MYCOBACTERIUM TUBERCULOSIS: GENETIC AND PHENOTYPIC COMPARISON

Samantha Leigh Sampson

Dissertation presented for approval for the degree of Doctor of Philosophy in Medical Sciences (Medical Biochemistry) at the University of Stellenbosch

Promoter: Dr. R. M. Warren

Co-Promoter: Prof. P.D. van Helden

March 2002

Declaration

I, the undersigned, hereby declare that the work contained in this dissertation is my own original work, and has not, to my knowledge, previously in its entirety or in part been submitted at any university for a degree.

S. L. Sampson (student number: 9030980) (Date)

Summary

This study exploits the *Mycobacterium tuberculosis* H37Rv genome sequence data in the context of *M. tuberculosis* clinical isolates, to elucidate genetic variation, and examine the phenotypic and molecular epidemiological implications thereof.

The study was initiated by investigation of the insertion sequence IS*6110*, the primary DNA fingerprinting probe for the molecular epidemiology of tuberculosis. The transposable element is present in variable copy number and chromosomal location in clinical isolates of *M. tuberculosis* strains, giving rise to extensive genetic diversity. At the inception of this study, little was known about this element in terms of the genetic identity of its surrounding regions, its chromosomal distribution, and the mechanisms contributing to genetic diversity. These shortcomings were therefore addressed by a number of approaches.

Firstly, to establish their genetic identity and chromosomal distribution, IS*6110* insertion sites from clinical isolates of *M. tuberculosis* were cloned and sequenced. This data was examined in conjunction with available genome sequence data. The results demonstrated that the majority of insertions occurred within coding regions. Furthermore, the element was shown to have a non-random chromosomal distribution, and a number of preferential integration sites were identified. Secondly, the stability of chromosomal domains flanking IS*6110* elements was investigated by utilizing the insertion site clones as hybridization probes against clinical isolates. This allowed the identification of extensive genetic variation associated with these chromosomal domains, arising from IS*6110* transpositions, deletions and point mutations. These events were expressed in terms of a phylogenetic tree which demonstrated ongoing genome evolution associated with IS*6110*. Thirdly, to investigate the hypothesis that IS*6110*-mediated deletions occur via homologous recombination between adjacent elements, deletion junctions were mapped and sequenced in clinical isolates representing predecessor and descendant strains. While these results support the involvement of IS*6110* as a mediator of genetic deletion, they suggest either alternative mechanisms or the existence of unidentified intermediates.

The investigation of IS*6110* flanking regions identified the disruption of a number of members of the PPE gene family, leading to the second main area of investigation. The PPE gene family was newly identified as a result of the *M. tuberculosis* genome sequencing project, and its products are speculated to be of antigenic importance. However, at the commencement of this study very little data was available regarding the biological role of PPE proteins. Therefore, to explore the phenotypic implications of PPE gene disruption, various aspects of the gene family were investigated.

Firstly, phylogenetic relationships between members of the PPE family were elucidated, which suggested an evolutionary progression, and highlighted the possibility that there may be functional subdivisions within the gene family. Secondly, the extent and mechanisms of PPE gene variation were analyzed by a combination of hybridization, PCR and sequence analysis. This approach revealed extensive variation associated the gene family, although different members of the family exhibit different levels of variation. Of special interest was the discovery that long tandem repeat regions (≥69 bp) found within 3 members of the gene family demonstrate variation in the numbers of these tandem repeats. A third avenue of investigation focused on *in vitro* and *in vivo* PPE gene expression profiles. RT-PCR was utilized to demonstrate *in vitro* expression of PPE genes, while RNA:RNA *in situ* hybridization demonstrated the expression of PPE genes in human tissue samples. Intriguingly, *in situ* hybridization suggests that there is variable PPE gene expression within the human granuloma. The final approach reported here focused on the subcellular localization of one member of the PPE family, Rv1917c. A combination of cell fractionation and whole-cell antibody binding experiments suggest that the Rv1917c protein is a cell wall-associated, surface exposed molecule.

In summary, the results obtained have potential implications for the interpretation of molecular epidemiological data, support the role of IS*6110* as an agent of genome evolution, and emphasize the potential for IS*6110* to impact on strain phenotype. Investigation of the PPE family demonstrated that this gene family contributes to genetic variation, is expressed *in vitro* and *in vivo* and that at least one protein encoded by the gene family is cell wall associated. Together, the results obtained support the hypothesis that selected members of the PPE gene family may encode products involved in antigenic variation.

Opsomming

Dié studie maak gebruik van die *Mycobacterium tuberculosis* H37Rv genoom volgorde data in die konteks van *M. tuberculosis* kliniese isolate, om genetiese variasie toe te lig en die fenotipiese en molekulêre epidemiologiese implikasies daarvan te ondersoek.

Die studie het 'n aanvang geneem deur die ondersoek van die inset-volgorde *IS6110*, wat die primêre DNS vingerafdruk pylfragment vir die molekulêre epidemiologie van tuberkulose is. Hierdie transponerende element is in wisselende kopiegetal en chromosomale posisies teenwoordig in kliniese isolate van *M. tuberculosis* stamme, en gee so oorsprong aan omvangryke genetiese afwisseling. Met die aanvang van hierdie studie was min bekend omtrent hierdie element betreffende die genetiese identiteit van die areas wat die insetsels omring, die chromosomale distribusie van insetsels, asook die meganismes wat bydra tot genetiese afwisseling. Hierdie gebreke is dus deur 'n aantal benaderings aangespreek.

Eerstens is IS*6110* insettingsetels van kliniese *M. tuberculosis* isolate gekloneer en hul nukleotiedvolgorde bepaal om sodoende hul genetiese identiteit en chromosomale verspreiding vas te stel. Hierdie data is in oorleg met beskikbare genomiese volgorde data geanaliseer. Die resultate het gewys dat die meerderheid van insetsels binne koderende gebiede plaasgevind het. Verder is gewys dat hierdie element nie na willekeur deur die chromosoom versprei is nie, en 'n aantal gebiede waar insetting by voorkeur plaasvind, is geïdentifiseer. Tweedens is die stabiliteit van die chromosomale gebiede wat IS*6110* elemente flankeer ondersoek deur die insettingsetel klone as pylfragmente te gebruik in hibridisasie van kliniese isolate. Dit het die identifisering van omvangryke genetiese afwisseling binne hierdie chromosomale gebiede, wat ontstaan deur IS*6110* transposisies, delesies en puntmutasies, tot gevolg gehad. Hierdie afwisselings is uitgedruk as 'n filogenetiese boom waarin die voortdurende genomiese evolusie wat geassosieer word met IS*6110* gewys word. Derdens, om die teorie dat IS*6110*-gedrewe delesies deur middel van homoloë rekombinasie tussen naasliggende elemente plaasvind te ondersoek, is die grense van delesies gekarteer en die nukleotiedvolgorde daarvan bepaal in kliniese isolate wat voorganger- en afstammelingstamme verteenwoordig. Alhoewel die resultate die betrokkenheid van IS*6110* as 'n bemiddelaar van genetiese delesie ondersteun, stel dit ook die bestaan van of alternatiewe meganismes of van onbekende intermediêre vorme voor.

Ondersoek van die IS*6110*-flankerende gebiede het gelei tot die ontdekking van ontwrigting van 'n aantal gene wat behoort tot die PPE geenfamilie, en het so gelei tot die tweede hoof ondersoek tema. Die PPE geenfamilie is ontdek as gevolg van die *M.*

tuberculosis genoomprojek, en dit word gespekuleer dat die produkte van hierdie gene van antigeniese belang mag wees. Daar was egter met die aanvang van hierdie studie baie min data beskikbaar omtrent die biologiese rol van die PPE proteïene. Om die fenotipiese implikasies van ontwrigting van PPE gene te ondersoek is daar dus ondersoek ingestel na verskeie aspekte van hierdie geenfamilie.

Eerstens is filogenetiese verwantskappe tussen lede van die PPE familie bepaal, wat gedui het op 'n evolusionêre progressie en wat ook aangedui het dat daar moontlik funksionele onderverdelings binne hierdie geenfamilie mag bestaan. Tweedens is die omvang en meganismes van PPE geenvariasie geanaliseer deur 'n kombinasie van hibridisasie, PKR en nukleotiedvolgorde analise. Hierdie benadering het omvangryke afwisseling binne hierdie geenfamilie getoon, alhoewel verskillende lede van die familie verskillende vlakke van afwisseling demonstreer. Wat veral interessant was, was die ontdekking dat lang tandem herhalingsvolgordes (≥69 bp) wat in 3 lede van hierdie geenfamilie voorkom, variasie toon in die getalle van hierdie tandem herhalingsvolgordes. 'n Derde been van ondersoek het gefokus op *in vitro* en *in vivo* PPE geen uitdrukkingsprofiele. RT-PKR is gebruik om te toon dat PPE gene *in vitro* uitgedruk word, terwyl RNA:RNA *in situ* hibridisasie getoon het dat PPE gene ook in menslike weefsel uitgedruk word. Interessant genoeg dui *in situ* hibridisasie daarop dat daar wisselende PPE geen uitdrukking binne die menslike granuloom voorkom. Die laaste benadering wat hier gerapporteer word fokus op die sub-sellulêre lokalisering van een lid van die PPE familie, Rv1917c. 'n Kombinasie van selfraksionering en heel-sel antiliggaam-bindingseksperimente dui daarop dat Rv1917c 'n selwand-geassosieerde molekuul is wat aan die oppervlak blootgestel word.

Ter opsomming het die resultate wat bereik is potensiële implikasies vir die interpretasie van molekulêr-epidemiologiese data, dit ondersteun die rol van IS*6110* as 'n bemiddelaar van genoom evolusie en beklemtoon die potensiaal vir IS*6110* om 'n invloed te hê op die fenotipe van die stam. Ondersoek van die PPE familie het getoon dat hierdie geenfamilie bydra tot genetiese afwisseling, dat dit uitgedruk word beide *in vitro* en *in vivo* en dat ten minste een lid van hierdie geenfamilie geassossieer word met die selwand. Tesame ondersteun hierdie resultate die teorie dat geselekteerde lede van die PPE geenfamilie wel produkte enkodeer wat betrokke is by antigeniese variasie.

Publications and Presentations

Portions of this thesis have been published as:

1) Disruption of coding regions by IS*6110* insertion in *Mycobacterium tuberculosis* clinical isolates. (1999) Sampson, S.L., Warren, R.M., Richardson, M., Van Der Spuy, G.D., Van Helden, P.D. *Tubercle and Lung Disease* 79(6):349-359. (Chapter 2)
2) Mapping of IS*6110* flanking regions in clinical isolates of *Mycobacterium tuberculosis* demonstrates genome plasticity. (2000) Warren, R.M., Sampson, S.L., Richardson, M., Van Der Spuy, G.D., Lombard, C.J., Victor, T.C., Van Helden, P.D. *Molecular Microbiology* 37(6):1405-1416. (Chapter 3)
3) IS*6110* insertions in *M. tuberculosis*: predominantly into coding regions. (2001) Sampson, S.L., Warren, R.M., Richardson, M., Van Der Spuy, G.D., Van Helden, P.D. *Journal of Clinical Microbiology* (letter) 39:3423-3424. (Addendum to Chapter 2)
4) Expression, characterization and subcellular localization of the *Mycobacterium tuberculosis* PPE gene Rv1917c. S. Sampson, P. Lukey, R. Warren, P. van Helden, M. Richardson, M. Everett 81:305-17. (Chapter 7)

The following manuscripts are in preparation:

1) IS*6110*-mediated deletion polymorphism in the DR region of clinical isolates of *Mycobacterium tuberculosis*. Sampson, S.L., Warren, R.M., Victor, T.C., Richardson, M., Jordaan, A.M., Van Der Spuy, G.D., Van Helden, P.D. (Chapter 4)
1) The PPE gene family of *Mycobacterium tuberculosis*: extent and mechanisms of variation in clinical isolates. Sampson, S.L., Warren, R.M., Richardson, M., Van Der Spuy, G.D., Van Helden, P.D. (Chapter 5)
2) Differential expression of the PPE gene family of *Mycobacterium tuberculosis*. Sampson, S.L., Warren, R.M., Moses, L., Fenhalls, G., Stevens, L., Van Helden, P.D. (Chapter 6)

Oral presentations of work contained in this thesis have been made by the author on the following occassions:

1) Action TB conference, Medical Research Council, Cape Town, South Africa, 1998
2) Acid Fast Club Meeting, St. Georges Hospital, London, United Kingdom, 1999
3) Astra Zeneca Joint UCT, UWC, US and MRC 2nd Medical Research Day, University of the Western Cape, Cape Town, South Africa, 1999
3) Astra Zeneca Joint UCT, UWC, US and MRC 3rd Medical Research Day, University of Stellenbosch Faculty of Health Sciences, Cape Town, South Africa, 2000

4) 44th Academic Year Day, University of Stellenbosch Faculty of Health Sciences, Cape Town, South Africa, 2000

6) Gene Quantification Training Course, Integrated DNA Technologies, Coralville, IA, USA, 2001

7) Various departmental lab talks both at the Department of Medical Biochemistry, University of Stellenbosch Faculty of Health Sciences, Cape Town, South Africa, and at the Glaxo SmithKline Medicines Research Centre, Stevenage, United Kingdom. (1998-2001)

The author has contributed to the following publications (not presented here):

1) Warren, R., Richardson, M., Sampson, S., Hauman, H.J., Beyers, N., Donald, P., van Helden, P.D. (1996) Genotyping of *M. tuberculosis* with additional markers enhances accuracy in epidemiological studies. *Journal of Clinical Microbiology*. 34:2219-2224.

2) Van Helden, P., Warren, R., Richardson, M., Sampson, S., Hauman, J., Beyers, N., Donald, P., van Rie, A., Classen, C. (1998) Molecular epidemiology of TB: lessons from a high-incidence community. In: *Clinical Mycobacteriology*, M. Casal (ed), 23-29. (Book Chapter)

3) Warren, R., Richardson, M., van der Spuy, G., Victor, T., Sampson, S., Beyers, N., van Helden, P. (1999) DNA fingerprinting and molecular epidemiology of tuberculosis: use and interpretation in an epidemic setting. *Electrophoresis* 20:1807-1812.

4) Van Rie, A., Warren, R.M., Beyers, N., Gie, R.P., Classen, C.N., Richardson, M., Sampson, S.L., Victor, T.C., van Helden, P.D. (1999) Transmission of a multidrug-resistant *Mycobacterium tuberculosis* strain resembling “strain W” among non-institutional, HIV seronegative patients. *Journal of Infectious Diseases*. 180:1608-1615.

5) Warren, R.M., Richardson, M.. Sampson, S.L., van der Spuy. G.D., Bourn, W., Hauman, J.H., Heersma, H., Hide, W., Beyers, N., van Helden, P.D. (2001) Molecular evolution of *Mycobacterium tuberculosis*: Phylogenetic reconstruction of clonal expansion. *Tuberculosis*. 81:291-302.

6) Upton, A., Mushtaq, A., Victor, T., Sampson, S., Smith, D.-M., van Helden, P., Sim, E. (2001) Arylamine N-acetyltransferase of *Mycobacterium tuberculosis* is a polymorphic enzyme and a site of Isoniazid Metabolism. *Molecular Microbiology*. 42:309-317.

Acknowledgements

I take this opportunity to extend my deepest gratitude to the following people who have contributed in various ways to the end product:

My family, for love and laughs: especially my mom, who has shown us all what strength is; my dad, for the yellow gremlin ears and unwavering belief in me; and Ryan, a brother in a million.

To special friends: Debbie, Grant, Lee-Anne, Nora, Rentia, Tanja, and others for varying doses of tissues, cake, sympathy, bullying, and many, many laughs. Francois, for sharing the "plateaux of sanity in the mountains of madness", I would never have made it without you!

Colleagues in the Department of Medical Biochemistry, Tygerberg: Especially Dr. Rob Warren, my promoter, whose insight, dedication and understanding continue to be an example. Thanks also to Prof. Paul van Helden, my department head and co-promoter, for support and encouragement; Madalene Richardson, for her solid support and friendship; Annemie Jordaan for generating the spoligotype data; Gian van der Spuy for answering my many database queries; Lorraine Moses, Gael Fenhalls, and Liesel Stevens-Muller for their contribution of the RNA:RNA *in situ* hybridization work.

Members of the Immunopathology and Cellular Biochemistry Units at Glaxo Smithkline, Stevenage, UK, for making my stay there a pleasant and productive one, especially Dr. Martin Everett, Dr. Ken Duncan, Dr. Selwyn Quan, Dr. Ruth McAdam and Dr. Pauline Lukey.

Also, a special thanks to all those who have read and re-read multiple versions of this thesis, providing constructive criticism and helpful discussions: Dr. Rob Warren, Dr. Martin Everett, Prof. Paul van Helden, Dr. Nora Carroll, Ms. Madalene Richardson, Dr. John Haumann, Dr. Tommie Victor, Mr. Nico Gey van Pittius and anonymous reviewers of chapters 2, 3 and 7. Thanks too, Hendrieka, Rentia, Manda, Hanlie and Tanja for help with Afrikaans and German translations.

This project was supported financially by the Glaxo Smithkline Action TB Initiative.

