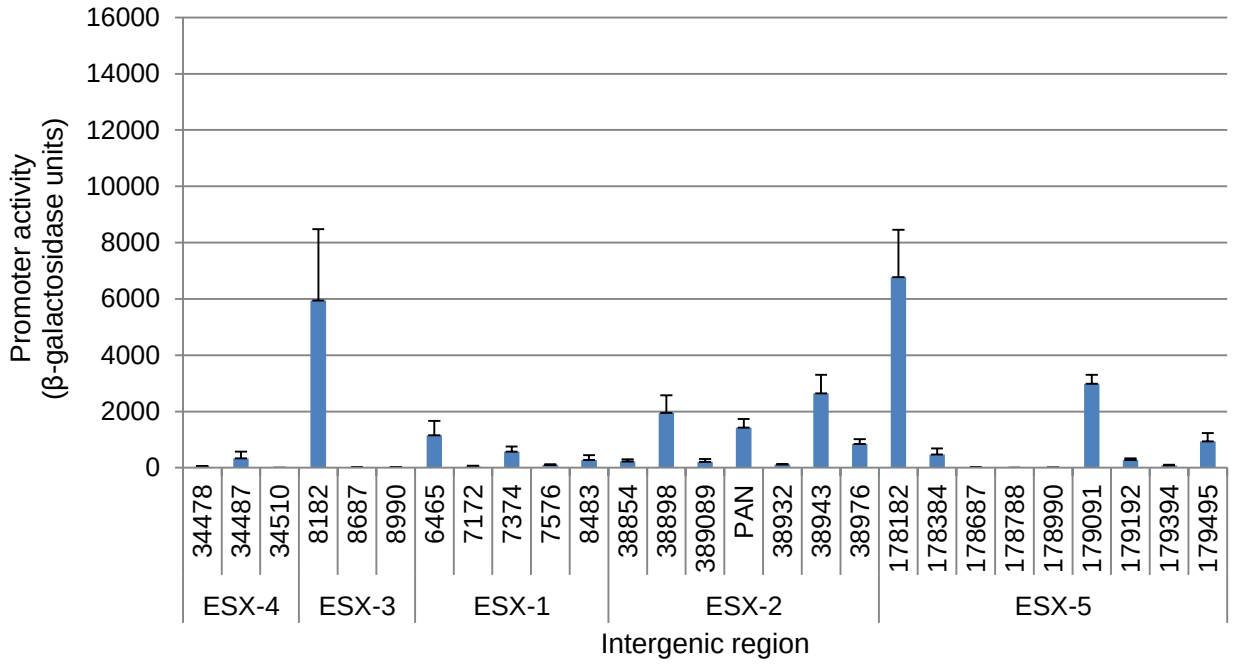


Figure 2.1. Selected intergenic regions of the *M. tuberculosis* ESX gene clusters. Intergenic regions greater than 40bp, which were selected for analysis, are indicated as arrows representing the direction in which the DNA was cloned into pJEM15. White block arrows represent genes outside of the ESX gene clusters. Spaces are not representative of the real distances between genes.

Promoter activity was assayed quantitatively on cultures grown in Kirchener's medium using a β-galactosidase assay. The intergenic region Rv8182 containing the ESX-3 promoter, P_{ESREG3}, the P_{AN} promoter containing region from ESX-2 and the P_{esat-6} promoter containing region Rv7374, from ESX-1, were used as positive controls, while the promoterless pJEM15 vector was used as a negative control for promoter activity.

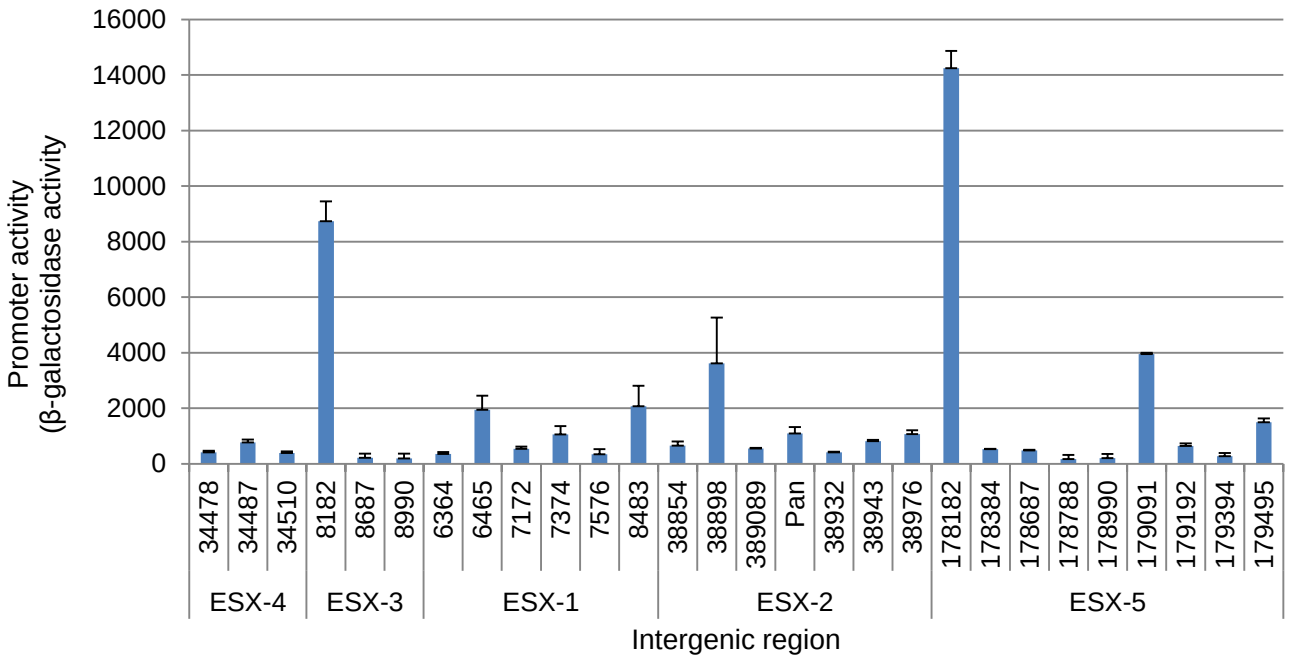
Promoter activity was classified as very strong, greater than 3000 β-galactosidase units; strong, 1000 to 3000 β-galactosidase units; moderate, 400 to 1000 β-galactosidase units and negligible, less than 400 β-galactosidase units. Relative promoter activities obtained in *M. smegmatis* cultures are given in Figure 2.2. Very strong promoter activities were obtained for the intergenic regions Rv8182 and Rv178182, upstream of the ESX-3 and ESX-5 gene cluster regions, respectively. Strong promoters were identified in the intergenic region upstream of ESX-1 (Rv6465), upstream of the *pe*/*ppe* (Rv38943), *esxD-C* (P_{AN} promoter) and *espI* (Rv38898) genes in ESX-2, and upstream of *pe19* (Rv179091) in ESX-5. Promoter activity was also identified in the intergenic regions Rv7374, upstream of esxA and esxB in ESX-1, upstream of Rv3896, the gene preceding ESX-2 (Rv38976) and upstream of Rv1795 (Rv179495) in ESX-5. No significant promoter activity was identified for ESX-4.

The promoter activities of the intergenic regions were also determined using *M. bovis* BCG as a host organism, to account for any differences resulting from the differences in transcriptional apparatus in *M. smegmatis*, as opposed to *M. tuberculosis* and other *M. tuberculosis* complex species (Figure 2.3). The assays were done as described for *M. smegmatis*. Promoter activities of the intergenic regions could be assigned as very strong, strong, moderate or negligible promoter activities which correspond to those obtained in *M. smegmatis*. The exception to this was the Rv8483 intergenic region which occurs on the complementary strand downstream of ESX-1, and showed greater promoter activity in *M. bovis* BCG, than in *M. smegmatis,* and could be classified as a strong promoter in *M. bovis* BCG. The β-galactosidase activities obtained in *M. bovis* BCG were, however, higher than those obtained in *M. smegmatis.*



ESX cluster	Intergenic region	β-gal activity	standard deviation
ESX-4	**34478**	35	23
	34487	332	244
	34510	5	5
ESX-3	**8182**	5941	2543
	8687	8	8
	8990	14	18
ESX-1	**6364**	171	50
	6465	1158	508
	7172	52	26
	7374	580	173
	7576	103	20
	8483	282	173
ESX-2	**38854**	223	74
	38898	1949	631
	389089	215	96
	P_{AN}	1427	311
	38932	117	14
	38943	2639	662
	38976	849	165
ESX-5	**178182**	6774	1677
	178384	472	218
	178687	25	7
	178788	-2	8
	178990	17	4
	179091	2980	322
	179192	283	57
	179394	80	27
	179495	943	294

Figure 2.2. Promoter activities of the intergenic regions of the *M. tuberculosis* ESX gene cluster regions in *M. smegmatis*. Results are the mean of 3 independent experiments.



ESX cluster	Intergenic region	β-gal activity	standard deviation
ESX-4	**34478**	406	60
	34487	771	98
	34510	394	48
ESX-3	**8182**	8731	717
	8687	211	149
	8990	197	169
ESX-1	**6364**	350	77
	6465	1945	511
	7172	538	82
	7374	1056	299
	7576	344	178
	8483	2075	736
ESX-2	**38854**	653	152
	38898	3613	1651
	389089	551	25
	***P*$_{AN}$**	1094	224
	38932	411	21
	38943	818	40
	38976	1068	137
ESX-5	**178182**	14247	615
	178384	531	3
	178687	480	24
	178788	167	151
	178990	204	152
	179091	3949	49
	179192	646	94
	179394	275	118
	179495	1498	134

Figure 2.3. Promoter activities of the intergenic regions of the *M. tuberculosis* ESX gene cluster regions in *M. bovis* BCG. Results are the mean of 2 independent experiments.

2.2.3.Sequence analysis

Two online promoter prediction programs, Softberry BPROM and BDGP Neural Network Promoter Prediction were used to identify possible promoter motifs in the intergenic regions which possess promoter activity. Furthermore, as functional sequences are expected to be more conserved than non-functional regions, the intergenic regions which showed promoter activity were aligned with the corresponding sequences from other mycobacterial species to identify potential promoter motifs. Conserved sequences which resemble consensus mycobacterial promoter sequence motifs were identified as putative promoter sequences. Putative promoters and associated DNA sequences were also manually searched for any regulatory domains and specific sigma factor binding motifs which may highlight potential regulation of transcription of the gene cluster and indicate possible functions. (See Addendum A: Newton-Foot, M, Gey van Pittius, NC, The complex architecture of mycobacterial promoters, Tuberculosis (2012), http://dx.doi.org/10.1016/j.tube.2012.08.003; for a review of the mycobacterial transcriptional apparatus and its regulation).

The intergenic regions which showed promoter activity are presented in Figure 2.4, in which putative promoter sequence motifs identified using multiple sequence alignments of the corresponding regions in related mycobacteria, the Softberry BPROM Promoter Prediction and BDGP Neural Network Promoter Prediction programs, and published promoter sequences are presented. Where promoter motifs were not clearly identified by promoter prediction programs and sequence alignments, promoter sequences are proposed based on identity to the mycobacterial and *E. coli* consensus promoter sequences.

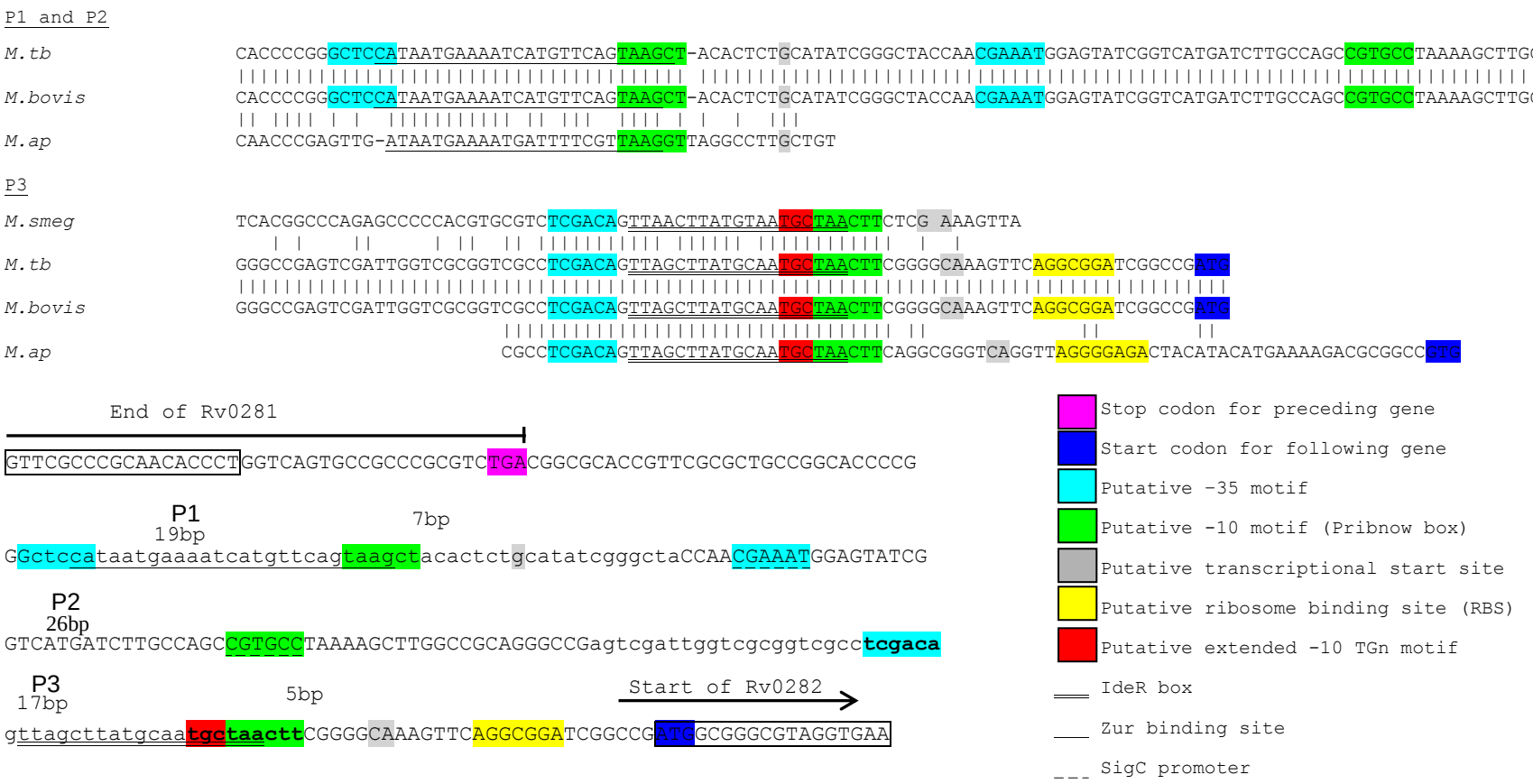


Figure 2.4 A. The P_{ESREG3} ESX-3 promoters in intergenic region Rv8182. The intergenic region Rv8182 contains 3 published promoters. P1 has a Zur binding motif in the interhexameric region overlapping the -10 and -35 motifs,[50] P2 contains the signature sequence motifs for the binding of SigC[153] and P3 has a IdeR-box overlapping its -35 and -10 motifs.[114] This correlates with data which shows that expression of *M. tuberculosis* ESX-3 is repressed in the presence of high concentrations of iron and zinc, and is dependent on Zur and IdeR. P3 has an extended -10 TGn motif which may be responsible for the high level of promoter activity in this region. The P1 promoter is not conserved in *M. smegmatis*, and it has been shown that ESX-3 expression is not regulated by Zur or zinc concentration in *M. smegmatis,* although the consistent promoter activities observed in *M. smegmatis* and *M. bovis* BCG demonstrate that the regulatory mechanisms may have been retained. The P2 promoter sequence is not conserved outside of the *M. tuberculosis* complex species and expression from this promoter may be restricted to these species. SigC is not conserved throughout all mycobacteria, and is absent in *M. smegmatis*, and this may be responsible for the lack of conservation of this sequence. SigC is involved in various stress conditions and its expression of ESX-3 may facilitate survival in the host. Promoter motifs predicted by BPROM are given in bold, and by BDGP in small letters. *M.tb - M. tuberculosis; M.ap - M. avium paratuberculosis; M.smeg - M. smegmatis*

Figure 2.4 B. The putative promoter in the intergenic region Rv6465 of ESX-1. The Rv6465 intergenic region is not well conserved across non *M. tuberculosis* complex slow-growers, or the fast growing mycobacteria, and may indicate the divergent expression from this promoter in different species. The proposed promoter motifs were identified based on identity to consensus promoter motifs. *M.tb - M. tuberculosis; M.smeg - M. smegmatis*

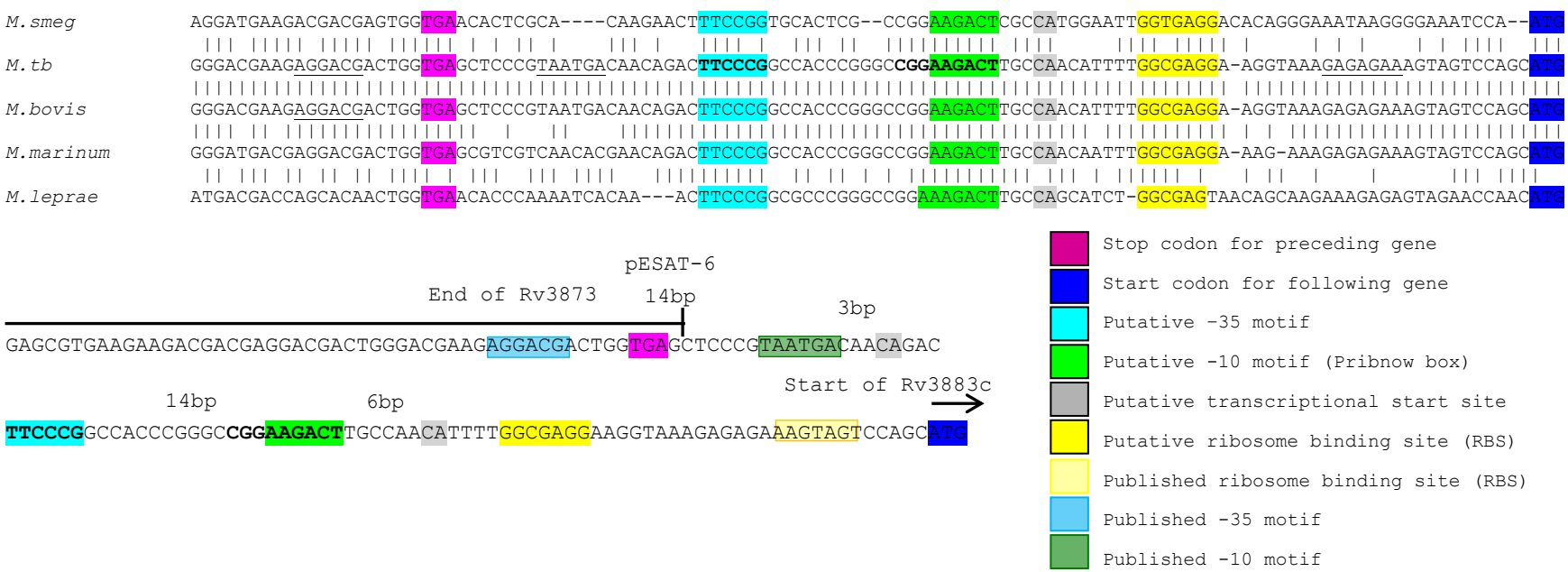


Figure 2.4 C. The P_{esat-6} promoter occurs in the Rv7374 intergenic region of ESX-1. The P_{esat-6} promoter has previously been identified and its role in the expression of ESAT-6 and CFP-10 described.[36] The multiple sequence alignment shows that these previously proposed sequence motifs (underlined) are not well conserved across the mycobacteria. Alternate promoter motifs predicted by BPROM (in bold) are more conserved and proposed as more likely promoter motifs in this region. *M.tb - M. tuberculosis; M.smeg - M. smegmatis*

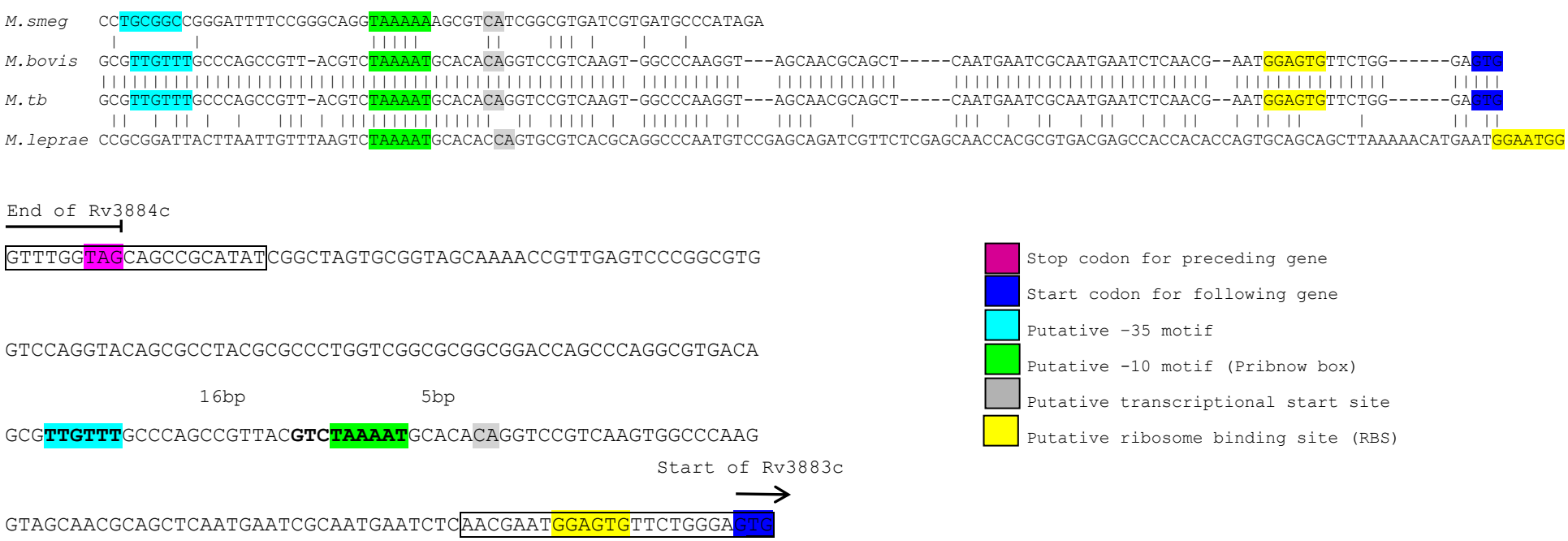


Figure 2.4 D. The putative promoter in the intergenic region Rv8483 of ESX-1. The Rv8483 intergenic region is not well conserved across the mycobacteria and no promoter motifs were predicted for this region. However a highly conserved putative -10 hexamer was identified and a -35 motif proposed. The variation in sequence between the *M. smegmatis* and *M. tuberculosis* -35 regions may represent differences in the specificities of the relevant sigma factor of these two species, which may account for the difference in promoter activity of this intergenic region when assayed in *M. smegmatis* and *M. bovis* BCG. Promoter motifs predicted by BPROM are given in bold. *M.tb - M. tuberculosis; M.smeg - M. smegmatis*

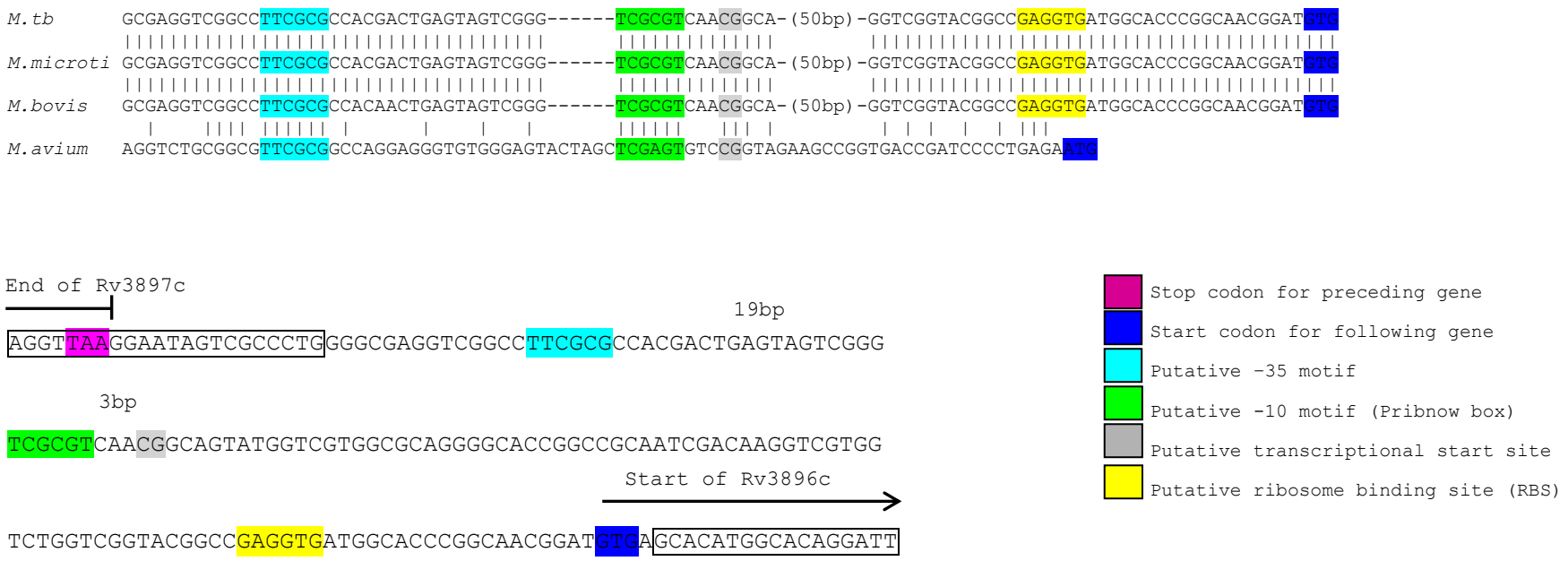


Figure 2.4 E. The putative promoter in the intergenic region Rv38976 of ESX-2. The proposed p38976 promoter was identified based on sequence identity within the slow-growing mycobacteria. Two conserved hexamers were identified as the -10 and -35 hexamers. *M.tb - M. tuberculosis*

Figure 2.4 F. The putative promoter in the intergenic region Rv38943 of ESX-2. The p38943 promoter motifs were identified by BPROM (bold) although the specific motifs are not well conserved outside of the *M. tuberculosis* complex. Expression of ESX-2 may therefore involve different mechanisms in different species. *M.tb - M. tuberculosis*

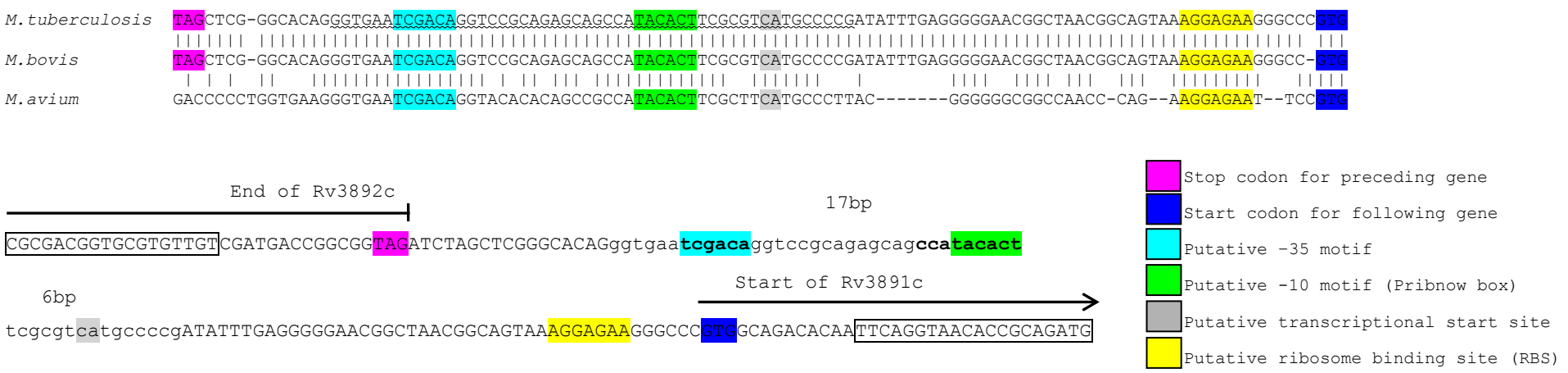


Figure 2.4 G. The P_{AN} promoter in ESX-2. The P_{AN} promoter was identified between Rv3892 and Rv3891 in *M. tuberculosis.*[154] The published promoter sequences correspond to the promoter region identified by the BDGP Neural Network Promoter Prediction program (small letters), BPROM (bold) and with conserved regions within the slow growing mycobacteria. This promoter sequence is very similar to the mycobacterial consensus SigA promoter sequence.