Table of contents

Detailed contents

List Of Figures

Chapter 1. Introduction

Chapter 2. IS*6110* insertions: chromosomal location

Chapter 3. IS*6110* and genome evolution

Chapter 4. IS*6110*-mediated deletion polymorphism in the DR region of clinical isolates of *Mycobacterium tuberculosis*

Chapter 5. The PPE gene family: mechanisms and extent of variation

Chapter 6. Analysis of *in vitro* and *in vivo* PPE gene expression

Chapter 7. Expression, Characterization and Subcellular Localization of the PPE gene Rv1917c

Appendix

List Of Tables

Chapter 2. IS*6110* insertions: chromosomal location

Chapter 3. IS*6110* and genome evolution

Chapter 4. IS*6110*-mediated deletion polymorphism in the DR region of clinical isolates of *Mycobacterium tuberculosis*

Chapter 5. The PPE gene family: mechanisms and extent of variation

Chapter 6. Analysis of *in vitro* and *in vivo* PPE gene expression

Chapter 7. Expression, Characterization and Subcellular Localization of the PPE Gene Rv1917c

Abbreviations Used In The Text

µg	microgram
µl	microlitre
/	per
°C	degrees Celsius
^{32}P	phosphorous radioisotope P-32
A	Adenine
aa	amino acid
ADC	Albumin-dextrose-catalase mycobacterial growth supplement
AFB	Acid-fast bacilli
AP	Alkaline phosphatase
APC	Antigen presenting cell
APS	ammonium peroxidosulphate $((NH_4)_2S_2O_8)$
ATP	adenosine triphosphate
BAC	Bacterial artificial chromosome
BCIP	5-bromo-4-chloro-3-indolyl-phosphate)
BCG	Bacille Calmette-Guérin
BLASTN	Basic Local Alignment Search Tool (for oligonucleotides)
BLASTP	Basic Local Alignment Search Tool (for oligopeptides)
BLASTX	Basic Local Alignment Search Tool (for oligopeptides and oligonucleotides)
bp	base pair
BSA	Bovine serum albumin
C	Cytosine
$CaCl_2$	calcium chloride
CDC	Center for Disease Control
cm	centimetre
CMI	cell-mediated immunity
ConA	Concanavelin A
dATP	deoxyadenosine triphosphate
dCTP	deoxycytosine triphosphate
ddATP	dideoxyadenosine triphosphate
ddCTP	dideoxycytosine triphosphate
ddGTP	dideoxyguanosine triphosphate
ddNTP	dideoxynucleoside triphosphate
ddTTP	dideoxythymidine triphosphate
dGTP	deoxyguanosine triphosphate
DEPC	diethylpyrocarbonate
DIG	Digoxigenin
DMF	dimethylformamide (H_3H_7NO)
DMSO	dimethysulfoxide
DNA	deoxyribonucleic acid
DOTS	directly observed treatment, short course
dNTP	deoxynucleoside triphosphate
DR	Direct Repeat
DTH	delayed type hypersensitivity
dTTP	deoxythymidine triphosphate
ECL	enhanced chemiluminescence
EDTA	ethylene diamine tetra-acetic acid
ELISA	Enzyme-linked immunosorbent assay

f	faraday
FACS	Fluorescent antibody cell sorter
ftp	file transfer protocols
G	Guanine
g	gram
GAR-PE	goat anti-rabbit phycoerythrin conjugated Immunoglobulin G
GFP	green fluorescent protein
h	hour
HIV	human immunodeficiency virus
H_2O	water
HCl	hydrochloric acid
HRP	Horseradish peroxidase
http	hyper text transfer protocol
IFN-γ	Interferon gamma
IgG	Immunoglobulin G
IL	Interleukin
ipl	preferential integration locus
IPTG	isopropyl-ß-D–thiogalactopyranoside
IR	Inverted repeat
IS	Insertion sequence
IS-3'	IS*6110* DNA fingerprinting probe, 3' side
IS-5'	IS*6110* DNA fingerprinting probe, 5' side
IS*6110*	Insertion sequence *6110*
ISH	*in situ* hybridization
ISL	IS*6110* insertion locus clone
KAc	potassium acetate (CH_3COOK)
kb	kilobase
KCl	potassium chloride
kDa	kilodalton
kV	kilovolt
l	liter
LB	Luria Bertani medium
LBGTW	Luria Bertani medium supplemented with 0.5 % glucose and 0.05 % Tween-80
LJ	Lowenstein Jensen
LTR	Long tandem repeat
M	molar
M63	M63 medium
MAC	*Mycobacterium avium* complex
MADCTW	Middlebrooks 7H9 mycobacterial culture medium with 10 % ADC enrichment and 0.05 % Tween-80
Mb	Megabase
mg	milligram
$MgCl_2$	magnesium chloride
$MgSO_4$	magnesium sulfate
MHC	Major histocompatability complex
min	minute
ml	milliliter
mM	millimolar
mm	millimetre
MPTR	Major Polymorphic Tandem Repeat
MRC	Medical Research Council

mRNA	messenger ribonucleic acid
MW	molecular weight
N/A	not applicable
$Na_2HPO_4.2H_2O$	disodium hydrogen orthophosphate dihydrate (Sörensen's salt)
NaAc	sodium acetate (CH_3COONa)
NaCl	sodium chloride
$NaH_2PO_4.2H_2O$	sodium dihydrogen phosphate dihydrate
NaI	sodium iodide
NaOH	sodium hydroxide
NBT	nitroblue-tetrazolium
NCBI	National Center for Biotechnology Information
ng	nanogram
NH_4Ac	ammonium acetate (CH_3COONH_4)
nm	nanometre
OADC	Oleic acid-albumin-dextrose-catalase mycobacterial growth supplement
OD	optical density
ORF	open reading frame
PAGE	polyacrylamide gel electrophoresis
PAUP* 4.0	Phylogenetic analysis using parsimony (* and other methods)
PBS	phosphate-buffered saline
PBS-T	phosphate buffered saline with 0.1% Tween-20
PCR	Polymerase chain reaction
PE	Proline glutamic acid
PPD	purified protein derivative
PPE	Proline proline glutamic acid
PGRS	polymorphic guanine-cytosine rich sequence
pmol	picomol
PMSF	phenylmethysulfonyl fluoride
POD	Horseradish peroxidase
PVP	polyvinyl pyrollidine
RFLP	Restriction fragment length polymorphism
RNA	ribonucleic acid
RNAse	ribonuclease
RNasin	ribonuclease inhibitor
rpm	revolutions per minute
rRNA	ribosomal ribonucleic acid
RT-PCR	Reverse transcription polymerase chain reaction
SAP	shrimp alkaline phosphatase
SDS	sodium dodecyl sulphate
SSC	sodium choride-sodium citrate buffer
ssDNA	single-stranded deoxyribonucleic acid
SSM	slipped strand mispairing
SSPE	sodium chloride-sodium phosphate-ethylene diamine tetracetic acid buffer
STE/TEN	sodium chloride-tris-ethylene diamine tetracetic acid buffer
T	Thymine
T4 PNK	T4 polynucleotide kinase
TAE	tris-acetic acid-ethylene diamine tetracetic acid buffer
TBE	tris-boric acid-ethylene diamine tetracetic acid buffer
T_c	cytotoxic T lymphocytes
TE	tris-ethylene diamine tetracetic acid buffer

TEMED	N, N, N’, N’, tetramethylene diamine
T_h	helper T lymphocytes
TIGR	The Institute for Genome Research
Tn	Transposon
TNF-α	Tumor necrosis factor alpha
Tris	Tris(hydroxymethyl)aminomethane
tRNA	transfer ribonucleic acid
TTE	tris-taurine-ethylene diamine tetracetic acid buffer
2XTY	tryptone-yeast medium
U	Units
UCT	University of Cape Town
UK	United Kingdom
UPGMA	Unweighted pair group method using arithmetic averages
URL	Uniform Resource Location
US	University of Stellenbosch
USA	United States of America
UWC	University of the Western Cape
UV	Ultraviolet
V	Volts
vs	versus
Vti	vertical titanium
v/v	volume per volume
wt	wild type
WHO	World Health Organization
w/v	weight per volume
www	world wide web
X-gal	5-bromo-4-chloro-3-indolyl-ß-D-galactoside
ZN	Ziehl-Neelsen

Amino Acid Abbreviations

A	**Ala**	**Alanine**	**L**	**Leu**	**Leucine**
R	**Arg**	**Arginine**	**K**	**Lys**	**Lysine**
N	**Asn**	**Asparagine**	**M**	**Met**	**Methionine**
D	**Asp**	**Aspartic acid**	**F**	**Phe**	**Phenylalanine**
C	**Cys**	**Cysteine**	**P**	**Pro**	**Proline**
Q	**Gln**	**Glutamine**	**S**	**Ser**	**Serine**
E	**Glu**	**Glutamic acid**	**T**	**Thr**	**Threonine**
G	**Gly**	**Glycine**	**W**	**Trp**	**Tryptophan**
H	**His**	**Histidine**	**Y**	**Tyr**	**Tyrosine**
I	**Ile**	**Isoleucine**	**V**	**Val**	**Valine**

It is good to have an end to journey towards,
but it is the journey that matters, in the end

Ursula Le Guin

CHAPTER 1

INTRODUCTION

1.1 Study Overview

It is well known that *Mycobacterium tuberculosis* is the etiologic agent of the disease tuberculosis. Despite the widespread use of the *Mycobacterium bovis* Bacille Calmette-Guérin (BCG) vaccine and relatively effective chemotherapy, the disease continues to claim approximately 2 million lives worldwide per annum (Dye *et al.*, 1999). Factors contributing to the continuing disease burden include the variable efficacy of BCG vaccination (Colditz *et al.*, 1994; Fine, 1995), the rising incidence of human immunodeficiency virus (HIV) infection (Barnes *et al.*, 1991; Schulzer *et al.*, 1992; Hass and Des Prez, 1994), widespread drug resistance (WHO, 1997) and treatment compliance failure (Brudney and Dobkin, 1991; Chaulet *et al.*, 1995), which is narrowly linked to the lengthy duration of chemotherapy regimens (Webb and Davies, 1999). There is an urgent need to introduce new chemotherapeutic agents, reduce the length of treatment regimens, and to develop a more effective vaccine. An improved understanding of the pathogen, the host response to infection and the host-pathogen interaction is essential for these endeavours. The recent completion of the genome sequence of the reference strain *M. tuberculosis* H37Rv (Cole *et al.*, 1998) has added great impetus to the field of tuberculosis research, particularly with regards to the pathogen itself.

In this study, the *M. tuberculosis* H37Rv genome sequence data is exploited in the context of clinical isolates, with the broad aim of elucidating the phenotypic significance of observed genome polymorphism. Initial investigation focused on the identity and stability of genome regions surrounding copies of the transposable element IS*6110*. This lead to the identification of gene disruption and polymorphisms associated with a newly-described gene family, namely the PPE family. This gene family was investigated further to establish the possible phenotypic significance of these observations.

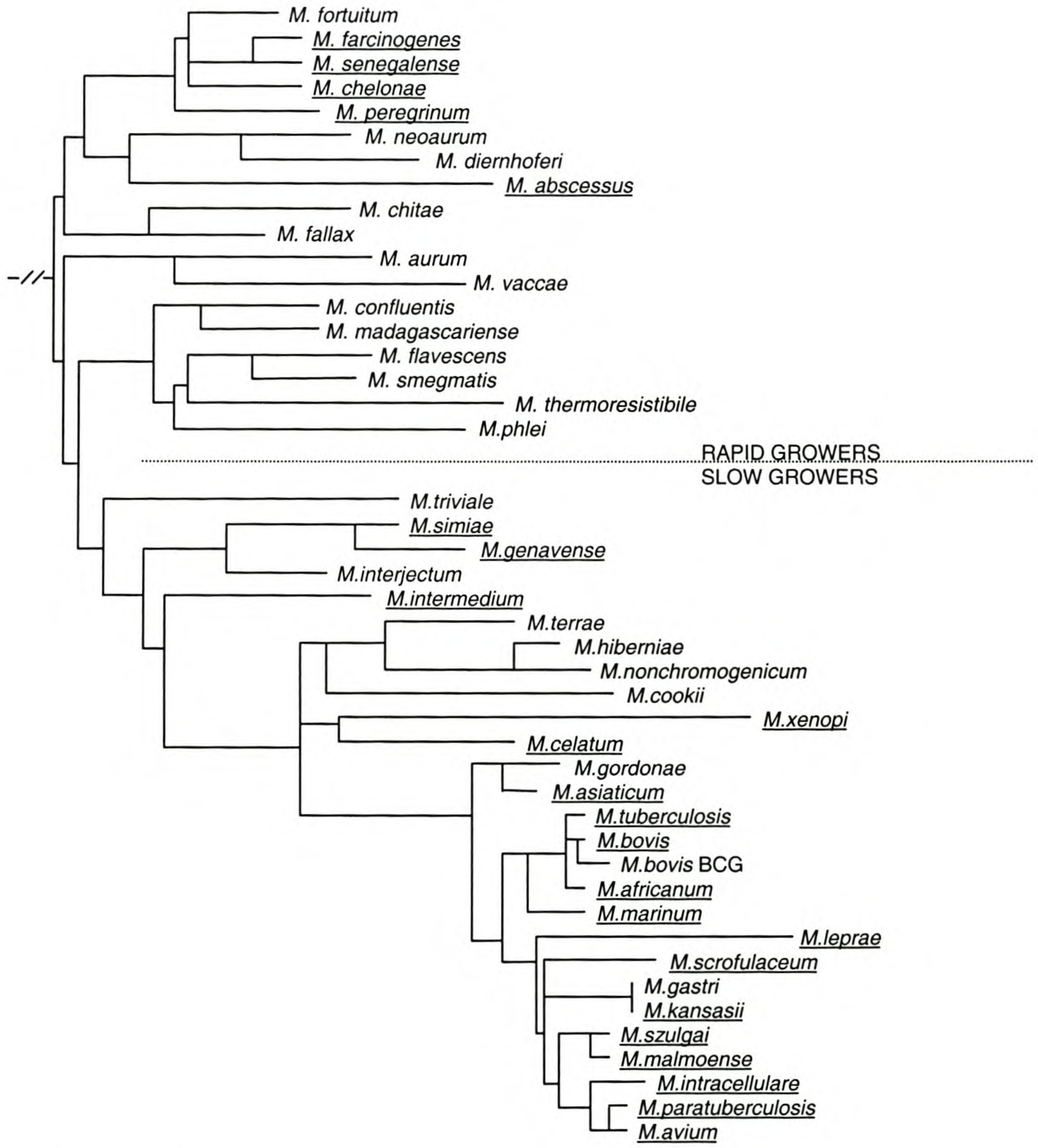
1.2 The causative agent, *Mycobacterium tuberculosis*

The causative agent of tuberculosis is *M. tuberculosis*, a member of the genus *Mycobacterium*. Over 70 mycobacterial species have been identified, and these are commonly grouped according to the time required for colonies to form on solid medium (Fig. 1.1) (Shinnick and Good, 1994). Fast growers form colonies within 7 days, and are usually (with some exceptions) non-pathogenic for humans or animals (Shinnick and Good, 1994). Slow growers require more than 7 days to form colonies on solid medium, and include many of the pathogenic mycobacterial species (Shinnick and Good, 1994).

In terms of human disease, the most well known mycobacteria are the slow growing members of the *M. tuberculosis* and *Mycobacterium avium* complexes, and *Mycobacterium leprae* (Fig. 1.1). Historically, members of the *M. tuberculosis* complex, namely *M. tuberculosis*, *Mycobacterium bovis, Mycobacterium africanum* and *Mycobacterium microti*, are responsible for the majority of mycobacterial infections in humans (Shinnick and Good, 1994). However, members of the *M. avium* complex (MAC), encompassing *M. avium*, *Mycobacterium paratuberculosis* and *Mycobacterium intracellulare*, frequently infect immunocompromised individuals (Havlir, 1994), and there is consequently a rising incidence of MAC infections associated with the HIV epidemic. The chronic neurological disease leprosy is caused by infection with *M. leprae*, with over half a million new cases reported per annum (WHO, 1998a).

M. tuberculosis was first isolated by Robert Koch in 1882 (Koch, 1882), and is a rod-shaped, gram-positive, acid-fast, facultative intracellular pathogen, with a complex, lipid-rich cell wall. Acid-fastness, or the ability to retain a basic fuchsin dye when treated with an acidified alcohol solution, is the basis of the commonly-used diagnostic procedure known as Ziehl-Neelsen (ZN) staining (Smithwick, 1976, see section 1.3.2).

Figure 1.1 Phylogenetic tree of members of the genus *Mycobacterium*. A phylogenetic reconstruction based on analysis of 16S rRNA genes of mycobacteria. Pathogenic species are underlined, and the division between fast and slow growing mycobacteria is indicated. (Reproduced with kind permission of Mr. Nico Gey van Pittius, Dept. of Medical Biochemistry, University of Stellenbosch Faculty of Health Sciences; originally adapted from Shinnick and Good, 1994)



The slow growth of *M. tuberculosis*, the tendency of mycobacteria to clump in culture and the resistance of mycobacterial cells to lysis has hindered investigation of the pathogen (Parish and Stoker, 1999a). Until recently, the shortage of suitable genetic tools has also hampered research efforts. However, recent progress in this field includes the development of a number of transposon mutagenesis systems, which have been applied to generate transposon mutant libraries (Guilhot *et al.*, 1994; McAdam *et al.*, 1995; Bardarov *et al.*, 1997; Pelicic *et al.*, 1997). In addition, improved techniques for homologous recombination have facilitated targeted gene replacement (Reyrat *et al.*, 1995; Azad *et al.*, 1996; Balasubramanian *et al.*, 1996). Another significant advance in the fight against the disease was the completion of the genome sequence of the reference strain *M. tuberculosis* H37Rv (Cole *et al.*, 1998).

A common characteristic of mycobacteria is the high guanine and cytosine (G + C) content of their genomes, and accordingly, the 4.4 Mb circular chromosome of *M. tuberculosis* H37Rv demonstrated an average G + C content of 66% (Cole *et al.*, 1998). Not surprisingly, in view of the complex and lipid-rich cell wall of the pathogen, analysis of the genome sequence also revealed the presence of an unusually large number of genes encoding components of fatty acid metabolism pathways (Cole *et al.*, 1998). The abundance of fatty acid degradation enzymes is thought to reflect the pathogen's ability to utilize host lipids as an energy source (Cole *et al.*, 1998; Bishai, 2000). Of relevance to this study was (i) the identification of 56 loci with homology to insertion sequences, including 16 copies of IS*6110*, a commonly used DNA fingerprinting probe (Cole *et al.*, 1998; Gordon *et al.*, 1999b), and (ii) the unexpected identification of two large multigene families which occupy approximately 10% of the genome. These are the PE and PPE gene families, named for the characteristic proline-glutamic acid and proline-proline-glutamic acid motifs found at the respective N-termini of their encoded proteins, and are speculated to be of immunological importance (Cole *et al.*, 1998).

1.3 Tuberculosis: History, Disease Pathology and Diagnosis

1.3.1 History of tuberculosis

It is theorized that the progenitor of the *Mycobacterium tuberculosis* complex arose from a soil bacterium, which spread to animals, and later to humans following the domestication of cattle (Daniel *et al.*, 1994; Stead, 1997). Tuberculosis remained a rare and endemic disease in man for many centuries (Stead *et al.*, 1995), until the establishment of large urban settlements and the advent of the Industrial Revolution provided the conditions ideal for the epidemic spread of the disease (Bates and Stead, 1993).

Following the identification of *M. tuberculosis* in 1882 (Koch, 1882), work began on the development of a tuberculosis vaccine. This culminated in the isolation of an attenuated strain of *M. bovis*, Bacille Calmette-Guérin (BCG), which was first administered to humans in 1921 (Weill-Hallé and Turpin, 1925). The BCG vaccine continues to be widely used today, despite concerns over its variable efficacy (Tripathy, 1979; Rodrigues and Smith, 1990; Colditz *et al.*, 1994; Fine, 1995; Smith *et al.*, 2000).

The first anti-tuberculosis drug, streptomycin, was introduced in the 1940's (Pfuetze *et al.*, 1955). This was followed shortly afterwards by other commonly used frontline chemotherapies, namely isoniazid, rifampin, pyrazinamide and ethambutol (Webb and Davies, 1999). Standard tuberculosis chemotherapy consists of an intensive phase and continuation phase of treatment (Webb and Davies, 1999), and usually lasts for a minimum of 6 months. World Health Organization guidelines recommend adherence to the directly observed treatment short-course (DOTS) strategy (Enarson, 1991). Today, the initial success of chemotherapeutic regimes is seriously threatened by emerging drug resistance (WHO, 1997), coupled with the fact that no new anti-tuberculosis agent has been introduced since pyrazinamide in the 1970's (Dickinson, 1977). The HIV epidemic is further cause for concern, and is a major contributing factor to the current resurgence of the tuberculosis epidemic (Barnes *et al.*, 1991; Schulzer *et al.*, 1992; Hass and Des Prez, 1994).

1.3.2 Disease Pathology and Diagnosis

Infection with *M. tuberculosis* occurs by the inhalation of aerosol droplets containing the bacilli (Wells, 1955; Riley *et al.*, 1959). The bacteria enter the lung, where they are phagocytosed, and often killed, by alveolar macrophages. However, the bacterium is sometimes able to survive and replicate within the macrophage, leading to the initiation of the cell-mediated immune response (CMI) (Chan and Kaufman, 1994; Barnes *et al.*, 1994; Schluger and Rom, 1998). Investigation of disease progression in a variety of animal models and in human subjects has established that the control of infection is largely dependent on the T lymphocyte response (Barnes *et al.*, 1994; Zhang *et al.*, 1995). Of particular importance are those of the $CD4^+$ subgroup (Mossman and Coffman, 1989), otherwise known as T helper (T_h) cells, although the cytotoxic T cell (T_c; $CD8^+$) subgroup also plays a role (Flynn *et al.*, 1992; Chan and Kaufman, 1994; Barnes *et al.*, 1994; Murray, 1999). Although a dominantly T_h1-type response is thought to favor a successful disease outcome, the relative contribution and roles of the various cytokines secreted by T_h1 and T_h2 lymphocytes is a complex (and often controversial) subject (Chan and Kaufman, 1994; Barnes *et al.*, 1994; Schluger and Rom, 1998, Wangoo *et al.*, 2001). Successful containment of mycobacterial infection is mediated by interferon-gamma (IFN-γ) secretion by T_h type I (T_h1) cells (Newport *et al.*, 1996; Condos *et al.* 1997; Sugawara *et al.*, 1998), although a finely regulated network of other cytokines is also involved, including interleukin-12 (IL-12) (Flynn and Bloom, 1996; Altare *et al.*, 1998) and tumor necrosis factor-alpha (TNF-α) (Smith *et al.*, 1997; Mohan *et al.*, 2001).

The secretion of an assortment of cytokines and chemokines leads to the activation and recruitment of various immune cells to the site of infection, resulting in the establishment of a granulomatous lesion, where bacteria may persist for many years (Robertson, 1933). The granuloma is an organised structure in which a network of immune cells are closely positioned in order to effectively interact and perform their respective functions (Saunders and Cooper, 2000). This creates an environment that is hostile to bacterial growth, and acts as a physical barrier preventing bacterial dissemination (North and Izzo, 1993; Doenhoff, 1997; Saunders *et al.*, 1999). The granuloma classically consists of a central region rich in macrophages, some of which may contain mycobacteria, surrounded by a dense margin consisting predominantly of lymphocytes (Canetti, 1955; Dannenberg and Rook, 1994; Saunders and Cooper, 2000). However, this is a dynamic structure, and after its initial formation, may resolve by calcification and fibrin deposition (Canetti, 1955; Dannenberg and Rook, 1994). Alternatively, the central region of the granuloma may undergo caseous necrosis, characterized by cellular breakdown and low bacterial numbers (Dannenberg and Rook, 1994; Saunders and Cooper, 2000). Following this, liquefaction of the granuloma center may occur, providing an ideal environment for bacterial growth (Canetti, 1955; Dannenberg and Rook, 1994). The liquefied granuloma may eventually rupture and expel its bacteria-laden contents into nearby airways, leading to disease transmission, and the cavitatory lesions characteristic of advanced disease (Canetti, 1955; Dannenberg and Rook, 1994).

M. tuberculosis infection mainly causes pulmonary disease, although bacterial dissemination can cause other sites to be affected (Hopewell, 1994). Clinical features of pulmonary tuberculosis include coughing, fever, weight loss, tiredness and in advanced disease, hemoptysis (coughing blood) (Hopewell, 1994). Many of these symptoms are related to elevated levels of host cytokines such as the pro-inflammatory cytokine TNF-α (Seah *et al.*, 2000; van Crevel, 2000; Wangoo *et al.*, 2001), rather than bacterial toxicity.

Diagnosis of pulmonary tuberculosis is frequently by radiographic examination (Hopewell, 1994). Another diagnostic tool is the tuberculin skin test, based on an observation first made by Koch (1891), of skin reactivity to culture filtrate proteins of *M. tuberculosis*. This involves measuring the delayed-type hypersensitivity (DTH) reaction following subcutaneous injection of a purified protein derivative (PPD) from tuberculosis culture filtrates (Seibert and Glenn, 1941). This technique is prone both to false-negative (Comstock *et al.*, 1981) and false-positive (Howard *et al.*, 1970) results, which can be due to anergy or exposure to environmental mycobacteria, respectively, and cannot distinguish between active and latent disease. In many countries, diagnosis is primarily by Ziehl-Neelsen (ZN) staining of sputum smears (Smithwick, 1976), although this technique is influenced by quality of the patient sample, can only detect active cavitatory pulmonary disease, also detects non-tuberculous and non-viable mycobacteria and is prone to laboratory error (WHO, 1998b). Current research efforts are focused on the development of improved immuno-diagnostic and molecular diagnostic tools, with a higher specificity and sensitivity (Bothamley *et al.*, 1999; Eisenach, 1999; Lalvani *et al.*, 2001). Molecular diagnostic tools rely largely on the sensitive polymerase chain reaction (PCR) (Noordhoek *et al.*, 1994), which can be adapted to distinguish between viable and non-viable bacilli (Jou *et al.*, 1997; Hellyer *et al.*, 1999).

1.4 The transposable element, IS*6110*

1.4.1 Insertion sequences

Insertion sequences are part of a family of mobile genetic elements known as transposable elements. Transposable elements were first discovered in the 1940's (McClintock, 1948) and have since been identified in a wide range of procaryotic and eucaryotic organisms. Over 500 diverse bacterial insertion sequences have been identified, and are distributed across many species (Galas and Chandler, 1989; Mahillon and Chandler, 1998). Common features are their small size (< 2,5 kb), self-encoded transposase function and terminal inverted repeats which range between 10 and 40 bp in length (Galas and Chandler, 1989). The majority of insertion sequences generate short (2-14 bp) directly repeated sequences of the target on insertion, the length of which is characteristic for a given element (Galas and Chandler, 1989).

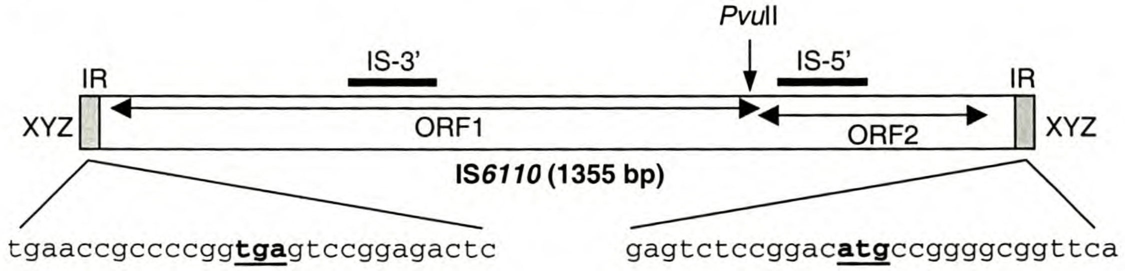
The most important feature of insertion sequences is that they have one or more open reading frames encoding a transposase, which is an enzyme that affords the insertion sequence the ability to move from one chromosomal location to another. Numerous studies have demonstrated the ability of insertion sequences to potentially (i) disrupt genes by insertion (Collins and Gutman, 1992; Hammerschmidt *et al.*, 1996), (ii) up- or down-regulate gene expression by a promoter effect (Charlier *et al.*, 1982; Podglajen *et al.*, 1994; Hubner and Hendrickson, 1997), or (iii) to provide a substrate for genomic rearrangements (Ishiguro and Sato, 1984; Galas and Chandler, 1989; Fang *et al.*, 1999a). These findings demonstrate that insertion sequences are important agents of genome evolution, with the capacity to impact on the phenotype of the host organism.

1.4.2 The insertion sequence IS*6110* in the pathogen, *Mycobacterium tuberculosis*

Numerous insertion sequences have been identified in mycobacterial species (Collins and Stephens, 1991; Dale, 1995; Fang *et al*., 1999b; Gordon *et al*., 1999b), with the insertion element IS*6110* the best known of these, due to its extensive application as a diagnostic marker and molecular probe in epidemiological studies. However, the biological significance of IS*6110* has only recently attracted research attention.

The presence of the repetitive element was first suggested by results obtained using a *Mycobacterium fortuitum* plasmid to probe *M. tuberculosis* genomic DNA (Zainuddin and Dale, 1989). The element was subsequently cloned and sequenced by 3 groups, and termed IS*986* (McAdam *et al.,* 1990), IS*987* (Hermans *et al.,* 1990) and IS*6110* (Thierry *et al.*, 1990a). These initial reports suggested the existence of a number of iso-elements, varying by only a few nucleotides. However, it has since been demonstrated that these variations were in fact due to sequencing errors (Dale *et al*., 1997), and the element will subsequently be referred to here as IS*6110*. IS*6110* is a member of the IS3 family of insertion sequences, and is 1355 bp in length, with 28 bp inverted repeats at its ends (Dale, 1995) (Fig. 1.2). The transposable element produces 3-4 bp duplications on insertion (Dale, 1995), but these are not conserved, as the element apparently demonstrates little sequence specificity of insertion.

Figure 1.2 Schematic representation of the insertion sequence IS*6110*. The figure illustrates the positions and sequences of the imperfect 28 bp inverted repeats (IRs) flanking IS*6110* (gray boxes). The underlined text indicates the trinucleotide variation between the two IRs. "XYZ" represents the duplicated target repeat sequence. ORF1 and ORF2 are the transposase ORFs which overlap by 1 bp. The relative positions of the internal *Pvu*II site, the IS-3' (Van Embden *et al*., 1993) and IS-5' probes (Warren *et al*., 2000) are shown (black bars).



Although the IS*6110* element is present in all members of the tuberculosis complex (*M. tuberculosis, M. bovis, M. microti* and *M. africanum*), its copy number varies considerably among these species. *M. tuberculosis* has 0-25 copies of the element, whereas *M. bovis* has from 1 to 6 (usually one), and *M. bovis* BCG typically has one or two copies (Dale, 1995). Analysis of sequences flanking the single insertion found in most *M. bovis* BCG strains demonstrated that this insertion was always located within the direct repeat (DR) region of the chromosome (Hermans *et al.*, 1991). (Although in those *M. bovis* BCG strains (of Brazilian, Japanese and Russian origin) which have been identified with two copies of IS*6110,* the second copy is situated just upstream of Rv0757 (Fomukong *et al.,* 1994)). In *M. tuberculosis* and *M. bovis*, there is in most cases, at least one copy also positioned within the DR region. The DR region was originally described as a "hotspot" for the insertion of an IS*6110* element (Hermans *et al.*, 1991). It was speculated that this could reflect true preferential insertion or reduced frequency of excision of the element once inserted into this region (Hermans *et al.*, 1991). Alternatively, this region may represent the ancestral IS*6110* integration site in the *M. tuberculosis* and *M. bovis* genomes, with the current chromosomal arrangement of the various copies of IS*6110* in *M. tuberculosis* reflecting outward migration from this position (Dale, 1995; Philipp *et al.*, 1996).

In vitro studies have demonstrated that the IS*6110* element encodes an active transposase (Fomukong and Dale, 1993; Wall *et al.*, 1999). These findings are supported by changes in fingerprints observed in serial isolates, which indicate *in vivo* transposition (Yeh *et al.*, 1998; Niemann *et al.*, 2000). The ability of the element to transpose is proposed to contribute to the generation of genetic diversity detected with this element. A report on the sequencing of numerous IS*6110* elements demonstrated that there was no sequence variation between the copies, suggesting that they all encoded a functional transposase (Dale *et al.*, 1997). Early reports suggested that, in a similar fashion to other members of the IS*3* family of insertion sequences (Fayet *et al.*, 1990; Polard *et al.*, 1991; Chandler and Fayet, 1993), a frameshift would be required to synthesize the IS*6110* transposase enzyme from its two ORFs (McAdam *et al.*, 1990; McAdam *et al.*, 1994). However, a recent study failed to detect frameshifting despite demonstrating transposition (Ghanekar *et al.*, 1999). This and other aspects of the mechanism of IS*6110* transposition and factors regulating transposition frequency have yet to be elucidated.

1.4.3 Application of IS*6110* to genotype *M. tuberculosis*

Although the IS*6110* element is able to transpose, it is assumed that DNA fingerprints obtained with the element are sufficiently stable over the time-period of most molecular epidemiological studies to be reliably used as a marker of recent transmission. The apparent stability of IS*6110*, in combination with the observed diversity in IS*6110* genotypes, has lead to the extensive application of the element as a DNA fingerprinting probe, according to a standardized methodology (Van Embden *et al.*, 1993) for the molecular epidemiology of tuberculosis (Hermans *et al.*, 1990; Thierry *et al.*, 1990b; Cave *et al.*, 1991; Van Soolingen *et al.*, 1991; Mazurek *et al.*, 1991; Chevrel-Dellagi *et al.*, 1993; Yang *et al.*, 1994; Goyal *et al.*, 1994; Huh *et al.*, 1995; Warren *et al.*, 1996a; O'Brien *et al.*, 1997).

When analyzing IS*6110* DNA fingerprints, two or more strains with identical fingerprints are designated as part of a cluster, and clustering is assumed to indicate recent transmission (Small and Moss, 1993; Alland *et al.*, 1994). Unique, or non-clustered strains are assumed to be epidemiologically unrelated, and are thought to arise as a result of reactivation of latent infection (Small and Moss, 1993; Alland *et al.*, 1994). Alternatively, unique strains may be due to immigration from a geographically distinct area of patients harboring other strains of *M. tuberculosis*; or failure to identify the index case (Glynn *et al.*, 1999). The percentage of recent transmission is calculated as a function of clustering according to the (n-1) formula (Small and Moss, 1993):

$$\text{\% Recent transmission} = \frac{\text{Number of samples in clusters} - \text{Number of clusters}}{\text{Total number of samples}}$$

The ratio between transmission and reactivation is interpreted as a measure of the status of an epidemic (Small *et al*., 1994). This data can be applied to assess disease flow, determine risk factors for transmission, identify contacts, and ultimately to monitor the efficacy of control or intervention programs (Smith and Moss, 1994; Van Soolingen and Hermans, 1995; Braden, 1997; Glynn *et al*., 1999).

Despite the international acceptance of IS*6110* DNA fingerprinting as a powerful epidemiological tool, there are certain limitations associated with its use, and the interpretation of results obtained. Firstly, there is some controversy in the literature regarding the assumption of stability of IS*6110* DNA fingerprints (Yeh *et al*.,1998; Alito *et al*., 1999; De Boer *et al*., 1999; Niemann *et al*., 1999a; Niemann *et al*., 1999b; Niemann *et al*., 2000). The rate at which IS*6110* DNA fingerprint patterns change has yet to be accurately defined. This information is an important consideration when determining whether closely related strains form part of a chain of recent transmission.

A second confounding factor in the interpretation of IS*6110* DNA fingerprint patterns is data which supports the existence of numerous preferential insertion loci for IS*6110*. Preferential integration into specific loci could result in convergent evolution of IS*6110* DNA fingerprints, and it has been suggested that this could lead to a subsequent over-estimation of strain relatedness based on IS*6110* DNA fingerprinting (McHugh and Gillespie, 1998; Gillespie *et al*., 2000). However, insertions that are sufficiently offset from each other, or mutational events in surrounding chromosomal domains will counteract such an overestimation of strain relatedness (Sampson *et al.*, 1999; Warren *et al.*, 2000). It is therefore important to gain an understanding of the extent of preferential integration, and the stability of chromosomal regions surrounding IS*6110* elements.

IS*6110* DNA fingerprints cannot be directly correlated with strain phenotype, and no relationship between IS*6110* fingerprints and virulence could be identified (Yang *et al.*, 1995). Numerous studies have also failed to correlate drug resistance profiles with particular fingerprint patterns (Godfrey-Faussett *et al.*, 1993; Rigouts and Portaels, 1994; Williams *et al.*, 1994; Yang *et al.*, 1995). Studies have identified dominant strain families (where a family is defined by percent relatedness according to IS*6110* DNA fingerprinting) in some communities (Fomukong *et al.*, 1994; Hermans *et al.*, 1995; Huh *et al.*, 1995; Van Soolingen *et al.*, 1995; Warren *et al.*, 1996a; Bifani *et al.*, 1996; Rhee *et al.*, 1999; Warren *et al.*, 1999 Zhang *et al.*, 1999). However, it has not been established whether insertions into particular loci in different strain families play a role in imparting a more or less successful strain phenotype, or whether the similar IS*6110* fingerprints are simply a coincidental marker.

1.4.4 IS*6110*: Investigation of genetic identity and evolution of insertion loci

In summary, the current interpretations of IS*6110* DNA fingerprint data fail to provide insight into the relationship between the insertion sequence and the genetic environment in which it resides. In this study, this shortcoming is addressed by investigating three primary aspects of the interplay between IS*6110* and the *M. tuberculosis* genome. Firstly, at the inception of this study, limited data was available regarding the chromosomal location of IS*6110* insertions, which is essential to the understanding of the potential phenotypic impact of IS*6110* insertion events. Therefore, this study has analyzed DNA sequences flanking IS*6110* insertion loci in the genome of clinical isolates, in conjunction with data extracted from the literature and DNA sequence databases. Secondly, while IS*6110* had been proposed as an important agent in the evolution of the apparently stable *M. tuberculosis* genome (Sreevatsan *et al*., 1997), this had not been systematically investigated. Knowledge of the rate of IS*6110*-mediated genome evolution is fundamental to the accurate interpretation of IS*6110* DNA fingerprinting data, and has been addressed in this study. Finally, mechanistic aspects of IS*6110*-mediated deletion events have been considered.

1.5 The polymorphic PPE gene family

1.5.1 Introduction

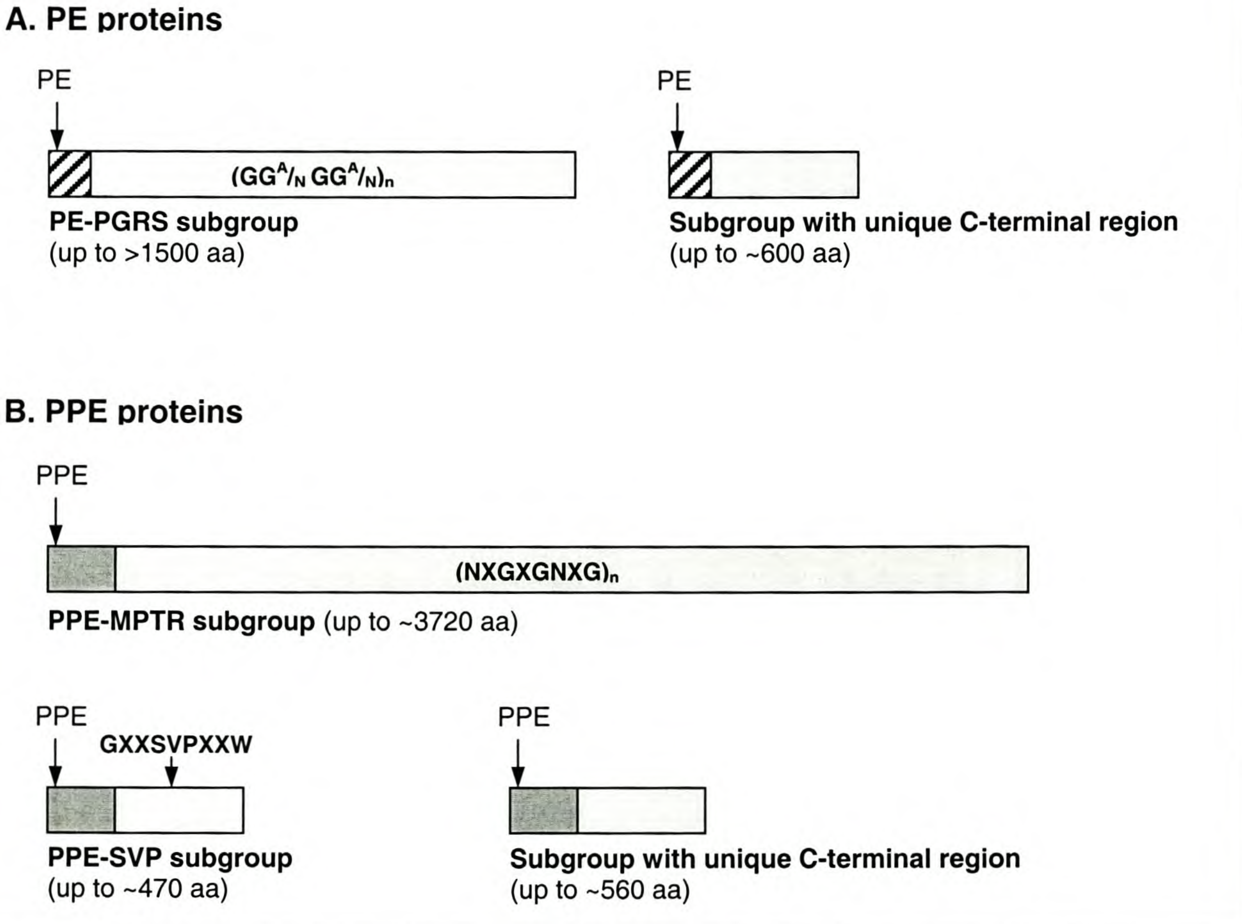
The investigation of IS*6110* as described earlier focused attention on the PPE gene family of *Mycobacterium tuberculosis*. Firstly, a number of IS*6110*-mediated PPE gene disruptions were identified. Secondly, PPE gene polymorphism was detected by hybridization with IS*6110* flanking sequences. To better understand the significance of the observed gene disruptions and polymorphism, the PPE gene family was investigated further.