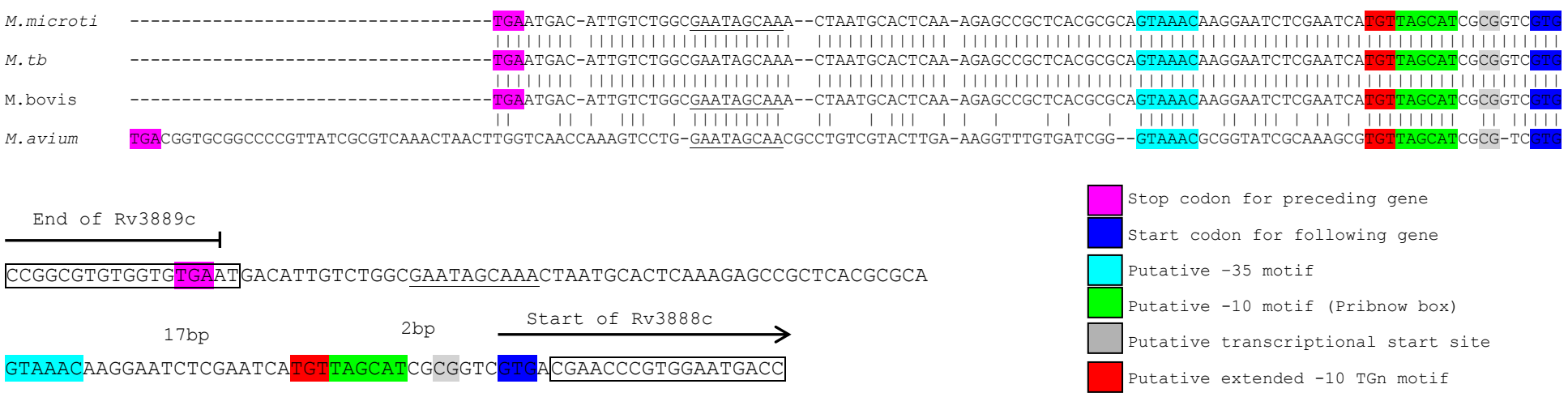


Figure 2.4 H. The putative promoter in the intergenic region Rv38898 of ESX-2. No promoters were predicted using the online promoter tools, however an extended -10 TGn motif and putative -10 and -35 are predicted based on sequence identity amongst the slow growers. The presence of the TGn motif corresponds to the high promoter activity exhibited by this promoter in the promoter assays. An additional conserved motif which may be important for transcription, possibly transcription factor binding, is underlined. *M.tb - M. tuberculosis*

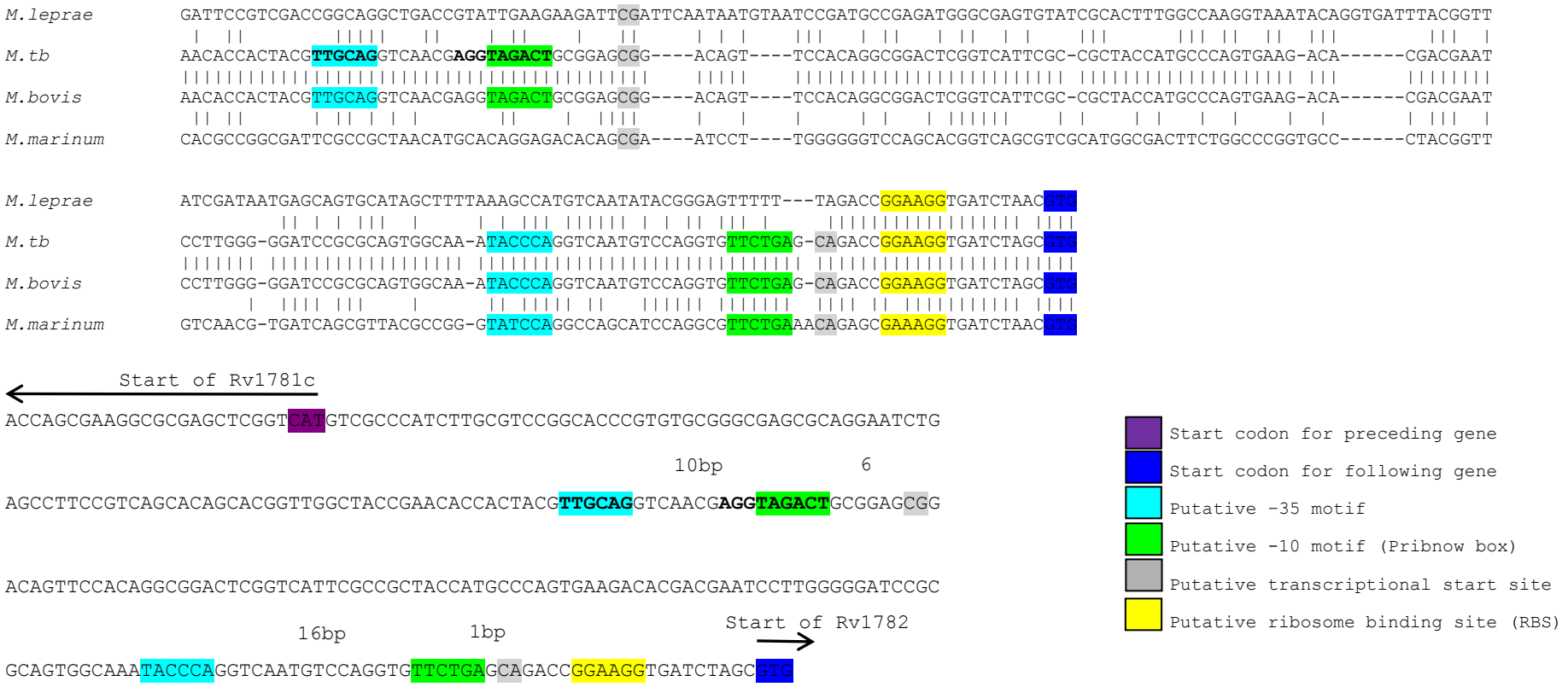


Figure 2.4 I. The putative promoter in the intergenic region Rv178182 of ESX-5. A possible promoter was identified between Rv1781c and Rv1782 by BROM (bold), however this region showed low sequence identity outside of the *M. tuberculosis* complex species. and alternative promoter motifs, closer to the start codon are based on sequence identity and similarity to mycobacterial consensus promoter sequences. Promoter motifs predicted by BPROM are given in bold. *M.tb - M. tuberculosis*

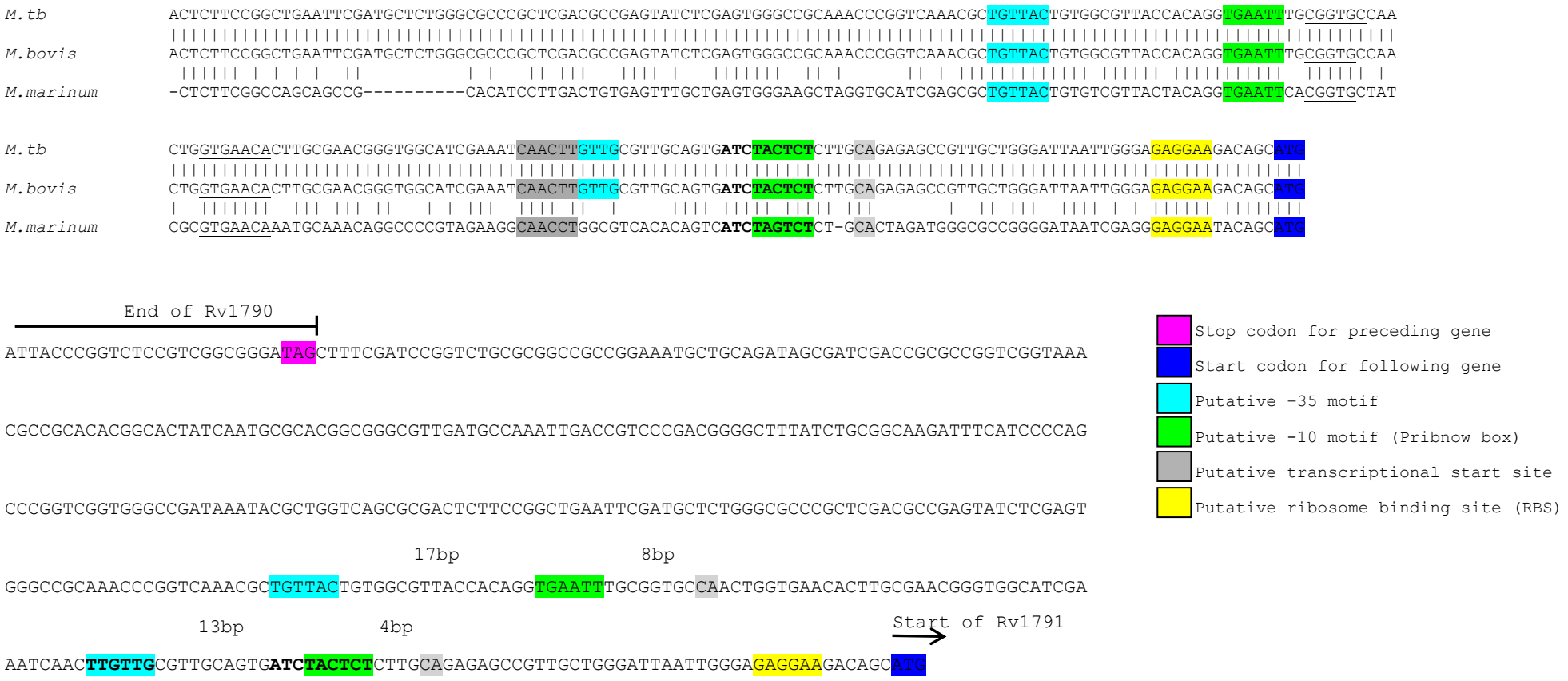


Figure 2.4 J. The putative promoter in the intergenic region Rv179091 of ESX-5. BPROM identified promoter motifs (bold) in the intergenic region between Rv1790 and Rv1791, however these sequences are not very conserved and another possible promoter was identified based on identity and similarity to the consensus promoter sequences.Additional conserved regions which may be important for transcription initiation, or the regulation thereof are underlined. *M.tb - M. tuberculosis*

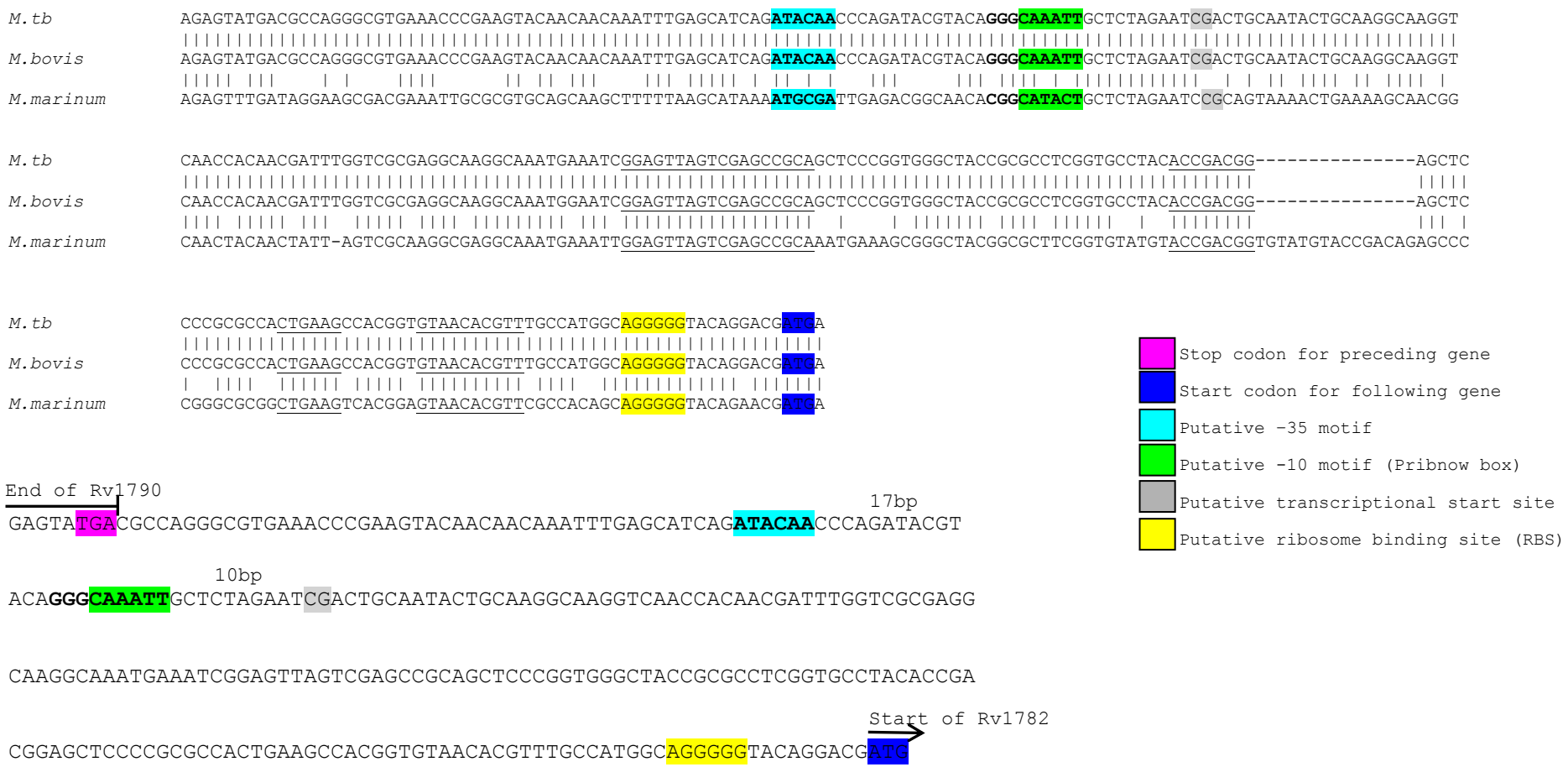


Figure 2.4 K. The putative promoter in the intergenic region Rv179091 of ESX-5. BPROM identified a possible promoter sequence in the intergenic region between Rv1794 and Rv1795 (bold). This region is not highly conserved outside of the *M. tuberculosis* complex, and more conserved areas in the intergenic region (underlined), may contain alternate promoter motifs, although they were not identified.

2.3. Discussion

The five *M. tuberculosis* ESX gene cluster regions each contain six conserved genes as well as several additional genes which occur in one or more of the regions.[26, 27] The arrangement of the constituent genes of each gene cluster region is different. The ESX-3 and ESX-2 gene components occur on a single strand and in close proximity, although the sequence and constituents vary. In the ESX-1 and ESX-5 gene cluster regions, most of the more conserved genes occur on a single strand, however gene insertions occur on the complementary strand and precipitate an inversion of two conserved genes in ESX-1. The protein products of each ESX gene cluster region are predicted to be components and substrates of dedicated protein secretion systems. This functional relationship is supported by the arrangement of the genes in clusters and the conservation of the clusters between bacterial species.[156] Although the precise functions of these five secretion systems remain to be resolved, they are unable to complement each other,[30] suggesting that their functions are distinct, and possibly that the conditions of their expression are different. Indeed, expression of some ESX gene clusters is dependent on specific environmental conditions, e.g. expression of ESX-3 is dependent on environmental iron and zinc concentration.[50, 114, 116] ESX-1 and ESX-5 have, however, been shown to influence one another, as deletion of EsxA and EsxB from ESX-1 results in increased secretion of the ESX-5 substrate PPE41,[59] and deletion of ESX-5 increases the secretion of ESX-1 substrates.[30] This suggests that the functions of one ESX may influence that of another. In this chapter we have identified the promoter-containing intergenic regions of the *M. tuberculosis* ESX gene clusters as a first step in determining the signals which allow expression of these ESX secretion systems. The promoter activities of the intergenic regions of the *M. tuberculosis* ESX gene clusters were compared using a β-galactosidase assay, and promoter containing regions identified (Figure 2.5).

No promoter activity was detected in the intergenic regions of ESX-4. It has been suggested that ESX-4 may not be expressed, or functional, in *M. tuberculosis* and *M. smegmatis*, as the EsxU protein has not been detected in culture supernatants, and exogenously expressed EsxU is not secreted.[119, 120] However, transcription of ESX-4 and its components, *esxT* and *esxU,* has been shown to be dependent on WhiB5 and SigM in *M. tuberculosis*.[112, 122] WhiB5 has been shown to be important for reactivation of *M. tuberculosis* after persistence and this may indicate that expression of ESX-4 occurs only under specific conditions. Furthermore, ESX-4 is conserved throughout the *Mycobacteria* and

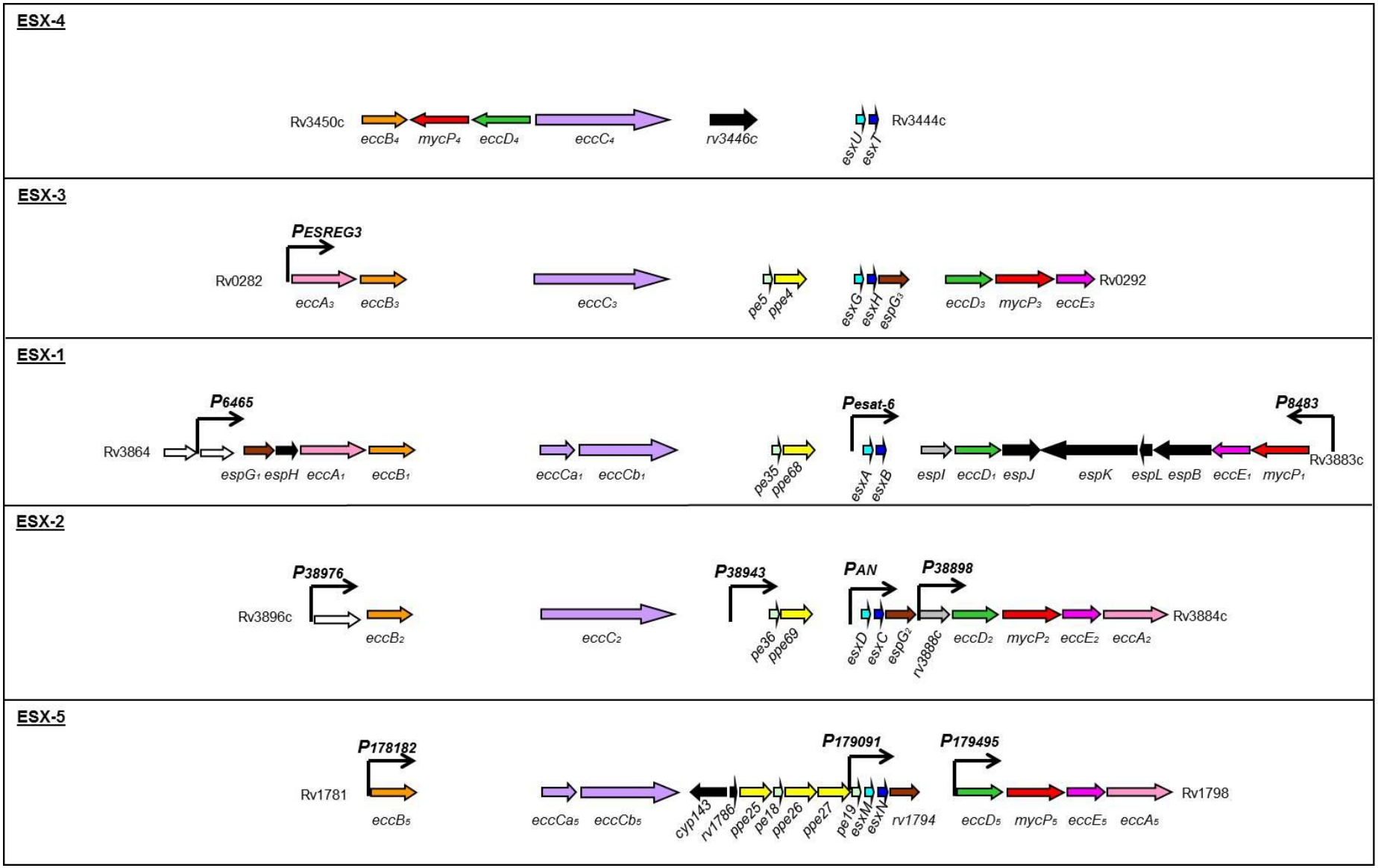


Figure 2.5. The promoters driving expression of the ESX gene clusters of *M. tuberculosis,* investigated in this study. Intergenic regions which showed promoter activity in this study are indicated as black arrows.

various other actinomycetes, suggesting that the region serves a biological function and it may only be expressed and functional under specific conditions and growth states, and not in normal *in vitro* growth.

The other 4 ESX gene clusters all have a promoter-containing intergenic region directly upstream, or within a few genes upstream of the cluster, with various additional promoters within the gene clusters (Figure 2.5). A single promoter-containing region, P_{ESREG3}, was identified upstream of ESX-3. Three promoters have been identified in this intergenic region, one regulated by IdeR in response to iron concentration,[50] one regulated by Zur in response to zinc concentration,[114] and one which contains the SigC promoter consensus sequence.[153] The various promoters are likely responsible for differential expression under various conditions to tightly regulate the levels of ESX-3 proteins in the cell.

In ESX-1 we identified three promoter-containing intergenic regions. One of these are located upstream of the gene cluster, directly upstream of *espF* (P_{6465}). P_{6465} is likely responsible for expression of the ESX-1 genes on the forward strand. *espF* (Rv3865) occurs directly upstream of espG and espH and is an ESX-1 secreted protein. Another promoter (P_{8483}) occurs downstream of the ESX-1 gene cluster, 5' to the *myc*P_1 gene, which occurs on the complementary strand, and may express the reverse-oriented portion of the ESX-1 gene cluster. This promoter was, however, only identified in the *M. bovis* BCG assay, indicating that this promoter may require a sigma factor or other transcription factors which are specific to *M. tuberculosis*, or the slow-growing mycobacteria in general, however no promoter motifs which are specific to the slow-growing mycobacteria could be identified in this promoter region. This may indicate the requirement of specific transcription factors which are only present in the slow-growing mycobacteria. Promoter activity was also identified upstream of the *esx*B-*esx*A operon (P_{7374}) and this promoter has previously been shown to express *esx*B and *esx*A as an operon.[36] These promoters can explain expression of the entire ESX-1 gene cluster, on both strands, with potential inducible, or increased, expression of *esx*A and *esx*B.

ESX-2 contains three internal promoter-containing intergenic regions and another is situated one gene upstream of the gene cluster (P_{38976}). The upstream gene, Rv3896, is a conserved hypothetical gene of unknown function and its co-expression with ESX-2 may imply that it is involved in ESX-2 function. The internal promoters occur upstream of the *pe/ppe* gene pair (P_{38943}), the esxD-esxC gene pair (P_{AN})[154] and the *esp*I-like gene Rv3888c (P_{38898}). The single strand layout of ESX-2 might suggest that this entire cluster is expressed in an operon from P_{38976}, and that the internal promoters regulate differential expression of various components as they are required under different conditions.

A promoter-containing intergenic region also occurs directly upstream of ESX -5, with two additional internal intergenic regions which exhibit promoter activity. One occurs directly upstream of *ecc*D_5 and the other upstream of the *pe*19 gene directly upstream of the *esx*M-*esx*N gene pair.

Clearly each of the ESX duplications subsequent to ESX-4 is expressed from an upstream promoter and in ESX-1 an additional promoter occurs upstream of the gene components on the complementary strand. An additional promoter occurs upstream of the *esx* gene pair in each of the ESX-1, -2 and -5 gene clusters, although in ESX-5 a *pe* gene is inserted between the promoter and the *esx* gene pair. In ESX-3 a promoter has been identified upstream of the esxG-esxH gene pair in *M. smegmatis*, although no promoter activity has been detected in *M. tuberculosis*.[115] The promoters upstream of the *esx* gene pairs, and other internal promoters, may increase expression of specific components or substrates in response to specific environmental cues.

The sequences of the various promoters, both previously identified and proposed in this study (Table 2.1), show little conservation between and within ESX gene clusters. This is consistent with the proposed distinct functions and expression profiles of the ESX secretion systems. Mycobacterial promoters are not very well characterised and extensive sequence variation, particularly in the -35 region of the promoter, is common, probably due to the large number of sigma factors and other transcription factors employed in the mycobacteria. (Mycobacterial promoters and their regulatory motifs are described in Addendum A: Newton-Foot, M, Gey van Pittius, NC, The complex architecture of mycobacterial promoters, Tuberculosis (2012), http://dx.doi.org/10.1016/j.tube.2012.08.00) This is consistent with what is seen in the ESX gene clusters, where although there is a level of similarity between the -10 hexamer sequences, there is little conservation in the -35 hexamer sequences. The two promoter prediction programs used to identify promoter motifs in the intergenic regions which exhibited promoter activity, are optimised for the identification of general bacterial promoter motifs (the analysis was done for prokaryotes on BDGP). Due to the extensive variation and lack of conservation of mycobacterial promoters with other bacterial motifs, these programs were not efficient in the identification of motifs in the promoter-containing intergenic regions. Furthermore, with the exception of the previously published SigC promoter and Zur and Fur binding sites in ESX-3, no specific sigma factor or regulatory binding sequences could be discerned. However, two extended -10 TGn motifs occur upstream of ESX promoters, one upstream of ESX-3 and another upstream of EspI in ESX-2 (Table 2.1, Figure 2.4). Extended -10 promoters are typically very strong promoters, as observed by the very strong promoter activity

Table 2.1. The proposed promoter motifs of the *M. tuberculosis* ESX gene cluster promoters.

REGION		PROPOSED PROMOTER										REFERENCE
ESX	**PROMOTER**	**-35**	bp	**TGn**	**-10**	bp	**TSP**	bp	**RBS**	bp	**START CODON**	
ESX-3	P_{ESREG3}	GCTCCA	19		TAAGCT	7	G				ATG	[50]
		CGAAAT	26		CGTGCC				AGGCGGA		ATG	[153]
		TCGACA	17	TGC	TAACTT	5	CA	6		7	ATG	[114]
ESX-1	P_{6465}	TCTGTG	19		TAGCGA	13	CG	3	AGGAGGG	8	ATG	This study
	P_{7374}	AGGACG	14		TAATGA	3	CA	58	GAGAAA	10	ATG	[36]
		TTCCCG	14		AAGACT	6	CA	4	GGCGAGG	26	ATG	This study
	P_{8483}	TTGTTT	16		TAAAAT	5	CA	42	GGAGTG	8	GTG	This study
ESX-2	P_{38976}	TTCGCG	19		TCGCGT	3	CG	66	GAGGTG	19	GTG	This study
	P_{38943}	TGTACG	20		TATCCT	5	CG	163	GGAGGGG	5	ATG	This study
	P_{38898}	GTAAAC	19	TGT	TAGCAT	2	CG	3			GTG	This study
	P_{AN}	TCGACA	17		TACACT	6	CA	38	AGGAGAA	6	GTG	[154]
ESX-5	P_{178182}	TTGCAG	10		TAGACT	6	CG	123	GGAAGG	9	GTG	This study
		TACCCA	16		TTCTGA	1	CA	4	GGAAGG	9	GTG	This study
	P_{179192}	TGTTAC	17		TGAATT	8	CA	97	GAGGAA	6	ATG	This study
		TTGTTG	13		TACTCT	4	CA	27	GAGGAA	6	ATG	This study
	P_{179495}	ATACAA	17		CAAATT	10	CG	169	AGGGGG	9	ATG	This study
Mycobacterial consensus		TTGACG/A	17±1	TGn	TATRMT	5-9	A/G		AGGAGG	4-7	GTG/ATG/TTG	[157]
***M. tuberculosis* consensus**		TTGACR	16-19	TGn	TAYgAT	5-9	A/G		AGGAGG	4-7	GTG/ATG/TTG	[158]

of the ESX-3 promoter (P_{ESREG3}). P_{38898} is also a strong promoter, although to a lesser degree, and may be regulated by other factors.

2.4. Conclusion

The five ESX gene clusters of *M. tuberculosis* encode functionally distinct secretion systems. We have identified promoter containing intergenic regions from the ESX gene cluster regions 3, 1, 2 and 5, but no promoter activity was detected for the ancestral region ESX-4. This study was however limited to the detection of intergenic promoters under *in vitro* growth in minimal culture medium and any promoters which are not active under these conditions may have been missed. Promoter activity was assayed using both *M. smegmatis* and *M. bovis* BCG as host organisms to account for any differences in the transcriptional machineries between the fast-growing saphrophytic, and slow-growing pathogenic mycobacteria. The results were consistent for all of the intergenic regions which were included in this study, with the exception of Rv8483, the region upstream of the ESX-1 components on the reverse strand, although no specific promoter motifs could be identified which would distinguish between these activities.

The positions of the promoters in the ESX duplications appears to be fairly conserved, with promoters occurring upstream of each gene cluster, with internal promoters occurring at similar positions, particularly upstream of the esx gene pairs. The proposed promoter sequences show little sequence identity, which may be indicative of the differential expression patterns of the different ESX gene clusters and their distinct functions. Identification of these promoters will assist in the investigation of the expression patterns and functional roles of the ESX secretion systems.

3. The functions of the mycobacterial ESX secretion systems and their substrates

3.1. Introduction

The ESX secretion systems have been implicated in the survival and virulence of *M. tuberculosis,* however their precise biological functions have yet to be clarified. The fast-growing, non-pathogenic mycobacterium, *M. smegmatis*, is frequently used as a model organism for the investigation of *M. tuberculosis* biological systems. *M. smegmatis* contains three ESX gene clusters, ESX-1, ESX-3 and ESX-4, which are similar in arrangement and composition to their orthologs in *M. tuberculosis*,[27] some of which have been shown to be functionally equivalent.[105, 107, 110]

To investigate the functions and substrates of the mycobacterial ESX secretion systems a library of ΔESX knockout strains of *M. smegmatis* was generated using allelic exchange. Each of the ESX gene clusters was deleted individually to create the ESX knockout strains ΔESX-1_{ms}, ΔESX-3_{ms} and ΔESX-4_{ms}, which were subsequently used to create the double and triple mutant strains ΔESX-1ΔESX-3_{ms}, ΔESX-1ΔESX-4_{ms}, ΔESX-3ΔESX-4_{ms} and ΔESX-1ΔESX-3ΔESX-4_{ms}. All deletions were confirmed by selective PCR amplification and sequencing across the deleted region. The successful deletion of all three *M. smegmatis* ESX gene clusters indicates that all three secretion systems are non-essential for *in vitro* growth of the organism; furthermore the generation of double- and triple- ESX deletion strains indicates that ESX secretion in general is dispensable for *M. smegmatis* growth *in vitro*. To investigate the specific biological changes associated with ESX-1, ESX-3 and ESX-4, ΔESX-1_{ms}, ΔESX-3_{ms} and ΔESX-4_{ms} were used in comparative metabolomic and proteomic analyses with wild type *M. smegmatis*. The differential protein and metabolite profiles of the bacteria in the presence and absence of each ESX was used to investigate the functions of the ESX secretion systems.

3.1.1. Whole cell lysate proteomic analysis of the M. smegmatis ESX secretion systems.

Proteomic analysis of the whole cell lysates of ΔESX-1_{ms}, ΔESX-3_{ms}, and ΔESX-4_{ms}, in comparison to that of WT_{ms}, was used to identify differences in the protein profiles of the mutant and wild type organisms resulting from the presence or absence of the ESX gene cluster. Whole cell lysate protein fractions were prepared from cultures grown in routine 7H9 culture medium in the absence of bovine serum albumin, and fractionated by SDS-PAGE to enable maximum sensitivity. LC-MS/MS analysis was done on the LTQ Orbitrap Velos to identify and quantify the proteins present each sample. Samples were analysed in technical duplicates of biological duplicates (4 replicates), using a single WT_{ms} experiment as reference

for all three ΔESX_{ms} strains. Proteins were identified using Thermo Proteome Discoverer 1.3 by automated database searching (Mascot and Sequest) of the tandem mass spectra against the *M. smegmatis* database at TB Database (TBDB). Proteins were searched and quantified using the Andromeda function from Maxquant 1.2.2.5. The peptide peak intensities were used to identify proteins with different abundances in the two strains. A two fold difference in protein abundance, with a significance level of 0.05 was considered statistically relevant.

3.1.2. Culture filtrate proteomic analysis of the M. smegmatis ESX secretion systems.

Proteomic analyses of the culture filtrates of $\Delta ESX\text{-}1_{ms;}$ $\Delta ESX\text{-}3_{ms;}$ $\Delta ESX\text{-}4_{ms}$ were done to identify novel substrates of the ESX secretions systems by identifying proteins which are secreted in WT_{ms} but not in the various knock out strains (Section 4), and to identify other proteomic changes in the culture filtrates resulting from the presence or absence of the ESX secretion systems. Culture filtrate fractions were prepared from WT_{ms}, $\Delta ESX\text{-}1_{ms}$, $\Delta ESX\text{-}3_{ms}$ and $\Delta ESX\text{-}4_{ms}$ grown to early log phase (OD_{600}=~ 0.3) in Middlebrook 7H9 medium, in the absence of BSA, catalase and Tween-80, and the constituent proteins were identified by LC-MS/MS analysis on the LTQ Orbitrap Velos. Samples were analysed in technical duplicates of biological duplicates (4 replicates), using a single WT_{ms} experiment as reference for all three ΔESX_{ms} strains. Proteins were identified and relative quantification done as for the whole cell lysate proteome. Protein concentrations in the culture filtrate fractions may be influenced directly by ESX secretion, by increased expression levels and by additional biological processes which are influenced by the presence or absence of the ESX secretion systems.

3.1.3. Metabolomic analysis of the M. smegmatis ESX secretion systems.

Metabolomics is an approach which aims to examine all the metabolites in an organism, with specific interest in those metabolites originating from, or influenced by, altered gene expression. This approach was used to analyse the changes in metabolites resulting from the absence and presence of the ESX-3 and ESX-4 gene clusters in *M. smegmatis*. Metabolites were extracted from *M. smegmatis* cultures and GC-MS analysis was done on an Agilent 7890A gas chromatograph coupled to an Agilent 5975B mass selective detector, in collaboration with Prof. Du Toit Loots at the Centre for Human Metabonomics, North West University, South Africa. PCA, PLS-DA and univariate statistical analyses were done on the metabolite data to determine whether the *ΔESX* strains group separately from wild type *M. smegmatis* and to identify metabolites which contribute most to the variation between the strains.

3.2. The functions of the M. smegmatis ESX-1 secretion system and its substrates

3.2.1. Introduction

ESX-1 has been implicated in virulence in *M. tuberculosis*; however this gene cluster is found throughout most of the fast growing mycobacteria, including *M. smegmatis,* and has been deleted from a number of the disease-causing slow growers. Several studies have shown that the *M. smegmatis* and *M. tuberculosis* ESX-1 secretion systems are, to at least some degree, functionally equivalent.[105, 107, 110] This suggests that ESX-1 has a non-essential function which is unrelated or indirectly related to virulence. This study investigates the role of ESX-1 in *M. smegmatis* on a proteomic level by comparing a Δ*ESX-1* deletion strain to the parental *M. smegmatis* strain, in order to identify possible biological functions of this secretion system.

3.2.2.Results

3.2.2.1. Whole cell lysate proteomic analysis of ESX-1 function

The whole cell lysate fractions of mid log phase cultures of ΔESX-1_{ms} and WT_{ms} were analysed using LC-MS/MS on the LTQ Orbitrap Velos in order to define their proteomes. In total, 2952 proteins were identified between the two strains (Supplementary Table S3.1). Protein abundances were compared and used to investigate the biological functions of the ESX-1 secretion system. Sixty-six proteins were differentially abundant between the two strains (Table 3.2.1), representing approximately 2.2% of the detected proteome. Ten proteins were only detected in ΔESX-1_{ms} and 14 only in WT_{ms}, while a further 19 and 23 proteins were detected at greater or lower levels in ΔESX-1_{ms} than in WT_{ms}, respectively.

Most of the ESX-1 protein components were detected in WT_{ms}, however only eight are given in Table 3.2.1, as the variation in abundance of other proteins was not considered significant due to the variation in the levels in which they were detected, or due to their detection in only one or two of the WT_{ms} replicates. Four of the ESX-1 proteins were detected in ΔESX-1_{ms}, all of which were detected at significantly reduced levels (between 5 and 240 fold less in ΔESX-1_{ms}). $EspG_1$ (MSMEG_0057) is only partially deleted from ΔESX-1_{ms}, and the presence of truncated $EspG_1$ accounts for the detection of this protein in ΔESX-1_{ms}. $EccA_1$ (MSMEG_0059) was detected in a single ΔESX-1_{ms} replicate, in which only three of thirteen unique peptides detected in WT_{ms} were identified. MSMEG_0071/0074 was also detected in only a single replicate of ΔESX-1_{ms}, and only three of the eight unique peptides detected in

Table 3.2.1. Proteins with different abundances in ΔESX-1_{ms} and WT_{ms} WCLs, classified according to biological function.

MSMEG_	Product	Differential abundance*	p value	Biological function	H37Rv ortholog
0061	FtsK/SpoIIIE family protein, $EccC_1a$	only WT_{ms}	0.007	ESX-1	Rv3870
0066	early secretory antigenic target, 6 kDa, ESAT-6, EsxA	only WT_{ms}	0.012	ESX-1	Rv3875
0076	antigen MTB48, EspB	only WT_{ms}	0.023	ESX-1	Rv3881c
0083	membrane-anchored mycosin $MycP_1$	only WT_{ms}	0.009	ESX-1	Rv3883c
0057	hypothetical protein, $EspG_1$	-11.7	0.021	ESX-1	Rv3866
0059	ATPase, AAA family protein, $EccA_1$	-5.1	0.023	ESX-1	Rv3868
0067	hypothetical protein, $EspI_1$	-15.7	0.000	ESX-1	Rv3876
0071/0075	hypothetical protein, ESX-1 encoded	-242.9	0.038	ESX-1	
3807	TetR-type regulator	only Δ$ESX\text{-}1_{ms}$	0.022	Transcriptional regulation	
0179	transcriptional regulator, GntR family protein	2.1	0.038	Transcriptional regulation	
0983	two-component system response regulator	2.5	0.014	Transcriptional regulation	
1397	transcriptional regulator, TetR family protein	2.2	0.005	Transcriptional regulation	Rv0681
4953	putative transcriptional regulator	2.6	0.006	Transcriptional regulation	
5673	transcriptional regulator	2.0	0.025	Transcriptional regulation	
6300	transcriptional regulator, GntR family protein	3.0	0.008	Transcriptional regulation	
1915	anti-sigma factor, family protein	-4.7	0.007	Transcriptional regulation	Rv3221A
2793	sensor-type histidine kinase PrrB	only WT_{ms}	0.013	Transcriptional regulation	
0087	glucitol operon repressor	3.1	0.052	Transcriptional regulation	
6096	Bvg accessory factor family protein	2.1	0.039	Transcriptional regulation	
2437	16S rRNA processing protein RimM	3.3	0.030	Translation	Rv2907c
0832	peptide deformylase	only Δ$ESX\text{-}1_{ms}$	0.033	Translation	Rv0429c
6911	ABC transporter, ATP-binding protein GluA	Only WT_{ms}	0.019	Amino acid metabolism	
3208	imidazole glycerol phosphate synthase, glutamine amidotransferase, hisH	2.6	0.051	Amino acid metabolism	Rv1602
3173	L-asparaginase	-3.1	0.036	Amino acid metabolism	Rv1538c
3207	imidazoleglycerol-phosphate dehydratase, HisB	-2.1	0.004	Amino acid metabolism	Rv1601
3878	precorrin-6Y C5,15-methyltransferase, CobL	only Δ$ESX\text{-}1_{ms}$	0.013	Cofactor metabolism	Rv2072c
2597	aldehyde dehydrogenase	only Δ$ESX\text{-}1_{ms}$	0.030	Reduction/oxidation	Rv2858c
6874	aldehyde dehydrogenase	only Δ$ESX\text{-}1_{ms}$	0.003	Reduction/oxidation	
1157	short chain dehydrogenase	only Δ$ESX\text{-}1_{ms}$	0.000	Reduction/oxidation	
0884	glyoxalase family protein	2.5	0.032	Reduction/oxidation	
5862	short chain dehydrogenase	4.5	0.035	Reduction/oxidation	Rv0765c
2206	3-oxoacyl-(acyl-carrier-protein) reductase, FabG	-2.5	0.036	Reduction/oxidation	Rv1350
5837	glutathione peroxidase family protein	-2.0	0.011	Reduction/oxidation	
0985	sugar transporter family protein	only Δ$ESX\text{-}1_{ms}$	0.000	Carbohydrate metabolism	Rv3476c
5676	citrate synthase	2.0	0.034	Carbohydrate metabolism	Rv0889c
5345	glycosyl hydrolases family protein 16	-3.1	0.039	Carbohydrate metabolism	

1676	adenosine deaminase	-2.1	0.001	DNA/RNA metabolism	Rv3313c
6927	MutT/nudix family protein	-2.5	0.041	DNA/RNA metabolism	Rv3908
6278	metallo-beta-lactamase superfamily protein	only WT_{ms}	0.028	Antibiotic	
0819	N-carbamoyl-L-amino acid amidohydrolase	only $\Delta ESX\text{-}1_{ms}$	0.021	Other	
0946	NAD dependent epimerase/dehydratase family	only WT_{ms}	0.009	Other	Rv0501
2579	zinc metalloprotease	only WT_{ms}	0.014	Other	Rv2869c
5516	phosphoribosylglycinamide formyltransferase	only WT_{ms}	0.024	Other	Rv0956
0203	IS1096, tnpR protein	2.7	0.014	Other	
2201	ZbpA protein	2.5	0.052	Other	Rv0130
3430	SAM-dependent methyltransferase	2.0	0.036	Other	
4704	acyltransferase family protein	8.2	0.006	Other	Rv2483c
0434	aminoglycoside 2'-N-acetyltransferase	-2.3	0.021	Other	Rv0262c
3655	ABC transporter, permease/ATP-binding protein	-2.0	0.013	Other	Rv1819c
4475	LppP protein	-3.3	0.023	Other	Rv2330c
0546	hypothetical protein	only $\Delta ESX\text{-}1_{ms}$	0.032	Hypothetical protein	
5087	hypothetical protein	only $\Delta ESX\text{-}1_{ms}$	0.003	Hypothetical protein	Rv1205
0840	hypothetical protein	only WT_{ms}	0.016	Hypothetical protein	
2032	hypothetical protein	only WT_{ms}	0.005	Hypothetical protein	
2679	hypothetical protein	only WT_{ms}	0.008	Hypothetical protein	
3204	hypothetical protein	only WT_{ms}	0.033	Hypothetical protein	Rv1598c
2445	hypothetical protein	2.5	0.022	Hypothetical protein	
6901	hypothetical protein	6.5	0.008	Hypothetical protein	Rv0049
2498	hypothetical protein	-2.5	0.012	Hypothetical protein	
3592	hypothetical protein	-2.3	0.038	Hypothetical protein	
3613	hypothetical protein	-7.2	0.017	Hypothetical protein	
4219	hypothetical protein	-2.0	0.021	Hypothetical protein	Rv2147c
5816	hypothetical protein	-2.4	0.037	Hypothetical protein	
6207	hypothetical protein	-2.3	0.026	Hypothetical protein	

*Differential abundances represent the fold change in abundance of a protein in $\Delta ESX\text{-}1_{ms}$ relative to WT_{ms}. Negative values indicate a lower abundance in $\Delta ESX\text{-}1_{ms}$ than WT_{ms}. Where indicated proteins were detected in only one strain.

WT_{ms} were identified, at dramatically reduced intensities. $EspI_1$ (MSMEG_0067) was detected in all four replicates, however only four or five peptides were identified in any ΔESX-1_{ms} replicate, relative to eleven in WT_{ms}. The detection of these peptides in ΔESX-1_{ms} is incongruent with the absence of the proteins expressed from just within the borders of the deletion and throughout the deleted region, and might be attributed to peptide contamination or the presence of homologous proteins. The reduced abundances and numbers of peptides from these proteins detected in ΔESX-1_{ms}, together with the complete absence of other ESX-1 proteins, however, support the deletion of ESX-1 from ΔESX-1_{ms} and indicates that ESX-1 is expressed under these *in vitro* conditions in $WT_{ms.}$

Despite the significant expression levels of the ESX-1 proteins, and significant changes in abundance of various proteins, no cellular pathways appear to be considerably influenced by the deletion of ESX-1. The only protein grouping which shows noteworthy differences between the two strains is transcriptional regulation, where there is a general increase in the abundances of various transcriptional regulators. Other differences include proteins involved in oxidation and reduction reactions, carbohydrate and amino acid metabolism, and a number of unclassified and hypothetical proteins.

3.2.2.2. Culture filtrate proteomic analysis of ESX-1 function

Proteins were identified from the culture filtrate fractions of WT_{ms} and ΔESX-1_{ms} cultures, using LC-MS/MS analysis on the LTQ Orbitrap Velos. A total of 1783 proteins were detected in the two strains (Supplementary table S3.2). Protein abundances were compared and thirty-six proteins found to vary in abundance between the two strains (Table 3.2.2.) Thirteen were present only in ΔESX-1_{ms} and 7 only in WT, while 9 were more abundant in ΔESX-1_{ms} and a further 7 more abundant in WT_{ms}. The deviant proteins could be classified into much the same groups as in the WCL proteomic analysis, however a reduction in divergent transcription factors may indicate that enrichment for secreted proteins was successful.

Table 3.2.2. Proteins with different abundances in ΔESX-1_{ms} and WT_{ms} culture filtrates, classified according to biological function.

MSMEG_	Product	Differential abundance*	p value	Biological function	H37Rv ortholog
0059	ATPase, AAA family protein, large subunit	only WT_{ms}	0.008	ESX-1	Rv3868
3142	HTH-type transcriptional repressor AcnR	only WT_{ms}	0.000	Transcriptional regulation	Rv1474c
5512	magnesium chelatase	3.3	0.001	Transcriptional regulation	Rv0958
3778	phenylalanyl-tRNA synthetase, alpha subunit	only WT_{ms}	0.002	Translation	Rv1649
6102	dihydroneopterin aldolase, folB	only ESX-1	0.049	Cofactor metabolism	Rv3607c
1603	IMP dehydrogenase family protein	only ΔESX-1_{ms}	0.000	Reduction/oxidation	Rv3410c
4320	alkyl hydroperoxide reductase	only ΔESX-1_{ms}	0.003	Reduction/oxidation	Rv3328c
5029	alkanal monooxygenase alpha chain	only ΔESX-1_{ms}	0.004	Reduction/oxidation	
2529	glyoxylate reductase	only ΔESX-1_{ms}	0.014	Reduction/oxidation	
6916	short chain dehydrogenase	only ΔESX-1_{ms}	0.048	Reduction/oxidation	
6362	quinone oxidoreductase	2.1	0.017	Reduction/oxidation	Rv3777
1573	carbohydrate kinase family protein	only WT_{ms}	0.012	Carbohydrate metabolism	Rv3433c
0313	phosphogluconate dehydratase	-2.5	0.041	Carbohydrate metabolism	
2362	DNA ligase, NAD-dependent	-2.0	0.002	DNA/RNA metabolism	Rv3014c
6278	metallo-beta-lactamase superfamily protein	only WT_{ms}	0.013	Antibiotic	
5245	universal stress protein family protein	2.0	0.049	Stress	Rv2028c/Rv3134c
0635	putative conserved exported protein	-2.0	0.007	Secreted protein	Rv0309
6815	secreted protein	-2.0	0.043	Secreted protein	
5746	gas vesicle protein	only ΔESX-1_{ms}	0.002	Other	
1852	selenide, water dikinase, SelD	only ΔESX-1_{ms}	0.013	Other	
0529	probable serine/threonine-protein kinase PknK	only ΔESX-1_{ms}	0.024	Other	Rv3080
6906	putative hydrolase	only ΔESX-1_{ms}	0.027	Other	Rv0045c
1108	hydrolase, alpha/beta fold family protein	only ΔESX-1_{ms}	0.049	Other	Rv0554
5243	helix-turn-helix motif	3.5	0.000	Other	
4200	peptidase M20	2.3	0.003	Other	Rv2141c
5136	helix-turn-helix motif	2.6	0.004	Other	Rv0080
3127	conserved protein, DUF59	2.7	0.029	Other	Rv1466
2073	CAIB/BAIF family protein	2.4	0.044	Other	
6179	acetyl-coenzyme A synthetase	-2.1	0.002	Other	Rv3667
3645	hypothetical protein	only ΔESX-1_{ms}	0.012	Hypothetical protein	Rv1825
4923	hypothetical protein	only ΔESX-1_{ms}	0.013	Hypothetical protein	Rv1321
3905	hypothetical protein	only ΔESX-1_{ms}	0.004	Hypothetical protein	Rv2117
3016	hypothetical protein	only WT_{ms}	0.014	Hypothetical protein	Rv2566
1340	hypothetical protein	2.1	0.007	Hypothetical protein	Rv0635
2739	hypothetical protein	-2.2	0.006	Hypothetical protein	Rv2721c
5629	hypothetical protein	-2.0	0.047	Hypothetical protein	

*Differential abundances represent the fold change in abundance of a protein in ΔESX-1_{ms} relative to WT_{ms}. Negative values indicate a lower abundance in ΔESX-1_{ms} than WT_{ms}. Where indicated proteins were detected in only one strain.

3.2.3.Discussion

The proteomic analysis of ΔESX-1_{ms} indicates that ESX-1 is expressed under normal culture conditions in *M. smegmatis*. A total of 15 ESX-1 proteins were detected in WT_{ms}, all of which were either not detected in ΔESX-1_{ms}, or present at a much lower abundance. However, no cellular pathways were considerably impacted in the absence of ESX-1 and approximately 98% of the detected proteome showed no significant variation between strains. The greatest distinction between ΔESX-1_{ms} and WT_{ms} is the altered abundances of transcription factors, most of which are more abundant in ΔESX-1_{ms}. The functions of these transcription factors are, however unknown, making the signal(s) and response(s) difficult to identify, particularly as the additional proteomic variation does not highlight any pathways. The absence of concerted pathway deviation on a proteomic or secretomic level in ΔESX-1_{ms} suggests that ESX-1 may function independently of the mycobacterial cellular metabolic processes. Whole metabolomic analyses of ΔESX-1_{ms} and WT_{ms} are currently underway to investigate the effects of the deletion of ESX-1 on the metabolite profile of the bacterium (In collaboration with Prof. Loots, North West University, South Africa). These results my corroborate these findings or highlight pathways which are influenced by the deletion of ESX-1.

The proteomic and secretomic results do not correlate directly with the previously identified functions of ESX-1, most notably conjugal DNA transfer[105, 106] and mycolic acid synthesis by $EccA_1$.[109] No significant changes to proteins involved in cell wall biosynthesis or in DNA binding proteins, other than the previously mentioned transcription factors, were seen. If ESX-1 is involved in the transfer of DNA between cells via conjugal DNA transfer, this may be expected not to impact on other basic cellular metabolic processes. $EccA_1$ is however highly expressed in WT_{ms} under these conditions, therefore it could be expected that the absence of this protein in ΔESX-1_{ms} would affect other proteins involved in cell wall and mycolic acid biosynthesis. It is possible that under these conditions *M. smegmatis* is not undergoing conjugal DNA transfer and that $EccA_1$ is only involved in mycolic acid metabolism during this process, to enable the transfer of the proteins involved in conjugation through the cell wall and mycomembrane. $EccA_1$ was also identified in the culture filtrate and whole cell lysate fractions of WT_{ms}. There are conflicting reports of the localisation of this protein in *M. tuberculosis* where it has been identified in the membrane fraction,[159] or only in the cytosolic fraction[160] in two separate studies. The presence of this protein in the culture filtrates in this study may correlate with membrane localisation, and subsequent release into the culture medium. The function of $EccA_1$ in mycolic acid biosynthesis

corresponds to its membrane localisation. However, $EccA_1$ has also been predicted to be involved in protein-protein interactions and the secretion of various substrates, suggesting an intracellular role.[73, 75]

3.2.4.Conclusion

The proteomic and secretomic analyses of *M. smegmatis* ESX-1 fail to highlight specific mechanisms or pathways whereby it functions. ESX-1 is highly expressed in *M. smegmatis* under these conditions, but shows very little variation on a proteomic level. Protein detection is, however, limited to those proteins which can be extracted successfully and in sufficient abundance using the current methodology. Amongst others, membrane proteins may be underrepresented and alternate extraction and enrichment methods may indicate increased proteomic variation in ΔESX-1_{ms}. The absence of ESX-1 clearly impacts on the transcriptional regulome of *M. smegmatis*. The lack of collective response by pathways of proteins, however, deters identification of signal and response resulting from the deletion of ESX-1. ESX-1 has been shown to regulate conjugal DNA transfer in *M. smegmatis* and to be involved in virulence and mycolic acid biosynthesis in *M. tuberculosis,* although the correlation between these functions in the two organisms remains unclear.

3.3. The functions of the M. smegmatis ESX-3 secretion system and its substrates

3.3.1. Introduction

The ESX-3 secretion system is essential for the *in vitro* survival of *M. tuberculosis*,[54, 113] where its expression has been shown to be regulated by IdeR and Zur in response to iron and zinc availability, respectively.[50, 114] ESX-3 is required for iron acquisition via the mycobactin mediated iron uptake system, and presumably zinc uptake via an additional pathway.[113] ESX-3, though non-essential and only regulated by iron in *M. smegmatis,*[115] is involved in similar pathways in this non-pathogenic, fast growing species. We have created an ESX-3 knockout strain of *M. smegmatis,* ΔESX-3_{ms}, which serves as a useful tool for the investigation of ESX-3 function. In this study ΔESX-3_{ms} was used in comparative metabolomic, proteomic and secretomic analyses with the wild-type *M. smegmatis* parental strain, WT_{ms}, to investigate the role of the ESX-3 secretion system in the biology of the organism.

3.3.2. Results

3.3.2.1. Metabolomic analysis of ESX-3 function

Twelve replicates of WT_{ms} and ΔESX-3_{ms} cultures, grown to mid log phase (OD_{600}= 0.6-0.8) in standard Middlebrook 7H9 culture medium, were subjected to GC-MS metabolomic analysis. The PCA scores plot of the metabolite data showed that the individual samples from each of the two strains clustered independently (Figure 3.3.1). This can be attributed to consistent differences in analysed metabolite profiles of the individual samples of each strain. Due to the natural separation achieved in the PCA, a PLS-DA was constructed to confirm and identify additional metabolites to those identified using the PCA, possibly supplying additional compound information for explaining ESX-3 gene functionality. The important metabolites which best describe the variation between WT_{ms} and ΔESX-3_{ms}, identified using the PCA, PLS-DA and univariate statistics, are compounds which are either entirely unique to a particular strain, or common to both groups, but exhibit a constant variation in concentration between the two strains (Tables 3.3.1a and b; analysis was done by Mr Reinart J. Meissner-Roloff in Prof. du Toit's group). The metabolites which vary most between WT_{ms} and ΔESX-3_{ms} are predominantly amino acids and sugars. Histidine, proline, alanine, cysteine, methionine, isoleucine, aspartic acid, tryptophan, valine, threonine, phenylalanine, serine, leucine, glucose, galactose and mannose, were all reduced, and lysine, oxy-proline, glycine, citrulline, glutamine, glutamate and tyrosine were elevated in ΔESX-3_{ms} relative to

WT_{ms}. Although lactic acid, aspartic acid, threonine, tryptophan, serine, phenylalanine and valine were not identified as top markers for either the PCA or the PLS-DA analyses, their relative concentrations were also determined for all samples, averaged and included in Tables 3.3.1a and b, due to their roles as intermediates in the metabolism of the previously mentioned metabolite markers.

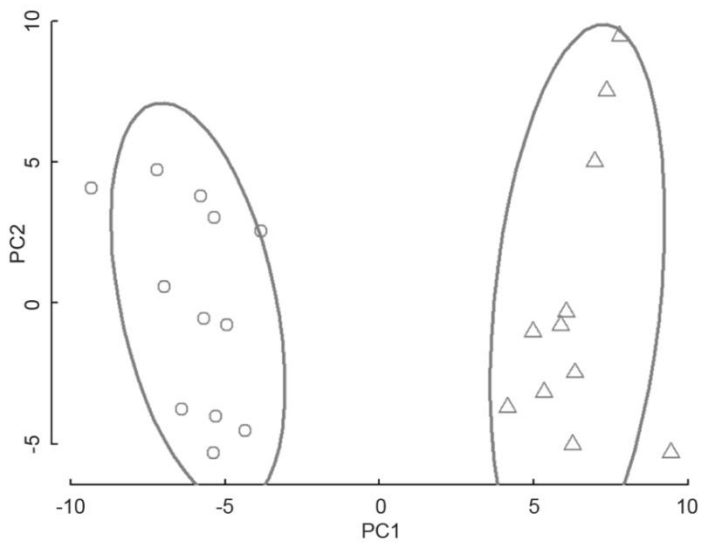


Figure 3.3.1. PCA scores plot of individually cultured ΔESX-3_{ms} (circles) **and WT_{ms}** (triangles). The individual samples from each strain cluster separately. PC1 and PC2 explain 33.8% of the total variation in the data. Separation of the two groups is due to differences in the metabolite profiles in each of the sample groups. The PCA plot was generated by Mr Reinart J. Meissner-Roloff in Prof. du Toit's group.

Table 3.3.1a. Metabolites which occur at higher concentrations in ΔESX-3_{ms} than in WT_{ms}

Compound name	Concentration (µg/mg) ±standard deviation (% of samples with detected compound)		Effect sizes
	WT_{ms}	ΔESX-3_{ms}	
Citrulline	0 (0)	0.258 ±0.058 (100)	>0.5
Lysine	0.012±0.096 (100)	0.24+0.096(100)	>0.5
Oxy-proline	28.193±8.998 (100)	33.838±5.3.1 (100)	>0.8
Glycine	0 (0)	0.06±.0.02 (100)	>0.8
Tyrosine	0.098 ±0.125 (75)	1.387 ±0.875 (100)	>0.8
Glutamate	0 (0)	0.0363±0.0318 (100)	>0.5
Acryl glycine	0 (0)	0.044±0.04 (63)	>0.8
Glutamine	0.027±0.025 (83)	0.478±0.081 (100)	>0.8
Parabanic acid	0.014±0.003 (100)	0.051±0.008 (100)	>0.8
Butanal	0 (0)	0.192±0.137 (73)	>0.5
2,4 bispyrimidine	0.113 ±0.145 (83)	4.128 ±0.944 (100)	>0.8
Phosphenodiimidic amine	0 (0)	0.431 ±0.226 (100)	>0.5
Succinic semialdihyde	0.005 ±0.050 (100)	0.016 ±0.181 (100)	>0.5
Hexanedioc acid	0.027±0.02 (75)	0.478±0.177 (100)	>0.8
9H-purine	0 (0)	0.21±0.009 (90)	>0.8
*Lactic acid	0.005±0.004 (75)	0.011±0.008 (82)	>0.5
Xylose	0.293±0,184 (64)	0.634±0.172 (82)	>0.5
Erythro-pentose	0 (0)	0.145±0.033 (100)	>0.5
Myo-inositol	0 (0)	0.145±0.033 (100)	>0.8

* Compounds which were not identified as top markers for either the PCA or the PLS-DA analyses.
Values in brackets represent the percentage of samples in which the compound was detected.

Table 3.3.1b. Metabolites which occur at higher concentrations in WT_{ms} than in ΔESX-3_{ms}

Compound name	Concentration (µg/mg) ±standard deviation (% of samples with detected compound)		Effect sizes
	WT_{ms}	ΔESX-3_{ms}	
Histidine	0.407 ±0.157 (100)	0 (0)	>0.8
Proline	1.452 ±0.594 (100)	0.005 ±0.005 (73)	>0.8
Alanine	8.377 ±1.225 (100)	3.571 ±0.867 (100)	>0.8
Methionine	1.169±0.324 (100)	0.039±0.052 (63)	>0.8
Isoleucine	2.895±0.501 (100)	1.467±0.374 (100)	>0.8
Leucine	1.585±0.721 (100)	0.288±0.032 (100)	>0.8
Cysteine	0.18±0.033 (100)	0.057±0.026 (100)	>0.8
*Aspartic acid	0.44±0.128 (100)	0.379±0.102 (100)	>0.8
*Tryptophane	0.205±0.198 (75)	0.199±0.184 (64)	>0.5
*Valine	0.371±0.099 (83)	0.227±0.132 (90)	>0.5
*Threonine	0.245±0.065 (100)	0.166±0.081 (100)	>0.8
*Phenylalanine	0.127±0.103 (100)	0.076±0.054 (100)	>0.8
*Serine	0.219±0.079 (100)	0.137±0.082 (100)	>0.8
Galactose	0.212 ±0.160 (100)	0 (0)	>0.8
Glucose	10±5.705 (100)	3.373±5.301 (100)	>0.8
Mannose	1.678±0.912 (75)	0.67±1.765 (64)	>0.8
Altro Heptulose	3.427±1.432 (100)	1.825±1.765 (73)	>0.5
Turanose	0.573±0.096 (100)	0.416±0.083 (100)	>0.8
Galactose oxime	7.366±2.405 (100)	3.686±1.603 (100)	>0.8
Ribofuranoside	0.081±0.037 (100)	0 (0)	>0.8
Butanoic acid	0.071 ±0.011 (100)	0.030 ±0.006 (100)	>0.8
Mannonic acid	0.117 ±0.04 (100)	0.0136 ±0.012 (82)	>0.8
Dodecanoic acid	1.378 ±0.228 (75)	0.366 ±0.083 (100)	>0.8
Phosphoric acid	0.515±0.106 (100)	0.337±0.101 (100)	>0.5
Gluconic acid	0.117±0.083 (100)	0.0136±0.0113 (82)	>0.5
2-Piperidone	1.476 ±0.248 (100)	0.405 ±0.109 (100)	>0.8
Bis-thiodiglycol	0.106 ±0.026 (100)	0.021 ±0.014 (73)	>0.5

* Compounds which were not identified as top markers for either the PCA or the PLS-DA analyses.
Values in brackets represent the percentage of samples in which the compound was detected.

3.3.2.2. Whole cell lysate proteomic analysis of ESX-3 function

The proteomes of ΔESX-3_{ms} and WT_{ms} cultures were analysed using LC-MS/MS on the LTQ Orbitrap Velos. A combined total of 2845 proteins were identified in the *M. smegmatis* ΔESX-3_{ms} and WT_{ms} strains (Supplementary Table S3.3). Protein abundances were compared and used to investigate the biological functions of the ESX-3 secretion system. Almost 95% of the detected proteins showed no significant variation between strains. Eight proteins were detected in ΔESX-3_{ms} but not in $WT_{ms,}$ while 63 proteins were present in WT_{ms}, but not detected in ΔESX-3_{ms}. A further 38 and 43 proteins were present at a significantly greater or lower abundance, respectively, in ΔESX-3_{ms} relative to WT_{ms} (Table 3.3.2). Two of the proteins which were completely absent in ΔESX-3_{ms}, but present in WT_{ms}, MSMEG_0615 ($EccA_3$) and MSMEG_0626 ($EccE_3$), are encoded by the first and last genes in ESX-3, confirming the deletion of ESX-3 from ΔESX-3_{ms}. Two additional ESX-3 encoded proteins, MSMEG_0617 and MSMEG_0623, were detected at approximately 5 and 30 times lower levels, respectively, in ΔESX-3_{ms} than in $WT_{ms.}$ The detection of these proteins in ΔESX-3_{ms} may be attributed to the presence of their paralogues encoded in ESX-4 and ESX-1, while their greatly reduced abundances support deletion of these genes in ΔESX-3_{ms}.

Table 3.3.2. Proteins with different abundances in ΔESX-3_{ms} and WT_{ms} WCLs, classified according to biological function.