The existence of two large gene families of unknown function, namely the PE and PPE gene families, was highlighted by analysis of the complete genome sequence of *M. tuberculosis* H37Rv (Cole *et al*., 1998). These gene families comprise about 10% of the genome and are named after the Pro-Glu and Pro-Pro-Glu motifs at the N-termini of their encoded proteins (Cole *et al*., 1998). The function of the glycine-rich PE and PPE proteins has not been established, although numerous hypotheses exist. Interestingly, it has been suggested that they may be of immunological significance, either (i) by providing a source of antigenic variation in an otherwise genetically homogenous bacterium, or (ii) by inhibiting antigen processing by host cells (Cole *et al*., 1998).

1.5.2 The PE gene family

The PE gene family consists of about 100 members, grouped by a relatively conserved 110 aa N-terminal region (Fig. 1.3). The encoded proteins vary considerably in length, attaining up to 1400 aa. The gene family can be divided into a number of subgroups, the largest of which is the PE-PGRS subgroup. The PE-PGRS subgroup is named for the characteristic expansions of the polymorphic GC-rich sequence (PGRS) found within its members. The PGRS consensus repeat is CGGCGGCAA, and stretches of the repeat encode

Figure 1.3 Schematic representation of proteins encoded by the PE and PPE gene families of *M. tuberculosis*. (A) PE proteins are characterized by an N-terminal "PE" motif, with conserved 110 aa N-terminal region (diagonal fill). The PE-PGRS subgroup is characterized by the PGRS sequence, encoding glycine-rich amino acid repeat regions as shown. (B) PPE proteins contain an N-terminal "PPE" motif, within a conserved 180 aa N-terminal region (dark fill). The PPE-MPTR subgroup is characterized by the MPTR sequence, encoding asparagine- and glycine-rich amino acid repeat regions as shown. (Figure adapted from Cole *et al.*, 1998).



multiple tandem repeats of glycine-rich motifs (often Gly-Gly-Ala or Gly-Gly-Asn). DNA fingerprints generated with probes containing PGRS repeats demonstrate considerable polymorphism in clinical isolates of *M. tuberculosis* (Ross *et al.*, 1992; Van Soolingen *et al.*, 1993; Warren *et al.*, 1996b; Yang *et al.*, 1996; Braden *et al.*, 1997; Strässle *et al.*, 1997), clearly demonstrating the variable nature of these sequences. The observed polymorphism has been exploited in a molecular epidemiological context, where the PGRS-containing regions are utilized as informative secondary DNA fingerprinting probes (Ross *et al.*, 1992; Van Soolingen *et al.*, 1993; Yang *et al.*, 1996; Warren *et al.*, 1996b; Braden *et al.*, 1997; Strässle *et al.*, 1997).

Recently, interest has focused on the important question of whether PE genes are expressed during infection. There is evidence that mice infected with *M. tuberculosis* demonstrate a strong humoral response to the PGRS region of PE-PGRS proteins (Delogu and Brennan, 2001). Two independent studies have shown recognition of recombinant PE-PGRS proteins by patient sera (Espitia *et al.*, 1999; Singh *et al.*, 2001). One of these studies also identified fibronectin-binding properties associated with one PE-PGRS gene (Espitia *et al.*, 1999), corroborating earlier results obtained for another member of the gene family (Abou-Zeid *et al.*, 1991). More intriguingly, it was recently demonstrated that two PE-PGRS genes expressed in *M. marinum* were essential for replication in macrophages and bacterial persistence in granulomas, implicating the PE-PGRS proteins in virulence (Ramakrishnan *et al.*, 2000).

1.5.3 The PPE gene family

The PPE family consists of 68 members that have a relatively conserved 180 amino acid N-terminal region in common (Fig. 1.3; Cole *et al.*, 1998). As with the PE gene family, the PPE family can be divided into at least 3 subgroups (Cole *et al.*, 1998). One of these is the PPE-MPTR (major polymorphic tandem repeat) subgroup. The second is a class characterized by a conserved motif (Gly-X-X-Ser-Val-Pro-X-X-Trp) at around position 350 of the proteins, referred to here as the PPE-SVP subgroup. The third group of PPE proteins are apparently unrelated except for the conserved 180 aa N-terminal region.

The PPE-MPTR subgroup consists of genes containing extensive stretches of the MPTR core consensus repeat GCCGGTGTTG, separated by 5 bp spacers (Cole and Barrell, 1998). The repeats encode multiple copies of the motif Asn-X-Gly-X-Gly-Asn-X-Gly. MPTR repeats were first described in an early report detailing the characterization of a novel repetitive DNA sequence in *M. bovis* (Doran *et al.*, 1992). On the basis of structural analogy to known virulence-associated proteins in other pathogens, the authors speculated that these repeats occurred within a family of genes encoding functionally related proteins, which were likely to be involved in host cell receptor binding (Doran *et al.*, 1992). The term MPTR was subsequently coined by Hermans *et al.* (1992), who investigated the utility of the repeat as a DNA fingerprinting probe. The MPTR repeats have found less extensive application in the molecular epidemiology of *M. tuberculosis* than the PGRS repeats, as they appear to be less variable (Hermans *et al.*, 1992).

A subset of the PPE-MPTR group of genes contain other repetitive features consisting of longer tandem repeats (69 – 75 bp repeats), and polymorphisms have been identified in these regions. This was first identified during an investigation of the region upstream of the *katG* gene (Zhang and Young, 1994; Goyal *et al.*, 1994), although at the time it was not known that this polymorphism was associated with a gene family. It has also been shown that these longer tandem repeats are a source of variation between *M. tuberculosis* H37Rv and *M. bovis* BCG (Cole *et al.*, 1998). A recent study has identified extensive polymorphism in *M. bovis* associated with 69 bp tandem repeats in one particular PPE-MPTR gene, Rv1917c (O' Brien *et al.*, 2000). PE and PPE gene polymorphisms are potentially significant, as it has been suggested that *M. tuberculosis* has an otherwise very stable genome, and could indicate that these genes are responsive to host immune pressures (Sreevatsan *et al*., 1997). The extensive polymorphism provides support for the hypothesis that PE and PPE proteins provide the bacterium with a potential source of antigenic variation (Cole *et al*., 1998).

At the inception of this study, very little was known in terms of expression profiles, subcellular location, gene stability and biological significance of members of the PPE gene family. Indirect evidence for *in vivo* expression of the PPE genes was provided by a study of a serine-rich antigen from *Mycobacterium leprae* (Vega-Lopéz *et al*., 1993), which was recognized by serum from both leprosy and tuberculosis patients. The protein was shown to have homology to a 51 kDa hypothetical antigen (identified upstream of the 65 kDa antigen) in *M. tuberculosis* (Shinnick, 1987), which is now known to be a member of the PPE gene family. Further evidence for the expression of PPE genes was provided by a study which demonstrated that the PPE gene Rv0755c, a member of the MPTR subgroup, is upregulated in H37Ra relative to H37Rv, as determined on the basis of differential display PCR (Rivera-Marrero *et al*., 1998). Similar methodology was utilized to show the upregulation of the PPE gene Rv2770c in H37Rv compared to H37Ra (Rindi *et al*., 1999). Recently, it was demonstrated that the PPE gene Rv2123 is upregulated under low iron conditions, leading to the suggestion that the gene product is involved in iron acquisition via siderophore uptake (Rodriguez *et al*., 1999). This differs considerably from the hypothesis that PPE gene products may function as cytosolic storage proteins for the relatively rare amino acid asparagine (Cole, 1999).

Only recently have studies focused on the biological significance of the PPE gene family. Intriguingly, it has been shown that a transposon mutant of the PPE gene Rv3018c is attenuated for growth in macrophages, potentially implicating the gene as a virulence factor (Camacho *et al*., 1999). Serological expression cloning has demonstrated that the PPE genes

Rv1196 (Dillon *et al*., 1999) and Rv0915c (Skeiky *et al*., 2000) encode proteins which stimulate a T-cell response, and induce protection in a mouse model of tuberculosis.

1.5.4 The PPE gene family: Investigation of extent and mechanisms of genetic variation, expression profiles and subcellular location

The emerging data described above justifies further exploration of the PPE gene family. This study has therefore exploited the available genome sequence data in the investigation of some basic aspects of the biology of this potentially important gene family. A previous study of the PE gene family investigated phylogenetic relationships between various PE-PGRS genes (Espitia *et al*., 1999), but speculated little on the significance of their result. In this study, a similar analysis of the PPE gene sequences from *M. tuberculosis* H37Rv was carried out to explore the phylogenetic relationships among the members of the PPE gene family. The sequence data has also been compared to other available genome sequence data, including that of *M. tuberculosis* CDC1551, *M. bovis* BCG and *M. leprae*, and the significance of the results is discussed.

The *M. tuberculosis* H37Rv sequence data has been utilized further to identify and investigate tandem repeat regions in the PPE-MPTR subgroup of genes. While a number of studies have hinted at the existence of PPE tandem repeat polymorphism (Zhang and Young, 1994; Cole *et al*., 1998; O'Brien *et al*., 2000), they have not investigated the extent and mechanisms of these polymorphisms. This question is important in evaluating the possible role of PPE proteins as a source of antigenic variation, and has therefore been addressed in this study. Furthermore, to expand the very limited knowledge of PPE gene expression, this study examines *in vitro* and *in vivo* expression profiles of selected PPE genes. Finally, the subcellular location of one member of the PPE gene family has been investigated, an important question in attempting to establish a biological role for the protein family.

1.6 Study Aims

The aims of this study were briefly described earlier and are set out here for ease of referral.

1.6.1 IS*6110* insertion loci were investigated in order to:

(i) Characterize the genomic location of IS*6110* insertions (Chapter 2),

(ii) Infer strain phylogeny (Chapter 3), and

(iii) Identify mechanisms of genome diversity associated with IS*6110* insertion loci (Chapters 3 and 4).

1.6.2 The PPE gene family was investigated in order to:

(i) Determine the phylogenetic structure of the PPE gene family (Chapter 5),

(ii) Establish the extent and mechanisms of PPE gene polymorphism (Chapter 5),

(iii) Examine the *in vitro* and *in vivo* expression profiles of PPE genes (Chapter 6),

(iv) Identify the subcellular location of a PPE gene product (Chapter7).

1.7 Thesis design

The 8 chapters contained in this thesis consist of an introductory overview chapter (Chapter 1), 6 chapters dealing with particular aspects of either IS*6110* insertion loci (Chapters 2 to 4) or the PPE gene family (Chapters 5 to 7), followed by a general discussion and conclusions chapter (Chapter 8). Chapters 2 to 7 are presented as manuscripts that have been published (Chapters 2 and 3), are currently under review (Chapter 7), or are in preparation (Chapters 4 to 6). As such, there may be some repetition of introductory material and methodology. All cited literature has been compiled into a single list at the end of the thesis for ease of reference.

CHAPTER 2

IS*6110* INSERTIONS: CHROMOSOMAL LOCATION

NOTE: The results presented in the following chapter were published in full as: "**Disruption of coding regions by IS*6110* insertion in *Mycobacterium tuberculosis* clinical isolates.** Sampson, S.L., Warren, R.M., Richardson, M., Van Der Spuy, G.D., Van Helden, P.D. *Tubercle and Lung Disease* (1999) 79(6):349-359."

(The style of the text and numbering of sections has been altered, and all references to the M. tuberculosis *strain "CSU93" have been replaced with "CDC1551", to conform to the style of this thesis. Literature cited in the text takes the form of author name and year of publication as opposed to the number format specified by* Tubercle and Lung Disease. *All cited literature is compiled into a single list at the end of the thesis for ease of reference. No other changes have been made to the body of the text.)*

2.1 Introduction

The *Mycobacterium tuberculosis* insertion sequence IS*6110* was first described in 1990 (McAdam *et al.*, 1990; Thierry *et al.*, 1990a), and has since found extensive application in the molecular epidemiology of tuberculosis (Hermans *et al.*, 1990) where it is used as a DNA fingerprinting probe according to a standardized methodology (Van Embden *et al.*, 1993). IS*6110*-based strain classification relies on the presence of variable numbers of the element at differing chromosomal locations, which remain stable over the time period of most epidemiological studies (Hermans *et al.*, 1990). The DNA fingerprint obtained is a representation of IS*6110* elements within the genome, where their chromosomal location is measured in terms of electrophoretic mobility.

A large diversity of IS*6110* DNA genotypes is evident from international studies (Cave *et al.*, 1991; Chevrell-Dellagi *et al.*, 1993; Small *et al.*, 1994; Yang *et al.*, 1994; Hermans *et al.*, 1995; Warren *et al.*, 1996a). However, a study of comparative sequence analysis of 26 structural genes from clinical isolates of *M. tuberculosis* demonstrated little variation, suggesting overall genome stability (Sreevatsan *et al.*, 1997). The transposable element has therefore been proposed as an important agent in the evolution of the *M. tuberculosis* genome (Sreevatsan *et al.*, 1997). The ability of the element to transpose has been demonstrated both *in vitro* (Fomukong and Dale, 1993; Wall *et al.*, 1999) and *in vivo* (single band changes in serial isolates) (Yeh *et al.*, 1998). Furthermore, genomic rearrangements, such as deletions, may be promoted by the presence of multiple copies of the element, as has recently been proposed (Fang *et al.*, 1999a). Therefore the potential exists that gene disruptions or knockouts can be mediated by IS*6110* insertion. The impact of insertion sequences on gene function has been demonstrated in other organisms, with experimental confirmation of gene knockouts (Collins and Gutman, 1992), genomic rearrangements (Ishiguro and Sato, 1984) and up- or down-regulation of gene expression (Hubner and Hendrickson, 1997; Hall, 1998) mediated by insertion sequences.

The potential for insertion sequences to exert a deleterious effect on the host organism has earned them the reputation of "selfish DNA" (Doolittle and Sapienza, 1980), only concerned with their own survival. The "selfish DNA" theory proposes that if an element is able to ensure its own survival, by restricting transposition to levels tolerable to its host, yet sufficient to ensure their own survival, then no other explanation is required for its continued existence (Orgel and Crick, 1980). However, the "adaptive theory" (Campbell, 1981) suggests that if the insertion sequences did not offer some advantage to the host organism, they would eventually be eliminated from its genome, as their presence would promote at least some harmful mutations.

The basis of understanding whether an adaptive mutation is mediated by a particular transposition event, is the characterization of the flanking regions of the insertion sequence at the nucleotide level. The first detailed investigation of an IS*6110* insertion locus described the apparent preferential insertion of the element into the direct repeat (DR) locus in *Mycobacterium bovis* BCG (Hermans *et al.*, 1991). The insertion sequence is found at an identical position within the DR locus in *M. tuberculosis* and *M. bovis*, and it has been proposed that the element was inserted into this position prior to their divergence (Dale, 1995) approximately 15,000 years ago (Kapur *et al.*, 1994). Other examples of preferential insertion have subsequently been investigated, using either direct methods, for example PCR amplification and sequencing (Fang and Forbes, 1997; Fang *et al.*, 1999b), or indirect, computer-based analysis (McHugh and Gillespie, 1998). Further investigations of insertion loci have focused on evolutionary questions, for example the evolutionary origins of low vs high copy number strains (Fomukong *et al.*, 1998). An assessment of strain phylogeny based on DR region analysis has also been reported (Fang *et al.*, 1998). However, prior to this study, little research effort was focused on characterizing the functional identity of IS*6110* insertion loci. Although two studies have reported the insertion of IS*6110* into a coding region (Fang *et al.*, 1999b; Vera-Cabrera *et al.*, 1997), analysis of the genome sequence of the reference strain, *M. tuberculosis* H37Rv, suggests that the majority of insertions in this isolate occur within non-coding regions (Cole *et al.*, 1998).

To determine whether the element inserts into other predicted coding regions, we focused on the characterization of IS*6110* insertion loci identified in clinical isolates originating from a high incidence study community in the Western Cape, South Africa. Thirty-three insertion loci were cloned from 8 clinical isolates and 29 of these were characterized by sequence analysis. In addition, 43 insertion sites identified in the available literature (Fomukong *et al.*, 1998; Vera-Cabrera *et al.*, 1997; Mendiola *et al.*, 1992) and DNA sequence databases were analyzed. *M. tuberculosis* H37Rv genome sequence data was utilized to map the genomic location of the insertion sites, as well as to identify "natural insertional mutants" created by IS*6110* insertion. The predicted impact of IS*6110* transpositions on phenotype, and the implications of the results obtained for the interpretation of molecular epidemiology are discussed.

2.2 Methods

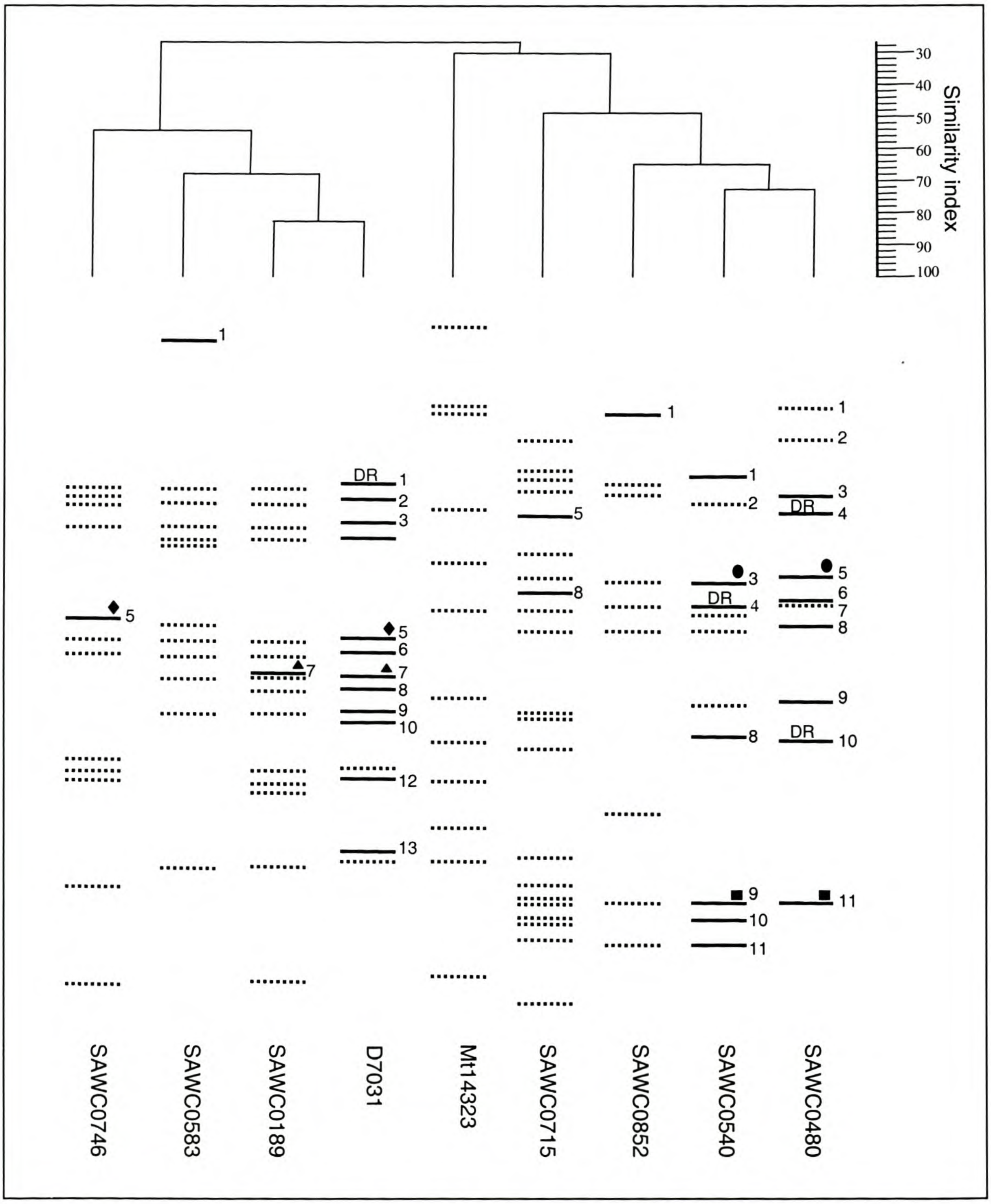
2.2.1 Study setting and strain selection

An ongoing molecular epidemiology study of tuberculosis in a high incidence community in the Western Cape, South Africa has produced a database of more than 1700 *M. tuberculosis* isolates from approximately 800 patients. These isolates have been typed by DNA fingerprinting with the probe IS*6110* according to an internationally standardized protocol (Van Embden *et al.*, 1993). Eight high copy number (9 to 21 IS*6110* copies) clinical isolates (SAWC0189, SAWC0540, SAWC0715, SAWC0746, SAWC0852, SAWC0583, SAWCO480 and D7031 (Fig. 2.1)), which represent 5 strain families commonly found within the study community, were selected for the cloning of insertion loci (where a family of strains is defined as > 65% related in terms of IS*6110* DNA fingerprinting, as determined by GelCompar (Applied Maths, Kortrijk, Belgium)).