MSMEG_	Product	Differential abundance*	p value	Biological function	H37Rv ortholog
0615	ATPase, AAA family protein, $EccA_3$	only WT_{ms}	0.007	ESX-3	Rv0282
0626	hypothetical protein, $EccE_3$	only WT_{ms}	0.003	ESX-3	Rv0292
0617	ftsk/spoiiie family protein, $EccC_3$	-4.9	0.009	ESX-3	Rv0284
0623	secretion protein Snm4, $EccD_3$	-30.8	0.018	ESX-3	Rv0290
0454	GntR-family protein transcriptional regulator	only WT_{ms}	0.001	Transcriptional regulation	Rv0454
6580	transcriptional regulator family protein	only WT_{ms}	0.003	Transcriptional regulation	Rv3095
5566	transcriptional regulator	only WT_{ms}	0.051	Transcriptional regulation	Rv1404
5422	transcriptional regulator, MazG family protein	only WT_{ms}	0.001	Transcriptional regulation	Rv1021
5040	transcriptional regulator, TetR family protein	only WT_{ms}	0.043	Transcriptional regulation	Rv1255c
3646	transcriptional regulator, MerR family protein	2.1	0.049	Transcriptional regulation	Rv1828
2248	two-component system sensor kinase	only Δ$ESX\text{-}3_{ms}$	0.039	Transcriptional regulation	
0592/6022	putative rhamnose catabolism operon/ xylose repressor, ROK-family protein	2.4	0.005	Transcriptional regulation	
1473	ribosomal protein L30, rpmD	only WT_{ms}	0.023	Translation	Rv0722
5432	peptidyl-tRNA hydrolase, pth	3.0	0.001	Translation	Rv1014c
6191	translation initiation inhibitor	2.5	0.016	Translation	Rv3678c
1439	ribosomal protein L2, rplB	-2.5	0.048	Translation	Rv0704
1442	ribosomal protein S3, rpsC	-2.8	0.035	Translation	Rv0707
1485	methionine aminopeptidase, type I, map	-9.6	0.034	Translation	Rv0734
2519	ribosomal protein S2, rpsB	-2.7	0.007	Translation	Rv2890c
2649	tRNA pseudouridine synthase B, truB	-2.1	0.001	Translation	Rv2793c
4571	ribosomal protein S20, rpsT	-4.7	0.013	Translation	Rv2413
4624	ribosomal protein L27, rpmA	-2.9	0.004	Translation	Rv2441c
1033/2313	ribonucleoside-diphosphate reductase, beta	-2.2	0.019	Translation	
0438	periplasmic binding protein (High affinity iron ion transport)	only WT_{ms}	0.013	Divalent cation homeostasis	
2615	chelatase	only WT_{ms}	0.000	Divalent cation homeostasis	Rv2850c
2652	iron repressor protein, SirR	only WT_{ms}	0.038	Divalent cation homeostasis	Rv2788
5014	copper-translocating P-type ATPase	-2.2	0.002	Divalent cation homeostasis	Rv0969
5006	phosphohistidine phosphatase	only WT_{ms}	0.005	Amino acid metabolism	Rv1276c
4538	cysteine desulphurase, SufS	only WT_{ms}	0.035	Amino acid metabolism	
1041/2321	phosphoserine phosphatase (glycine biosynthesis/L-serine biosynthesis), SerB	2.0	0.042	Amino acid metabolism	
3033	3-dehydroquinate synthase, aroB	2.5	0.002	Amino acid metabolism	Rv2538c
3183	threonine dehydratase (pyruvate family amino acid biosynthesis/isoleucine biosynthesis from threonine and pyruvate), ilvA	2.0	0.018	Amino acid metabolism	Rv1559

3317	dihydrodipicolinate reductase, N-terminus domain (lysine biosynthesis aspartate family amino acid biosynthesis), dapB	2.3	0.041	Amino acid metabolism	
3648	glycine cleavage system H protein, gcvH	2.0	0.001	Amino acid metabolism	Rv1826
1671	succinate dehydrogenase hydrophobic membrane, sdhD	-2.2	0.018	Amino acid metabolism	Rv3317
2111	chorismate mutase	-2.2	0.034	Amino acid metabolism	Rv1885c
5524	succinyl-CoA synthetase, alpha subunit, sucD	-2.1	0.014	Amino acid metabolism	Rv0952
2684	dihydrodipicolinate synthase, dapA	-2.1	0.021	Amino acid metabolism	Rv2753c
1668	pyridoxamine 5'-phosphate oxidase family	only WT_{ms}	0.050	Cofactor metabolism	
2653	riboflavin biosynthesis protein RibF	only WT_{ms}	0.036	Cofactor metabolism	Rv2786c
1822	biotin-[acetyl-CoA-carboxylase] ligase	only WT_{ms}	0.007	Cofactor metabolism	
0964	pyridoxamine 5'-phosphate oxidase family	2.7	0.010	Cofactor metabolism	
2201	ZbpA protein, Z-domain biotinylation protein	2.1	0.003	Cofactor metabolism	Rv0130
4829	cytochrome P450	only ΔESX-3_{ms}	0.013	Reduction/oxidation	
6858	epoxide hydrolase 1	only ΔESX-3_{ms}	0.035	Reduction/oxidation	
4685	oxidoreductase	only WT_{ms}	0.020	Reduction/oxidation	
2275	hydrogenase expression/formation protein HypD	only WT_{ms}	0.011	Reduction/oxidation	
6031	carveol dehydrogenase	only WT_{ms}	0.025	Reduction/oxidation	
0127	oxidoreductase, zinc-binding dehydrogenase	2.2	0.018	Reduction/oxidation	Rv0162c
0835	copper/zinc superoxide dismutase, sodD	-2.6	0.014	Reduction/oxidation	Rv0432
1011/2291	short chain dehydrogenase	-3.3	0.045	Reduction/oxidation	
2351	electron transfer flavoprotein, beta subunit, etfB	-2.1	0.036	Reduction/oxidation	Rv3029c
5287	dehydrogenase	2.0	0.025	Reduction/oxidation	
4119	3-hydroxybutyryl-CoA dehydratase	only WT_{ms}	0.039	Reduction/oxidation	
6239	1,3-propanediol dehydrogenase	only WT_{ms}	0.015	Reduction/oxidation	
3422	D-isomer specific 2-hydroxyacid dehydrogenase,	2.1	0.012	Reduction/oxidation	
1363	glucokinase	only WT_{ms}	0.003	Carbohydrate metabolism	Rv0650
1695	phosphoglucomutase/phosphomannomutase	only WT_{ms}	0.022	Carbohydrate metabolism	Rv3308
2512	lactate 2-monooxygenase	only WT_{ms}	0.024	Carbohydrate metabolism	
0361	glycosyl hydrolase family protein 3	only WT_{ms}	0.031	Carbohydrate metabolism	Rv0237
3269	putative sugar ABC transporter ATP-binding	only WT_{ms}	0.002	Carbohydrate metabolism	
6804	sugar ABC transporter substrate-binding protein	only WT_{ms}	0.013	Carbohydrate metabolism	
0049	pirin domain protein	only WT_{ms}	0.002	Carbohydrate metabolism	
4183	phosphoglycolate phosphatase, chromosomal	only WT_{ms}	0.016	Carbohydrate metabolism	
3262	L-xylulose reductase	only WT_{ms}	0.016	Carbohydrate metabolism	
6645	2-methylcitrate dehydratase 2	only ΔESX-3_{ms}	0.000	Carbohydrate metabolism	Rv1130
0947	acyltransferase	only WT_{ms}	0.050	Fatty acid metabolism	Rv0502

1886	fatty acid desaturase	only WT_{ms}	0.024	Fatty acid metabolism	Rv3229c
3139	enoyl-CoA hydratase/isomerase	only WT_{ms}	0.045	Fatty acid metabolism	Rv1472
3812	acyl-CoA thioesterase	only WT_{ms}	0.019	Fatty acid metabolism	
5886	enoyl-CoA hydratase	only WT_{ms}	0.002	Fatty acid metabolism	
4833	putative acyl-CoA dehydrogenase	only WT_{ms}	0.007	Fatty acid metabolism	
4118	acyl-CoA dehydrogenase	2.0	0.031	Fatty acid metabolism	
5276	enoyl-CoA hydratase	3.1	0.021	Fatty acid metabolism	Rv1071c
2393	NAD-dependent glycerol-3-phosphate, gpsA	-4.4	0.020	Fatty acid metabolism	Rv2982c
1284	PIN domain protein, single strand nuclease	only WT_{ms}	0.024	DNA/RNA metabolism	Rv0624
2389	DNA-binding protein HU, hup/hlp, stabilisation	-2.0	0.038	DNA/RNA metabolism	Rv2986c
2630	DHH family protein	-2.4	0.009	DNA/RNA metabolism	
5278	conserved membrane-spanning protein (biosynthesis and degredation of peptidoglycan)	only $\Delta ESX\text{-}3_{ms}$	0.012	Cell wall synthesis	Rv1069c
6935	N-acetylmuramoyl-L-alanine amidase, (degradation of peptidoglycan)	2.0	0.035	Cell wall synthesis	Rv3915
1826	dTDP-RhA:a-D-GlcNAc-diphosphoryl polyprenol,	-2.5	0.028	Cell wall synthesis	Rv3265c
0380	MmpS4 protein	2.9	0.044	Cell membrane and cell surface	Rv0457c
0143	probable conserved mce associated membrane	2.2	0.011	Cell membrane and cell surface	
0640	oligopeptide transport ATP-binding protein OppD	-2.7	0.007	Cell membrane and cell surface	
0919	heparin-binding hemagglutinin	-3.0	0.038	Cell membrane and cell surface	Rv0475
2695	35 kDa protein	-2.7	0.000	Cell membrane and cell surface	Rv2744c
4239	LppM protein	-2.6	0.011	Cell membrane and cell surface	Rv2171
3887	twin arginine-targeting protein translocase, tatA	-6.0	0.048	Cell membrane and cell surface	Rv2094c
6109	LpqG protein	only WT_{ms}	0.020	Cell membrane and cell surface	
4228	cell division protein FtsW	only WT_{ms}	0.005	Cell division	Rv2154c
1811	septum formation protein Maf	only WT_{ms}	0.048	Cell division	Rv3282
0095	methyltransferase	only $\Delta ESX\text{-}3_{ms}$	0.016	Methyltransferase	
2773	hypothetical RNA methyltransferase	WT	0.048	Methyltransferase	Rv2689c
6278	metallo-beta-lactamase superfamily protein,	only WT_{ms}	0.019	Antibiotic	
6339	metallo-beta-lactamase family protein	only WT_{ms}	0.009	Antibiotic	Rv3762c
6086	antibiotic biosynthesis monooxygenase domain	2.2	0.018	Antibiotic	Rv3592
6190	metallo-beta-lactamase family protein	2.8	0.042	Antibiotic	Rv3677c
6575	beta-lactamase	2.2	0.003	Antibiotic	Rv3775
3958	glyoxalase/bleomycin resistance	2.2	0.006	Antibiotic	
1134	putative protease HtpX	-5.1	0.011	Stress protein	Rv0563
5733	universal stress protein family protein	2.2	0.013	Stress protein	
6202	secreted protein	only WT_{ms}	0.022	Secreted protein	Rv3683

6218	secreted protein	only WT_{ms}	0.007	Secreted protein	Rv3691
5015	secreted protein	only WT_{ms}	0.019	Secreted protein	
5664	binding protein	2.2	0.026	Other	
4495	CBS domain protein	-5.6	0.025	Other	
5162	cinnamoyl ester hydrolase	only WT_{ms}	0.013	Other	
5436	4-diphosphocytidyl-2C-methyl-D-erythritol, ispE	only WT_{ms}	0.006	Other	Rv1011
6423	glycerophosphoryl diester phosphodiesterase	only WT_{ms}	0.008	Other	Rv3842
0884	glyoxalase family protein	3.6	0.018	Other	
1603	IMP dehydrogenase family protein	2.0	0.007	Other	Rv3410
1835	TobH protein	2.0	0.024	Other	Rv3256c
2042	phosphotransferase enzyme family protein	2.4	0.040	Other	Rv3168
2932	HIT family protein hydrolase	2.2	0.038	Other	Rv2613c
3457	DoxD family protein/pyridine nucleotide-disulfide oxidoreductase	3.3	0.049	Other	
5680	glyoxalase family protein	2.1	0.012	Other	Rv0887c
3859	glycosyl transferase, group 2 family protein	-2.0	0.010	Other	
5182	2-nitropropane dioxygenase	-3.9	0.032	Other	
5789	putative thiosulfate sulfurtransferase	-2.3	0.015	Other	Rv0815c
6114	inorganic pyrophosphatase, ppa	-2.0	0.026	Other	Rv3628
6554	hypothetical ABC transporter ATP-binding	-4.6	0.045	Other	Rv1348
3168	phenolpthIocerol synthesis type-i polyketide ppse	only WT_{ms}	0.046	Other	
0074/1865/ 2824/3433 /5181	IS1549, transposase	only WT_{ms}	0.028	Other	
2746	conserved hypothetical alanine and leucine rich	only $\Delta ESX\text{-}3_{ms}$	0.004	Other	Rv2714
3026	hypothetical protein	only $\Delta ESX\text{-}3_{ms}$	0.000	Hypothetical protein	Rv2554c
0971	hypothetical protein	only WT_{ms}	0.043	Hypothetical protein	Rv0526
1891	hypothetical protein	only WT_{ms}	0.028	Hypothetical protein	Rv3226c
2032	hypothetical protein	only WT_{ms}	0.002	Hypothetical protein	
2197	hypothetical protein	only WT_{ms}	0.034	Hypothetical protein	
2779	hypothetical protein	only WT_{ms}	0.000	Hypothetical protein	Rv2680
3027	hypothetical protein	only WT_{ms}	0.015	Hypothetical protein	Rv2553c
3229	hypothetical protein	only WT_{ms}	0.000	Hypothetical protein	
6081	hypothetical protein	only WT_{ms}	0.010	Hypothetical protein	Rv3587c
6373	hypothetical protein	only WT_{ms}	0.003	Hypothetical protein	
4693	hypothetical protein	only WT_{ms}	0.019	Hypothetical protein	Rv3623
4141	hypothetical protein	only WT_{ms}	0.034	Hypothetical protein	

5279	hypothetical protein	only WT_{ms}	0.002	Hypothetical protein	Rv1066
6080	hypothetical protein	2.8	0.007	Hypothetical protein	Rv3586
6159	hypothetical protein	2.1	0.035	Hypothetical protein	Rv3648c
1136	hypothetical protein	2.0	0.007	Hypothetical protein	
1734	hypothetical protein	2.8	0.004	Hypothetical protein	Rv3304
3004	conserved hypothetical protein	2.7	0.000	Hypothetical protein	
4284	hypothetical protein	2.0	0.005	Hypothetical protein	Rv2216
0368	hypothetical protein	-2.2	0.016	Hypothetical protein	
1513	hypothetical protein	-2.2	0.044	Hypothetical protein	
3081	hypothetical protein	-2.0	0.020	Hypothetical protein	Rv2569c
3261	hypothetical protein	-2.7	0.008	Hypothetical protein	
4199	hypothetical protein	-5.8	0.005	Hypothetical protein	Rv2140c
4322	hypothetical protein	-2.2	0.044	Hypothetical protein	Rv2240c
6129	hypothetical protein	-4.3	0.010	Hypothetical protein	
6386	membrane protein	-2.0	0.029	Hypothetical protein	Rv3792
6502	hypothetical protein	-3.7	0.024	Hypothetical protein	Rv2366

*Differential abundances represent the fold change in abundance of a protein in ΔESX-3_{ms} relative to WT_{ms}. Negative values indicate a lower abundance in ΔESX-3_{ms} than $WT_{ms.}$ Where indicated proteins were detected in only one strain.

The ΔESX-3_{ms} and WT_{ms} strains exhibit clear differences in protein profiles, indicative that ESX-3 plays a biological role in *M. smegmatis* physiology under these conditions. The differentially expressed proteins are involved in multiple cellular functions including amino acid and cofactor metabolism, fatty acid and carbohydrate metabolism, redox homeostasis, transcriptional regulation, translation and cell wall synthesis. ESX-3 clearly impacts on multiple aspects of the cellular biology of *M. smegmatis*, some of which, including amino acid, carbohydrate and fatty acid metabolism and oxidative stress, correlate with the variation detected in the metabolomics experiment.

3.3.2.3. Culture filtrate proteomic analysis of ESX-3

Culture filtrate fractions were prepared from WT_{ms} and ΔESX-3_{ms} and the constituent proteins identified by LC-MS/MS analysis on the LTQ Orbitrap Velos. One thousand eight hundred and three proteins were identified in the two strains (Supplementary table S3.4). Protein abundances were compared to identify exported proteins which are influenced by the presence or absence of ESX-3 (Table 3.3.3). No ESX-3 proteins were detected in the CF fractions of either strain. Similar pathways to those identified in the WCL proteome analysis were found to be influenced by the deletion of ESX-3 in the CF fraction, including amino acid biosynthesis, vitamin/cofactor biosynthesis, translation and transcriptional regulation and cell wall biosynthesis, although most proteins detected in different abundances appear to be cytosolic, and not extracellular proteins.

3.3.2.4. Divalent metal cation regulation of ESX-3 expression

Regulation of expression of ESX-3 by various divalent metal cations was investigated using a β-galactoside promoter assay. The activity of the *M. smegmatis* ESX-3 promoter in the absence and presence of various metal cations was determined and compared to that of the *M. tuberculosis* ESX-3 promoter (Figure 3.3.2). The *M. smegmatis* ESX-3 promoter activity is only repressed by the presence of iron, while the *M. tuberculosis* ESX-3 promoter is repressed by zinc, and further repressed in the presence of both iron and zinc. Promoter activities were not determined for cobalt in *M. smegmatis* and for copper in *M. bovis* BCG as these metals had a toxic effect on the respective species. No regulation was observed for any other metal cations.

Table 3.3.3. Proteins with different abundances in ΔESX-3_{ms} and WT_{ms} culture filtrates, classified according to biological function.

MSMEG_	Product	Differential abundance*	p value	Biological function	H37Rv ortholog
6077	transcriptional regulator, CarD family protein	only WT_{ms}	0.013	Transcriptional regulation	Rv3583c
6451	transcriptional regulator, ArsR family protein	only WT_{ms}	0.026	Transcriptional regulation	Rv0081
2916	DNA-binding response regulator, PhoP family	only Δ$ESX\text{-}3_{ms}$	0.003	Transcriptional regulation	Rv1033c
4070	transcriptional regulator, TetR family protein,	only Δ$ESX\text{-}3_{ms}$	0.022	Transcriptional regulation	
4242	transcriptional regulatory protein	-2.3	0.017	Transcriptional regulation	Rv2175c
6628	transcriptional regulator, TetR family protein	-3.6	0.024	Transcriptional regulation	
1345	transcription antitermination protein NusG	-2.4	0.001	Transcriptional regulation	Rv0639
5774	tRNA-dihydrouridine synthase, putative	only Δ$ESX\text{-}3_{ms}$	0.001	Translation	Rv0823c
1365	ribosomal protein L7/L12, rplL	-3.2	0.020	Translation	Rv0652
1657	tryptophanyl-tRNA synthetase, trpS	-2.9	0.033	Translation	Rv3336c
3035	translation elongation factor P, efp	-2.3	0.003	Translation	Rv2534c
3792	ribosomal protein L35, rpmI	-3.9	0.009	Translation	Rv1642
6897	ribosomal protein S6 , rpsF	-2.5	0.001	Translation	Rv0053
4291	D-tyrosyl-tRNA(Tyr) deacylase, dtd	9.9	0.046	Translation	Rv1897c
0020	periplasmic binding protein - iron ion transporter	2.2	0.002	Divalent cation homeostasis	
2750	iron-dependent repressor IdeR	-2.2	0.022	Divalent cation homeostasis	Rv2711
3636	ferric iron-binding periplasmic protein of ABC transporter	-2.0	0.014	Divalent cation homeostasis	
3648	glycine cleavage system H protein, gcvH	2.1	0.021	Amino acid metabolism	Rv1826
3173	L-asparaginase (asparagine to aspartate)	4.0	0.006	Amino acid metabolism	Rv1538c
1496	3-hydroxyisobutyrate dehydrogenase , mmsB (valine, leucine, isoleucine degradation)	34.5	0.030	Amino acid metabolism	Rv0751c
4976	isochorismatase hydrolase	-2.9	0.028	Amino acid metabolism	
3872	precorrin-8X methylmutase	only WT_{ms}	0.046	Cofactor metabolism	Rv2065
3873	cobalamin biosynthesis protein cobIJ	only Δ$ESX\text{-}3_{ms}$	0.021	Cofactor metabolism	Rv2066
2937	pyridoxine biosynthesis protein	2.0	0.003	Cofactor metabolism	Rv2606c
2198	ThiF family protein	only Δ$ESX\text{-}3_{ms}$	0.039	Cofactor metabolism	Rv1355c
1416	pyridine nucleotide-disulphide oxidoreductase	-2.0	0.013	Reduction/oxidation	Rv0688
3199	quinolinate synthetase complex, A subunit, nad	-3.3	0.001	Reduction/oxidation	Rv1594c
2923	dehydrogenase/reductase SDR family protein	only WT_{ms}	0.002	Reduction/oxidation	
5319	alkylhydroperoxidase AhpD core	only WT_{ms}	0.001	Reduction/oxidation	
2566	3-alpha-(or 20-beta)-hydroxysteroid dehydrogenase	only Δ$ESX\text{-}3_{ms}$	0.025	Reduction/oxidation	
4305	phosphoglycerate mutase	2.5	0.026	Carbohydrate metabolism	Rv2228c
2119	N-acetylglucosamine-6-phosphate deacetylase, nagA	only WT_{ms}	0.008	Carbohydrate metabolism	Rv3332

3598	periplasmic sugar-binding proteins	only ΔESX-3$_{ms}$	0.009	Carbohydrate metabolism	
4696	alpha-amylase family protein	only ΔESX-3$_{ms}$	0.004	Carbohydrate metabolism	Rv2471
0603	putative acyl-CoA dehydrogenase	only ΔESX-3$_{ms}$	0.000	Fatty acid metabolism	Rv0271c
6001	enoyl-CoA hydratase	2.1	0.028	Fatty acid metabolism	Rv3550
0986	RemM protein (DNA binding)	-2.3	0.007	DNA/RNA metabolism	
2399	uracil-DNA glycosylase	only WT$_{ms}$	0.003	DNA/RNA metabolism	Rv2976c
1900	D-alanyl-D-alanine carboxypeptidase family	2.1	0.002	Cell wall synthesis	
0928	UDP-N-acetylenolpyruvoylglucosamine reductase, murB	-3.5	0.049	Cell wall synthesis	Rv0482
6354	serine esterase, cutinase family protein	-3.4	0.008	Cell wall synthesis	Rv1758
3790	RNA methyltransferase, TrmH family protein	only ΔESX-3$_{ms}$	0.003	Methyltransferase	Rv1644
6784	beta-lactamase	only ΔESX-3$_{ms}$	0.001	Antibiotic	
6815	secreted protein	-2.1	0.034	Secreted protein	
5245	universal stress protein family protein	2.1	0.009	Stress	Rv2028c/3134c
0760	thioesterase family protein	only WT$_{ms}$	0.001	Other	Rv0356c
1078	hydrolase	only WT$_{ms}$	0.000	Other	
1835	TobH protein	only WT$_{ms}$	0.004	Other	Rv3256c
6425	rhodanese-like domain protein	2.1	0.024	Other	
0865	bacterial membrane flanked domain family	2.5	0.044	Other	
0408	type I modular polyketide synthase	2.5	0.006	Other	
2410	putative serine-threonine protein kinase	-2.1	0.035	Other	Rv2969c
2769	155 TrkB protein	-2.1	0.012	Other	Rv2692
6092	Lsr2 protein	-2.1	0.026	Other	Rv3597c
6282	KanY protein	-2.0	0.018	Other	Rv3718c
4999	bacterial extracellular solute-binding protein,	-2.2	0.005	Other	Rv1280c
6783	integral membrane protein	-2.3	0.001	Other	
5224	4-hydroxy-3-methylbut-2-enyl diphosphate reductase, ispH	only WT$_{ms}$	0.002	Other	Rv1110
4043	amidohydrolase 2	only ΔESX-3$_{ms}$	0.001	Other	
5017	lipoprotein	only ΔESX-3$_{ms}$	0.029	Other	
6035	nitrilotriacetate monooxygenase component B	only ΔESX-3$_{ms}$	0.000	Other	Rv3567c
6788	dihydroxyacetone kinase	only ΔESX-3$_{ms}$	0.007	Other	
5827	glyoxalase family protein	only ΔESX-3$_{ms}$	0.001	Other	Rv0801
5913	dioxygenase	only ΔESX-3$_{ms}$	0.020	Other	
4114	naphthoate synthase, menB	-2.2	0.009	Other	
0369	hypothetical protein	only WT$_{ms}$	0.038	Hypothetical protein	
1581	hypothetical protein	only WT$_{ms}$	0.009	Hypothetical protein	

1976	hypothetical protein	only WT_{ms}	0.000	Hypothetical protein	
2763	hypothetical protein	only WT_{ms}	0.003	Hypothetical protein	Rv2699c
2861	hypothetical protein	only WT_{ms}	0.011	Hypothetical protein	
3024	hypothetical protein	only WT_{ms}	0.024	Hypothetical protein	Rv2556c
3843	hypothetical protein	only WT_{ms}	0.003	Hypothetical protein	Rv1628c
5634	hypothetical protein	only WT_{ms}	0.048	Hypothetical protein	Rv0910
6579	hypothetical protein	only WT_{ms}	0.014	Hypothetical protein	
3114	hypothetical protein	only ΔESX-3_{ms}	0.003	Hypothetical protein	
3253	hypothetical protein	only ΔESX-3_{ms}	0.013	Hypothetical protein	
5865	hypothetical protein	only ΔESX-3_{ms}	0.019	Hypothetical protein	Rv0762c
5932	hypothetical protein	only ΔESX-3_{ms}	0.012	Hypothetical protein	Rv3531c
0692	hypothetical protein	2.8	0.011	Hypothetical protein	
2434	hypothetical protein	2.5	0.034	Hypothetical protein	Rv2910c
1042/2322	hypothetical protein	4.2	0.006	Hypothetical protein	
0423	hypothetical protein	-2.4	0.001	Hypothetical protein	
3171	hypothetical protein	-2.1	0.011	Hypothetical protein	
3674	hypothetical protein	-2.1	0.009	Hypothetical protein	
4399	hypothetical protein	-3.7	0.037	Hypothetical protein	
4689	hypothetical protein	-3.2	0.024	Hypothetical protein	Rv1690/3354
4784	hypothetical protein	-2.3	0.005	Hypothetical protein	
6454	hypothetical protein	-2.1	0.034	Hypothetical protein	
2027	conserved hypothetical protein TIGR00026	-2.2	0.001	Hypothetical protein	

*Differential abundances represent the fold change in abundance of a protein in ΔESX-3_{ms} relative to WT_{ms}. Negative values indicate a lower abundance in ΔESX-3_{ms} than $WT_{ms.}$ Where indicated proteins were detected in only one strain.

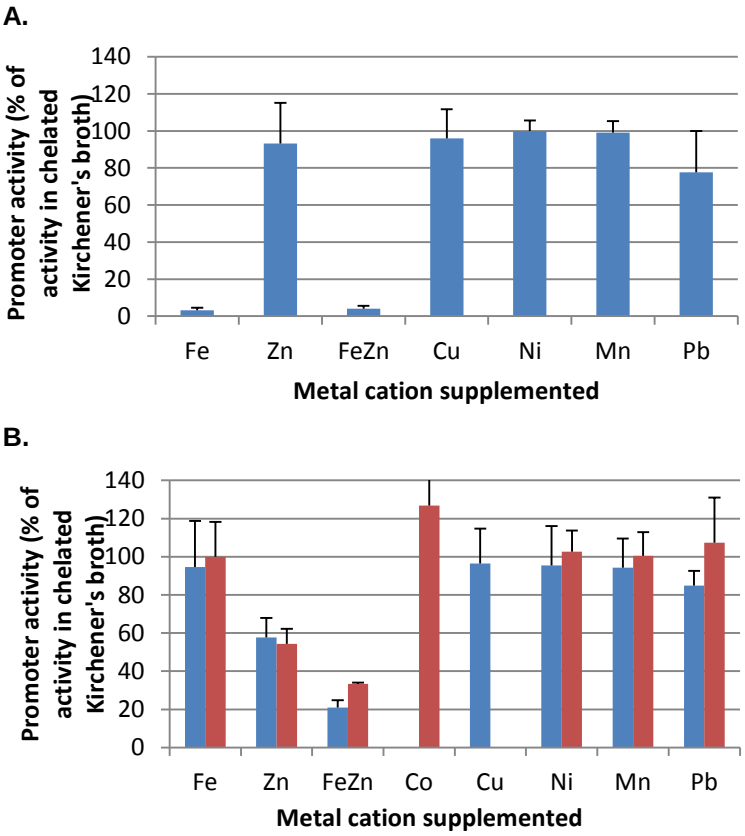


Figure 3.3.2. Metal cation regulation of the *M. smegmatis* (A) and *M. tuberculosis* (B) ESX-3 promoter. Cultures were grown in chelated Kirchener's broth and Kirchener's broth supplemented with the metal cations, iron (Fe, 100uM); zinc (Zn, 100uM); iron and zinc (FeZn,100uM ea.); cobalt (Co, 7.5uM); copper (Cu, 50uM); nickel (Ni, 35uM); manganese (Mn, 0.75uM) and lead (Pb, 10uM) and the promoter activities expressed as a percentage of the un-supplemented promoter activity. Promoter activities assayed in *M. smegmatis* are represented by blue bars, and in *M. bovis* BCG by red bars. The *M. smegmatis* promoter was only regulated by iron, while the *M. tuberculosis* promoter was regulated by zinc, and further repressed by the co-supplementation with iron and zinc.

3.3.3. Discussion

The ESX-3 secretion system has been implicated in divalent metal cation homeostasis, and specifically iron and zinc uptake. Iron is essential for all organisms, in which it functions as a cofactor for many enzymes and forms structural scaffolds for various proteins. Excessive intracellular iron is, however, toxic to cells, and iron levels are tightly regulated by mechanisms which control iron uptake and storage.[161] Zinc is also an important cofactor for many enzymes and its uptake is tightly regulated.[162] The acquisition of these metal cations, and the tight regulation thereof is important for *M. smegmatis*. The involvement of ESX-3 in iron and zinc homeostasis is well established.[50, 113-116] In this study we have investigated the

biological impact of the absence and presence of ESX-3 on the physiology of *M. smegmatis* by investigating the metabolomics and proteomic differences between *M. smegmatis* wild type and Δ*ESX-3* strains.

3.3.3.1. The metabolomic analysis of the *M. smegmatis* ESX-3 knockout strain indicates altered amino acid and energy metabolism and increased oxidative stress

Amongst the compounds which contribute most to the variation between the metabolomes of WT_{ms} and $\Delta ESX\text{-}3_{ms}$ are various amino acids, sugars and fatty acids (Figure 3.3.3). Notably, most amino acids are reduced in $\Delta ESX\text{-}3_{ms}$, suggesting a general reduction in amino acid synthesis. Of the amino acids which are more prevalent in $\Delta ESX\text{-}3_{ms}$, several are derived directly from glutamate, which is the only amino acid provided directly in the culture medium. Therefore the increased levels of glutamine, citrulline and oxy-proline may stem from increased uptake of glutamate in response to generalised amino acid depletion. Furthermore, influx of glutamate may increase the activity of the glyoxylate shunt, increasing glycine production.

Many of the metabolite differences between WT_{ms} and $\Delta ESX\text{-}3_{ms}$ can be attributed to the involvement of ESX-3 in iron and zinc homeostasis. The synthesis of histidine, which was completely absent from $\Delta ESX\text{-}3_{ms}$, is dependent on two enzymes, HisG and HisE, which are coexpressed, but expression of which is repressed by IdeR in the presence of iron.[50, 51] In addition, expression of PheA and TrpE2, which are required for the synthesis of phenylalanine and tryptophan, is also regulated by IdeR in response to iron,[50, 163, 164] and synthesis of threonine and isoleucine is upregulated during iron starvation.[165] Considering these data, the reduced concentrations of these amino acids in $\Delta ESX\text{-}3_{ms}$ may indicate an increase in iron within $\Delta ESX\text{-}3_{ms}$ relative to $WT_{ms.}$

In addition to the glutamate supplied in the culture medium, glutamate synthase activity is dependent on iron in many organisms,[165] therefore increased iron concentrations within $\Delta ESX\text{-}3_{ms}$ may contribute to the increased levels of glutamate and glutamine and additional glutamate-metabolism linked compounds such as oxyproline, citrulline, succinic semialdehyde, butanoic acid and the purines and pyrimidines, all of which were detected at increased concentrations in $\Delta ESX\text{-}3_{ms.}$

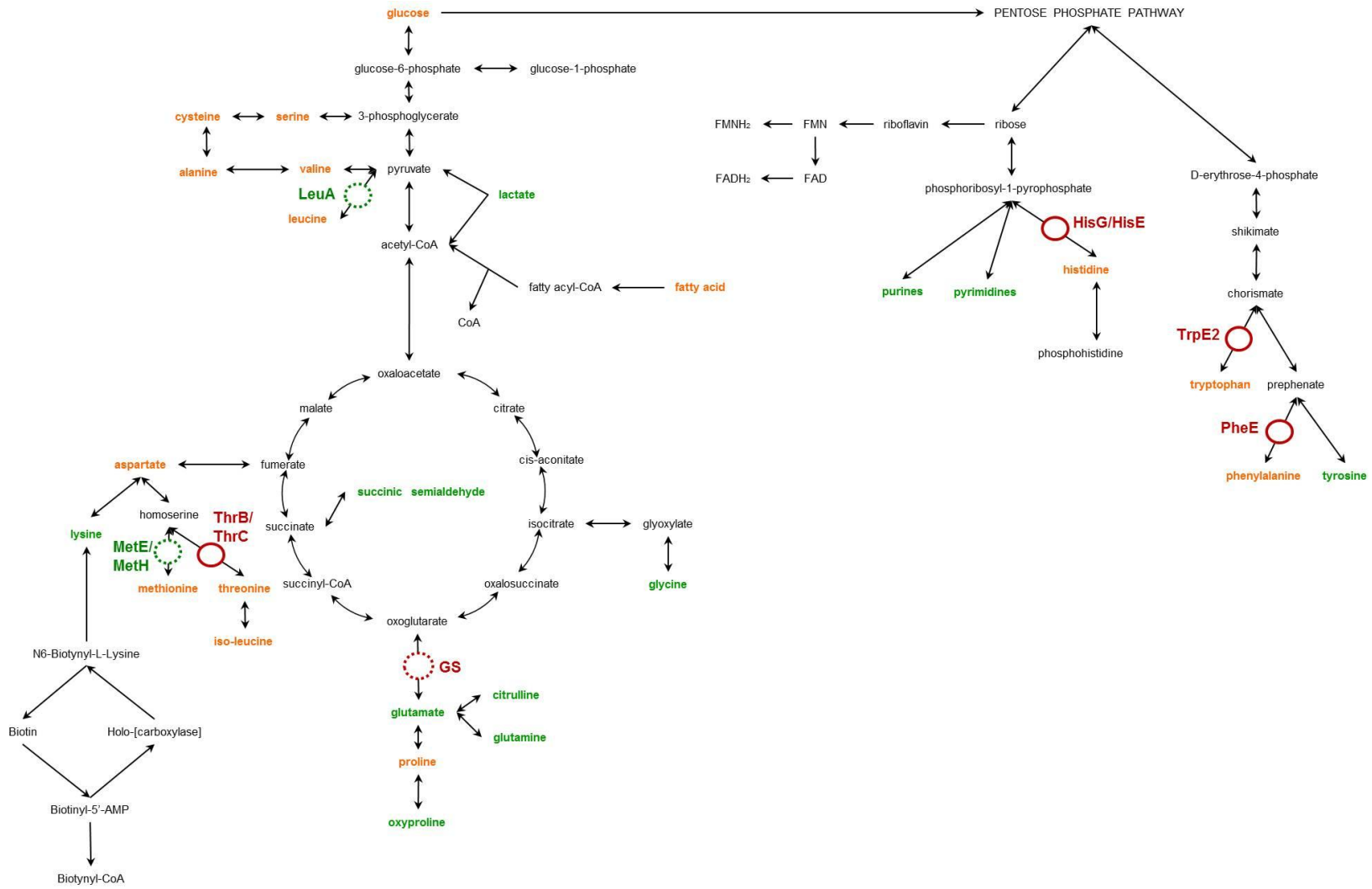


Figure 3.3.3. The metabolites influenced by the deletion of ESX-3 from *M. smegmatis*. Metabolites in green are present in increased concentrations in ΔESX-3_{ms}, and metabolites in orange in reduced concentrations in ΔESX-3_{ms}, relative to WT_{ms}. Solid red circles indicate enzymes which are repressed under high iron conditions, interrupted red circles represent enzymes which are iron-dependent and interrupted green circles represent enzymes which are zinc-dependent, according to the literature, as described in the text. Metabolic pathways were adapted from KEGG.

Several amino acid biosynthesis pathways are also influenced by zinc, which functions as a cofactor for many enzymes. MetE and MetH are cobalamin independent and cobalamin dependent methionine synthases, respectively, both of which require zinc to enable the biosynthesis of methionine.[166] Methionine was detected at significantly reduced levels in ΔESX-3_{ms}, possibly indicating that activity of these enzymes is reduced, pointing to lower zinc levels in this strain. LeuA, responsible for leucine synthesis, is also a zinc-dependent enzyme,[167] and leucine levels are also reduced in ΔESX-3_{ms}, lending support to the theory of reduced zinc levels.

Disregulation of iron and zinc homeostasis in ΔESX-3_{ms} may influence other metabolic pathways. Excess iron is a significant source of oxidative stress in all organisms, due to the Fenton reactions in which ferrous iron and hydrogen peroxide produce ferric iron, hydroxide anions and hydroxide radicals.[168] The oxidative response in *M. bovis* BCG involves an increase in amino acid transport, increased expression of some amino acid synthesis genes, decreased lipid transport and metabolism and increased membrane synthesis.[169] This correlates with the metabolic phenotype of ΔESX-3_{ms}, which exhibits altered amino acid metabolism as well as a reduction in fatty acids and increased lactic acid concentrations. Oxidative stress also results in inhibition of the electron transport chain, increasing the catabolism of simple sugars via glycolysis and the glyoxylate shunt.[170] This may account for the reduced levels of simple sugars in ΔESX-3_{ms}, as well as the reduced concentrations of valine, alanine and leucine which are derived from pyruvate, a substrate of the glyoxylate cycle, and the excess of glycine, a product of the glyoxylate shunt.

ESX-3 has been implicated in the acquisition of iron and zinc, and as such we anticipated that the absence of ESX-3 would result in iron and zinc starvation. However, the metabolic profile of ΔESX-3_{ms} relative to wild type *M. smegmatis* suggests that in the absence of ESX-3, the organism experiences iron excess. It has previously been shown that repression of ESX-3 in *M. bovis* BCG results in upregulation of expression of mycobactin and other iron-uptake genes.[116] This was ascribed to a greater perceived iron starvation, resulting in overcompensation for the repression of ESX-3. *M. smegmatis* possesses an alternative iron acquisition system, the exochelin system which may be induced to acquire additional iron in the absence of ESX-3, resulting in the greater iron levels and subsequent metabolic changes observed in ΔESX-3_{ms}. Alternatively, ESX-3 may be involved in the export of excess iron resulting in iron build-up in ΔESX-3_{ms}. ESX-3 has not previously been associated with zinc homeostasis in *M. smegmatis* where its expression is regulated only by iron. However, here we describe metabolic changes which implicate ESX-

3 in the uptake of zinc, suggesting that in *M. smegmatis* ESX-3 may play a role in the homeostasis of metal cations which do not directly impact on its expression.

This metabolomic analysis has clearly distinguished the *M. smegmatis* wildtype and ESX-3 knockout strains based on their metabolite profiles. Furthermore, the variation in metabolite concentrations in the two strains can be correlated with the role of ESX-3 in iron and zinc uptake and/or homeostasis.

3.3.3.2. The proteomic profile of the *M. smegmatis* ESX-3 knockout strain does not indicate altered expression of the IdeR and Zur regulons.

The proteomic variation between ΔESX-3_{ms} and WT_{ms} correlates with the metabolic differences observed. Differentially expressed proteins are involved in the pathways and biological processes which influence the levels of metabolites which varied in the metabolomics study (Figure 3.3.4). However, the abundances of the enzymes which were presumed to be differentially expressed in the two strains, to explain the metabolomic variation, remained constant. Furthermore, the protein profiles of ΔESX-3_{ms} and WT_{ms} do not show the proposed disregulation of iron and zinc homeostasis; neither the *M. smegmatis* zinc (Zur) regulon[114] nor the iron (IdeR) regulon,[164] is varied between the two strains. Although some of the same metabolic pathways are implicated in both the metabolomics and proteomic results, differential abundance of the specific proteins used to explain the metabolomic variation, was not observed in the proteomic study.

Recently it has been shown that *M. smegmatis* acquires iron by passive diffusion through the Msp family of outer membrane porins and that high-affinity iron acquisition systems (mycobactins and exochelins) are obsolete under these conditions.[171] In iron concentrations of 150uM ferric citrate, as is present in 7H9 culture medium, the IdeR regulon is repressed, inhibiting expression of the high affinity iron uptake systems.[171] At iron concentrations greater than 5uM, expression of IdeR-regulated genes was minimal, suggesting that amongst other genes, ESX-3 is repressed under the conditions of this study. This is supported by the promoter activity assay in which expression from the *M. smegmatis* ESX-3 promoter is almost completely repressed in the presence of 100uM Ferric ammonium citrate. This supports the proteomic results which show no variation in the IdeR regulon protein abundances. Low levels of some ESX-3 proteins were detected in WT_{ms}, indicating that these proteins are being expressed, albeit at a low level, under these conditions; and the metabolomic and proteomic variation between the WT_{ms} and ΔESX-3_{ms} clearly indicate that this expression has significant effects on the physiology of *M. smegmatis*.

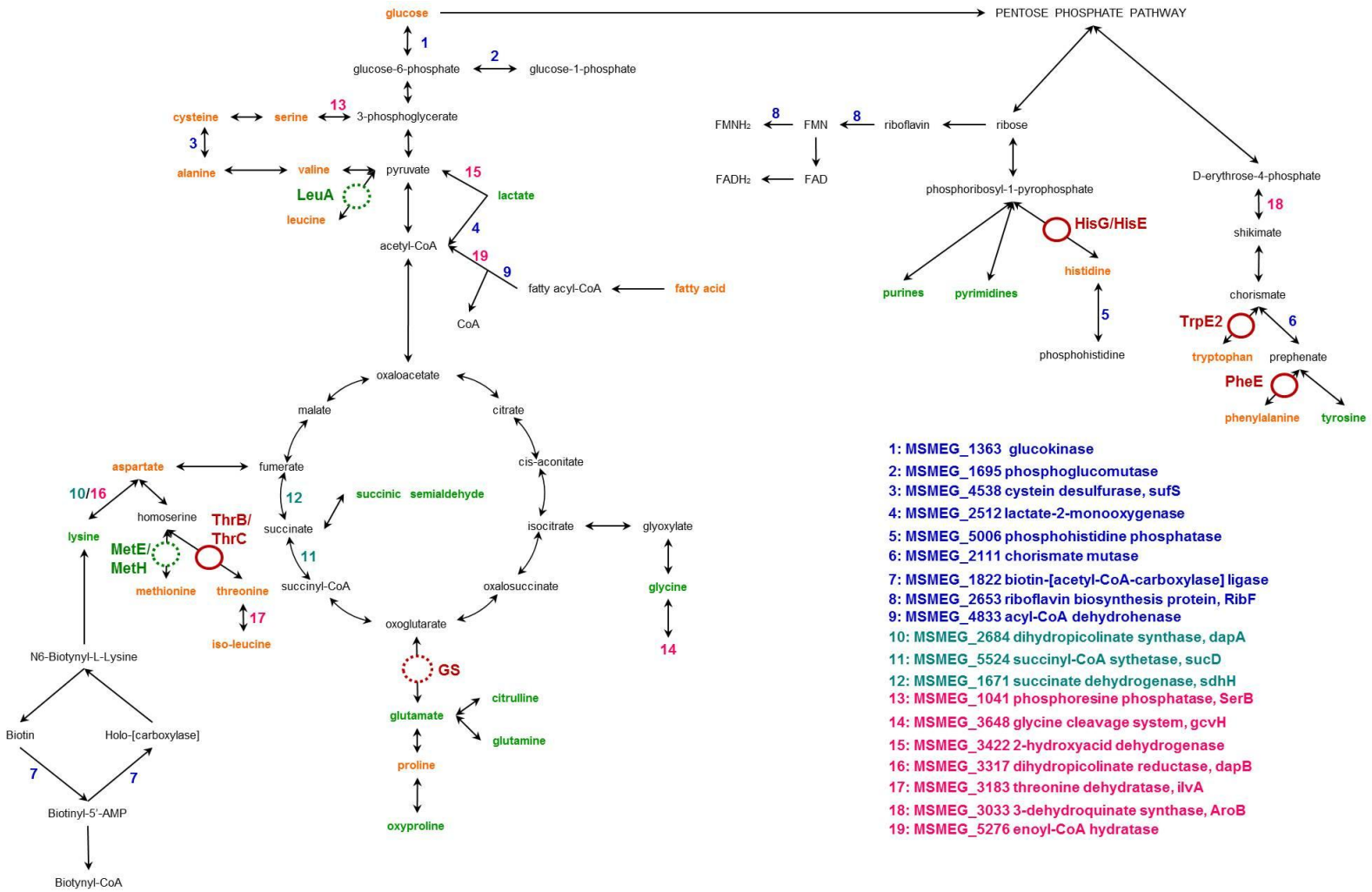


Figure 3.3.4. The enzymes influenced by the deletion of ESX-3 from *M. smegmatis*. Several enzymes involved in energy and amino acid metabolism were found to be differentially abundant in WT_{ms} and ΔESX-3_{ms} in this study. Enzymes detected only in WT_{ms} are indicated in blue, only in ΔESX-3_{ms} in pink, and at higher levels in WT_{ms} than ΔESX-3_{ms}, in teal. Although these enzymes correlate with the pathways involved in determining the altered levels of metabolites, they do not correlate with those which are regulated by IdeR in response to iron levels. Metabolic pathways were adapted from KEGG.

We do observe some disregulation of divalent metal cation-associated proteins in the WCLs and CFs of the two strains. The four differentially abundant proteins identified in the WCL proteomic analysis include a periplasmic binding protein (MSMEG_0438), thought to be involved in iron ion transport, a chelatase (MSMEG_2615), and a putative iron repressor protein (SirR, MSMEG_2652) which were only detected in WT_{ms}; and a copper-translocating P-type ATPase (MSMEG_5014), which is twice as abundant in WT_{ms} as in $\Delta ESX\text{-}3_{sm}$. Two additional periplasmic iron binding proteins are found in different abundances in the CFs of the two strains; MSMEG_0020 is twice as abundant in WT_{ms}, while MSMEG_3636 is twice as abundant in $\Delta ESX\text{-}3_{sm}$. IdeR (MSMEG_2750) is also twice as abundant in the CF of WT_{ms} than of $\Delta ESX\text{-}3_{ms}$. WT_{ms} contains increased abundances of iron-binding proteins, chelatases and Cu-translocators, as well as increased amounts of the iron repressor proteins SirR and IdeR. SirR is closely related to IdeR and other iron-repressors, but no role has been identified for SirR in iron homeostasis in mycobacteria,[172] and this regulator may be involved in the regulation of transcription in response to a different signal, potentially that of an additional metal cation. The significance of the increased abundance of IdeR in the CF of WT_{ms} is also unclear, as there is no difference in the abundance of this protein in the WCLs of the two strains. It is important to consider that it is not increased levels of these regulators which elicit a response, but the binding of the metal cation to the repressor which results in repression of transcription. These proteins provide evidence of some disregulation of divalent metal cation homeostasis in $\Delta ESX\text{-}3_{ms}$. The reduction in metal-cation-binding and -regulator proteins in $\Delta ESX\text{-}3_{ms}$ may indicate an increased iron concentration relative to $WT_{ms,}$ although this variation may be above the threshold for IdeR or may be indicative of a disregulation in the homeostasis of other divalent metal cations.

3.3.3.3. The proteome of $\Delta ESX\text{-}3_{ms}$ shows signs of disrupted polyamine metabolism

Polyamines are small, positively charged aliphatic hydrocarbons with quaternary nitrogen groups. Putrescine and spermidine are the most common polyamines found in bacteria.[173] Putrescine is synthesised from arginine, via agmatine or ornithine. S-adenosylmethionine, generated from methionine, is used in the conversion of putrescine to spermidine (Figure 3.3.5.).[174] Positively charged polyamines interact with negatively charged nucleic acids to stabilise and maintain their optimum conformation, and are therefore required for optimal DNA replication, transcription and translation.[175] They have also been linked to outer membrane functionality, physiological stress responses, microbial pathogenicity, biofilm formation, cell membrane and cell wall stabilisation and the synthesis of molecules involved in iron acquisition.[173, 175, 176]

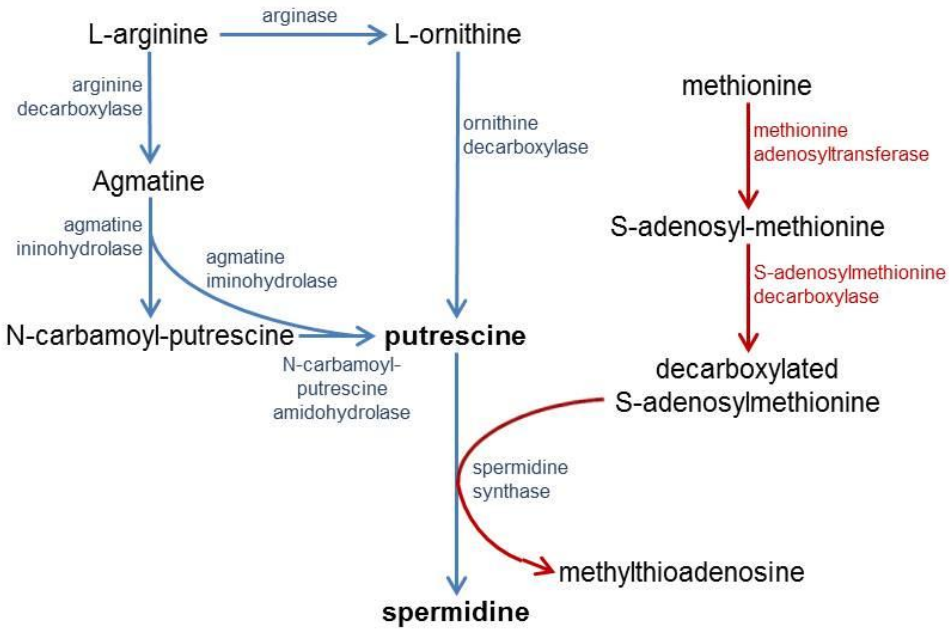


Figure 3.3.5. Bacterial polyamine biosynthesis. Putrescine is synthesised from arginine and ornithine. Spermidine biosynthesis from putrescine is dependent on decarboxylated S-adenosylmethionine which is derived from methionine.

Polyamine metabolism has not been well studied in mycobacteria, however the WCL and CF proteomic variation observed in ΔESX-3_{ms} closely resembles the transcriptomic and proteomic changes seen in *Plasmodium falciparum* when S-adenosylmethionine (SAM) decarboxylase, an essential enzyme in the production of spermidine, is inhibited,[177] and in polyamine deficient mutants of *E. coli*.[178] The reduced levels of ribosomal proteins, DNA stabilisation proteins and proteins involved in cell division, as well as altered levels of transcriptional regulators in ΔESX-3_{ms} correlate with the functions of polyamines in nucleic acid stabilisation and their requirement for optimal DNA replication, transcription and translation. The changes in protein levels in ΔESX-3_{ms} correlate with those observed upon inhibition of SAM decarboxylase (in *P. falciparum*), suggesting that ΔESX-3_{ms} is deficient in polyamines, specifically spermidine and spermine. Methionine was detected at considerably lower levels in ΔESX-3_{ms} than WT_{ms}. As methionine is a precursor of SAM, and carboxy-SAM, which is required for the synthesis of spermidine, and hence spermine, this may account for the apparent deficiency of polyamines ΔESX-3_{ms}.

Variations in abundance of other proteins including methyltransferases, fatty acid metabolism- and glycolysis- associated enzymes, proteins involved in oxidation-reduction reactions and enzymes involved in the biosynthesis and metabolism of cofactors (pyridoxamine, riboflavin, biotin and cobalamin) also correlate with the changes seen under SAM decarboxylase inhibition in *P. falciparum*. In addition to the

effects of polyamines on DNA and RNA regulatory processes they are also involved in oxidative stress responses, cell signalling and membrane and cell wall stabilisation, to which they may be integrated or attached.[175] These functions also concur with proteomic changes observed in ΔESX-3_{ms}.

Furthermore, polyamines have been implicated in the biosynthesis of siderophores and the intracellular transport of chelators, as well as in biofilm formation[173, 175], which in *M. smegmatis*, has been shown to be dependent on iron.[179] Due to their positive charge, polyamines can bind to chelators in the cell to prevent their association with metal cations during their transport from the cell. Conversely they can compete with metal cation binding to chelators, interfering with metal cation uptake and homeostasis. Polyamines have also been shown to exert direct effects on several membrane ion channels and outer membrane porins of various species, altering their charge and pore size, and thereby affecting the permeability of the channels to various molecules, affecting membrane permeability to amongst others, various metal cations.[173, 180, 181] As described earlier, under iron-rich conditions, *M. smegmatis* relies on passive uptake of iron though outer membrane pores, and altered polyamine levels could affect the permeability of the membrane to iron, and other metal cations. Therefore, the disturbed polyamine metabolism of ΔESX-3_{ms} can be linked back to the proposed role of ESX-3 in divalent cation homeostasis, in addition to the proteomic changes observed in the mutant strain. It is also possible that a reduced zinc availability in ΔESX-3_{ms} results in inhibition of MetE and MetH, preventing sufficient production of methionine, preventing polyamine synthesis. Polyamines remain understudied in mycobacteria and this study highlights a new avenue of research into the function of ESX-3, which may, in addition to its role in iron acquisition under iron-scarce conditions in *M. smegmatis* be linked to polyamine-regulation and its role in mycobacterial physiology.

3.3.4. Limitations

The major limitation of this study involves the high iron concentration in the culture medium, which represses transcription of ESX-3 in *M. smegmatis,* reducing the effect of the absence of the secretion system in ΔESX-3_{ms}. The strain variation detected despite this reduced effect highlights a possible additional role of ESX-3 beyond its direct role in high affinity iron uptake. Metabolomic and proteomic investigations under conditions of iron (and other metal cation) starvation may identify additional mechanisms and pathways in which ESX-3 functions and may be more effective in identifying the physiological role of ESX in divalent cation homeostasis.

3.3.5.Conclusion

The importance of ESX-3 in divalent metal cation homeostasis in *M. smegmatis*, and *M. tuberculosis*, has been well established, however its precise biological function remains elusive. In this study we have used a metabolomic and proteomic approach to investigate the physiological changes resulting from the presence and absence of ESX-3 in *M. smegmatis*. The variation in cellular metabolite concentrations could clearly distinguish between the two strains. The absence of ESX-3 resulted in significant reductions in the abundances of amino acids and simple sugars, and increases in fatty acid metabolism. We endeavoured to explain these results in the context of ESX-3's role in iron and zinc uptake/homeostasis and subsequent downstream effects on oxidative stress, suggesting that the loss of ESX-3 resulted in iron excess and zinc starvation. However, although the proteomic variation between WT_{ms} and $\Delta ESX\text{-}3_{ms}$ correlated with pathways contributing to the metabolite changes, there was no proteomic evidence of disturbed iron or zinc homeostasis, as neither the IdeR nor the Zur regulon was altered. The proteomic variation does, however resemble changes resulting from disturbed polyamine metabolism, affecting DNA stability, transcriptional regulation, translation as well as several secondary pathways. The disruption to polyamine metabolism may be explained by the substantially lower levels of methionine in ΔESX-3, as S-adenosyl-methionine is required for the synthesis of the polyamine, spermidine. Due to their positive charge, polyamines can influence the permeability of plasma membranes to various molecules, including metal cations, by influencing the charge and size of porins and other membrane channels. Polyamines have also been associated with the transport of metal chelators within the cell, with competing with metal cations in protein-binding, and in the production of siderophores, once again associating ESX-3 with divalent cation homeostasis

In these experiments *M. smegmatis* was exposed to high levels of iron, under which conditions it has been shown to acquire iron via passive porin mediated uptake, and high affinity iron uptake is repressed. Passive iron uptake is sufficient to repress the IdeR regulon, however other mechanisms, possibly the influence of disturbed polyamine metabolism on these porins, may influence the iron levels within the cells, above the threshold of the IdeR regulon, but still exerting an effect of bacterial metabolism. Although ESX-3 is not required for iron acquisition under these conditions, the low level expression of this secretion system under these conditions, together with the substantial metabolomic and proteomic effects of its deletion suggest that this secretion system performs additional physiological functions. The proteome variation highlights polyamine metabolism as a novel avenue of research into ESX-3 function.

3.4. The functions of the M. smegmatis ESX-4 secretion system and its substrates

3.4.1. Introduction

The ESX-4 gene cluster is the most ancestral ESX,[27] and is present in the genomes of all mycobacteria, with the exception of *M. leprae,* and in other actinomycetes. Although most ESX-4 component proteins have not been detected in proteomic analyses, and it is not essential for *in vitro* or *in vivo* growth, the maintenance and conservation of this gene cluster suggests that it performs an important biological function. Investigation of ΔESX-4_{ms} in comparison to WT_{ms} on a metabolomic and proteomic level may highlight pathways or mechanisms in which it may function.

3.4.2.Results

3.4.2.1. Metabolomic analysis of ESX-4 function

Five replicates of WT_{ms} and ΔESX-4_{ms} cultures, grown to mid log phase (OD_{600}= 0.6-0.8) in standard Middlebrook 7H9 culture medium, were subjected to GC-MS metabolomic analysis. The two dimensional PCA scores plots of the metabolite data, based on combinations of PC1, PC2 and PC3 did not show significant separation of the two strains (data not shown), however a three dimensional PCA scores plot based on PC1, PC2 and PC3 showed some clustering of the individual samples from each of the two strains (Figure 3.4.1). A PLS-DA was constructed and the PCA and PLS-DA analyses combined to compile a list of compounds which best represent the variation between ΔESX-4_{ms} and WT_{ms} (Table 3.4.1; analysis was done by Prof. Loots's group). The list of variable compounds includes several amino acids and fatty acids. However, the variation between individual samples of the same strain is high, with no significant variation between strains; as suggested by the poor discrimination between strains in the PCA scores plots.

3.4.2.2. Whole cell lysate proteomic analysis of ESX-4 function

The whole cell lysate proteomes of ΔESX-4_{ms} and WT_{ms} cultures were analysed using LC-MS/MS on the LTQ Orbitrap Velos. A combined total of 2884 proteins were identified in the *M. smegmatis* ΔESX-4_{ms} and WT_{ms} strains (Supplementary Table S3.5). Protein abundances were compared and used to investigate the biological functions of the ESX-4 secretion system (Table 3.4.2). Ninety seven percent of the detected proteome showed no significant variation in protein abundance. Seven proteins were detected in ΔESX-

4_{ms} but not in $WT_{ms,}$ while 27 proteins were present in WT_{ms}, but not detected in ΔESX-3_{ms}. A further 17 and 33 proteins were present at a significantly greater or lower abundance, respectively,in ΔESX-3_{ms} relative to WT_{ms}. No ESX-4 proteins were detected in either of the strains, however two ESX-3 proteins, $EccC_3$ and $EccD_3$, had a two-fold reduction in abundance in ΔESX-4_{ms} compared to WT_{ms}. Two additional ESX-3 proteins $EccA_3$ and $EccE_3$ also had reduced abundances in ΔESX-4_{ms} although this was not significant.

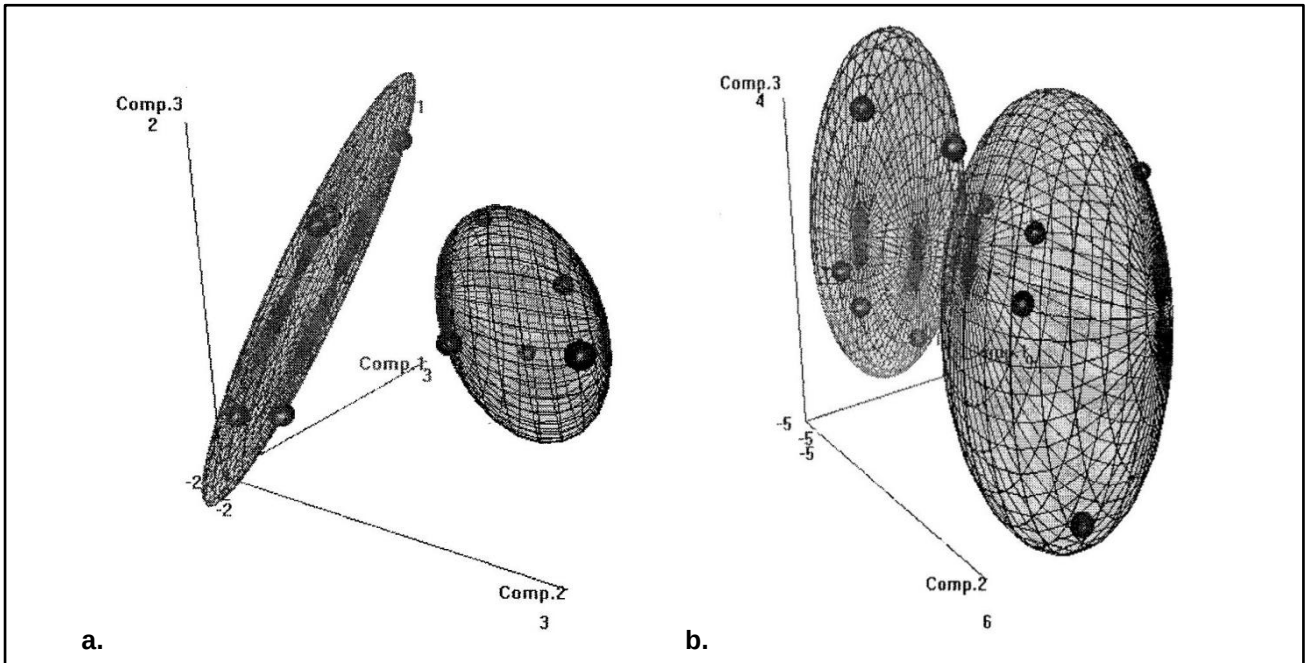


Figure 3.4.1. Three dimensional PCA scores plot of individually cultured ΔESX-4_{ms} and WT_{ms} cultures representing non-parametric (a) and logarithmic (b) scaling. The individual samples from the two strains could be separated based on PC1, PC2 and PC3, which explain 81.7% of the variation between the two groups. Separation of the two groups is due to differences in the metabolite profiles in each of the sample groups. The PCA plot was generated by in Prof. du Loots's group.

Table 3.4.1. Differential abundance of the metabolites best representing the variation between ΔESX-4_{ms} and WT_{ms}

Compound name	Concentration (µg/mg) ±standard deviation	
	WT_{ms}	ΔESX-4_{ms}
Serine	0.451 ± 0.069	0.633 ± 0.091
Glycine	0.771 ± 0.403	0.957 ± 0.476
Threonine	0.417 ± 0.087	0.475 ± 0.086
Sarcosine	0.020 ± 0.016	0.067 ± 0.021
Valine	3.406 ± 0.684	4.119 ± 0.746
Leucine	2.125 ± 0.431	2.795 ± 0.541
Isoleucine	0.933 ± 0.197	1.122 ± 0.193
Alanine	0.186 ± 0.233	0.345 ± 0.423
Phenylalanine	0.018 ± 0.008	0.025 ± 0.005
Octadecanoic acid	1.741 ± 1.594	4.496 ± 2.257
Nonadecanoic acid	4.282 ± 2.087	0.680 ± 0.927
5-oxo-proline	19.690 ± 6.106	14.204 ± 8.383
Propanoic acid	0.142 ± 0.024	0.331 ± 0.360
Benzoic acid	0.033 ± 0.013	0.055 ± 0.019

Proteins that were differentially expressed in the two strains include multiple transcriptional regulators, proteins involved in cofactor metabolism, specifically folate and molybdopterin synthesis, amino acid and carbohydrate metabolism and several hypotheticals. However, these differentially abundant proteins only reflect a very small proportion of the proteome and therefore this minor variation may have little or no physiological effect.

3.4.2.3. Culture filtrate proteomic analysis of ESX-4

Culture filtrate fractions were prepared from WT_{ms} and $\Delta ESX\text{-}4_{ms}$ and the constituent proteins identified by LC-MS/MS analysis on the LTQ Orbitrap Velos. One thousand eight hundred and nine proteins were identified in the two strains (Supplementary Table S3.6). Protein abundances were compared and proteins with a minimum of 2-fold variance with a statistical significance of $p<0.05$, as well as proteins which were present only in the culture filtrate fractions of one strain ($p<0.05$) were considered to be influenced by the presence or absence of ESX-4. Six proteins were detected only in WT_{ms} and another six only in $\Delta ESX\text{-}4_{ms}$, while 29 and 25 proteins were present in higher and lower abundances in $\Delta ESX\text{-}4_{ms}$ than $WT_{ms.}$

No ESX proteins were detected in the CF fractions. Most proteins identified in the culture filtrates appear to be cytosolic proteins, many involved in transcriptional regulation, DNA metabolism, redox metabolism and pathways implicated in the WCL proteome analysis (Table 3.4.3).

3.4.3.Discussion

Our understanding of the ancestral ESX-4 secretion system is poor, and this study used comparative whole metabolomic and proteomic analyses to investigate the influence of the absence and presence of ESX-4 on the physiology of *M. smegmatis* in an attempt to identify potential functions or cellular processes in which it may be involved. ESX-4 is found throughout the most of the mycobacteria and the components and arrangement of ESX-4 in *M. smegmatis* and *M. tuberculosis* are conserved, making $\Delta ESX\text{-}4_{ms}$ a useful tool in the investigation of this secretion system.

Table 3.4.2. Proteins with different abundances in ΔESX-4_{ms} and WT_{ms} WCLs, classified according to biological function.