2.2.2 Insertion locus cloning

Insertion loci were randomly selected for cloning from the 8 clinical isolates. DNA from the isolates was restricted with *Pvu*II at 37°C for 3 hours, and electrophoretically fractionated in 1% agarose (1X TBE pH 8.0, 2 V/cm, overnight), with Molecular Weight Marker X (Roche) in adjacent lanes. Gel lanes containing the restricted DNA were cut transversally into 5 mm strips, and the agarose gel slices containing *Pvu*II fragments corresponding to the IS*6110* hybridizing band were selected according to the molecular weight marker. DNA extracted from the agarose gel slices (using the Cleanmix DNA isolation kit (Talent, Italy)) was ligated into *Eco*RV-linearized, shrimp alkaline phosphatase (SAP) (Amersham) dephosphorylated pBluescript utilizing T4 DNA ligase (Roche) at 16°C, overnight. Ligated DNA was transformed into electrocompetent XL-1 blue cells, which were plated onto LB-agar plates (containing 50 μg/ml ampicillin, 0.1 mM X-gal and 0.2 mM IPTG) and incubated at 37°C overnight. Colony lifts onto Hybond N+ membrane were performed (Sambrook *et al.*, 1989), and the membranes were subsequently hybridized with IS*6110* (labeled using the ECL system (Amersham)) according to a standardized protocol (Van Embden *et al.*, 1993). Positive colonies (visualized by exposure to autoradiographic film following ECL detection) were picked into LB medium containing ampicillin (50 µg/ml), and incubated overnight at 37°C with shaking. Plasmid DNA was isolated using the Wizard Plus Minipreps DNA purification system (Promega). The DNA was restricted with *Xho*I and *Eco*RI to excise and map insert DNA. After electrophoretic fractionation (in 1% agarose, 1X TBE pH 8.0, 2V/cm, overnight), insert DNA was isolated from the gel fragment using the Cleanmix DNA isolation kit (Talent, Italy). The purified insert was used to back-hybridize to the original *Pvu*II-restricted DNA, to determine whether the insertion site of choice had been cloned. Clones representing insertion into the DR locus were not sequenced, but all other cloned insertion loci were sequenced using a primer specific to the 3' end of IS*6110* (5'-TTCAACCATCGCCGCCTCTACC-3'), on an ABI automated sequencer. Sequence data (representing IS*6110* 3' flanking sequence only) was deposited into GenBank using the BankIt sequence submission tool at the National Centre for Biotechnology Information (NCBI) at www.ncbi.nlm.nih.gov (See Table 2.1 for accession numbers).

Figure 2.1. GelCompar representation of IS*6110* DNA fingerprints of strains utilized in this study, and the reference strain Mt14323 (Van Embden *et al.*, 1993). Solid bands indicate loci cloned from South African clinical isolates, and those not cloned are depicted by dotted lines. The numbering to the right of the cloned loci indicates the numerical designation of a particular band, assigned in terms of molecular weight, from highest to lowest in each lane. The symbols ▲, ■, • and ♦ indicate insertions into identical positions shared by different strains. Bands representing DR region insertions are indicated by "DR".



2.2.3 Nomenclature

Cloned insertion loci were named according to their strain of origin, and band position, where bands were ranked numerically according to their molecular weight (from highest to lowest), and given the prefix "ISL". (For example, the designation "ISL0583.1" refers to the highest molecular weight band cloned from strain SAWC0583 (Fig. 2.1, lane 2).) Insertion loci in H37Rv are identified by the Rv numbers of their coding regions (Table 2.2), while loci from CDC1551 are referred to by the contig in which they are present (Table 2.3). Insertion loci identified in the literature and DNA sequence databases are referred to by the prefixes DK (Fomukong *et al.*, 1998), V-C (Vera-Cabrera *et al.*, 1997), M (Mendiola *et al.*, 1992), or by their GenBank accession numbers (Table 2.4).

2.2.4 Sequence analysis

The ENTREZ browser (provided by the National Center for Biotechnology Information (NCBI) at www.ncbi.nlm.nih.gov) was used to search the GenBank DNA sequence database (maintained by the NCBI) to identify the IS*6110* insertion loci in the *M. tuberculosis* H37Rv reference strain (Cole *et al.*, 1998), as well as other *M. tuberculosis* IS*6110* insertion loci deposited there. Sequence data for IS*6110* insertion loci from the strain CDC1551 was obtained through early release from The Institute for Genomic Research (TIGR) at www.tigr.org. In addition, available literature was analyzed to identify any further IS*6110* insertion loci not described in the DNA sequence databases (Fomukong *et al.*, 1998; Vera-Cabrera *et al.*, 1997; Mendiola *et al.*, 1992). Sequence data from the insertion loci cloned from South African clinical isolates was used in conjunction with the BLASTN search tool to search against the *M. tuberculosis* H37Rv genome sequence data (Cole *et al.*, 1998) deposited in GenBank to determine the genomic location of the IS*6110* insertion loci. Other insertion loci identified (in the literature and DNA sequence databases) were mapped in the same way. H37Rv genome sequence annotations were analyzed to determine the identity of disrupted regions.

2.3 Results

2.3.1 Sources of insertion loci data

Eight high copy number clinical isolates (9-21 copies of IS*6110*) were identified from the strain database generated from an ongoing molecular epidemiology study in a high incidence community. These strains were classified into 5 strain families on the basis of GelCompar (Applied Maths, Kortrijk, Belgium) analysis, with a cutoff of 65% relatedness defining a family. These strains were chosen to maximize the chances of identifying unique loci when cloning randomly selected IS*6110*-hybridizing bands. A total of 33 clones, representing 26 distinct insertion loci (including 1 DR insertion) were generated from the 8 clinical isolates (Fig. 2.1, Table 2.1). In addition, fourteen H37Rv cosmids containing the 16 IS*6110* insertion sites found in this strain were identified using ENTREZ software, and retrieved from the GenBank DNA sequence database for further analysis (Table 2.2). Sequence data representing 4 insertion loci from the CDC1551 strain (TIGR) was also retrieved (Table 2.3).

An ENTREZ search of the GenBank DNA sequence database using the keyword "IS*6110*" identified a further 49 hits (excluding H37Rv sequences, and sequences deposited by our group). Thirty-six of these hits were excluded from further analysis for the following reasons: (a) Sequence deposits represented identical insertion loci duplicated by data from another source, already included in analysis (n=11); (b) Sequence deposited represented cloning vector (n=1); (c) Sequence was obtained from mycobacteria other than *M. tuberculosis* (n=4); (d) Only IS*6110* sequence and no flanking region was present (n=4); (e) No IS*6110* sequence was present (n=4); (f) Sequences represented the ancestral DR region insertion (identical to the H37Rv DR IS*6110* insertion locus) (n=6); (g) Fourteen sequences deposited by Fang *et al.*, specifically dealing with insertions into the *ipl* locus, represented 8 unique insertion loci. Single sequence deposits representing one example of each point of insertion were selected (Table 2.4), while the remainder (n=6) were excluded to simplify subsequent BLASTN comparisons.

Of the 18 insertion loci identified in the literature, 10 were included for further analysis (Table 2.4). The reasons for excluding 8 loci were as follows: (a) Sequence deposits represented identical insertion loci duplicated by data from another source, already included in analysis (n=3); (b) IS*6110* flanking sequence too short to generate significant blast hits (n=2); (c) Duplication of H37Rv DR region insertion (n=3).

2.3.2 Functional identity of insertion loci

Analysis of the total of 76 insertion loci identified from various sources demonstrated that these represented 66 unique insertion loci, including 9 unique insertions into the *ipl* locus (with one of these *ipl* insertions newly identified by our group). Of the insertion loci cloned from South African clinical isolates, 4 representing DR region insertions, and 4 representing insertion into identical positions in different strains were excluded. Two CDC1551 insertions, into identical positions to the H37Rv DR region insertion and the H37Rv insertion represented by Rv1756c/Rv1757c were also excluded.) Further analysis was confined to unique insertion loci only, to avoid over-representation of specific sites. Fifty-nine of the unique insertion loci could be mapped to the H37Rv genome. Sixty-four percent (42/66) of the unique loci mapped to coding regions of the H37Rv genome, while 26% mapped to intergenic regions (17/66). Exclusion of insertions into the *ipl* locus (n=9) from this calculation resulted in 58% of the unique loci mapping to coding regions (33/57) and 30% mapping to intergenic regions (17/57).

Precise functions have been assigned to 40% of the predicted proteins in the *M.*

tuberculosis genome by homology comparisons (Cole *et al.*, 1998). The remainder of the open reading frames (ORFs) either have some similarity (44%), or do not show any resemblance (16%) to known proteins (Cole *et al.*, 1998). The predicted functions (where these have been assigned) of products of the disrupted ORFs are diverse, and include: 1) cell wall components, such as cutinases and a conserved small membrane protein (*mmp*S1); 2) proteins involved in energy metabolism, including a putative polyketide synthase (*pks*9); 3) macromolecule metabolism, such as phospholipase c (*plc*C), 4) detoxification, for example cytochrome p450 and bromoperoxidase (*bpo*A); and 5) regulatory functions, exemplified by a serine-threonine-kinase (*pkn*J). Interestingly, analysis of the collated data reveals 8 examples of disruption of 5 members of a group of multiple copy genes, namely the PPE (Proline-Proline-Glutamine) gene family (Cole *et al.*, 1998). Disruptions of a number of hypothetical genes with no homology to known proteins (n = 9) have also been identified (Tables 2.1 and 2.4).

Table 2.1 Insertion loci cloned from South African clinical isolates. Insertion sites cloned and characterized in this study are given with their GenBank accession numbers. DR region insertion loci, indicated by *, were not sequenced. The symbols ▲, ■, • and ♦ indicate insertions into identical positions shared by different strains. The positions of insertion loci clones that mapped to the H37Rv genome are given in terms of H37Rv gene designations of the regions disrupted. The identity of IS*6110*-disrupted genes is indicated. (N/A = not applicable.)

Locus name	Accession number	Rv designation of region disrupted	Gene identified
ISL0189.7 ▲	U60565	Rv1319c	Adenylate cyclase-like ORF
ISL0480.3	AF077727	Rv1917c	PPE
ISL0480.4 *	DR insert	DR region insertion	N/A
ISL0480.5 •	AF077949	Rv2818c	Unknown
ISL0480.6	AF077948	Rv3125c	PPE
ISL0480.8	AF077947	Rv3327	IS*1547*
ISL0480.9	AF077946	Rv1664	*pks*9
ISL0480.10 *	DR insert	DR region insertion	N/A.
ISL0480.11 ■	AF077945	Rv1755c	Partial *plc*D
ISL0540.1	AF126475	Rv1928c	Similar to 7-alpha-hydroxysteroid dehydrogenase
ISL0540.3 •	U60566	Rv2818c	Unknown
ISL0540.4 *	DR insert	DR region	N/A.
ISL0540.8	AF126474	Rv0594	Unknown
ISL0540.9 ■	AF126473	Rv1755c	Partial *plc*D
ISL0540.10	AF126477	Rv1234-Rv1235 intergenic	N/A.
ISL0540.11	AF126476	Rv3346c-Rv3347c intergenic	N/A.
ISL0583.1	AF086633	Rv1917c	PPE
ISL0715.5	U60568	Rv2435c	Unknown
ISL0715.8	U60567	Rv0001-Rv0002 intergenic	N/A.
ISL0746.5 ♦	U60569	Does not map to H37Rv	N/A.
ISL0852.1	U60570	Does not map to H37Rv	N/A.
ISL7031.1 *	DR insert	DR region insertion	N/A.
ISL7031.2	AF092218	Rv1754c	Unknown
ISL7031.3	AF126479	Rv 2352c	PPE
ISL7031.4	AF126478	Rv3113	Similar to phosphoglycolate phosphatase
ISL7031.5 ♦	AF092217	Does not map to H37Rv	N/A.
ISL7031.6	AF092216	Rv0835-Rv0836 intergenic	N/A.
ISL7031.7 ▲	AF092214	Rv1319c	Adenylate cyclase-like ORF
ISL7031.8	AF092213	Rv1758	Possible cutinase
ISL7031.9	AF092215	Rv2015c	Unknown
ISL7031.10	AF092212	Rv1765c-Rv1766 intergenic	N/A.
ISL7031.12	AF092211	Rv1777	Probable cytochrome p450
ISL7031.13	AF092210	Rv2353c	PPE

Table 2.2 H37Rv insertion loci. The Rv gene designations of the 16 IS*6110* elements in the *M. tuberculosis* H37Rv genome are given. The identity and Rv designations of IS*6110*-disrupted genes are indicated. (Note: The IS*6110* element represented by the Rv1756c/Rv1757c ISL clone is annotated as disrupting both its flanking ORFs.)

Transposase ORFs:	Rv designation of region disrupted	Gene identified
Rv0795; Rv0796	Intergenic	N/A.
Rv1369c; Rv1370c	Intergenic	N/A.
Rv1756c; Rv1757c	Rv1755c Rv1758	Partial *plc*D Possible cutinase
Rv1763; Rv1764	Rv1765c	Unknown
Rv2105; Rv2106	Intergenic	N/A.
Rv2167c; Rv2168c	Intergenic	N/A.
Rv2278; Rv2279	Rv2277c	Possible glycerol phosphodiesterase
Rv2354; Rv2355	Rv2353c	PPE
Rv2479; Rv2480	Intergenic	N/A.
Rv2648; Rv2649	Intergenic	N/A.
Rv2814c; Rv2815c	DR region insertion	N/A.
Rv3184; Rv3185	Intergenic	N/A.
Rv3186; Rv3187	Intergenic	N/A.
Rv3325; Rv3326	Intergenic	N/A.
Rv3380; Rv3381	Rv3379c	Transketolase-like protein
Rv3474; Rv3475	Rv3473c	Probable *bpo*A

Table 2.3. CDC1551 insertion loci. IS*6110*-containing contigs in the *M. tuberculosis* reference strain CDC1551. The positions of insertion loci relative to the H37Rv genome are given in terms of H37Rv gene designations of the regions disrupted. The identity of IS*6110*-disrupted genes is indicated. (Note the DR region insertion, and the insertion into Rv1758 are in identical positions to the H37Rv insertions in the corresponding regions.)

CDC1551 contig	Rv designation of region disrupted	Gene identified
3737	Rv0403c	*mmp*S1
3737	DR region insertion	N/A.
3661	Rv3018c	PPE
3721	Rv1758	Possible cutinase

Table 2.4. Insertion loci identified in the literature and DNA sequence databases. Insertion loci identified in literature and databases are given with their respective reference and/or GenBank accession number. The positions of insertion loci that mapped to the H37Rv genome are given in terms of H37Rv gene designations of the regions disrupted. The identity of IS*6110*-disrupted genes is indicated. The symbol * indicates a unique DR region insertion.

Locus name	Accession number/ Reference	Rv designation of region disrupted	Gene identified
DK3	(Fomukong *et al.*, 1998)	Does not map to H37Rv	N/A.
DK4	(Fomukong *et al.*, 1998)	DR region insertion*	N/A.
DK6	(Fomukong *et al.*, 1998)	Rv1762c	Unknown
DK9	(Fomukong *et al.*, 1998)	Rv1777	Probable cytochrome P450
DK10	(Fomukong *et al.*, 1998)	Rv2088	*pkn*J
DK12	(Fomukong *et al.*, 1998)	Rv2808	Unknown
DK13	(Fomukong *et al.*, 1998)	Does not map to H37Rv	N/A.
DK14	(Fomukong *et al.*, 1998)	Does not map to H37Rv	N/A.
M4	(Mendiola *et al.*, 1992)	Does not map to H37Rv	N/A.
V-C	L11868 (Vera-Cabrera *et al.*, 1997)	Rv2351c	*plc*A
-	S76966	Rv0756c-Rv0757 intergenic	N/A.
-	X94956	Rv3018c	PPE
-	X94959	Rv1371	Unknown
*ipl*1	X98149	The *ipl* locus is found within the insertion sequence IS*1547,* which maps to two positions on the H37Rv genome (Fang and Forbes, 1997; Fang *et al.*, 1999b), corresponding to Rv0797 and Rv3327	IS*1547*
*ipl*2	X98151		
*ipl*3	X98153		
*ipl*4	X98154		
*ipl*5	X98156		
*ipl*6	X98158		
*ipl*7	Y14613		
*ipl*8	Y14614		
-	Y15749	Rv0794c	*lpd*B
-	Y17220	Does not map to H37Rv	N/A.

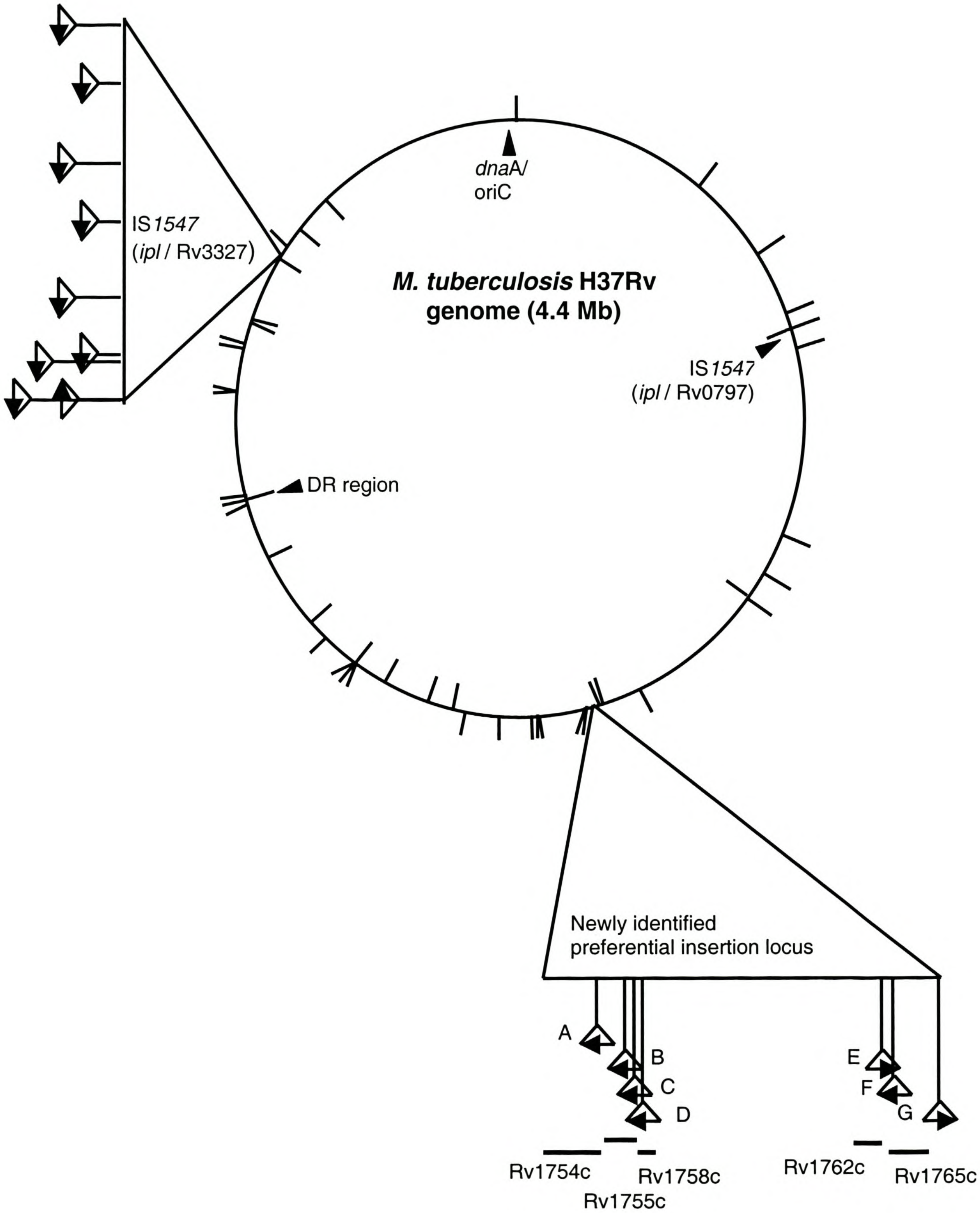
2.3.3 Genomic distribution of insertion loci

The genomic positions of the 66 unique insertion loci identified in this study were plotted on a genome map of H37Rv, relative to the positions of the IS*6110* insertions in this reference strain. DNA sequence analysis failed to map the genomic locations of 7 unique insertion loci, suggesting that these loci are absent in H37Rv. The insertion "DK4" (Table 2.4) represented a unique insertion point within the DR region (Hermans *et al.*, 1991; Fomukong *et al.*, 1998).