MSMEG_	Product	Differential abundance*	p value	Biological function	H37Rv ortholog
0617	ftsk/spoiiie family protein, EccC3	-2.2	0.003	ESX-3	Rv0284
0623	secretion protein Snm4, EccD3	-2.0	0.011	ESX-3	Rv0290
1015/2295	transcriptional regulator, TetR family protein	only Δ$ESX\text{-}4_{ms}$	0.024	Transcriptional regulation	Rv3055
2476	MarR family protein transcriptional regulatory	only Δ$ESX\text{-}4_{ms}$	0.017	Transcriptional regulation	
6829	transcriptional regulatory protein	only Δ$ESX\text{-}4_{ms}$	0.036	Transcriptional regulation	
1842	transcriptional regulator, TetR family protein	only WT_{ms}	0.016	Transcriptional regulation	Rv3248c
4517	TetR-type transcriptional regulator of sulfur	only WT_{ms}	0.028	Transcriptional regulation	
6139	putative HTH-type transcriptional regulator	only WT_{ms}	0.022	Transcriptional regulation	Rv2912c
1831	Transcription factor WhiB	4.3	0.037	Transcriptional regulation	
0663	TetR-family protein transcriptional regulator	-2.3	0.034	Transcriptional regulation	
1915	anti-sigma factor, family protein	-3.2	0.017	Transcriptional regulation	Rv3221A
3628	ComA operon protein 2	2.1	0.027	Transcriptional regulation	Rv1847
5696	'cold-shock' DNA-binding domain protein	-2.2	0.016	Transcriptional regulation	Rv0871
2426	nitrogen regulatory protein P-II	-2.5	0.042	Transcriptional regulation	Rv2919c
2793	sensor-type histidine kinase PrrB	-2.4	0.012	Transcriptional regulation	
1445	30S ribosomal protein S17	-2.1	0.001	Translation	Rv0710
2400	ribosomal protein L28, rpmB	-2.3	0.038	Translation	
3483	mosc domain protein (molybdenum ion binding)	only WT_{ms}	0.042	Divalent cation homeostasis	Rv1864c
4903	glutamate racemase, MurI	only WT_{ms}	0.048	Amino acid metabolism	Rv1338
5536	chorismate mutase	only WT_{ms}	0.003	Amino acid metabolism	Rv0948c
2726	glutamate permease	-2.2	0.016	Amino acid metabolism	
6307	glutamine-binding periplasmic protein/glutamine transport system permease	-2.2	0.002	Amino acid metabolism	
6911	ABC transporter, ATP-binding protein GluA	-2.0	0.036	Amino acid metabolism	
5085	dihydropteroate synthase, FolP	only WT_{ms}	0.036	Cofactor metabolism	Rv1207
5470	molybdopterin biosynthesis protein MoeA 1	only WT_{ms}	0.022	Cofactor metabolism	Rv0994
5703	molybdenum cofactor biosynthesis protein C, MoaC	2.0	0.010	Cofactor metabolism	Rv0864
0974	cytochrome c-type biogenesis protein CcsB	only WT_{ms}	0.007	Reduction/oxidation	Rv0529
4263	ubiquinol-cytochrome c reductase cytochrome b	-2.0	0.002	Reduction/oxidation	Rv2196
2201	ZbpA protein	2.1	0.001	Reduction/oxidation	Rv0130
0086	1-phosphofructokinase	only WT_{ms}	0.031	Carbohydrate metabolism	
0655	glucose 1-dehydrogenase, putative	only WT_{ms}	0.003	Carbohydrate metabolism	
0452	inner membrane permease YgbN (gluconate transport)	only WT_{ms}	0.031	Carbohydrate metabolism	
4183	phosphoglycolate phosphatase, chromosomal	only WT_{ms}	0.007	Carbohydrate metabolism	
0784	acetate kinase	2.3	0.034	Carbohydrate metabolism	Rv0409
6021	xylose isomerase, XylA	3.2	0.017	Carbohydrate metabolism	
1671	succinate dehydrogenase hydrophobic membrane, SdhD	-3.2	0.023	Carbohydrate metabolism	Rv3317

0389	glycosyl transferase	-2.9	0.050	Carbohydrate metabolism	
6367	glycosyl transferase, group 2 family protein	-2.1	0.049	Carbohydrate metabolism	Rv3782
3739	cytidylate kinase	only WT_{ms}	0.047	DNA/RNA metabolism	Rv1712
3473	uracil phosphoribosyltransferase	only ΔESX-4_{ms}	0.044	DNA/RNA metabolism	
2683	thymidylate synthase, flavin-dependent, ThyX	2.4	0.015	DNA/RNA metabolism	Rv2754c
6276	mur ligase family protein	only WT_{ms}	0.011	Cell wall biosynthesis	Rv3712
4236	MraZ protein (cell division/cell wall biosynth)	only WT_{ms}	0.048	Cell wall biosynthesis	Rv2166c
1661	D-alanyl-D-alanine carboxypeptidase	-2.9	0.001	Cell wall biosynthesis	Rv3330
0928	UDP-N-acetylenolpyruvoylglucosamine reductase, murB	only ΔESX-4ms	0.038	Cell wall biosynthesis	Rv0482
4228	cell division protein FtsW	only WT_{ms}	0.003	Cell division	Rv2154c
4236	MraZ protein (cell division/cell wall biosynth)	only WT_{ms}	0.048	Cell division	Rv2166c
2688	antibiotic biosynthesis monooxygenase domain	12.2	0.000	Antibiotic	Rv2749
3848	para-nitrobenzyl esterase	2.1	0.006	Antibiotic	Rv2045c
6202	secreted protein	only WT_{ms}	0.001	Secreted protein	Rv3683
5412	immunogenic protein MPT63	-2.3	0.007	Secreted protein	
2377	P49 protein	only WT_{ms}	0.024	Other	Rv2997
0203/0398	IS1096, tnpR protein	only WT_{ms}	0.009	Other	
2490	decarboxylase	only WT_{ms}	0.005	Other	
2579	zinc metalloprotease	only WT_{ms}	0.041	Other	Rv2869c
2696	putative conserved membrane alanine rich	only WT_{ms}	0.001	Other	Rv2743c
3280	polyamine-binding lipoprotein	only WT_{ms}	0.036	Other	
4908	endo-type 6-aminohexanoate oligomer hydrolase	only WT_{ms}	0.014	Other	Rv1333
5162	cinnamoyl ester hydrolase	only WT_{ms}	0.012	Other	
6710	hydrolase, alpha/beta fold family protein	only WT_{ms}	0.025	Other	
0884	glyoxalase family protein	3.1	0.005	Other	
1640	ATP/GTP-binding protein	2.0	0.011	Other	Rv3362c
2409	alpha/beta hydrolase fold family protein	2.1	0.005	Other	Rv2970c
6386	membrane protein	-2.3	0.030	Other	Rv3792
0403	integral membrane protein	-2.4	0.048	Other	
2100	peptidase family protein M20/M25/M40	-3.2	0.037	Other	
2441	signal peptidase I, LepB	-2.9	0.032	Other	Rv2903c
4250	putative membrane transport ATPase	-3.8	0.036	Other	Rv2184c
5257	channel protein, hemolysin III family protein	-2.0	0.000	Other	Rv1085
5368	ectoine/hydroxyectoine ABC transporter binding protein, EhuB	-2.2	0.036	Other	
5827	glyoxalase family protein	-2.7	0.015	Other	Rv0801
6239	1,3-propanediol dehydrogenase	-3.6	0.010	Other	
3026	hypothetical protein	only ΔESX-4_{ms}	0.039	Hypothetical protein	Rv2554c
6601	hypothetical protein	only ΔESX-4_{ms}	0.004	Hypothetical protein	
1434	hypothetical protein	only WT_{ms}	0.003	Hypothetical protein	
0971	hypothetical protein	only WT_{ms}	0.033	Hypothetical protein	Rv0526
3514	hypothetical protein	only WT_{ms}	0.027	Hypothetical protein	

4569	hypothetical protein	only WT_{ms}	0.000	Hypothetical protein	Rv2410c
1734	hypothetical protein	2.6	0.007	Hypothetical protein	Rv3304
4324	hypothetical protein	2.7	0.004	Hypothetical protein	Rv2242
5614	hypothetical protein	2.0	0.031	Hypothetical protein	
6159	hypothetical protein	2.3	0.015	Hypothetical protein	Rv3648c
6280	hypothetical protein	8.2	0.014	Hypothetical protein	Rv3716c
0368	hypothetical protein	-2.2	0.017	Hypothetical protein	
0833	hypothetical protein	-3.0	0.047	Hypothetical protein	
1076	hypothetical protein	-2.0	0.008	Hypothetical protein	
1243	hypothetical protein	-2.0	0.005	Hypothetical protein	
1981	hypothetical protein	-2.1	0.031	Hypothetical protein	
5223	hypothetical protein	-2.8	0.004	Hypothetical protein	Rv1111c

*Differential abundances represent the fold change in abundance of a protein in ΔESX-4_{ms} relative to WT_{ms}. Negative values indicate a lower abundance in ΔESX-4_{ms} than WT_{ms}. Where indicated proteins were detected in only one strain.

Table 3.4.3. Proteins with different abundances in ΔESX-4_{ms} and WT_{ms} culture filtrates, classified according to biological function.

MSMEG_	Product	Differential abundance*	p value	Biological function	H37Rv ortholog
3319	repressor protein	only ΔESX-4_{ms}	0.009	Transcriptional regulation	
1305	TetR family protein regulatory protein	-2.2	0.024	Transcriptional regulation	
1611	transcriptional regulator, TetR family protein,	-14.2	0.028	Transcriptional regulation	
3980	transcriptional regulator, GntR family protein	-2.1	0.032	Transcriptional regulation	
1364	50S ribosomal protein L10	-2.4	0.010	Translation	Rv0651
1444	ribosomal protein L29	-2.2	0.048	Translation	Rv0709
5432	peptidyl-tRNA hydrolase	3.1	0.029	Translation	Rv1014c
2768	OB-fold nucleic acid binding domain protein	only WT_{ms}	0.000	DNA/RNA metabolism	Rv2694c
3078	excinuclease ABC, C subunit	2.1	0.003	DNA/RNA metabolism	Rv1420
1901	DNA-binding HTH domain containing protein	28.4	0.016	DNA/RNA metabolism	
6941	R3H domain-containing protein	2.3	0.023	DNA/RNA metabolism	Rv3920c
4724	oligoribonuclease	-2.2	0.044	DNA/RNA metabolism	Rv2511
5243	helix-turn-helix motif	-4.2	0.044	DNA/RNA metabolism	
2724	regulatory protein RecX	-4.0	0.050	DNA/RNA metabolism	Rv2736c
1182	imidazolonepropionase	only ΔESX-4_{ms}	0.003	Amino acid metabolism	
1496	3-hydroxyisobutyrate dehydrogenase	8.0	0.001	Amino acid metabolism	Rv0751c
4905	cysteine synthase B	2.0	0.008	Amino acid metabolism	Rv1336
3397	acetylornithine deacetylase	4.3	0.044	Amino acid metabolism	
2937	pyridoxine biosynthesis protein	2.4	0.005	Cofactor metabolism	Rv2606c
2787	putative riboflavin biosynthesis protein RibD	4.2	0.044	Cofactor metabolism	Rv2671
5319	alkylhydroperoxidase AhpD core	only WT_{ms}	0.002	Reduction/oxidation	
6362	quinone oxidoreductase	2.5	0.000	Reduction/oxidation	Rv3777
0408	type I modular polyketide synthase	2.4	0.009	Reduction/oxidation	
0604	glyoxylate reductase	4.7	0.028	Reduction/oxidation	Rv0728c
2536	3-oxoacyl-[acyl-carrier-protein] reductase	-2.0	0.017	Reduction/oxidation	
6636	[Mn] superoxide dismutase	-2.0	0.017	Reduction/oxidation	Rv3846
0502	glucosidase	6.4	0.014	Carbohydrate metabolism	
4305	phosphoglycerate mutase	2.6	0.050	Carbohydrate metabolism	Rv2228
2512	lactate 2-monooxygenase	-2.6	0.030	Carbohydrate metabolism	
1820	phosphoribosylaminoimidazole carboxylase	-2.6	0.039	Carbohydrate metabolism	Rv3275c
1952	ATP-dependent DNA helicase	-2.2	0.041	Carbohydrate metabolism	Rv3198c
2349	glycosyl hydrolase, family protein 57	-2.8	0.044	Carbohydrate metabolism	Rv3031
2070	acyl-CoA dehydrogenase family protein	3.1	0.009	Fatty acid metabolism	
6388	probable arabinosyltransferase B	only WT_{ms}	0.001	Cell wall biosynthesis	Rv3794
2395	D-alanine--D-alanine ligase	2.3	0.002	Cell wall biosynthesis	Rv2891c

1580	O-sialoglycoprotein endopeptidase	only ΔESX-4_{ms}	0.002	Other	Rv3419c
1703	amidohydrolase	only ΔESX-4_{ms}	0.002	Other	Rv3305c
0939	Ppx/GppA phosphatase family protein	only ΔESX-4_{ms}	0.003	Other	Rv0496
4362	universal stress protein family protein	only ΔESX-4_{ms}	0.013	Other	
4939	ATP synthase delta chain	3.5	0.014	Other	Rv1307
1835	TobH protein	3.4	0.026	Other	Rv3256c
4043	amidohydrolase 2	4.7	0.028	Other	
5017	lipoprotein	3.0	0.031	Other	
0270	aminoglycoside phosphotransferase	4.5	0.033	Other	
0518	ABC transporter, nucleotide binding/ATPase	4.3	0.036	Other	
5118	nudix hydrolase	4.1	0.049	Other	
4672	Clp protease	-2.0	0.014	Other	Rv2460c
0241	MmpL11 protein	-4.6	0.015	Other	Rv0202c
4975	flavin-nucleotide-binding protein	-2.3	0.018	Other	
5198	carnitinyl-CoA dehydratase	-2.1	0.018	Other	
1976	hypothetical protein	only WT_{ms}	0.000	Hypothetical protein	
2763	hypothetical protein	only WT_{ms}	0.001	Hypothetical protein	Rv2699c
6579	hypothetical protein	only WT_{ms}	0.001	Hypothetical protein	
1356	hypothetical protein	only ΔESX-4_{ms}	0.002	Hypothetical protein	
3910	hypothetical protein	only ΔESX-4_{ms}	0.005	Hypothetical protein	
1210	hypothetical protein	only ΔESX-4_{ms}	0.031	Hypothetical protein	
4365	hypothetical protein	6.4	0.000	Hypothetical protein	
1176	hypothetical protein	4.2	0.019	Hypothetical protein	
4128	hypothetical protein	5.1	0.019	Hypothetical protein	
2037	hypothetical protein	3.1	0.036	Hypothetical protein	
3661	hypothetical protein	2.7	0.049	Hypothetical protein	Rv1815
2679	hypothetical protein	2.5	0.050	Hypothetical protein	
4399	hypothetical protein	-14.0	0.000	Hypothetical protein	
5614	hypothetical protein	-2.8	0.000	Hypothetical protein	
3026	hypothetical protein	-2.2	0.006	Hypothetical protein	Rv2554c
2436	hypothetical protein	-3.4	0.007	Hypothetical protein	Rv2908c
1680	hypothetical protein	-2.1	0.012	Hypothetical protein	
3843	hypothetical protein	-4.7	0.026	Hypothetical protein	Rv1628c
3016	hypothetical protein	-4.2	0.044	Hypothetical protein	Rv2566

*Differential abundances represent the fold change in abundance of a protein in ΔESX-4_{ms} relative to WT_{ms}. Negative values indicate a lower abundance in ΔESX-4_{ms} than WT_{ms}. Where indicated proteins were detected in only one strain.

3.4.3.1. Expression of ESX-4 is repressed under standard culture conditions in *M. smegmatis*.

The whole cell lysate and culture filtrate proteomic analyses were unable to identify any of the ESX-4 proteins in any of the WT_{ms} replicates. This apparent lack of expression is supported by the absence of significant metabolomic variation between WT_{ms} and ΔESX-4_{ms}. These results correlate with several *in vitro* studies which have failed to identify ESX-4 proteins in various cellular fractions,[120, 182, 183] although $EccC_4$ and $MycP_4$ have been detected in the cell membrane and cytosolic fractions of *M. tuberculosis* respectively, grown in an alternate culture medium under different conditions.[184] On a protein level, however, significant variation in protein abundances was detected in the whole cell lysate and culture filtrate fractions. This may be attributed to very low level, undetectable, expression of ESX-4 which none-the-less influences cellular biology of the mycobacterium. These influences may result in insignificant changes to the metabolite profiles of the two strains, or influence processes which do not influence the metabolite levels. The lack of significant variation in metabolite concentrations might also be attributed to the small sample numbers in the metabolomic analyses, which influence statistical significance. It appears that under the conditions of this study expression of ESX-4 is repressed, but that low level, undetectable expression impacts various cellular processes leading to substantial proteomic variation, but no significant metabolomic distinction, between WT_{ms} and ΔESX-4_{ms}.

3.4.3.2. The absence of ESX-4 influences the expression of ESX-3

Interestingly, two ESX-3 proteins, $EccC_3$ and $EccD_3$, were only half as abundant in ΔESX-4_{ms} as in WT_{ms}. Two additional ESX-3 proteins, $EccA_3$ and $EccE_3$, were also underrepresented in ΔESX-4_{ms}, although not significantly so. Therefore it appears that the presence or absence of ESX-4 influences the expression of ESX-3 in *M. smegmatis*. Cross-talk between different ESX secretion systems has previously been shown for ESX-1 and ESX-5, where deletion of *esxA-esxB* results in increased secretion of PPE41 (ESX-5),[59] while deletion of ESX-5 increases the secretion of ESX-1 substrates.[30] Although the five ESX secretion systems are unable to complement one another, it has been suggested that partial complementation by some proteins may be possible.[30] Alternatively, the functions of one ESX secretion system may influence those of another.

Several pathways which were influenced by the deletion of ESX-3 also show variations in protein abundances in ΔESX-3_{ms}. These include proteins involved in carbohydrate and amino acid metabolism, redox homeostasis, cell wall biosynthesis, translation, secreted proteins and hypotheticals. The variations

in these proteins might be attributed to the reduced expression of ESX-3 rather than directly to the absence of ESX-4.

3.4.3.3. ESX-4 and cell division

The absence of ESX-4 resulted in the disregulation of numerous transcription factors, most of which are of unknown function, but which likely result in the other changes in protein expression. Notably, a WhiB transcriptional regulator, encoded by MSMEG_1831, is more than four times more abundant in ΔESX-4_{ms} than in WT_{ms}. MSMEG_1831 encodes WhmD, the *M. smegmatis* ortholog of *M. tuberculosis* WhiB2. The gene *whmD* is essential, and is involved in cell division, specifically in ensuring proper septum formation.[185, 186] Under-expression of WhmD results in filamentation with multiple arrested septa distributed unevenly thoughout the length, but does not affect biomass accumulation. Overexpression of WhmD resulted in small colony size, but only slightly reduced growth rate in liquid culture and a frequent occurrence of multiseptate bacteria. WhiB2 is expressed throughout log phase, but greatly reduced in late stationary phase,[187] indicating its requirement for active growth. WhiB2 expression is also upregulated when exposed to various antibiotics which inhibit cell wall biosynthesis.[187] Therefore it appears that WhmD may be involved in cell division by regulating the expression of proteins directly involved in septum formation or in cell division-associated cell wall biosynthesis.

$EccC_4$ was classified as a FtsK/SpoIIIE protein, proteins which are typically involved in cell division and peptidoglycan synthesis and modification, and have been shown to assist in the transfer of DNA through the septum.[188, 189] Rv3447c (*M. tuberculosis* $EccC_4$) was found to be upregulated when FtsI, a protein involved in peptide crosslinking at the septum, is inhibited.[121] Furthermore, inhibition of FtsI results in a similar filamentous phenotype as caused by WhmD underexpression. Therefore it appears that WhmD, FtsI and ESX-4 may all be linked to cell division and septum formation. There are several other proteins involved in cell division and cell wall biosynthesis which were found to be differentially abundant in WT_{ms} and $\Delta ESX4_{ms}$, including FtsW (involved in Z-ring stabilisation and septal peptidoglycan synthesis), MrzA (involved in cell wall synthesis and cell division) and some Mur family proteins (involved in cell wall biosynthesis), suggesting that these processes may be influenced by ESX-4.

Interestingly, ESX-4 expression has been shown to be induced by WhiB5, another WhiB transcriptional regulator, during reactivation of *M. tuberculosis* following dormancy and starvation.[112] Reactivation involves increased cell growth and division, further associating ESX-4 with the processes of cell division,

septum formation, or cell division-related cell wall biosynthesis. It is possible that ESX-4 is responsible for secreting specific proteins involved in these processes, and possibly the transport of DNA at the septum, as indicated by the FtsK/SpoIIIE motif in $EccC_4$. The involvement of ESX-4 in these processes may be limited to reactivation, during which its expression is induced, allowing normal cell growth and cell division of ΔESX-4_{ms} during standard *in vitro* culture.

3.4.3.4. ESX-4 and molybdenum metabolism

Interestingly three proteins which are involved in molybdopterin metabolism were differentially abundant in WT_{ms} and ΔESX-4_{ms}. Molybdopterin is a molybdenum containing cofactor which is required by various enzymes. MoeA and MoaC are involved in the biosynthesis of molybdopterin,[190] while the mosc domain protein requires a molybdopterin cofactor. Molybdenum is a transition metal and altered molybdopterin metabolism may be indicative that ESX-4 is involved in the homeostasis of molybdenum levels in the cell, and possibly the uptake thereof, as described for iron and zinc and ESX-3.

3.4.4.Conclusion

In this study we have attempted to define the metabolomic and proteomic changes resulting from the absence or presence of the ancestral ESX-4 secretion system in *M. smegmatis*, in order to identify possible functions and cellular processes in which it may be involved. Although the inability to detect any of the ESX-4 proteins in the wild-type *M. smegmatis* samples, and the lack of significant metabolomic differences between WT_{ms} and ΔESX-4_{ms} suggest that ESX-4 is not expressed under these conditions, significant differences were observed between the whole cell lysate and culture filtrate fractions from the two strains. This suggests that ESX-4 may rather be expressed at undetectable levels under these conditions. A reduced abundance of ESX-3 proteins was detected in ΔESX-4_{ms} relative to WT_{ms}, suggesting that ESX-4 influences the expression of ESX-3, and that their functions may be somehow linked. We have identified two pathways in which ESX-4 may be involved, namely cell division or cell division-associated cell wall biosynthesis and molybdopterin homeostasis. The functions of this inessential, but highly conserved, ESX-4 secretion system have remained elusive, and this study has identified novel directions for investigation of its functional role in mycobacterial physiology.

3.5. General conclusion

This study has used has used a “systems biology” approach, incorporating whole proteomic and metabolomic analyses to investigate the global biological changes resulting from a specific genomic modification. The effects on multiple cellular pathways and their impact on one another can be detected simultaneously, allowing for the identification of pathways affected by such a genomic modification. These global analyses provide a unique way of understanding bacterial physiology in response to various stimuli and stresses in addition to specific genome modifications and may be hugely beneficial in expanding our understanding of mycobacteria and their mechanisms of pathogenicity.

Individual ESX knockout strains of *M. smegmatis* were used to investigate the effect of the presence or absence of each of the *M. smegmatis* ESX gene clusters on the metabolomes and proteomes of the bacteria, in order to identify possible biological functions of the associated ESX secretion systems. No pathways were considerably altered by the deletion of ESX-1, despite the detection of most of the ESX-1 components in the WCL of WT_{ms}. The numbers of proteins which vary between ΔESX-1_{ms} and WT_{ms} on both a proteomic and secretomic level are much lower than those obtained in the ESX-3 and ESX-4 studies, despite the far higher expression level of ESX-1 under these conditions, suggesting that the function of ESX-1 may be more independent of general cellular functions. Deletion of ESX-3 resulted in major changes on a proteomic and metabolomic level, including altered amino acid and carbohydrate metabolism, translational processes and redox metabolism which might be ascribed to polyamine metabolism. No significant metabolomic variation resulted from the deletion of ESX-4. However, although no ESX-4 component proteins were detected in WT_{ms}, the proteome of ΔESX-4_{ms} was altered, possibly linking ESX-4 to cell division or molybdenum homeostasis. Interestingly the deletion of ESX-4 resulted in reduced the abundance of ESX-3 components, possibly linking the functions of these two secretion systems. The identification of pathways influenced by the absence or presence of the ESX gene clusters provides new directions for research into their functions and role in mycobacterial pathogenicity.

Supplementary Tables S3.1 to S3.6. The detected WCL and CF proteomes of ΔESX-1_{ms}, ΔESX-3_{ms} and ΔESX-4_{ms} with abundances relative to the detected WT_{ms} proteome.

4. The substrates of the *M. smegmatis* ESX secretion systems

4.1. Introduction

The mycobacterial ESX secretion systems are predicted to secrete the Esx, PE and PPE proteins encoded within them.[30] In *M. tuberculosis,* some ESX have further been shown to secrete additional proteins, including Esx, PE and PPE proteins which were duplicated from these gene clusters and situated in other parts of the genome, and various other proteins encoded within the ESX gene clusters and at other genome locations. Recently a YxxxD/E amino acid motif has been found to be a general ESX (Type-VII) secretion signal for *M. tuberculosis* ESX substrates.[72] The non-ESX encoded Esx, PE and PPE duplications and various other *M. tuberculosis* ESX substrates are not present in *M. smegmatis* and additional ESX substrates have not been identified in this organism. This study aimed to use the secretomic analyses (as described in section 3) to identify novel substrates of the *M. smegmatis* ESX-1, ESX-3 and ESX-4 secretion systems.

4.2. Results

The secretomes of ΔESX-1_{ms}, ΔESX-3_{ms} and ΔESX-4_{ms} were analysed and compared to that of WT_{ms}. Proteins which were detected in WT_{ms} but not in a specific ΔESX_{ms} strain are potential secreted substrates of that ESX secretion system (Tables 4.1 to 4.3). Proteins with a minimum of 4 times greater abundance in WT_{ms} than in the ΔESX strain were included to account for cell leakage or cell lysis which may result in some proteins in the culture filtrate.

The general Type-VII secretion signal, YxxxD/E was identified using *M. tuberculosis* and *M. marinum* proteins, [72] therefore to investigate the appropriateness of using the YxxxD/E motif for the confirmation of ESX substrates in *M. smegmatis*, the Esx and PE protein sequences from the three *M. smegmatis* ESX gene clusters were analysed. The CFP-10-like Esx proteins and the PE proteins from ESX-3 and ESX-1 in *M. smegmatis* contain conserved YxxxD motifs near their C-termini (Figure 4.1.a). This correlates with previous findings that these proteins contain the signal to allow secretion of the Esx-Esx and PE-PPE complexes. The motif was not identified at the C-terminal of EsxU or EsxT from ESX-4, however a YxxxE is present near the N-terminal of EsxT, and may substitute for the secretion signal.

The amino acid sequence of each possible *M. smegmatis* ESX substrate protein was analysed to identify a YxxxD/E motif. Five proteins containing this motif were identified for ESX-4, six for ESX-3 and one for

ESX-1 (Figure 4.1.b). Two proteins, MSMEG_5319 and MSMEG_1976, were common for both ESX-4 and ESX-3. No previously assumed or identified ESX substrates were identified, and only a single ESX-encoded protein, $EccA_1$, was identified in the culture filtrates.

Table 4.1. Possible *M. smegmatis* ESX-1 secreted proteins.

MSMEG_	Product	Differential abundance ΔESX-1_{ms} : WT_{ms}	p value	YxxxD/E motif	H37Rv ortholog
0059	ATPase, AAA family protein large subunit	only WT	0.008	YspdE	Rv3868
3142	HTH-type transcriptional repressor AcnR	only WT	0.000		Rv1474c
1573	carbohydrate kinase family protein	only WT	0.012		Rv3433c
3778	phenylalanyl-tRNA synthetase, alpha subunit	only WT	0.002		Rv1649
6278	metallo-beta-lactamase superfamily protein	only WT	0.013		
1108	hydrolase, alpha/beta fold family protein	only WT	0.049		Rv0554
3016	hypothetical protein	only WT	0.014		Rv2566

Table 4.2. Possible *M. smegmatis* ESX-3 secreted proteins.

MSMEG_	Product	Differential abundance ΔESX-$3_{ms:}$ WT_{ms}	p value	YxxxD/E motif	H37Rv ortholog
3872	precorrin-8X methylmutase	only WT	0.046		Rv2065
2119	N-acetylglucosamine-6-phosphate deacetylase, nagA	only WT	0.008		Rv3332
2923	dehydrogenase/reductase SDR family protein	only WT	0.002	YgitD	
5319*	alkylhydroperoxidase AhpD core	only WT	0.001	YftdE	
6077	transcriptional regulator, CarD family protein	only WT	0.013		Rv3583c
6451	transcriptional regulator, ArsR family protein	only WT	0.026		Rv0081
2399	uracil-DNA glycosylase	only WT	0.003		Rv2976c
0369	hypothetical protein	only WT	0.038		
1581	hypothetical protein	only WT	0.009	YryaE	
1976*	hypothetical protein	only WT	0.000	YsddE	
2763	hypothetical protein	only WT	0.003		Rv2699c
2861	hypothetical protein	only WT	0.011		
3024	hypothetical protein	only WT	0.024		Rv2556c
3843	hypothetical protein	only WT	0.003		Rv1628c
5634	hypothetical protein	only WT	0.048		Rv0910
6579	hypothetical protein	only WT	0.014		
0760	thioesterase family protein	only WT	0.001	YeagE	Rv0356c
1078	hydrolase	only WT	0.000		
1835	TobH protein	only WT	0.004		Rv3256c
5224	4-hydroxy-3-methylbut-2-enyl diphosphate reductase, ispH	only WT	0.002	YurhE	Rv1110

* absent in both ΔESX-3_{ms} and ΔESX-4_{ms} culture filtrates.

Table 4.3. Possible *M. smegmatis* ESX-4 secreted proteins.

MSMEG_	Product	Differential abundance ΔESX-4_{ms}: WT_{ms}	p value	YxxxD/E motif	H37Rv ortholog
2768	OB-fold nucleic acid binding domain protein	only WT	0.000		Rv2694c
5319*	alkylhydroperoxidase AhpD core	only WT	0.002	YftdE	
6388	probable arabinosyltransferase B	only WT	0.001	YlrdD	Rv3794
1976*	hypothetical protein	only WT	0.000	YsddE	
2763	hypothetical protein	only WT	0.001		Rv2699c
6579	hypothetical protein	only WT	0.001		
1611	transcriptional regulator, TetR family protein,	-14.2	0.028		
4399	hypothetical protein	-14.0	0.000		
3843	hypothetical protein	-4.7	0.026	YtddD	Rv1628c
0241	MmpL11 protein	-4.6	0.015	YcgrD	Rv0202c
5243	helix-turn-helix motif	-4.2	0.044		
3016	hypothetical protein	-4.2	0.044		Rv2566
2724	regulatory protein RecX	-4.0	0.050		Rv2736c

* absent in both ΔESX-3_{ms} and ΔESX-4_{ms} culture filtrates.

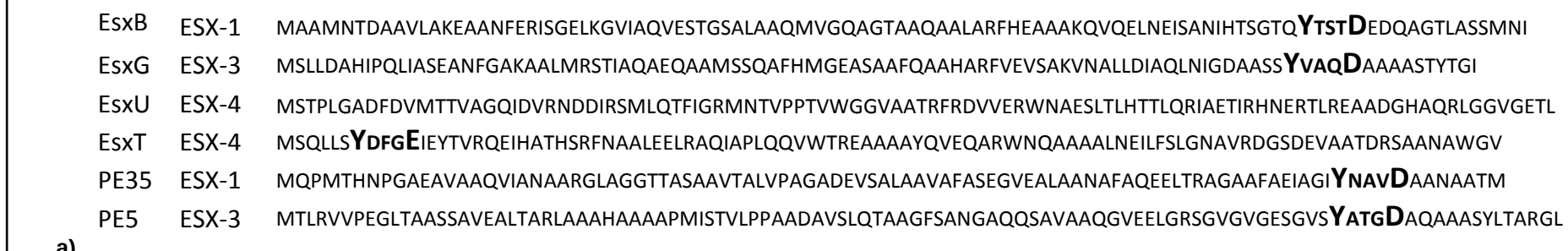


Figure 4.1 ESX secreted substrates contain a general mycobacterial Type VII secretion motif. a) The prototype substrates of ESX-1 and ESX-3 contain C-terminal YxxxD motifs, while the ESX-4 Esx substrate contains a YxxxE motif toward the N-terminal. **b)** Ten of the possible ESX secreted substrates contain YxxxD/E motifs. The region of the protein sequence shown is indicated, total amino acid length of each protein is in brackets and the YxxxD/E motifs are in bold.

4.3. Discussion

The ESX gene clusters encode dedicated Type-VII protein secretion systems, responsible for the export of specific proteins into the extracellular milieu. The secretion of Esx, PE and PPE proteins has been linked to their associated ESX secretion systems, and several additional secreted substrates have been identified in *M. tuberculosis*, predominantly for the ESX-1 and ESX-5 secretion systems. However, additional substrates of the *M. smegmatis* ESX have not been identified. Here we have used a comparative whole secretomic analysis to identify novel *M. smegmatis* ESX substrates by comparing the protein constituents of the culture filtrates of WT_{ms} with those of ΔESX-1_{ms}, ΔESX-3_{ms} and ΔESX-$4_{ms.}$ The absence of each specific secretion system prevents the secretion of the correlating protein substrates allowing their identification only from WT_{ms}. Several proteins were identified which were secreted in WT_{ms} but not in a ΔESX_{ms} strain, or had greatly reduced abundances in the ΔESX_{ms} strain than in WT_{ms}, and were defined as possible ESX substrates.

Protein concentrations in the culture filtrate fractions may be influenced directly by ESX secretion, by increased expression levels and by additional biological processes which are influenced by the presence or absence of the ESX secretion systems. Therefore not all proteins which are only detected in the culture filtrates of WT_{ms} can be directly associated with ESX secretion. A recently identified general secretion signal for mycobacterial ESX secretion, YxxxD/E,[72] was used to screen the possible substrates. Ten differentially secreted proteins, containing YxxxD/E motifs, were identified for all three ΔESX_{ms} strains, two of which were absent in both ESX-4 and ESX-3. This correlates with the proteomic findings which showed that ESX-3 abundance was lower in the absence of ESX-4, and suggests that these proteins are substrates of ESX-3, rather than ESX-4.

Only one possible protein substrate containing a YxxxD/E motif was identified for ESX-1, namely $EccA_1$, encoded by MSMEG_0059. $EccA_1$ is generally considered to be a cytosolic protein, although it has been detected in the membrane fraction of *M. tuberculosis*.[159] Recently $EccA_1$ has been associated with mycolic acid synthesis,[109] and this may link it functionally to the cell membrane or periplasmic space. However, our study shows that it does contain an ESX secretion system motif, indicating that it might, in fact, be a true ESX secreted protein.

Of the six possible secreted substrates with the general Type-VII secretion signal identified for ESX-3, four are involved in redox metabolism, and two are conserved hypotheticals. The metabolomic analysis of ΔESX-3_{ms} indicated that the organism experiences increased levels of oxidative stress, therefore the redox-related proteins in the culture filtrate of WT_{ms} may contribute to the reduction of this stress. Two of the proteins, MSMEG_2923 and MSMEG_1976, contain the YxxxD/E motif near to their C-termini, resembling those of the know substrates, while in the remaining four the motif occurs closer to the N-terminus.

The possible substrates of ESX-4, MSMEG_6388, MSMEG_3843 and MSMEG_0241, encode an arabinosyltransferase, a conserved hypothetical and a MmpL11 protein respectively, the homologs of which have all been detected in the membrane of *M. tuberculosis*.[159] Arabinosyltransferases are involved in cell wall synthesis, while the MmpL proteins are predicted to export mycolic acids and other lipids for cell wall and membrane synthesis.[191] This correlates with a possible involvement of ESX-4 in cell division-related cell wall synthesis. MSMEG_6388 and MSMEG_0241 are large proteins with a YxxxD/E motif toward the C-terminus, while MSMEG_3843 is a smaller protein which contains a secretion motif within 17 amino acids of the C-terminus, possibly presenting a better prospect as an ESX substrate (which are usually around 100 amino acids in length).

Proteins which contain a general Type-VII secretion signal which is not located at the N-terminus may still be substrates of an ESX secretion system, as some proteins have a YxxxD/E motif within an N-terminal domain of the protein which is cleaved from a functional C-terminus domain after secretion.[72] Furthermore, because most identified ESX substrates appear to be secreted as heterodimers, proteins without the YxxxD/E motif may be co-secreted with a Type-VII secretion motif-containing protein. However, while the presence of the YxxxD/E motif suggests that these proteins are exported through the membrane using the ESX secretion machinery, it is possible that the increased abundance of these proteins is rather indicative of another alteration to the bacterial metabolism resulting from the absence of the ESX.

None of the ESX-4 component proteins were detected in the WCL or CF of *M. smegmatis* and only a few of the ESX-3 component proteins were detected at low levels in the WCL. Furthermore, although ESX-1 was significantly expressed, only two of the prototype substrates, EsxA and PPE68 were detected in the WCL, and none in the CF. EsxA and PPE68 rely on EsxB and PE35 to provide the

signal for secretion.[72] Therefore low expression levels, export rates or inefficient concentration of the culture filtrate may preclude effective identification of known and novel substrates in this study. In addition, cytosolic proteins were abundant in the CF fraction, suggesting that culture filtrate fractionation was not sufficiently precise. Several of the possible substrates which were identified and contain secretion motifs are localised to the cell membrane and it remains unclear whether they are genuine substrates which are integrated into the membrane via the ESX secretion system, as has been seen for some PPE proteins, or whether their detection is due to a lack of specificity of the sample fractionation. Developing new methods for the fractionation of mycobacteria to improve specificity, and increasing expression and export by using more appropriate culture conditions will advance our ability to identify novel substrates.

4.4. Conclusion

In this study we have attempted to use comparative proteomic analyses of the culture filtrates of *M. smegmatis* wild type and ΔESX1$_{ms}$, ΔESX-3$_{ms}$ and ΔESX-4$_{ms}$ strains to identify novel secreted protein substrates of each *M. smegmatis* ESX secretion system. The secretome of a bacterium, containing the exported proteins, is predicted to contain approximately 10% of the annotated proteins, however, in our culture filtrate analysis, we identified approximately 1800 proteins. Although care was taken to prevent excessive lysis of the cells during culture, due to the low abundance of secreted proteins relative to cytosolic proteins, the levels of cytosolic protein contamination in the culture filtrates was high (data not shown). None of the prototype Esx, PE and PPE secreted substrates were identified in the culture filtrate of wild type *M. smegmatis*. Ten possible substrates of the *M. smegmatis* ESX secretions systems were identified at high confidence, using the presence of the YxxxD/E general mycobacterial Type-VII secretion signal as a validation. This study may serve as a fishing tool to highlight proteins to validate as substrates of the *M. smegmatis* ESX secretion systems.

CONCLUSION

Tuberculosis, caused by *M. tuberculosis,* continues to infect and kill millions of people annually,[16] despite the development and implementation of the BCG vaccine and several anti-tuberculosis drugs. Improving our understanding of the biology and mechanisms of pathogenicity of *M. tuberculosis* will enhance our ability to develop novel drugs and vaccines to eradicate the disease. The mycobacterial ESX secretion systems, recently defined as a novel Type-VII secretion machinery, have been linked to mycobacterial virulence and survival and as such present a promising subject for research in the field of tuberculosis. This study aimed to investigate the evolution, regulation, functions and substrates of the mycobacterial ESX secretion systems.

The evolutionary history of the ESX gene cluster was investigated using *in silico* methods and determined through phylogenetic analyses of sequences obtained from all mycobacterial species which have been sequenced, as well as several closely related actinomycetes and the WXG-FtsK clusters of other bacterial species. The ESX-4 gene cluster is the ancestral ESX which appears to have evolved from the WXG-FtsK cluster along with the evolution of the mycomembrane. ESX-4 incorporates several additional genes which may be involved in mycomembrane transport. ESX-3 was the first duplication from ESX-4, and marks the evolution of the genus *Mycobacterium*. An ancestral ESX-3 was identified in *N. farcinica*, which seems to have evolved divergently from the mycobacterial ESX-3. The ESX-1 duplication appears to have occurred next, followed by ESX-2 and ESX-5, which occur only in the slow growers. An ancestral ESX-2 was identified in *M. tusciae*, a slow-growing mycobacterium which clusters phylogenetically with the fast growers, and this may be an evolutionary link between the fast- and slow- growing mycobacteria. Five additional, previously unidentified, ESX gene clusters were identified in the mycobacteria. These were identified, or predicted to be located, on plasmid DNA, and were named ESX-P1 to -P5. The plasmid ESX appear to be evolutionary precursors of the genomic ESX-3, -1 and -5, suggesting that the duplication and evolution of the ESX gene clusters may have been, at least partly, plasmid-mediated.

Promoter-containing intergenic regions of the *M. tuberculosis* ESX gene clusters were identified using a promoter-probe assay. Promoters driving expression of ESX-3, -1, -2 and -5 were identified and promoter motifs proposed using *in silico* promoter prediction programs and sequence conservation. No promoters were identified for ESX-4. Identification of the promoters of the ESX, and the availability

of the promoter probes, will facilitate the investigation of the conditions and regulation of expression of the ESX secretion systems.

M. smegmatis, a fast growing non-pathogenic model organism used for *M. tuberculosis* research, contains three ESX gene clusters, ESX-4, -3 and -1, which are similar to those of *M. tuberculosis*. A library of ESX knockout strains of *M. smegmatis* was generated by allelic exchange, and includes strains from which each ESX gene cluster has been deleted individually, and in combination to create double and triple ESX knockouts. The ΔESX-4_{ms}, ΔESX-3_{ms} and ΔESX-1_{ms} strains were used in comparative proteomic, secretomic and metabolomic analyses, with wild type *M. smegmatis*, to investigate the functions and substrates of the *M. smegmatis* ESX secretion systems. This is a novel approach which aimed to identify global changes related to the functions of the ESX secretion systems.

The ESX-3 knockout showed significant variation from wild-type *M. smegmatis* on a metabolite level. The levels of most amino acids and sugars were reduced, while increases were detected in oxidative stress markers. The proteomic variation correlated with these results, with substantial variation in proteins involved in amino acid, carbohydrate and fatty acid metabolism and oxidative stress. However, although the metabolomic variation could be ascribed to perturbed iron and zinc homeostasis, the proteomic variation did not corroborate this, as no variation in the iron and zinc regulons was detected. ESX-3 promoter activity in *M. smegmatis* was shown to be repressed in high iron conditions, correlating with low levels of the ESX-3 protein components detected. Therefore ESX-3 is likely not involved in iron uptake in the presence of high iron concentrations. Rather, the proteomic variation indicated possible disturbed polyamine metabolism in the absence of ESX-3. Spermidine synthesis requires methionine, which was almost completely absent in ΔESX-3_{ms}, supporting this theory. Polyamines influence multiple cellular processes as indicated by the variation in the proteomic and metabolomic profiles of the two *M. smegmatis* strains.

No ESX-4 component proteins were detected in wild type *M. smegmatis*, and no significant variation was observed in the metabolomic analysis. This correlates with the lack of promoter activity obtained for ESX-4. The proteomes of ΔESX-4_{ms} and WT_{ms}, however, showed substantial variation. ESX-3 component proteins were found to be half as abundant in ΔESX-4_{ms} as WT_{ms}, suggesting that ESX-4 expression influences the expression of ESX-3, and that their functions might be linked. Much of the

proteomic variation correlated with pathways which were altered in ΔESX-$3_{ms,}$ and could be ascribed to the reduced expression of ESX-3. WhmD, a WhiB2 global transcriptional regulator of cell division, was 4 times more abundant in ΔESX-4_{ms}, reaffirming a previous link between ESX-4 and cell division. Furthermore, some other cell division and cell-division related cell wall biosynthetic proteins were undetected in the absence of ESX-4. Molybdopterin metabolism was also influenced by the absence of ESX-4, suggesting that ESX-4 may be involved in molybdenum homeostasis, linking ESX-3 and ESX-4 to metal cation homeostasis.

ESX-1 was the most highly expressed ESX in *M. smegmatis* under the conditions of this study. However, although most ESX-1 component proteins were detected in WT_{ms} and not in ΔESX-1_{ms}, the ΔESX-1_{ms} proteome showed the least variation relative to wild type *M. smegmatis*. Proteins with altered abundances are involved in a variety of functions and pathways, and no specific pathways showed substantial variation, suggesting that ESX-1 functions independently of many other cellular processes. ESX-1 has been linked to conjugal DNA transfer in *M. smegmatis* and mycolic acid synthesis in *M. tuberculosis*; however the proteomic variation does not correlate with these functions. The involvement of ESX-1 in conjugative DNA transfer, in light of the association of the ESX with plasmids, identified in this study, is intriguing, suggesting that the ESX may originally have facilitated conjugal transfer of plasmid DNA between mycobacterial cells. The absence of ESX-1 does result in changes in the abundance of several transcriptional regulators. This was also observed for ESX-3 and ESX-4, suggesting that the ESX secretion systems influence the regulome of the bacterium, in different ways, in order to maintain cellular homeostasis.

Proteins which were secreted in WT_{ms}, but not in any of the ΔESX_{ms} strains, were investigated as possible secreted substrates of the associated ESX secretion system. Several of the proteins contain a YxxxD/E general mycobacterial type-VII ESX secretion motif, and may be secreted substrates of these secretion systems. However, many of these proteins are membrane localised, and it is unclear whether these proteins are integrated into the membrane via the ESX, or if the altered abundance in the culture filtrate is a result of other cellular processes. Other possible substrates are hypothetical proteins, and no functional relevance could be determined. None of the prototype ESX secreted substrate proteins were detected in the culture filtrates. The current techniques for fractionation of mycobacterial cells are not specific, and proteins from other fractions are frequently found. Numerous

cytosolic and membrane proteins were detected in the culture filtrates, therefore improved fractionation methods will facilitate the identification of novel secreted substrates. The low expression levels of ESX-3 and ESX-4, under the conditions of this study, may also prevent the identification of novel substrates. Although multiple substrates of the ESX secretion systems have been identified for *M. tuberculosis* and *M. marinum*, homologs of these proteins are not present in *M. smegmatis*. Secretion of these additional substrates may be responsible for the functional variation in some ESX secretion systems between *M. tuberculosis* and *M. smegmatis*.

This study has described the evolutionary history of the mycobacterial ESX secretion systems, identified promoters driving their expression, identified novel pathways in which the *M. smegmatis* ESX secretion systems are involved, and highlighted possible substrates of the secretion systems. This study sets the groundwork for future work in understanding the functional roles and expression patterns of these ESX secretion systems and in using global proteomic and metabolomic analyses to understand cellular changes in response to specific signals or genomic changes. The ESX secretion systems are essential for virulence and survival in *M. tuberculosis*, and improved understanding of their expression, functions and mechanisms may lead to improved treatment and prevention strategies for tuberculosis.

FUTURE DIRECTIONS

In this study, we have investigated the evolutionary history of the mycobacterial ESX secretion systems, based on sequenced genomes. Evolutionary analyses of unsequenced mycobacteria, by PCR-sequencing and Southern-blotting, as well as future *in silico* analysis of the rapidly growing collection of mycobacterial whole genome sequences will fill in gaps in the evolutionary history of the ESX secretion systems.

The proteomic, secretomic and metabolomic variation between *M. smegmatis* wild type and ΔESX-1_{ms}, ΔESX-3_{ms} and ΔESX-4_{ms} strains was investigated under standard 7H9 culture conditions. No ESX-4 proteins and only low levels of some ESX-3 proteins were detected under these conditions. Similar studies investigating ESX-3 under conditions of iron and zinc deprivation may provide better insight into the role of ESX-3 in divalent metal cation homeostasis, and provide more marked differences between the strains (metabolomic analyses of WT_{ms} and ΔESX-3_{ms} in media containing different iron concentrations are currently underway, in collaboration with Prof. Loots at North West University, South Africa). The ESX-4 expression patterns should be determined, possibly by assaying promoter activity under various conditions, to determine the optimal conditions for expression. WhiB5 and SigM dependent expression has recently been shown for ESX-4, suggesting that ESX-4 may be expressed under starvation or reactivation. Similar studies under conditions in which ESX-4 is more abundantly expressed, will enhance the identification of changes resulting from its absence or presence. Metabolomic analyses of WT_{ms} and ΔESX-1_{ms} are currently underway, in collaboration with Prof. Loots at North West University, South Africa.

Identification of novel secreted substrates of the ESX secretion systems was hampered by inefficient fractionation and purification of the culture supernatant. Alternative culture conditions and fractionation methods should be investigated to optimise this process to prevent contamination with cytosolic and membrane proteins. The use of solid media has been investigated, but resulted in similar amounts of cytosolic contamination (Z. Fang, unpublished results). Increasing the expression of the ESX secretion systems, and thereby the rate of secretion, will also assist in the detection of the associated secreted substrates.

Transcriptomic analyses of the ΔESX_{ms} strains will further establish the functional role of the ESX secretion systems in *M. smegmatis*. Linking the transcriptomic, proteomic, secretomic and

metabolomic effects of the presence or absence of each ESX secretion system will generate a complete “systems biology” view of ESX functionality.

Complementation of the ESX deletions and whole genome sequencing of the ΔESX_{ms} strains will assist in confirming which changes are directly associated with the ESX deletion, and which may result from other genomic mutations or adaptive changes resulting from the allelic exchange process used to generate the ΔESX_{ms} strains.

Investigation of the transcriptomes, proteomes, secretomes and metabolomes of the double- and triple ΔESX_{ms} strains will help to further establish the functional roles of the *M. smegmatis* ESX secretion systems, the overall cellular function of ESX secretion and as well as the interactions between different ESX secretion systems.

Similar proteomic, secretomic and metabolomic analyses could be done on *M. tuberculosis* or the non-pathogenic *M. bovis* BCG to investigate the functions and substrates of these secretion systems in slow growing mycobacteria. This will help to identify functions and substrates of ESX-2, which remains largely unstudied, and of ESX-5. The promoter probe vectors could be utilised to identify optimal expression conditions for these analyses.

This study identified novel pathways associated with the ESX-3 and ESX-4 secretion systems. The involvement of the ESX secretion systems in these processes must be verified and the mechanisms involved established. The possible secreted substrates can be confirmed by tagged protein expression and western blotting detection.

In vivo studies, using proteomics and secretomics in infected THP-1 cell lines or macrophages, will investigate the roles of the ESX secretion systems during infection and establish host-pathogen interactions influenced by ESX secretion.

Reference List

1. Stathopoulos C, Hendrixson DR, Thanassi DG, Hultgren SJ, St Geme JW, Curtiss R. Secretion of virulence determinants by the general secretory pathway in gram-negative pathogens: An evolving story. *Microbes Infect* 2000;**2**:1061-1072.

2. van Wely KH, Swaving J, Freudl R, Driessen AJ. Translocation of proteins across the cell envelope of gram-positive bacteria. *FEMS Microbiol Rev* 2001;**25**:437-454.

3. Osborne AR, Rapoport TA, van den Berg B. Protein translocation by the Sec61/SecY channel. *Annu Rev Cell Dev Biol* 2005;**21**:529-550.

4. Driessen AJ, Manting EH, van der Does C. The structural basis of protein targeting and translocation in bacteria. *Nat Struct Biol* 2001;**8**:492-498.

5. Wickner W, Driessen AJ, Hartl FU. The enzymology of protein translocation across the *Escherichia coli* plasma membrane. *Annu Rev Biochem* 1991;**60**:101-124.

6. Berks BC, Palmer T, Sargent F. Protein targeting by the bacterial twin-arginine translocation (tat) pathway. *Curr Opin Microbiol* 2005;**8**:174-181.

7. Tjalsma H, Bolhuis A, Jongbloed JD, Bron S, van Dijl JM. Signal peptide-dependent protein transport in *Bacillus subtilis*: A genome-based survey of the secretome. *Microbiol Mol Biol Rev* 2000;**64**:515-547.

8. Holland IB, Schmitt L, Young J. Type 1 protein secretion in bacteria, the ABC-transporter dependent pathway. *Mol Membr Biol* 2005;**22**:29-39.

9. Johnson TL, Abendroth J, Hol WG, Sandkvist M. Type II secretion: From structure to function. *FEMS Microbiol Lett* 2006;**255**:175-186.

10. Cornelis GR. The type III secretion injectisome. *Nat Rev Microbiol* 2006;**4**:811-825.

11. Henderson IR, Navarro-Garcia F, Desvaux M, Fernandez RC, Ala'Aldeen D. Type V protein secretion pathway: The autotransporter story. *Microbiol Mol Biol Rev* 2004;**68**:692-744.

12. Mougous JD, Cuff ME, Raunser S, Shen A, Zhou M, Gifford CA, Goodman AL, Joachimiak G, Ordonez CL, Lory S, Walz T, Joachimiak A, Mekalanos JJ. A virulence locus of *Pseudomonas aeruginosa* encodes a protein secretion apparatus. *Science* 2006;**312**:1526-1530.

13. Pukatzki S, Ma AT, Sturtevant D, Krastins B, Sarracino D, Nelson WC, Heidelberg JF, Mekalanos JJ. Identification of a conserved bacterial protein secretion system in *Vibrio cholerae* using the Dictyostelium host model system. *Proc Natl Acad Sci U S A* 2006;**103**:1528-1533.

14. Christie PJ, Atmakuri K, Krishnamoorthy V, Jakubowski S, Cascales E. Biogenesis, architecture, and function of bacterial type IV secretion systems. *Annu Rev Microbiol* 2005;**59**:451-485.

15. Euzeby JP. List of Prokaryotic names with Standing in Nomenclature. 2012; http://www.bacterio.cict.fr/.

16. World Health Organisation. Tuberculosis global facts. 2011/2012; www.who.int/tb.

17. Sutcliffe IC. Cell envelope composition and organisation in the genus Rhodococcus. *Antonie Van Leeuwenhoek* 1998;**74**:49-58.

18. Minnikin DE, Kremer L, Dover LG, Besra GS. The methyl-branched fortifications of *Mycobacterium tuberculosis*. *Chem Biol* 2002;**9**:545-553.

19. Bayan N, Houssin C, Chami M, Leblon G. Mycomembrane and S-layer: Two important structures of *Corynebacterium glutamicum* cell envelope with promising biotechnology applications. *J Biotechnol* 2003;**104**:55-67.

20. Zuber B, Chami M, Houssin C, Dubochet J, Griffiths G, Daffe M. Direct visualization of the outer membrane of mycobacteria and corynebacteria in their native state. *J Bacteriol* 2008;**190**:5672-5680.

21. Hoffmann C, Leis A, Niederweis M, Plitzko JM, Engelhardt H. Disclosure of the mycobacterial outer membrane: Cryo-electron tomography and vitreous sections reveal the lipid bilayer structure. *Proc Natl Acad Sci U S A* 2008;**105**:3963-3967.

22. Pugsley AP. The complete general secretory pathway in gram-negative bacteria. *Microbiol Rev* 1993;**57**:50-108.

23. Braunstein M, Espinosa BJ, Chan J, Belisle JT, Jacobs WR,Jr. SecA2 functions in the secretion of superoxide dismutase A and in the virulence of *Mycobacterium tuberculosis. Mol Microbiol* 2003;**48**:453-464.

24. McDonough JA, McCann JR, Tekippe EM, Silverman JS, Rigel NW, Braunstein M. Identification of functional Tat signal sequences in *Mycobacterium tuberculosis* proteins. *J Bacteriol* 2008;**190**:6428-6438.

25. Saint-Joanis B, Demangel C, Jackson M, Brodin P, Marsollier L, Boshoff H, Cole ST. Inactivation of Rv2525c, a substrate of the twin arginine translocation (tat) system of *Mycobacterium tuberculosis*, increases beta-lactam susceptibility and virulence. *J Bacteriol* 2006;**188**:6669-6679.

26. Tekaia F, Gordon SV, Garnier T, Brosch R, Barrell BG, Cole ST. Analysis of the proteome of *Mycobacterium tuberculosis in silico. Tuber Lung Dis* 1999;**79**:329-342.

27. Gey Van Pittius NC, Gamieldien J, Hide W, Brown GD, Siezen RJ, Beyers AD. The ESAT-6 gene cluster of *Mycobacterium tuberculosis* and other high G+C gram-positive bacteria. *Genome Biol* 2001;**2**:RESEARCH0044.

28. Pym AS, Domenech P, Honore N, Song J, Deretic V, Cole ST. Regulation of catalase-peroxidase (KatG) expression, isoniazid sensitivity and virulence by FurA of *Mycobacterium tuberculosis*. *Mol Microbiol* 2001;**40**:879-889.

29. Brodin P, Majlessi L, Marsollier L, de Jonge MI, Bottai D, Demangel C, Hinds J, Neyrolles O, Butcher PD, Leclerc C, Cole ST, Brosch R. Dissection of ESAT-6 system 1 of *Mycobacterium tuberculosis* and impact on immunogenicity and virulence. *Infect Immun* 2006;**74**:88-98.

30. Abdallah AM, Gey van Pittius NC, Champion PA, Cox J, Luirink J, Vandenbroucke-Grauls CM, Appelmelk BJ, Bitter W. Type VII secretion - mycobacteria show the way. *Nat Rev Microbiol* 2007;**5**:883-891.

31. Bitter W, Houben EN, Bottai D, Brodin P, Brown EJ, Cox JS, Derbyshire K, Fortune SM, Gao LY, Liu J, Gey van Pittius NC, Pym AS, Rubin EJ, Sherman DR, Cole ST, Brosch R. Systematic genetic nomenclature for type VII secretion systems. *PLoS Pathog* 2009;**5**:e1000507.

32. Stanley SA, Raghavan S, Hwang WW, Cox JS. Acute infection and macrophage subversion by *Mycobacterium tuberculosis* require a specialized secretion system. *Proc Natl Acad Sci U S A* 2003;**100**:13001-13006.

33. Mahairas GG, Sabo PJ, Hickey MJ, Singh DC, Stover CK. Molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent *M. bovis. J Bacteriol* 1996;**178**:1274-1282.

34. Behr MA, Wilson MA, Gill WP, Salamon H, Schoolnik GK, Rane S, Small PM. Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* 1999;**284**:1520-1523.

35. Brosch R, Gordon SV, Buchrieser C, Pym AS, Garnier T, Cole ST. Comparative genomics uncovers large tandem chromosomal duplications in *Mycobacterium bovis* BCG Pasteur. *Yeast* 2000;**17**:111-123.

36. Berthet FX, Rasmussen PB, Rosenkrands I, Andersen P, Gicquel B. A *Mycobacterium tuberculosis* operon encoding ESAT-6 and a novel low-molecular-mass culture filtrate protein (CFP-10). *Microbiology* 1998;**144**:3195-3203.

37. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE,III, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream MA, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Squares S, Sulston JE, Taylor K, Whitehead S, Barrell BG. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 1998;**393**:537-544.

38. Gey van Pittius NC, Sampson SL, Lee H, Kim Y, van Helden PD, Warren RM. Evolution and expansion of the *Mycobacterium tuberculosis* PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions. *BMC Evol Biol* 2006;**6**:95.

39. Pallen MJ. The ESAT-6/WXG100 superfamily – and a new gram-positive secretion system? *Trends Microbiol* 2002;**10**:209-212.

40. Renshaw PS, Lightbody KL, Veverka V, Muskett FW, Kelly G, Frenkiel TA, Gordon SV, Hewinson RG, Burke B, Norman J, Williamson RA, Carr MD. Structure and function of the complex formed by the tuberculosis virulence factors CFP-10 and ESAT-6. *EMBO J* 2005;**24**:2491-2498.

41. Ilghari D, Lightbody KL, Veverka V, Waters LC, Muskett FW, Renshaw PS, Carr MD. Solution structure of the *Mycobacterium tuberculosis* EsxG.EsxH complex: Functional implications and comparisons with other *M. tuberculosis* esx family complexes. *J Biol Chem* 2011;**286**:29993-30002.

42. Arbing MA, Kaufmann M, Phan T, Chan S, Cascio D, Eisenberg D. The crystal structure of the *Mycobacterium tuberculosis* Rv3019c-Rv3020c ESX complex reveals a domain-swapped heterotetramer. *Protein Sci* 2010;**19**:1692-1703.

43. de Jonge MI, Pehau-Arnaudet G, Fretz MM, Romain F, Bottai D, Brodin P, Honore N, Marchal G, Jiskoot W, England P, Cole ST, Brosch R. ESAT-6 from *Mycobacterium tuberculosis* dissociates from its putative chaperone CFP-10 under acidic conditions and exhibits membrane-lysing activity. *J Bacteriol* 2007;**189**:6028-6034.

44. Camus JC, Pryor MJ, Medigue C, Cole ST. Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology* 2002;**148**:2967-2973.

45. Gordon SV, Eiglmeier K, Brosch R, Garnier T, Honoré N, Barrell BG, Cole ST. Genomics of *Mycobacterium tuberculosis* and *Mycobacterium leprae*. In: *Mycobacteria.* Blackwell Publishing Ltd; 1999; 2009. p 93-109.

46. Brennan MJ, Delogu G, Chen Y, Bardarov S, Kriakov J, Alavi M, Jacobs WR,Jr. Evidence that mycobacterial PE_PGRS proteins are cell surface constituents that influence interactions with other cells. *Infect Immun* 2001;**69**:7326-7333.

47. Banu S, Honore N, Saint-Joanis B, Philpott D, Prevost MC, Cole ST. Are the PE-PGRS proteins of *Mycobacterium tuberculosis* variable surface antigens? *Mol Microbiol* 2002;**44**:9-19.

48. Delogu G, Pusceddu C, Bua A, Fadda G, Brennan MJ, Zanetti S. Rv1818c-encoded PE_PGRS protein of *Mycobacterium tuberculosis* is surface exposed and influences bacterial cell structure. *Mol Microbiol* 2004;**52**:725-733.

49. Cole ST. Learning from the genome sequence of *Mycobacterium tuberculosis* H37Rv. *FEBS Lett* 1999;**452**:7-10.

50. Rodriguez GM, Voskuil MI, Gold B, Schoolnik GK, Smith I. *ideR*, an essential gene in *Mycobacterium tuberculosis*: Role of IdeR in iron-dependent gene expression, iron metabolism, and oxidative stress response. *Infect Immun* 2002;**70**:3371-3381.

51. Rodriguez GM, Gold B, Gomez M, Dussurget O, Smith I. Identification and characterization of two divergently transcribed iron regulated genes in *Mycobacterium tuberculosis. Tuber Lung Dis* 1999;**79**:287-298.

52. Abou-Zeid C, Garbe T, Lathigra R, Wiker HG, Harboe M, Rook GA, Young DB. Genetic and immunological analysis of *Mycobacterium tuberculosis* fibronectin-binding proteins. *Infect Immun* 1991;**59**:2712-2718.

53. Espitia C, Laclette JP, Mondragon-Palomino M, Amador A, Campuzano J, Martens A, Singh M, Cicero R, Zhang Y, Moreno C. The PE-PGRS glycine-rich proteins of *Mycobacterium tuberculosis*: A new family of fibronectin-binding proteins? *Microbiology* 1999;**145**:3487-3495.

54. Sassetti CM, Boyd DH, Rubin EJ. Genes required for mycobacterial growth defined by high density mutagenesis. *Mol Microbiol* 2003;**48**:77-84.

55. Li Y, Miltner E, Wu M, Petrofsky M, Bermudez LE. A *Mycobacterium avium* PPE gene is associated with the ability of the bacterium to grow in macrophages and virulence in mice. *Cell Microbiol* 2005;**7**:539-548.

56. Okkels LM, Brock I, Follmann F, Agger EM, Arend SM, Ottenhoff TH, Oftung F, Rosenkrands I, Andersen P. PPE protein (Rv3873) from DNA segment RD1 of *Mycobacterium tuberculosis:* Strong recognition of both specific T-cell epitopes and epitopes conserved within the PPE family. *Infect Immun* 2003;**71**:6116-6123.

57. Fortune SM, Jaeger A, Sarracino DA, Chase MR, Sassetti CM, Sherman DR, Bloom BR, Rubin EJ. Mutually dependent secretion of proteins required for mycobacterial virulence. *Proc Natl Acad Sci U S A* 2005;**102**:10676-10681.

58. Abdallah AM, Verboom T, Weerdenburg EM, Gey van Pittius NC, Mahasha PW, Jimenez C, Parra M, Cadieux N, Brennan MJ, Appelmelk BJ, Bitter W. PPE and PE_PGRS proteins of *Mycobacterium marinum* are transported via the type VII secretion system ESX-5. *Mol Microbiol* 2009;**73**:329-340.

59. Abdallah AM, Verboom T, Hannes F, Safi M, Strong M, Eisenberg D, Musters RJ, Vandenbroucke-Grauls CM, Appelmelk BJ, Luirink J, Bitter W. A specific secretion system mediates PPE41 transport in pathogenic mycobacteria. *Mol Microbiol* 2006;**62**:667-679.