The resulting map (Fig. 2.2), representing the collated insertion loci data, demonstrates distribution of insertion sites throughout the genome, although there is a concentration of these within the genome quadrants flanking the DR locus. The collated data also confirms the presence of preferential insertion loci. As reported by Fang *et al.*, the *ipl* locus, which forms part of the newly identified IS*1547* insertion sequence, maps to two genomic locations (Fang and Forbes, 1997; Fang *et al.*, 1999b), represented by Rv0797 and Rv3327 (Fig. 2.2). A unique insertion into the *ipl* locus is represented by ISL0480.8 (2 nucleotides 3' of the "*ipl*5" insertion point) (Fig. 2.2). Furthermore, seven insertion loci, 4 of which were cloned from South African clinical isolates, map to different positions in the Rv1754c to Rv1765c region of the genome (H37Rv cosmid MTCY28, GenBank accession number Z95890). The 7 insertion points span a total region of 9447 bp (excluding the 2 H37Rv IS*6110* elements) (Fig. 2.3). Five of the 7 insertion points are grouped in 2 "clusters" in this region. "Cluster 1" encompasses 3 insertion points within 271 bp of each other (ISL0480.11, ISL7031.8 and an H37Rv insertion) and in "cluster 2" 2 insertion points are separated by 295 bp (DK6, and a second H37Rv insertion) (Fig. 2.3). This data suggests the presence of a newly identified insertional hotspot, which is region-specific, rather than base-specific.

Figure 2.2 (Legend). The positions of the proposed origin of replication and the DR region provide reference points. The positions of H37Rv insertions are depicted inside the circle, while all other insertions are indicated on the outside. The insertion sequence IS*1547*, which contains the *ipl* locus, maps to two positions (Rv0797 and Rv3327), as indicated, with the positions of insertions occurring in this locus only depicted at the Rv3327 position. The position and orientations of the 9 unique insertions into the *ipl* locus are shown, with the → indicating IS*6110* in a 5' to 3' orientation. The newly identified preferential insertion region which spans 9459 bp, is shown in detail, with the position and orientation of 7 IS*6110* insertions in this region. The relative positions and Rv designations of disrupted genes are given. The insertions indicated are as follows: A= ISL7031.2 (Table 2.1); B = ISL0480.11 (Table 2.1); C = H37Rv IS*6110* element (Rv1756c/Rv1757c) (Table 2.2); D = ISL7031.8 (Table 2.1); E = DK6 (Fomukong *et al.*, 1998) (Table 2.4); F = H37Rv IS*6110* element (Rv1763/Rv1764) (Table 2.2) and G = ISL7031.10 (Table 2.1). Insertion A interrupts Rv1754c, B and C interrupt Rv1755c, C and D interrupt Rv1758, E interrupts Rv1762c, F interrupts Rv1765c and G is located within an intergenic region. (The two H37Rv insertions in this region are duplicated inside the circle, for reference purposes.)

Figure 2.2 Graphic representation of the *M. tuberculosis* genome, demonstrating the distribution of insertion loci identified in this study. (See opposite page for legend)



2.4 Discussion

The transposable element IS*6110*, which is routinely employed for the genotyping of *M. tuberculosis* isolates, has recently been proposed to be an agent of adaptive mutation in the *M. tuberculosis* genome (Sreevatsan *et al.*, 1997). To investigate this hypothesis, knowledge of the genetic identity of sequences flanking the element is essential. To address this, we have cloned and characterized 33 IS*6110* insertion loci from 8 clinical isolates of *M. tuberculosis* isolated from a high incidence South African community. The cloned insertion loci were investigated in conjunction with 43 insertion loci identified in the literature and DNA sequence databases. The sequence data was analyzed to determine the genetic identity of regions disrupted by IS*6110* transposition and the genomic distribution of unique insertion loci.

2.4.1 Genetic identity of regions disrupted by IS*6110* transposition

The impact of any insertion on strain phenotype will be dependent on the nature of the domain into which it inserts. Seventeen insertions into intergenic regions have been identified in this study. Investigation of transposition of other insertion sequences into non-coding regions has demonstrated the up-regulation of gene expression (Charlier *et al.*, 1982; Camarena *et al.*, 1998). However, no promoter sequences have been identified in IS*6110*. It is possible that IS*6110* insertion may disrupt existing promoter regions, leading to down-regulation of gene expression, but this remains speculative.

Forty-two of the 66 unique insertion loci analyzed demonstrate disruption of coding regions by IS*6110* insertion. Nine insertions occur within genes coding for hypothetical proteins of unknown function, 10 within hypothetical ORFs with a suggested function (by homology) and 23 occur within coding regions for which a function has been assigned (as defined by Cole *et al.*, 1998). Included in the latter group are insertions into the newly identified insertion sequence IS*1547*, as described by Fang *et al.* (1999b).

2.4.1 (a) Disruption of other insertion sequences:

Insertion sequences are non-essential regions of the *M. tuberculosis* genome. From the evolutionary viewpoint of IS*6110*, transposition of IS*6110* into other insertion sequences may allow the perpetuation of the IS*6110* element without reduction in fitness of the host bacillus. This may explain the occurrence of preferential insertion into the *ipl* locus, which contains the insertion sequence IS*1547* (GenBank accession number Y16254) (Fang and Forbes, 1997; Fang *et al.*, 1999b). We have identified a unique insertion point in the *ipl* locus, defined by the clone ISL0480.8.

2.4.1 (b) Disruption of genes with no resemblance to known proteins:

Nine insertion loci demonstrate disruption of ORFs with no predicted function. The fact that insertions have been identified within these ORFs indicates that the bacillus can continue to infect patients and cause disease without these gene products. However, with no knowledge of their function, it is not possible to predict the impact of disruption of these regions.

2.4.1 (c) Disruption of genes of known function, predicted by database comparisons:

Numerous examples of IS*6110* insertion into *M. tuberculosis* genes have been identified in this study. Disruption of essential single-copy genes will result in non-viable organisms, which will not be detected. However, knockout of non-essential genes will potentially result in viable organisms with an altered phenotype. The nature of any alteration of phenotype will depend on the function of the products of the genes involved. The 30 disrupted genes identified here code for a variety of proteins. Their suggested functions include cell wall components, proteins involved in energy metabolism, macromolecule metabolism, detoxification and regulation (see results and Tables 2.1-2.4). The potential impact of IS*6110* insertions on strain phenotype is therefore diverse. However, it should be noted that analysis of the *M. tuberculosis* genome demonstrates possible genetic redundancy (Cole *et al.*, 1998), which may suggest a mechanism whereby the bacillus is able to minimize the impact of deleterious mutation. The majority of genes disrupted by IS*6110*, as identified in this study, demonstrate functional redundancy. For example, multiple open reading frames coding for conserved small membrane proteins, phospholipase C, cytochrome p450, cutinases, kinases, polyketide synthases, adenylate cyclases and regulatory proteins have been identified in the *M. tuberculosis* genome (Cole *et al.*, 1998). This supports the notion that multiple copies of a gene could compensate for natural knockouts created by IS*6110* insertion, with unknown consequences for fitness of the organism.

Sequence analysis has revealed 8 examples of disruption of 5 members of the PPE gene family (see 2.3, and Tables 2.1-2.4). It has been suggested that this gene family, together with the PE gene family, could represent a source of antigenic variation in *M. tuberculosis* (Cole *et al.*, 1998). Although little evidence for the function of the PPE gene products exists, a recent report suggests expression of at least one member of this gene family (Rivera-Marrero *et al.*, 1998). Only with a clear understanding of the function of this intriguing gene family will it be possible to speculate on the impact of IS*6110* insertion. However, continuing on the theme of genetic redundancy, it should be noted that this is a large family of genes (68 members) (Cole *et al.*, 1998). Therefore, disruption of one of its members would not be expected to produce severely disadvantaged organisms.

This study has demonstrated that IS*6110* transposition into coding regions is more common than was previously suggested, on the basis of analysis of H37Rv genome sequence data alone. However, the issue of whether the adaptive or selfish DNA theories play a role in the evolution of the *M. tuberculosis* genome remains a complex one. The relatively long evolutionary history of association of IS*6110* with the genome suggests that the organism may gain some benefit from maintenance of the element in its chromosome, favouring the adaptive theory. However, a complicating factor is the proliferation of strains with no or low copy numbers of IS*6110* elements in some areas (Yuen *et al.*, 1993). One might argue that this suggests that IS*6110* is a purely selfish genetic element, as these strains do not appear to be disadvantaged by the lack thereof. To counter this argument is the hypothesis that low copy number and high copy number strains have arisen as separate lineages (Fomukong *et al.*, 1998), which might suggest that they have evolved distinct mechanisms of adaptation and evolution. Continued study of the impact of IS*6110* insertions on strain phenotype will be necessary to answer this question. However, to assess the potential role of IS*6110* in conferring any selective advantage, it will be necessary to consider a study design based on epidemiological data. The clonal proliferation of a strain family (> 65% related as determined by GelCompar analysis of IS*6110* DNA fingerprints) with a particular IS*6110* insertion may reflect the importance of that insertion in conferring a "more fit" phenotype. (Alternatively, the IS*6110* DNA fingerprint pattern may simply be a co-incidental marker of some other genomic change conferring a selective advantage.) To elucidate the impact, if any, of insertions on strain phenotype, it is proposed that future studies focus on the potential functional impact of specific insertions found in dominant or rapidly proliferating strain families within a community.

2.4.2 Genomic distribution of IS*6110* insertion loci

To investigate the distribution of IS*6110* insertion loci within the *M. tuberculosis* genome, the positions of 66 unique sites were mapped relative to the H37Rv complete genome sequence (Fig. 2.2). Seven insertion loci could not be mapped to the H37Rv genome, suggesting deletion of these regions from clinical isolates. This may implicate the IS*6110* element in genome rearrangements, consistent with the hypothesis of Fang *et al.* (1999a).

In agreement with previous studies of chromosomal distribution of IS*6110* (Cole *et al.*, 1998; Philipp *et al.*, 1996), a concentration of insertion loci in the genome quadrants flanking the DR region of the chromosome (within 90° of the DR locus) is observed (Fig. 2.2). However, where previous reports (based on the assessment of a limited number of insertion sites) have suggested that there is selection against insertion into particular regions of the genome (for example, the oriC region) (Cole *et al.*, 1998), the collated map reveals a distribution of insertion loci throughout the *M. tuberculosis* genome, challenging this assumption.

In agreement with other studies that have suggested the presence of insertional hotspots (Hermans *et al.*, 1991; Fang and Forbes, 1997; McHugh and Gillespie, 1998), the clustering of insertions within specific loci is observed. Previous studies of chromosomal distribution of IS*6110* (Cole *et al.*, 1998; Philipp *et al.*, 1996) have noted a concentration of insertion loci in the genome quadrants flanking the DR region of the chromosome (within 90° of the DR locus). This result is supported by the data illustrated in Figure 2.2. Other regions also demonstrate concentration of insertions, with some of these possibly representing preferential insertion loci. Preferential insertion loci have been defined here as relatively small genomic domains (<1000 bp) where multiple insertions occur at multiple positions, or at identical positions (with the *proviso* that the insertions in question are not inherited by descent). The definition of preferential insertion used here includes the *ipl* locus (Fang *et al.*, 1997). The identification of two clusters of IS*6110* insertion loci in the region of the genome spanned by Rv1754c to Rv1762 (Fig. 2.2) suggests the presence of further preferential insertion loci.

In summary, the results reported here demonstrate an unexpectedly high frequency of IS*6110* insertion into coding regions. This suggests the potential for the element to play a role in the evolution of *M. tuberculosis* genome, possibly towards a subtly improved phenotype, allowing continued survival of both *M. tuberculosis* and the insertion sequence. However, due to the presence of multi-gene families, it is predicted that the IS*6110*-mediated coding region disruptions identified here will have unknown impact on strain phenotype. Elucidation of the potential impact of IS*6110* insertion into coding regions of *M. tuberculosis* will require an in-depth functional analysis of the disrupted genes. To further elucidate the relationship between insertion into specific loci and the impact on strain phenotype, it will be necessary to identify and characterize specific insertion loci, selected on the basis of epidemiological data. The application of insertion site data to the investigation of IS*6110* insertions that are inherited by descent could provide an understanding of inter-strain evolutionary relationships. Finally, a useful tool for future studies of IS*6110* insertion sites would be a web-based genome map of all known IS*6110* insertion sites.

2.5 Addendum

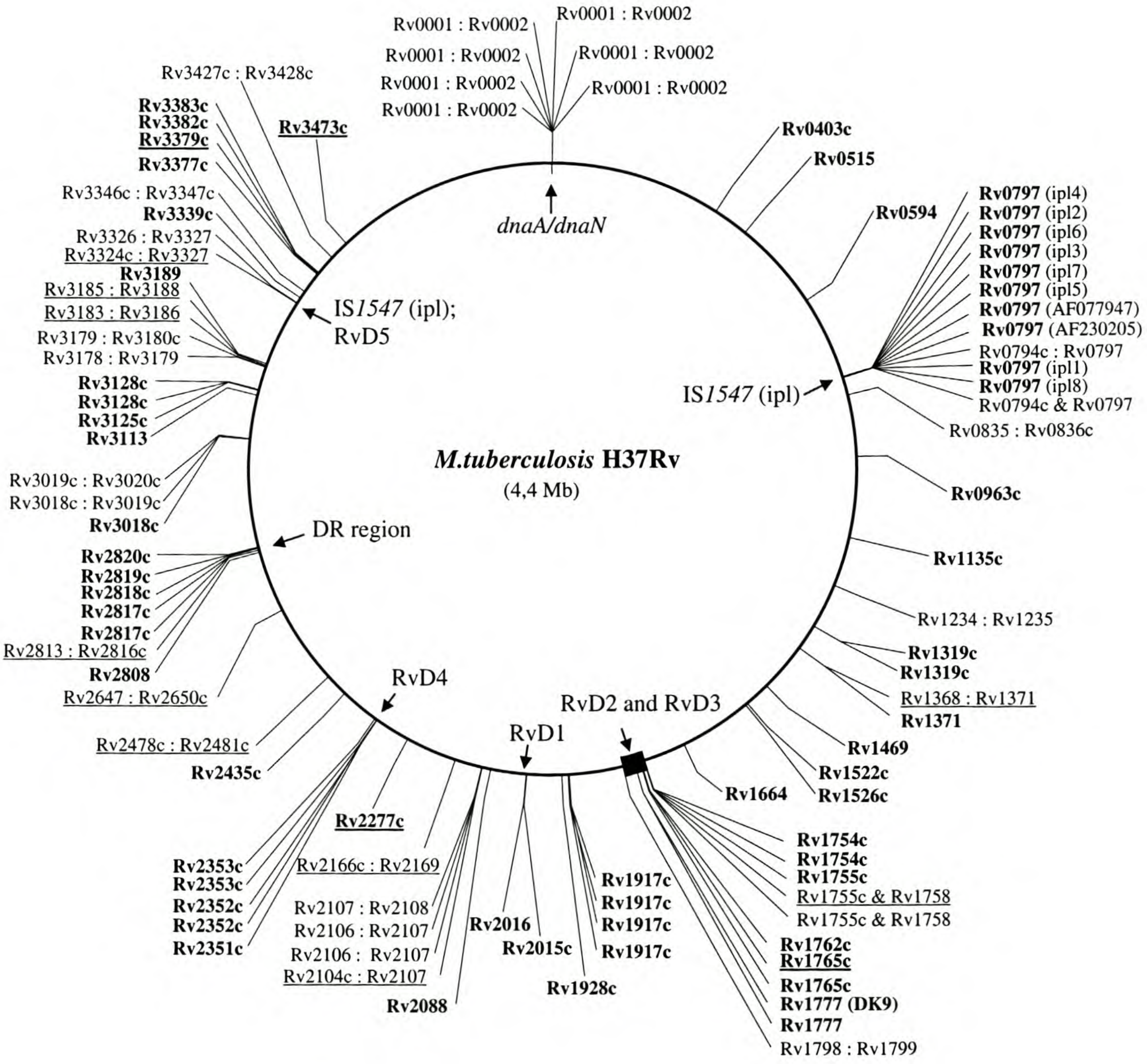
The following correspondence (which represents an update of the data described in the preceding chapter) was published in The Journal of Clinical Microbiology *in response to a report by Benjamin* et al. *(2001), as "IS*6110 *insertions in* Mycobacterium tuberculosis: *predominantly into coding regions. S. Sampson, R. Warren, M. Richardson, G. van der Spuy, P. van Helden, 32:3423-3434" The style of the text and numbering of sections has been altered to conform to the style of this thesis. Literature cited in the text takes the form of author name and year of publication as opposed to the number format specified by* The Journal of Clinical Microbiology. *All cited literature is compiled into a single list at the end of the thesis for ease of reference. No other changes have been made to the body of the text.*

We read with interest the recent publication by Benjamin *et al.*, (2001), regarding the characterization of IS*6110* insertion sites in the direct repeat (DR) region of *Mycobacterium tuberculosis*. This topical and relevant study described the dissection of a single molecular event leading to an altered DNA fingerprint pattern that was detected by two different strain genotyping systems, namely, IS*6110* DNA fingerprinting and spoligotyping. It was shown that a single IS*6110* transposition event in the DR region disrupted one of the spoligotyping primer regions, thereby resulting in simultaneous changes in the IS*6110* fingerprint and spoligotype pattern. The authors rightly recommended that some caution should be applied to the interpretation of similar, but non-identical, IS*6110* DNA fingerprint and spoligotype patterns. In addition, the results corroborated the occurrence of IS*6110* preferential integration loci, as previously identified by a number of investigators (Fang and Forbes, 1997; Sampson *et al.*, 1999; Warren *et al.*, 2000).

However, we would like to question the statement by the authors that "nearly all IS*6110* insertions are between open reading frames…" The assertion was not referenced, but we assume that this statement was occasioned by the original reporting of IS*6110* distribution in the reference strain *M. tuberculosis* H37Rv (Philipp *et al.*, 1996), which states that the majority of insertions occur within noncoding regions. However, this early conclusion was restricted to the analysis of a very limited number of insertion sites in *M. tuberculosis* H37Rv. Since this preliminary characterization, numerous investigators (ourselves included) have characterized IS*6110* integration loci in clinical isolates (Beggs *et al.*, 2000; Fang and Forbes, 1997; Sampson *et al.*, 1999; Warren *et al.*, 2000). Our analysis of insertion locus sequence data collated from various sources demonstrated that 33 of 57 discrete IS*6110* insertion sites (58%) occurred within coding regions of the *M. tuberculosis* genome (Sampson *et al.*, 1999). We performed an updated analysis (inclusive of data published in the literature and DNA sequence databases since our original report), the results of which demonstrated that of 95 discrete IS*6110* integration loci identified (excluding multiple insertions into the DR and *ipl* locus), 60% (57 of 95 sites) occur within coding regions, 33% (31 of 95 sites) occur within intergenic regions (Fig. 1), and 7% (7 of 95 sites) did not map to the *M. tuberculosis* H37Rv genome. This strengthens our conclusion (Sampson *et al.*, 1999) that IS*6110* frequently disrupts coding regions. In addition, the revised map (Fig. 1) of the genome distribution of IS*6110* insertions highlights the existence of numerous preferential integration sites in addition to those previously described (Benjamin *et al.*, 2001; Fang and Forbes, 1997; Sampson *et al.*, 1999; Warren *et al.*, 2000).

In the light of the limited structural gene diversity within *M. tuberculosis* strains, it has been suggested that IS*6110* may be capable of driving genome evolution (Sreevatsan *et al.*, 1997). Therefore, while the biological significance of our finding remains to be elucidated, it nonetheless deserves further attention, as it clearly demonstrates the potential for the IS*6110* element to impact on strain phenotype by gene disruption. Finally, our approach highlights the value of collation of data from various sources, and we advocate the establishment of a web-based genome map of IS*6110* integration sites to facilitate this endeavor.

Figure 2.3. Distribution of IS*6110* insertion loci relative to open reading frames (ORFs) of the *Mycobacterium tuberculosis* H37Rv genome. The ORF numbers of the disrupted genes are indicated in bold, while the intergenic insertions are represented by two ORF numbers separated by a colon. The *M. tuberculosis* H37Rv insertion sites are underlined. Multiple integration sites within a single gene or intergenic region are denoted by multiple entries with the same name. The relative positions of the DR region and the *ipl* locus are shown, and the positions of regions deleted from *M. tuberculosis* H37Rv relative to *M. bovis* are indicated by their RvD numbers.



CHAPTER 3

IS*6110* AND GENOME EVOLUTION

NOTE: The results presented in the following chapter were published in full as: "**Mapping of IS*6110* flanking regions in clinical isolates of *Mycobacterium tuberculosis* demonstrates genome plasticity.** Warren, R.M., Sampson, S.L., Richardson, M., Van Der Spuy, G.D., Lombard, C.J., Victor, T.C., Van Helden, P.D. *Molecular Microbiology* (2000) 37(6):1405-1416."

(The style of the text and numbering of sections has been altered, and all references to the M. tuberculosis *strain "CSU93" have been replaced with "CDC1551" to conform to the style of this thesis. All cited literature is compiled into a single list at the end of the thesis for ease of reference. No other changes have been made to the body of the text.)*

3.1 Introduction

The insertion element IS*6110* was first described in *Mycobacterium tuberculosis* in 1990 (McAdam *et al.*, 1990; Thierry *et al.*, 1990a). This element is also present in the genomes of other members of the *M. tuberculosis* complex (Dale, 1995), suggesting that *M. tuberculosis* inherited the IS*6110* insertion element from an ancestral precursor organism. The identification of an IS*6110* element within an identical position in the direct repeat (DR) region of the ancestrally related *M. tuberculosis* and *M. bovis* strains has been hypothesized to reflect the ancestral insertion (Dale, 1995). Analysis of the H37Rv genome sequence has revealed that the 16 IS*6110* insertions are not evenly distributed throughout the chromosome, leading to the assumption that the IS*6110* element has migrated outwards from the DR region by replicative transposition (Philipp *et al.*, 1996). The variability in IS*6110* copy number seen in clinical isolates together with the presumed random nature of IS*6110* transposition has led to the establishment of a standardized DNA fingerprinting method to classify clinical isolates of *M. tuberculosis* for epidemiological analysis (van Embden *et al.*, 1993; van Soolingen *et al.*, 1994). From these molecular epidemiological studies it is evident that extensive IS*6110* restriction length fragment polymorphism (RFLP) exists between clinical isolates (Alland *et al.*, 1994; Small *et al.*, 1994; Warren *et al.*, 1996a). These results contrast with comparative DNA sequence data which showed an extremely low level of synonymous substitution in 26 genes (Sreevatsan *et al.*, 1997). This has led to the proposal that IS*6110* transposition may play an important role in *M. tuberculosis* evolution by altering gene expression (Sreevatsan *et al.*, 1997).

In order to establish the possible impact of IS*6110* insertion on gene function, the chromosomal positions of inserted IS*6110* elements were identified in clinical isolates of *M. tuberculosis* by sequencing the 3' flanking regions (Sampson *et al.*, 1999). Comparison of these sequences with the annotated H37Rv sequence (Cole *et al.*, 1998) identified "natural knockouts" of predicted open reading frames (ORFs). The results demonstrated that 64% of IS*6110* insertions occur within ORFs (Sampson *et al.*, 1999). However, the impact of these insertions on phenotype remains unclear as many of the disrupted genes have a function similar to other genes present in the genome.

Sequencing of the chromosomal domains flanking IS*6110* elements has also identified chromosomal regions where more than one insertion has occurred (Fang and Forbes, 1997; Kurepina *et al.*, 1998; Sampson *et al.*, 1999). This has led to the hypothesis that these domains reflect preferential integration regions. To date, 3 preferential integration regions have been described, namely the *ipl* locus (Fang and Forbes, 1997), the *dna*A-*dna*N intergenic region (Kurepina *et al.*, 1998), and the region between Rv1754c and Rv1765c (Sampson *et al.*, 1999). However, the number of preferential integration regions in the chromosome of *M. tuberculosis* has not been established. Furthermore, no studies have been performed to map the structure of the chromosomal domains surrounding IS*6110* insertion elements, in different clinical isolates, to determine whether these domains have remained conserved over the evolutionary history of *M. tuberculosis.*

The aim of this study was to analyze the chromosomal domains flanking IS*6110* elements in clinical isolates of *M. tuberculosis*, to improve our understanding and interpretation of the genotyping patterns generated by IS*6110* hybridization. Probes homologous to the 3' flanking regions of IS*6110* elements identified in five clinical isolates (Sampson *et al.*, 1999), were back-hybridized onto *Pvu*II-digested DNA, from 34 clinical isolates of *M. tuberculosis*. The RFLP hybridization data generated by each probe were analyzed in conjunction with genome sequence data to determine the chromosomal position and orientation of inserted IS*6110* elements, as well as to identify chromosomal mutations and deletions. To establish the evolutionary history of the analyzed chromosomal domains, the RFLP data was subjected to parsimony analysis to reconstruct a phylogenetic tree. The results presented provide new insights into the evolution of *M. tuberculosis* and challenge previous assumptions of genome stability.

3.2 Methods

3.2.1 Purification of insertion locus probes

Each ISL clone, representing a different *Pvu*II fragment encompassing the 3' domain of the IS*6110* element and flanking chromosome cloned from different clinical isolates (Sampson *et al.*, 1999) was digested with *Xho*I and *Eco*RI and fractionated by preparative gel electrophoresis in 1% agarose (1x TBE, pH 8.3). The fragment corresponding to the 3' flanking chromosomal domain (including 89 bp of the 3' terminal domain of the IS6110 element) or a domain downstream to the IS*6110* element were identified under UV after staining with ethidium bromide, excised and purified using a Talent Cleanmix kit (Italy) (see Table 3.1).

3.2.2 RFLP analysis

DNA from 33 clinical isolates and the standard strain Mtb14323 (Van Embden *et al.*, 1993) was digested with *Pvu*II and electrophoretically fractionated together with 8.0 ng Marker X (Roche) in 0.8% agarose, as described previously (Warren *et al.*, 1996a). The DNA was Southern transferred onto a Hybond N$^+$ membrane (Amersham) and fixed by heating at 80°C for 2 h. Prior to each hybridization step, the membranes were incubated in 0.4 M NaOH at 50°C for 45 min. to denature all DNA hybrids. The membranes were then neutralized in 0.2 M Tris-HCl pH 8.0, 0.1x SSC and 0.1x SDS at 50°C for 15 min. Each membrane was sequentially hybridized, according to the protocols described by the manufacturer (Amersham) with the ECL-labeled IS-3' (van Embden *et al.*, 1993; van Soolingen *et al.*, 1994), IS-5' (complementary to the 5' IS*6110* domain between nucleotides 77 and 462), Marker X and ISL probes (n = 25). The resulting RFLP's were visualized by autoradiography. Each autoradiograph was scanned and the image was normalized according to the position of the internal molecular weight marker bands (Marker X), using GelCompar

Table 3.1. Probes used for Southern hybridization.

Probe	Size (bp)	Regions of homology* [cosmid and position (bp)]	ORF
480.3	3583[a]	**MTCY180** (39930-43424) MTY13e10 (24195-24599) MTCY3C7 (11832-12056)	Rv1917c[b] Rv0355c[b] Rv3533c[b]
480.6	991[a]	**MTCY164** (34221-35123)	Rv3125c; Rv3126c; Rv3127[b]
480.8	1466	**MTV016** (22454-23920) MTCI429B (198-927)	Rv3327[b] Rv0797[b]
480.9	579[a]	**MTCY275** (6895-7385)	Rv1664[b]
480.11	402[a]	**MTCY28** (19490-19803) MTCY98 (20647-20887)	Rv1755c[b] Rv2350c[b]
540.1	4186[a]	**MTCY9F9** (2855-6952)	Rv1924c; Rv1925; Rv1926c Rv1927; Rv1928c[b]
540.3	2514[a]	**MTCY16B7** (20783-23208)	Rv2818c; Rv2819c; Rv2820c[b]
540.8	544[a]	**MTCY19H5** (9050-9505)	Rv0594[b]
540.10	375[a]	**MTV006** (622-908)	Rv1234; Rv1235[b]
540.11	253[a]	**MTV004** (528-692)	Rv3347c[b]
715.5	191[a]	**MTCY428** (13322-13424)	Rv2435c[b]
715.8	260[a]	**MTV029** (1592-1759)	intergenic (Rv0001-Rv0002) [b]
852.1	381[a]	**MBO18605** (163-642)	RvD1-ORF1; RvD1-ORF2[c]
7031.2	2208	**MTCY28** (16055-18263)	Rv1753c; Rv1754c[b]
7031.3	2324[a]	**MTCY164** (23068-25303) MTV012 (37524-38805) MTCY10G2 (36016-37296) MTCY07A7 (19913-21194) MTCI364 (7578-8858)	Rv3113; Rv3114; Rv3115[b] Rv3032c[b] Rv1046c[b] Rv2512c[b] Rv1199c[b]
7031.4	1825	**MTCY98** (26394-27130) MTCY98 (29960-31081)	Rv2353c[b] Rv2356c[b]
7031.5	786[a]	**MTV016/MTMOAIS** (20528-20939; 1-286)	Rv3324c[b]; *moaA*[c]
7031.6	917[a]	**MTV043** (27682-28510)	Rv0835[b]
7031.7	989[a]	**MTCY130** (8127-9027) MTCY130 (6460-6624) not found in H37Rv	Rv1319c[b] Rv1318c[b] No homology found
7031.8	1496[a]	**MTU242907** (7574-8981)	RvD2-ORF3[c]; Rv1758[b]
7031.9	1407[a]	**MTV018/MTCY39** (1-485; 37667-38500) MTCY28 (29799-29976)	Rv2013; Rv2014; Rv2015c[b] Rv1765c[b]
7031.10	1268[a]	**MTCY28** (31138-32317)	Rv1766[b]
7031.11	1013[a]	**MTV014** (27991-28915)	Rv3182; Rv3183[b]
7031.12	948[a]	**MTC25C11** (2298-3157)	Rv1776c; Rv1777[b]
7031.13	579[a]	**MTCY98** (28483-28973) MTCY98 (31069-31166)	Rv2353c[b] Rv2356c[b]

a. Probe fragments include 89 bp of the 3' terminal domain of IS*6110*

b. Rv designation according to Cole *et al.*, 1998

c. ORF designation according to Gordon *et al.*, 1999a

d. According to Fang *et al.*, 1999a

(Cosmids in normal font represent duplicated chromosomal domains.)

4.0 software. The size of each hybridizing fragment was calculated according to the internal markers run within each lane using GelCompar 4.0 software. The reproducibility of this technique was calculated by comparing the IS-3' fragment lengths of the standard strain Mtb14323 (Van Embden *et al.*, 1993) which was run in the two outer lanes on each gel. The standard deviation of fragment lengths fractionated on different gels was calculated to be 25 bp for fragments in the molecular range 1000 to 8000 bp, while a standard deviation in fragment length within one gel was calculated to be 10 bp.

3.2.3 IS*6110* orientation

The ancestral chromosomal ISL hybridizing fragment was assigned based on the assumption that such a fragment would be electrophoretically conserved in most of the isolates analyzed and was not disrupted by IS*6110* insertion. ISL hybridizing bands which cohybridized (identical band position) to either the IS-3' and/or IS-5' hybridizing band/s were assigned as domains containing an IS*6110* element. The orientation of the IS*6110* element in relation to the cloned probe was determined by the IS (3' or 5') probe to which it cohybridized. The absence of a hybridization signal was inferred to reflect the deletion of the ISL hybridizing domain, while an electrophoretic shift of an ISL hybridizing band (not containing an IS*6110* element) was interpreted as chromosomal mutation (including insertions, deletions and point mutations). The identification of more than one hybridizing band implied genome duplication, which was confirmed by analysis of the H37Rv genome sequence using the BLASTN algorithm (http://www.ncbi.nlm.nih.gov). The duplicated chromosomal domains were differentiated either by hybridization signal intensity or by Southern hybridization with probes generated by PCR amplification of the domain adjacent to the duplicated sequence (see Table 3.2 for primers).

3.2.4 Mapping of IS*6110* insertion points

The fragment length for each hybridizing band was calculated using GelCompar 4.0 software. For those bands which cohybridized to either the IS-3' or IS-5' probes, the fragment lengths were adjusted by subtracting the IS*6110* contribution (895 bp and 459 bp for the IS-3' and IS-5' domains, respectively). The resulting fragment lengths indicate the approximate position of the IS*6110* element in relation to the terminal *Pvu*II restriction site of each locus. The accuracy of this technique was determined by comparing the ISL fragment lengths with the corresponding *Pvu*II fragments derived from the H37Rv sequence. In cases where the ISL probes hybridized to adjacent regions in the same chromosomal domain, the insertion point detected by both probes was only scored once. To confirm the insertion point in 8 of the highly polymorphic insertion loci, the domains between the inserted IS*6110* elements' terminal repeat and a specific chromosomal sequence within the insertion locus were PCR amplified using the IS-primer together with one of the ISL specific primers (Table 3.3). The PCR reactions were carried out in a total volume of 50 µl, containing 0.5 µg chromosomal DNA, enzyme buffer, 1.5 mM $MgCl_2$, 0.4 mM dNTP's, 50 pmol of each primer and 2.5 U *Taq* DNA polymerase (Promega). Amplification was performed for 35 cycles of 93°C for 1 min., 62°C (or as in Table 3.3) for 1 min., and 72°C for 1 min. After the last cycle the samples were incubated at 72°C for 10 min. Amplified products were electrophoretically fractionated in 5% polyacrylamide (Sambrook *et al.*, 1989), and visualized by staining with ethidium bromide. Identical IS*6110* insertion points were assigned on the basis of identical PCR fragment lengths, together with IS-3' and IS-5' cohybridization data (IS*6110* orientation). Chromosomal mutation in the domain surrounding an inserted IS*6110* was identified when the PCR products generated from different clinical isolates were of equal length, while the ISL RFLP hybridizing bands showed differing lengths.

Table 3.2. Oligonucleotide primer sequences used to PCR amplify chromosomal domains adjacent to duplicated domains detected by hybridization with the ISL probes.

Cosmid	Primer name	Primer sequence	Tm
MTCY10G2	10g2L	5'-TGCCACGTCGGTGAGATGT -3'	55°C
	10g2R	5'-CCTGGACGTAACTGCTGA-3'	
MTC1364	364L	5'-AACACCTGCCACGACGATG-3'	60°C
	364R	5'-TGCTTTCGCCAATGGTGTAG-3'	
MTV012	012L	5'-AGTCGCCGGGTTTCTACGAG-3'	62°C
	012R	5'-ATGGTCGTCCGCGCTATGT-3'	
MTCY28	28L	5'-GCTCTGCATCGCTGACATTG-3'	62°C
	28R	5'-AAGCCGATGCCCTCAAAGC-3'	

Table 3.3. Oligonucleotide primer sequences used to PCR amplify the domain between the IS*6110* element and the insertion locus.

Primer name	Primer sequence
IS primer	5'-GGACTCACCGGGGCGGTTC-3'
ISL480.8	5'-CCCAGATCAGCTCGAGGCCG-3'
ISL480.11	5'-CAGGCACCGTGACATAATCG-3'
ISL7031.2	5'-ATGAGTCCAATAGCGGCCGCC-3'
ISL7031.3	5'-CAGTTCTGCCAGCATCATGG-3'
ISL7031.4	5'-TAGGCGGAGCCGTTGAGG-3'
ISL7031.5	5'-CGCGGACAGGATGCCTAC-3'
ISL7031.8	5'-TATCGGATGGCGCGTGTCCC-3'
ISL7031.13	5'-TTTCGACCCGTCACCAAGC-3'

In the absence of cohybridization to either the IS-3' or IS-5' probes, changes in fragment lengths were calculated and compared to the H37Rv sequence using DNAMAN software to identify the possible mutational mechanisms. Mutation in the terminal *Pvu*II site was assigned if the increase in fragment length corresponded in length to the adjacent domain up to the downstream *Pvu*II site. If the change in fragment length corresponded to the length of repeat sequences within the ISL domain, mutation was assumed to have occurred via either expansion or contraction of these repeat sequences. In the absence of identified repeat sequences, a decrease in fragment length was assumed to reflect a mutation leading to the creation of an additional *Pvu*II site or small deletions.

ORF's situated within the deleted domains were identified by aligning the probe sequences with the annotated H37Rv (Cole *et al.*, 1998), RvD1, RvD2 (Gordon *et al.*, 1999a) and MTMOAIS (Fang *et al.*, 1999a) sequences. Similarly, insertions within predicted ORF's were identified by aligning the ISL fragment lengths against the sequence of H37Rv (Cole *et al.*, 1998), RvD1, RvD2 (Gordon *et al.*, 1999a), MTMOAIS (Fang *et al.*, 1999a) and CDC1551 (preliminary sequence data of CDC1551 was obtained from The Institute for Genomic Research website at http://www.tigr.org). Sequence comparison was carried out using the BLASTN algorithm (www.ncbi.nlm.nih.gov).