60. Daleke MH, Cascioferro A, de Punder K, Ummels R, Abdallah AM, van der Wel N, Peters PJ, Luirink J, Manganelli R, Bitter W. Conserved pro-glu (PE) and pro-pro-glu (PPE) protein domains target LipY lipases of pathogenic mycobacteria to the cell surface via the ESX-5 pathway. *J Biol Chem* 2011;**286**:19024-19034.

61. Bottai D, Di Luca M, Majlessi L, Frigui W, Simeone R, Sayes F, Bitter W, Brennan MJ, Leclerc C, Batoni G, Campa M, Brosch R, Esin S. Disruption of the ESX-5 system of *Mycobacterium tuberculosis* causes loss of PPE protein secretion, reduction of cell wall integrity and strong attenuation. *Mol Microbiol* 2012; 10.1111/j.1365-2958.2012.08001.x;

62. Doran TJ, Hodgson AL, Davies JK, Radford AJ. Characterisation of a novel repetitive DNA sequence from *Mycobacterium bovis. FEMS Microbiol Lett* 1992;**75**:179-185.

63. Sampson SL, Lukey P, Warren RM, van Helden PD, Richardson M, Everett MJ. Expression, characterization and subcellular localization of the *Mycobacterium tuberculosis* PPE gene Rv1917c. *Tuberculosis (Edinb)* 2001;**81**:305-317.

64. Le Moigne V, Robreau G, Borot C, Guesdon JL, Mahana W. Expression, immunochemical characterization and localization of the *Mycobacterium tuberculosis* protein p27. *Tuberculosis (Edinb)* 2005;**85**:213-219.

65. Pajon R, Yero D, Lage A, Llanes A, Borroto CJ. Computational identification of beta-barrel outer-membrane proteins in *Mycobacterium tuberculosis* predicted proteomes as putative vaccine candidates. *Tuberculosis (Edinb)* 2006;**86**:290-302.

66. Riley R, Pellegrini M, Eisenberg D. Identifying cognate binding pairs among a large set of paralogs: The case of PE/PPE proteins of *Mycobacterium tuberculosis*. *PLoS Comput Biol* 2008;**4**:e1000174.

67. Okkels LM, Andersen P. Protein-protein interactions of proteins from the ESAT-6 family of *Mycobacterium tuberculosis*. *J Bacteriol* 2004;**186**:2487-2491.

68. Strong M, Sawaya MR, Wang S, Phillips M, Cascio D, Eisenberg D. Toward the structural genomics of complexes: Crystal structure of a PE/PPE protein complex from *Mycobacterium tuberculosis*. *Proc Natl Acad Sci U S A* 2006;**103**:8060-8065.

69. Tundup S, Akhter Y, Thiagarajan D, Hasnain SE. Clusters of PE and PPE genes of *Mycobacterium tuberculosis* are organized in operons: Evidence that PE Rv2431c is co-transcribed with PPE Rv2430c and their gene products interact with each other. *FEBS Lett* 2006;**580**:1285-1293.

70. Teutschbein J, Schumann G, Mollmann U, Grabley S, Cole ST, Munder T. A protein linkage map of the ESAT-6 secretion system 1 (ESX-1) of *Mycobacterium tuberculosis. Microbiol Res* 2009;**164**:253-259.

71. Das C, Ghosh TS, Mande SS. Computational analysis of the ESX-1 region of *Mycobacterium tuberculosis*: Insights into the mechanism of type VII secretion system. *PLoS One* 2011;**6**:e27980.

72. Daleke MH, Ummels R, Bawono P, Heringa J, Vandenbroucke-Grauls CM, Luirink J, Bitter W. General secretion signal for the mycobacterial type VII secretion pathway. *Proc Natl Acad Sci U S A* 2012;**109**:11342-11347.

73. DiGiuseppe Champion PA, Champion MM, Manzanillo P, Cox JS. ESX-1 secreted virulence factors are recognized by multiple cytosolic AAA ATPases in pathogenic mycobacteria. *Mol Microbiol* 2009;**73**:950-962.

74. Champion PA, Stanley SA, Champion MM, Brown EJ, Cox JS. C-terminal signal sequence promotes virulence factor secretion in *Mycobacterium tuberculosis. Science* 2006;**313**:1632-1636.

75. Luthra A, Mahmood A, Arora A, Ramachandran R. Characterization of Rv3868, an essential hypothetical protein of the ESX-1 secretion system in *Mycobacterium tuberculosis. J Biol Chem* 2008;**283**:36532-36541.

76. Ogura T, Whiteheart SW, Wilkinson AJ. Conserved arginine residues implicated in ATP hydrolysis, nucleotide-sensing, and inter-subunit interactions in AAA and AAA+ ATPases. *J Struct Biol* 2004;**146**:106-112.

77. Brown GD, Dave JA, Gey van Pittius NC, Stevens L, Ehlers MR, Beyers AD. The mycosins of *Mycobacterium tuberculosis* H37Rv: A family of subtilisin-like serine proteases. *Gene* 2000;**254**:147-155.

78. Ohol YM, Goetz DH, Chan K, Shiloh MU, Craik CS, Cox JS. *Mycobacterium tuberculosis* MycP1 protease plays a dual role in regulation of ESX-1 secretion and virulence. *Cell Host Microbe* 2010;**7**:210-220.

79. van Soolingen D, van der Zanden AG, de Haas PE, Noordhoek GT, Kiers A, Foudraine NA, Portaels F, Kolk AH, Kremer K, van Embden JD. Diagnosis of *Mycobacterium microti* infections among humans by using novel genetic markers. *J Clin Microbiol* 1998;**36**:1840-1845.

80. Gordon SV, Brosch R, Billault A, Garnier T, Eiglmeier K, Cole ST. Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Mol Microbiol* 1999;**32**:643-655.

81. Brodin P, Eiglmeier K, Marmiesse M, Billault A, Garnier T, Niemann S, Cole ST, Brosch R. Bacterial artificial chromosome-based comparative genomic analysis identifies *Mycobacterium microti* as a natural ESAT-6 deletion mutant. *Infect Immun* 2002;**70**:5568-5578.

82. Hsu T, Hingley-Wilson SM, Chen B, Chen M, Dai AZ, Morin PM, Marks CB, Padiyar J, Goulding C, Gingery M, Eisenberg D, Russell RG, Derrick SC, Collins FM, Morris SL, King CH, Jacobs WR,Jr. The primary mechanism of attenuation of Bacillus Calmette-Guerin is a loss of secreted lytic function required for invasion of lung interstitial tissue. *Proc Natl Acad Sci U S A* 2003;**100**:12420-12425.

83. Lewis KN, Liao R, Guinn KM, Hickey MJ, Smith S, Behr MA, Sherman DR. Deletion of RD1 from *Mycobacterium tuberculosis* mimics Bacille Calmette-Guerin attenuation. *J Infect Dis* 2003;**187**:117-123.

84. Guinn KM, Hickey MJ, Mathur SK, Zakel KL, Grotzke JE, Lewinsohn DM, Smith S, Sherman DR. Individual RD1-region genes are required for export of ESAT-6/CFP-10 and for virulence of *Mycobacterium tuberculosis. Mol Microbiol* 2004;**51**:359-370.

85. Pym AS, Brodin P, Majlessi L, Brosch R, Demangel C, Williams A, Griffiths KE, Marchal G, Leclerc C, Cole ST. Recombinant BCG exporting ESAT-6 confers enhanced protection against tuberculosis. *Nat Med* 2003;**9**:533-539.

86. Stoop EJ, Schipper T, Huber SK, Nezhinsky AE, Verbeek FJ, Gurcha SS, Besra GS, Vandenbroucke-Grauls CM, Bitter W, van der Sar AM. Zebrafish embryo screen for mycobacterial genes involved in the initiation of granuloma formation reveals a newly identified ESX-1 component. *Dis Model Mech* 2011;**4**:526-536.

87. McLaughlin B, Chon JS, MacGurn JA, Carlsson F, Cheng TL, Cox JS, Brown EJ. A mycobacterium ESX-1-secreted virulence factor with unique requirements for export. *PLoS Pathog* 2007;**3**:e105.

88. Xu J, Laine O, Masciocchi M, Manoranjan J, Smith J, Du SJ, Edwards N, Zhu X, Fenselau C, Gao LY. A unique mycobacterium ESX-1 protein co-secretes with CFP-10/ESAT-6 and is necessary for inhibiting phagosome maturation. *Mol Microbiol* 2007;**66**:787-800.

89. Millington KA, Fortune SM, Low J, Garces A, Hingley-Wilson SM, Wickremasinghe M, Kon OM, Lalvani A. Rv3615c is a highly immunodominant RD1 (region of difference 1)-dependent secreted antigen specific for *Mycobacterium tuberculosis* infection. *Proc Natl Acad Sci USA* 2011;**108**:5730-5735.

90. Rosenberg OS, Dovey C, Tempesta M, Robbins RA, Finer-Moore JS, Stroud RM, Cox JS. EspR, a key regulator of *Mycobacterium tuberculosis* virulence, adopts a unique dimeric structure among helix-turn-helix proteins. *Proc Natl Acad Sci USA* 2011;**108**:13450-13455.

91. Chen JM, Boy-Rottger S, Dhar N, Sweeney N, Buxton RS, Pojer F, Rosenkrands I, Cole ST. EspD is critical for the virulence-mediating ESX-1 secretion system in *Mycobacterium tuberculosis*. *J Bacteriol* 2012;**194**:884-893.

92. Carlsson F, Kim J, Dumitru C, Barck KH, Carano RA, Sun M, Diehl L, Brown EJ. Host-detrimental role of ESX-1-mediated inflammasome activation in mycobacterial infection. *PLoS Pathog* 2010;**6**:e1000895.

93. Bottai D, Majlessi L, Simeone R, Frigui W, Laurent C, Lenormand P, Chen J, Rosenkrands I, Huerre M, Leclerc C, Cole ST, Brosch R. ESAT-6 secretion-independent impact of ESX-1 genes *espF* and *espG1* on virulence of *Mycobacterium tuberculosis*. *J Infect Dis* 2011;**203**:1155-1164.

94. Raghavan S, Manzanillo P, Chan K, Dovey C, Cox JS. Secreted transcription factor controls *Mycobacterium tuberculosis* virulence. *Nature* 2008;**454**:717-721.

95. Blasco B, Chen JM, Hartkoorn R, Sala C, Uplekar S, Rougemont J, Pojer F, Cole ST. Virulence regulator EspR of *Mycobacterium tuberculosis* is a nucleoid-associated protein. *PLoS Pathog* 2012;**8**:e1002621.

96. MacGurn JA, Cox JS. A genetic screen for *Mycobacterium tuberculosis* mutants defective for phagosome maturation arrest identifies components of the ESX-1 secretion system. *Infect Immun* 2007;**75**:2668-2678.

97. Samten B, Wang X, Barnes PF. *Mycobacterium tuberculosis* ESX-1 system-secreted protein ESAT-6 but not CFP10 inhibits human T-cell immune responses. *Tuberculosis (Edinb)* 2009;**89**:S74-6.

98. Smith J, Manoranjan J, Pan M, Bohsali A, Xu J, Liu J, McDonald KL, Szyk A, LaRonde-LeBlanc N, Gao LY. Evidence for pore formation in host cell membranes by ESX-1-secreted ESAT-6 and its role in *Mycobacterium marinum* escape from the vacuole. *Infect Immun* 2008;**76**:5478-5487.

99. Abdallah AM, Bestebroer J, Savage ND, de Punder K, van Zon M, Wilson L, Korbee CJ, van der Sar AM, Ottenhoff TH, van der Wel NN, Bitter W, Peters PJ. Mycobacterial secretion systems ESX-1 and ESX-5 play distinct roles in host cell death and inflammasome activation. *J Immunol* 2011;**187**:4744-4753.

100. Houben D, Demangel C, van Ingen J, Perez J, Baldeon L, Abdallah AM, Caleechurn L, Bottai D, van Zon M, de Punder K, van der Laan T, Kant A, Bossers-de Vries R, Willemsen P, Bitter W, van Soolingen D, Brosch R, van der Wel N, Peters PJ. ESX-1-mediated translocation to the cytosol controls virulence of mycobacteria. *Cell Microbiol* 2012; 10.1111/j.1462-5822.2012.01799.x.

101. Manzanillo PS, Shiloh MU, Portnoy DA, Cox JS. *Mycobacterium tuberculosis* activates the DNA-dependent cytosolic surveillance pathway within macrophages. *Cell Host Microbe* 2012;**11**:469-480.

102. Stanley SA, Johndrow JE, Manzanillo P, Cox JS. The type I IFN response to infection with *Mycobacterium tuberculosis* requires ESX-1-mediated secretion and contributes to pathogenesis. *J Immunol* 2007;**178**:3143-3152.

103. Koo IC, Wang C, Raghavan S, Morisaki JH, Cox JS, Brown EJ. ESX-1-dependent cytolysis in lysosome secretion and inflammasome activation during mycobacterial infection. *Cell Microbiol* 2008;**10**:1866-1878.

104. Mishra BB, Moura-Alves P, Sonawane A, Hacohen N, Griffiths G, Moita LF, Anes E. *Mycobacterium tuberculosis* protein ESAT-6 is a potent activator of the NLRP3/ASC inflammasome. *Cell Microbiol* 2010;**12**:1046-1063.

105. Flint JL, Kowalski JC, Karnati PK, Derbyshire KM. The RD1 virulence locus of *Mycobacterium tuberculosis* regulates DNA transfer in *Mycobacterium smegmatis*. *Proc Natl Acad Sci USA* 2004;**101**:12598-12603.

106. Coros A, Callahan B, Battaglioli E, Derbyshire KM. The specialized secretory apparatus ESX-1 is essential for DNA transfer in *Mycobacterium smegmatis*. *Mol Microbiol* 2008;**69**:794-808.

107. Wirth SE, Krywy JA, Aldridge BB, Fortune SM, Fernandez-Suarez M, Gray TA, Derbyshire KM. Polar assembly and scaffolding proteins of the virulence-associated ESX-1 secretory apparatus in mycobacteria. *Mol Microbiol* 2012;**83**:654-664.

108. Carlsson F, Joshi SA, Rangell L, Brown EJ. Polar localization of virulence-related ESX-1 secretion in mycobacteria. *PLoS Pathog* 2009;**5**:e1000285.

109. Joshi SA, Ball DA, Sun MG, Carlsson F, Watkins BY, Aggarwal N, McCracken JM, Huynh KK, Brown EJ. EccA1, a component of the *Mycobacterium marinum* ESX-1 protein virulence factor secretion pathway, regulates mycolic acid lipid synthesis. *Chem Biol* 2012;**19**:372-380.

110. Converse SE, Cox JS. A protein secretion pathway critical for *Mycobacterium tuberculosis* virulence is conserved and functional in *Mycobacterium smegmatis*. *J Bacteriol* 2005;**187**:1238-1245.

111. Tsolaki AG, Hirsh AE, DeRiemer K, Enciso JA, Wong MZ, Hannan M, Goguet de la Salmoniere YO, Aman K, Kato-Maeda M, Small PM. Functional and evolutionary genomics of *Mycobacterium tuberculosis*: Insights from genomic deletions in 100 strains. *Proc Natl Acad Sci USA* 2004;**101**:4865-4870.

112. Casonato S, Cervantes Sanchez A, Haruki H, Rengifo Gonzalez M, Provvedi R, Dainese E, Jaouen T, Gola S, Bini E, Vicente M, Johnsson K, Ghisotti D, Palu G, Hernandez-Pando R, Manganelli R. WhiB5: A transcriptional regulator contributing to *Mycobacterium tuberculosis* virulence and reactivation. *Infect Immun* 2012; **80**:3132-44.

113. Serafini A, Boldrin F, Palu G, Manganelli R. Characterization of a *Mycobacterium tuberculosis* ESX-3 conditional mutant: Essentiality and rescue by iron and zinc. *J Bacteriol* 2009;**191**:6340-6344.

114. Maciag A, Dainese E, Rodriguez GM, Milano A, Provvedi R, Pasca MR, Smith I, Palu G, Riccardi G, Manganelli R. Global analysis of the *Mycobacterium tuberculosis* Zur (FurB) regulon. *J Bacteriol* 2007;**189**:730-740.

115. Maciag A, Piazza A, Riccardi G, Milano A. Transcriptional analysis of ESAT-6 cluster 3 in *Mycobacterium smegmatis. BMC Microbiol* 2009;**9**:48.

116. Siegrist MS, Unnikrishnan M, McConnell MJ, Borowsky M, Cheng TY, Siddiqi N, Fortune SM, Moody DB, Rubin EJ. Mycobacterial ESX-3 is required for mycobactin-mediated iron acquisition. *Proc Natl Acad Sci U S A* 2009;**106**:18792-18797.

117. Skjot RL, Oettinger T, Rosenkrands I, Ravn P, Brock I, Jacobsen S, Andersen P. Comparative evaluation of low-molecular-mass proteins from *Mycobacterium tuberculosis* identifies members of the ESAT-6 family as immunodominant T-cell antigens. *Infect Immun* 2000;**68**:214-220.

118. Sweeney KA, Dao DN, Goldberg MF, Hsu T, Venkataswamy MM, Henao-Tamayo M, Ordway D, Sellers RS, Jain P, Chen B, Chen M, Kim J, Lukose R, Chan J, Orme IM, Porcelli SA, Jacobs WR,Jr. A recombinant *Mycobacterium smegmatis* induces potent bactericidal immunity against *Mycobacterium tuberculosis*. *Nat Med* 2011;**17**:1261-1268.

119. Rosenberger T, Brulle JK, Sander P. A beta-lactamase based reporter system for ESX dependent protein translocation in mycobacteria. *PLoS One* 2012;**7**:e35453.

120. Malen H, Berven FS, Fladmark KE, Wiker HG. Comprehensive analysis of exported proteins from *Mycobacterium tuberculosis* H37Rv. *Proteomics* 2007;**7**:1702-1718.

121. Slayden RA, Belisle JT. Morphological features and signature gene response elicited by inactivation of FtsI in *Mycobacterium tuberculosis*. *J Antimicrob Chemother* 2009;**63**:451-457.

122. Raman S, Puyang X, Cheng TY, Young DC, Moody DB, Husson RN. *Mycobacterium tuberculosis* SigM positively regulates esx secreted protein and nonribosomal peptide synthetase genes and down regulates virulence-associated surface lipid synthesis. *J Bacteriol* 2006;**188**:8460-8468.

123. Agarwal N, Woolwine SC, Tyagi S, Bishai WR. Characterization of the *Mycobacterium tuberculosis* sigma factor SigM by assessment of virulence and identification of SigM-dependent genes. *Infect Immun* 2007;**75**:452-461.

124. Abdallah AM, Savage ND, van Zon M, Wilson L, Vandenbroucke-Grauls CM, van der Wel NN, Ottenhoff TH, Bitter W. The ESX-5 secretion system of *Mycobacterium marinum* modulates the macrophage response. *J Immunol* 2008;**181**:7166-7175.

125. Weerdenburg EM, Abdallah AM, Mitra S, de Punder K, van der Wel NN, Bird S, Appelmelk BJ, Bitter W, van der Sar AM. ESX-5-deficient *Mycobacterium marinum* is hypervirulent in adult zebrafish. *Cell Microbiol* 2012; 10.1111/j.1462-5822.2012.01755.x.

126. Snapper SB, Melton RE, Mustafa S, Kieser T, Jacobs WR,Jr. Isolation and characterization of efficient plasmid transformation mutants of *Mycobacterium smegmatis*. *Mol Microbiol* 1990;**4**:1911-1919.

127. Kent B, Kubica G. Public health mycobacteriology: A guide for the level II laboratory. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control; 1985.

128. Timm J, Lim EM, Gicquel B. *Escherichia coli*-mycobacteria shuttle vectors for operon and gene fusions to lacZ: The pJEM series. *J Bacteriol* 1994;**176**:6749-6753.

129. Parish T, Stoker NG. Use of a flexible cassette method to generate a double unmarked *Mycobacterium tuberculosis tlyA plcABC* mutant by gene replacement. *Microbiology* 2000;**146**:1969-1975.

130. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403-410.

131. Gish W, States DJ. Identification of protein coding regions by database similarity search. *Nat Genet* 1993;**3**:266-272.

132. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;**22**:4673-4680.

133. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Clustal W and Clustal X version 2.0. *Bioinformatics* 2007;**23**:2947-2948.

134. Hall T. BioEdit: A user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. *Nucl Acids Symp Ser* 1999;**41**:95-98.

135. Swofford DL. PAUP*. phylogenetic analysis using parsimony (*and other methods). *Sinauer Associates, Sunderland, Massachusetts* 1998;**4**.

136. Page RD. TreeView: An application to display phylogenetic trees on personal computers. *Comput Appl Biosci* 1996;**12**:357-358.

137. Pardee AB, Prestidge LS. On the nature of the repressor of beta-galactosidase synthesis in *Escherichia coli. Biochim Biophys Acta* 1959;**36**:545-547.

138. Reese MG. Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. *Comput Chem* 2001;**26**:51-56.

139. Campbell DR, Chapman KE, Waldron KJ, Tottey S, Kendall S, Cavallaro G, Andreini C, Hinds J, Stoker NG, Robinson NJ, Cavet JS. Mycobacterial cells have dual nickel-cobalt sensors: Sequence relationships and metal sites of metal-responsive repressors are not congruent. *J Biol Chem* 2007;**282**:32298-32310.

140. Gordhan BG, Parish T. Gene replacement using pretreated DNA. *Methods Mol Med* 2001;**54**:77-92.

141. Jaki BU, Franzblau SG, Cho SH, Pauli GF. Development of an extraction method for mycobacterial metabolome analysis. *J Pharm Biomed Anal* 2006;**41**:196-200.

142. Meissner-Roloff R, Koekemoer G, Warren R, Loots DT. A metabolomics investigation of a hyper- and hypo-virulent phenotype of Beijing lineage *M. tuberculosis*. *Metabolomics* 2012: 10.1007/s11306-012-0424-6.

143. Kanani HH, Klapa MI. Data correction strategy for metabolomics analysis using gas chromatography-mass spectrometry. *Metab Eng* 2007;**9**:39-51.

144. Olivier I, Loots du T. A metabolomics approach to characterise and identify various mycobacterium species. *J Microbiol Methods* 2012;**88**:419-426.

145. Koekemoer G, Swanepoel J. A semi-parametric method for transforming data to normality. *Statistics and Computing* 2008;18:241-257.

146. Field A. Discovering statistics using SPSS. SAGE publications, England; 2005.

147. Reddy TB, Riley R, Wymore F, Montgomery P, DeCaprio D, Engels R, Gellesch M, Hubble J, Jen D, Jin H, Koehrsen M, Larson L, Mao M, Nitzberg M, Sisk P, Stolte C, Weiner B, White J, Zachariah ZK, Sherlock G, Galagan JE, Ball CA, Schoolnik GK. TB database: An integrated platform for tuberculosis research. *Nucleic Acids Res* 2009;**37**:D499-508.

148. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotech* 2008;**26**:1367-1372.

149. Kapopoulou A, Lew JM, Cole ST. The MycoBrowser portal: A comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis (Edinb)* 2011;**91**:8-13.

150. Burts ML, Williams WA, DeBord K, Missiakas DM. EsxA and EsxB are secreted by an ESAT-6-like system that is required for the pathogenesis of *Staphylococcus aureus* infections. *Proc Natl Acad Sci U S A* 2005;**102**:1169-1174.

151. Sutcliffe IC. New insights into the distribution of WXG100 protein secretion systems. *Antonie Van Leeuwenhoek* 2011;**99**:127-131.

152. Tortoli E, Kroppenstedt RM, Bartoloni A, Caroli G, Jan I, Pawlowski J, Emler S. *Mycobacterium tusciae* sp. nov. *Int J Syst Bacteriol* 1999;**49**:1839-1844.

153. Sun R, Converse PJ, Ko C, Tyagi S, Morrison NE, Bishai WR. *Mycobacterium tuberculosis* ECF sigma factor sigC is required for lethality in mice and for the conditional expression of a defined gene set. *Mol Microbiol* 2004;**52**:25-38.

154. Murray A, Winter N, Lagranderie M, Hill DF, Rauzier J, Timm J, Leclerc C, Moriarty KM, Gheorghiu M, Gicquel B. Expression of *Escherichia coli* beta-galactosidase in *mycobacterium bovis* BCG using an expression system isolated from *Mycobacterium paratuberculosis* which induced humoral and cellular immune responses. *Mol Microbiol* 1992;**6**:3331-3342.

155. Tuberculist *Institut Pasteur* 2008; http://genolist.pasteur.fr/TubercuList/.

156. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* 1999;**96**:2896-2901.

157. Mulder MA, Zappe H, Steyn LM. Mycobacterial promoters. *Tuber Lung Dis* 1997;**78**:211-223.

158. Bashyam MD, Kaushal D, Dasgupta SK, Tyagi AK. A study of mycobacterial transcriptional apparatus: Identification of novel features in promoter elements. *J Bacteriol* 1996;**178**:4847-4853.

159. Gu S, Chen J, Dobos KM, Bradbury EM, Belisle JT, Chen X. Comprehensive proteomic profiling of the membrane constituents of a *Mycobacterium tuberculosis* strain. *Mol Cell Proteomics* 2003;**2**:1284-1296.

160. de Souza GA, Wiker HG. A proteomic view of mycobacteria. *Proteomics* 2011;**11**:3118-3127.

161. Rodriguez GM, Smith I. Mechanisms of iron regulation in mycobacteria: Role in physiology and virulence. *Mol Microbiol* 2003;**47**:1485-1494.

162. Blencowe DK, Morby AP. Zn(II) metabolism in prokaryotes. *FEMS Microbiol Rev* 2003;**27**:291-311.

163. Gold B, Rodriguez GM, Marras SA, Pentecost M, Smith I. The *Mycobacterium tuberculosis* IdeR is a dual functional regulator that controls transcription of genes involved in iron acquisition, iron storage and survival in macrophages. *Mol Microbiol* 2001;**42**:851-865.

164. Yellaboina S, Ranjan S, Vindal V, Ranjan A. Comparative analysis of iron regulated genes in mycobacteria. *FEBS Lett* 2006;**580**:2567-2576.

165. Miethke M, Westers H, Blom EJ, Kuipers OP, Marahiel MA. Iron starvation triggers the stringent response and induces amino acid biosynthesis for bacillibactin production in *Bacillus subtilis*. *J Bacteriol* 2006;**188**:8655-8657.

166. Pejchal R, Ludwig ML. Cobalamin-independent methionine synthase (MetE): A face-to-face double barrel that evolved by gene duplication. *PLoS Biol* 2005;**3**:e31.

167. Koon N, Squire CJ, Baker EN. Crystal structure of LeuA from *Mycobacterium tuberculosis*, a key enzyme in leucine biosynthesis. *Proc Natl Acad Sci U S A* 2004;**101**:8295-8300.

168. Touati D. Iron and oxidative stress in bacteria. *Arch Biochem Biophys* 2000;**373**:1-6.

169. Jang HJ, Nde C, Toghrol F, Bentley WE. Microarray analysis of *Mycobacterium bovis* BCG revealed induction of iron acquisition related genes in response to hydrogen peroxide. *Environ Sci Technol* 2009;**43**:9465-9472.

170. Tang YJ, Sapra R, Joyner D, Hazen TC, Myers S, Reichmuth D, Blanch H, Keasling JD. Analysis of metabolic pathways and fluxes in a newly discovered thermophilic and ethanol-tolerant geobacillus strain. *Biotechnol Bioeng* 2009;**102**:1377-1386.

171. Jones CM, Niederweis M. Role of porins in iron uptake by *Mycobacterium smegmatis*. *J Bacteriol* 2010;**192**:6411-6417.

172. Manabe YC, Hatem CL, Kesavan AK, Durack J, Murphy JR. Both *Corynebacterium diphtheriae* DtxR(E175K) and *Mycobacterium tuberculosis* IdeR(D177K) are dominant positive repressors of IdeR-regulated genes in *M. tuberculosis*. *Infect Immun* 2005;**73**:5988-5994.

173. Kusano T, Berberich T, Tateda C, Takahashi Y. Polyamines: Essential factors for growth and survival. *Planta* 2008;**228**:367-381.

174. Tabor CW, Tabor H. Polyamines. *Annu Rev Biochem* 1984;**53**:749-790.

175. Wortham BW, Patel CN, Oliveira MA. Polyamines in bacteria: Pleiotropic effects yet specific mechanisms. *Adv Exp Med Biol* 2007;**603**:106-115.

176. Shah P, Swiatlo E. A multifaceted role for polyamines in bacterial pathogens. *Mol Microbiol* 2008;**68**:4-16.

177. Smit S. Functional consequences of the inhibition of malaria S-adenosylmethionine decarboxylase as a key regulator of polyamine and methionine metabolism. *[Dissertation]* 2010; University of Pretoria, South Africa.

178. Yoshida M, Kashiwagi K, Shigemasa A, Taniguchi S, Yamamoto K, Makinoshima H, Ishihama A, Igarashi K. A unifying model for the role of polyamines in bacterial cell growth, the polyamine modulon. *J Biol Chem* 2004;**279**:46008-46013.

179. Ojha A, Hatfull GF. The role of iron in *Mycobacterium smegmatis* biofilm formation: The exochelin siderophore is essential in limiting iron conditions for biofilm formation but not for planktonic growth. *Mol Microbiol* 2007;**66**:468-483.

180. Iyer R, Delcour AH. Complex inhibition of OmpF and OmpC bacterial porins by polyamines. *J Biol Chem* 1997;**272**:18595-18601.

181. Dela Vega AL, Delcour AH. Polyamines decrease *Escherichia coli* outer membrane permeability. *J Bacteriol* 1996;**178**:3715-3721.

182. Malen H, De Souza GA, Pathak S, Softeland T, Wiker HG. Comparison of membrane proteins of *Mycobacterium tuberculosis* H37Rv and H37Ra strains. *BMC Microbiol* 2011;**11**:18.

183. Malen H, Pathak S, Softeland T, de Souza GA, Wiker HG. Definition of novel cell envelope associated proteins in triton X-114 extracts of *Mycobacterium tuberculosis* H37Rv. *BMC Microbiol* 2010;**10**:132.

184. Mawuenyega KG, Forst CV, Dobos KM, Belisle JT, Chen J, Bradbury EM, Bradbury AR, Chen X. *Mycobacterium tuberculosis* functional network analysis by global subcellular protein profiling. *Mol Biol Cell* 2005;**16**:396-404.

185. Raghunand TR, Bishai WR. *Mycobacterium smegmatis* WhmD and its homologue *Mycobacterium tuberculosis* WhiB2 are functionally equivalent. *Microbiology* 2006;**152**:2735-2747.

186. Gomez JE, Bishai WR. WhmD is an essential mycobacterial gene required for proper septation and cell division. *Proc Natl Acad Sci U S A* 2000;**97**:8554-8559.

187. Geiman DE, Raghunand TR, Agarwal N, Bishai WR. Differential gene expression in response to exposure to antimycobacterial agents and other stress conditions among seven *Mycobacterium tuberculosis whiB*-like genes. *Antimicrob Agents Chemother* 2006;**50**:2836-2841.

188. Wang L, Lutkenhaus J. FtsK is an essential cell division protein that is localized to the septum and induced as part of the SOS response. *Mol Microbiol* 1998;**29**:731-740.

189. Wu LJ, Errington J. *Bacillus subtilis* SpoIIIE protein required for DNA segregation during asymmetric cell division. *Science* 1994;**264**:572-575.

190. Iobbi-Nivol C, Palmer T, Whitty PW, McNairn E, Boxer DH. The *mob* locus of *Escherichia coli* K12 required for molybdenum cofactor biosynthesis is expressed at very low levels. *Microbiology* 1995;**141**:1663-1671.

191. Domenech P, Reed MB, Barry CE,. Contribution of the *Mycobacterium tuberculosis* MmpL protein family to virulence and drug resistance. *Infect Immun* 2005;**73**:3492-3501.

Addendum A: Newton-Foot, M, Gey van Pittius, NC, The complex architecture of mycobacterial promoters, Tuberculosis (2012), http://dx.doi.org/10.1016/j.tube.2012.08.003. *A review of the mycobacterial transcriptional apparatus and its regulation.*