3.2.5 Phylogenetic analysis

RFLP position data for each ISL were transferred from GelCompar into a Microsoft Excel spreadsheet. The data were ordered as follows: Each isolate was assigned to a row, while the band positions were represented in columns. Fragments identified in different isolates and which showed identical mobility, were aligned in the same column, while bands with unique positions were assigned to individual columns as a function of eletrophoretic mobility. Bands which failed to cohybridize to either the IS-3' or IS-5' probe, and which were conserved in the majority of the isolates, were assigned as the ancestral state "0". Polymorphic variants of these fragments were assigned as the evolved state "1". Bands which cohybridized with either the IS-3' or IS-5' probes, were assigned as the evolved state "1", based on the assumption that the ancestral insertion occurred in only the direct repeat (DR) region (Dale, 1995). Each evolved state was dependent on both the orientation and position of the IS*6110* insertion element within the locus. However, if the ISL probe cohybridized to both the IS-3' and IS-5' bands (indicating an IS*6110* insertion into the central region of the ISL) the evolved state "1" was only recorded once, to ensure that all characters evolved independently. In addition, identical insertions or chromosomal mutations detected by different ISL probes were only scored once. Chromosomal mutations adjacent to an inserted IS*6110* (see above) were assigned as an evolved state "1" and recorded in a separate column. Similarly, the absence of a hybridization signal (deletion) was assigned as an evolved state ("1") and recorded in separate column.

Phylogenetic reconstructions were performed using the heuristic algorithm (PAUP* 4.0; version 4, Sinauer Associates) including the following assumptions: (1) chromosomal mutations and IS*6110* transpositions could be either gained or lost (0 ↔ 1), and (2) deletions are irreversible (0 → 1). Bootstrap analysis was performed to establish a confidence interval to support the internal topology of the tree (Felsenstein, 1985). A consensus tree was generated using the majority rule formula. All branches with a zero branch length were collapsed. The output files were analyzed to identify the mutational events represented at each node and to identify convergent evolutionary events (homoplasy index).

The sequences of the *katG* codon 463 and *gyrA* codon 95 were determined using a dot blot hybridization assay (Victor *et al.*, 1999).

3.2.6 Statistical analysis

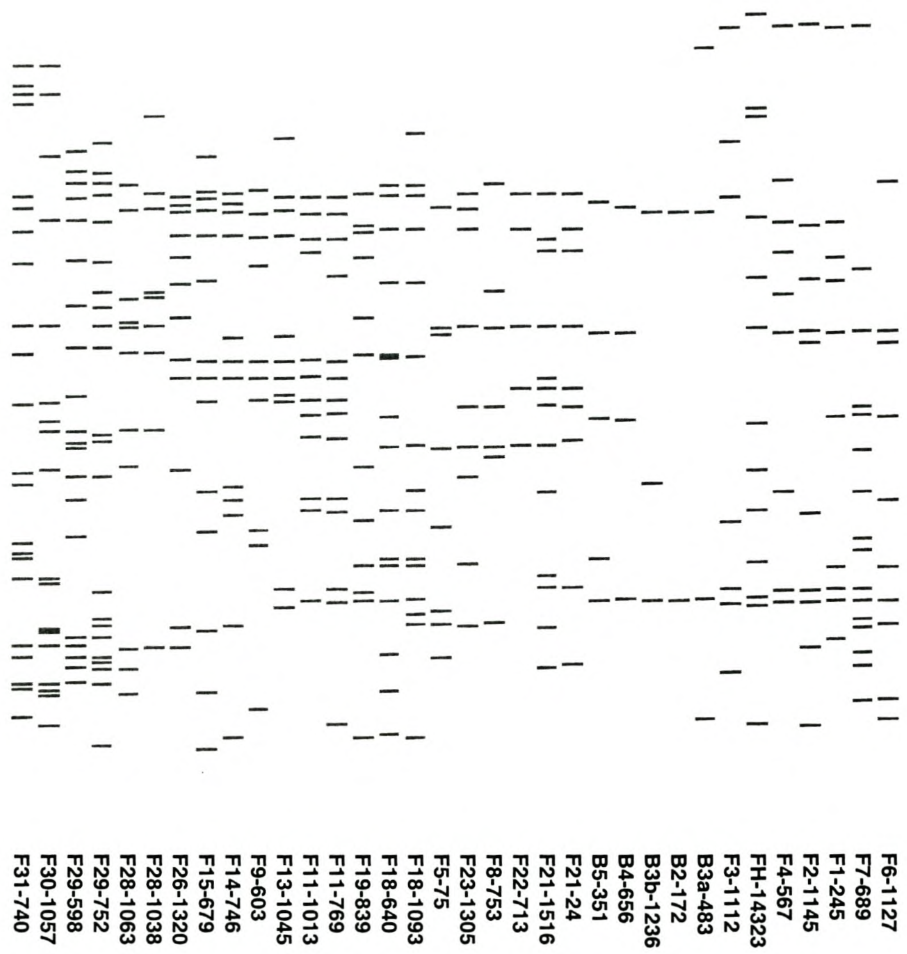
To assess the association between the different genetic factors, Spearman correlation coefficients were calculated and tested using exact inference of the test statistic. P-values for two-sided tests are reported and a value of 0.05 was used as the reference level for significance.

3.3 Results

3.3.1 Southern hybridization analysis

IS*6110* RFLP analysis of *M. tuberculosis* isolates collected from patients residing in two adjacent suburbs of Cape Town, South Africa, revealed that the isolates were clustered into a number of defined strain family groupings [grouped according to a similarity index of > 65%, as calculated using the Dice coefficient and UPGMA (unweighted pair group method using arithmetic averages) clustering method] (Warren *et al.*, 1996a; Warren *et al.*, 1999). Each strain grouping may represent an evolutionary lineage which has evolved independently from a common ancestral strain (Kremer *et al.*, 1999). Thirty-three clinical isolates representing these different strain groupings were selected for further Southern hybridization analysis using cloned chromosomal domains homologous to the 3' flanking regions adjacent to IS*6110* elements isolated from five clinical isolates [Insertion locus (ISL) probes] (Sampson *et al.*, 1999) (Table 3.1). Characterization of the 33 clinical isolates using the probes IS-3' (van Embden *et al.*, 1993; van Soolingen *et al.*, 1994) (Fig. 3.1) and IS-5' (complementary to the 5' IS*6110* domain between nucleotides 77 and 462) (data not shown), showed that the IS*6110* copy number for the selected isolates ranged between 2 and 22 insertions per genome. Five of the isolates represented low copy number strain family groupings (< 6 IS*6110* insertions), while 28 isolates represented 24 high copy number strain family groupings (≥ 6 IS*6110* insertions) (Fig. 3.1). Also included in the analysis was the reference strain Mtb14323, a clinical isolate which originates from the Netherlands (van Embden *et al.*, 1993; van Soolingen *et al.*, 1994).

Figure 3.1. IS-3' banding pattern of clinical isolates of *M. tuberculosis* generated by GelCompar software. Each lane represents a single isolate of *M. tuberculosis*. The isolate number is preceded by the strain family classification number (Warren *et al.*, 1999) (F: high copy number strain; B: low copy number strains).



3.3.2 RFLP mapping of chromosomal domains

The ISL probes (n = 25) were sequentially Southern hybridized onto *Pvu*II digested DNA isolated from the 34 strains (including the reference strain Mtb14323) to determine the electrophoretic mobility of the complementary chromosomal domains. Seventeen of the ISL probes hybridized to a single fragment, five of the ISL probes hybridized to two fragments, two of the ISL probes hybridized to three fragments, and one ISL probe hybridized to five fragments (Table 3.4). Hybridization of an ISL probe to more than one *Pvu*II fragment was interpreted to reflect duplication of chromosomal domains. Duplication was confirmed by comparing the ISL probe sequences with *M. tuberculosis* and *M. bovis* genome sequences (Table 3.1).

Figure 3.2A shows an example of the banding patterns generated by hybridization with the ISL7031.13 probe. Mutation driven by IS*6110* insertion was demonstrated by the cohyridization of the ISL hybridizing band and the IS-3' or IS-5' hybridizing bands (Fig. 3.2A). The orientation of the inserted IS*6110* element was established according to which of the IS (3' or 5') probes cohybridized to the ISL hybridizing band (Fig. 3.2A and 3.2C). ISL hybridizing fragments which were electrophoretically conserved in most of the isolates analyzed and which did not cohybridize to either of the IS (3' or 5') probes were assumed to represent the ancestral chromosomal fragment (non-mutated) (Fig. 3.2A, 4.7 kb fragment seen in the isolates 1127 and 1145). Mutation within the ancestral fragment was identified by a change in electrophoretic mobility of the ISL hybridizing band in the absence of cohybridization with either the IS-3' or IS-5' probe (Fig. 3.2A, and 3.2C, isolate 567). To differentiate between IS*6110* insertion into different positions within the chromosomal domain and mutation of the chromosomal domain flanking an inserted IS*6110* element, a PCR assay was developed to amplify the region between the IS*6110* terminal repeat and a downstream chromosomal sequence (see 3.2.4). Differences in the size of the amplification products, obtained from different clinical isolates, were indicative of different points of insertion (Fig. 3.2B), while an identical amplification product size (Fig 3.2B, isolates 640, 1013 and 1305) in association with a change in electrophoretic mobility detected by ISL probe hybridization (Fig. 3.2A, isolates 640, 1013 and 1305, respectively) was indicative of chromosomal mutation in the domain flanking the inserted IS*6110* element (Fig. 3.2C). Deletion of the duplicated chromosomal domain homologous to the ISL probe was identified by the absence of a hybridization signal (Fig. 3.2A, 2.7 kb band in isolate 1127 is absent). Each mutation was scored according to the mechanism leading to the observed polymorphism (Fig. 3.2C). Table 3.4 summarizes the mutational events identified in each chromosomal domain and for each of the clinical isolates analyzed. Of the 38 chromosomal domains analyzed, 35 were polymorphic. These polymorphisms resulted from 147 different mutational events, (1) IS*6110* insertion into different positions within the different chromosomal domains analyzed (n = 96), (2) different chromosomal mutations (n = 37), and (3) deletion of different chromosomal domains (n =14), however, the number of deletions detected by each probe may be an underestimate of the number of times each deletion event occurred, as it was not possible to define the extent of each deletion.

3.3.3 Mutational events occurring within each chromosomal domain

Preferential IS*6110* integration regions have been assigned to chromosomal domains of < 500 bp where different points of IS*6110* insertions have been identified (Fang and Forbes, 1997). According to this definition, the chromosomal domains analyzed can be grouped into two categories based on the number of different IS*6110* transposition events identified in each domain. Twenty-one domains evolved by a single insertional event, while 18 were found to have evolved by different insertional events, within a region of < 500 bp,

Figure 3.2. Mapping of chromosomal domain homologous to the ISL7031.13 probe. (A) RFLP banding patterns generated after Southern hybridization using the probe ISL7031.13. Bands labeled with 3' or 5' indicate cohybridization to either the IS-3' or IS-5' probes. The electrophoretically conserved band (2.7 kb) represents hybridization to a duplicated chromosomal domain. (B) PCR amplification of the region between the terminal repeat sequence of IS*6110* and an adjacent chromosomal sequence. PCR amplification was performed as described (see methods), and the products were electrophoretically fractionated in 5% polyacrylamide and visualized after staining with ethidium bromide. Only isolates containing an IS*6110* element in this chromosomal region generated a PCR product (PCR amplification reactions from isolates lacking an IS*6110* element are not shown). M – molecular weight marker.

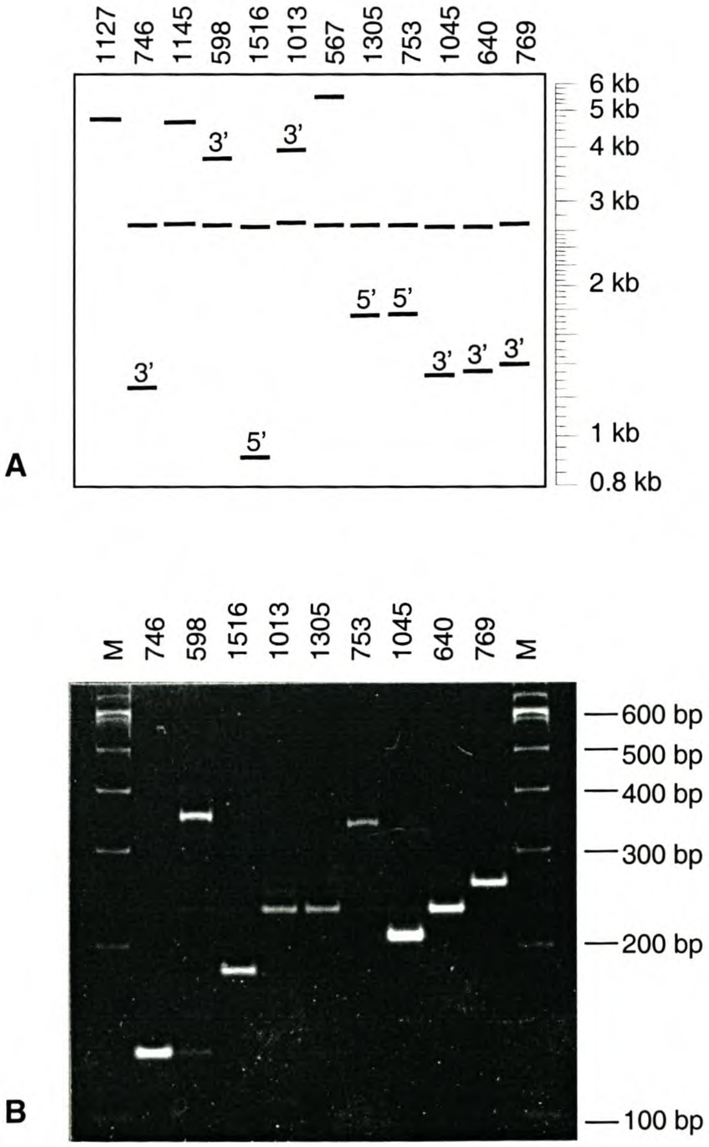


Figure 3.2 (continued) (C) Schematic diagram of the structure of the chromosomal domains. The probe ISL7031.13 hybridizes to two different domains as indicated by the grey boxes in region 1 and region 2 (shown on the H37Rv structure). Arrows indicate *Pvu*II restriction sites. Inserted IS*6110* elements are depicted by the open box, and the orientation of each IS*6110* element is indicated by the hatched region which reflects the 5' domain. Filled bar represents the size of the PCR amplification product (calculated from the gel depicted in Fig. 3.2B), while the size of the 3' flanking region up to the *Pvu*II site (excluding the IS*6110* domain) is shown below each diagram (calculated from the gel depicted in Figure 2A). The mechanisms of mutation leading to the observed polymorphisms within region 1 and region 2 were scored for each isolate analyzed (wt - indicates the non mutated domain, M - indicates chromosomal mutation, T - indicates mutation by IS*6110* insertion (numerical value indicates each different event) and D - indicates chromosomal deletion).

Figure 3.2C (see opposite page for legend)

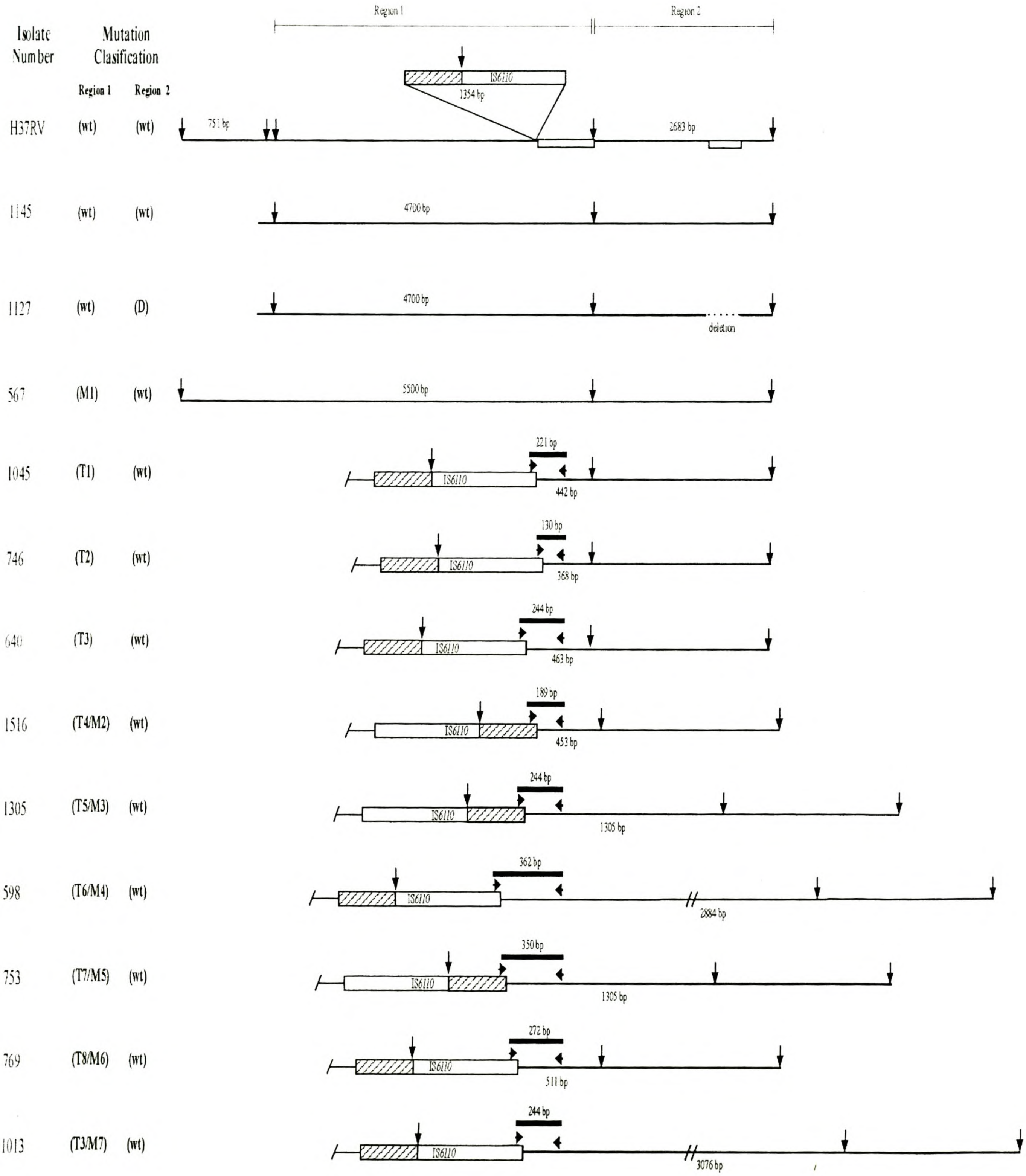


Table 3.4. Mutational events identified in specific chromosomal domains of clinical isolates of *M. tuberculosis*.

		Isolate number																																	
Probe	**Cosmid**	14323	1045	746	679	1093	640	839	24	1516	713	1305	1320	1038	1063	752	598	1057	740	483	172	1236	656	351	1112	75	245	1145	567	1127	689	753	603	769	1013
480.3	MTCY180													T1	T1	M1	M1/T2	T3	T4																
	MTY13e10																																		
	MTCY3c7																																		
480.6	MTCY164			T1									T2	T3												T4								T5	
480.8	MTV016	T1	T2	T3	D	T4	T4	T5	T6	T6	T6	T6	T7	T8	T8	T9	T9		T10						T11	T6	T1	T1	T1	T1	T1	T6	T12	T13	T13
	MTCI429B	T1				T2	T2		T3	T4	T4	T4	T5			T6	T6	D	D							T4		T7			T8	T4			
480.9	MTCY275													T1	T1																				
480.11	MTCY28			D	D	T1					D			T2	T2	D	D	T2	T2		D			D	D			T2			T3		D		
	MTCY98																	M1												D	D		D		
540.1	MTCY9F9		M1												T1																				
540.3	MTCY16B7	T1								M1				T2	T2	D	D											T1		T3					
540.8	MTCY19H5														T1																				
540.10	MTV006														T1																				
540.11	MTV004	T1	D												T2															M1	M1		D		
715.5	MTCY428															M1/T1	M1																		
715.8	MTV029		T1													T2	T2		T3																
852.1	MBO18605	T1		M1	M1	M2	M2	T2	M2	M2		M2		M3/T3	M3	M3	M3															M2			
7031.2	MTCY28	T1/M1	T2	T2	T2			T1		T3	T4		T2	M2	M2	T5	T5	M3	M3	M4	M4	M4	M5	M6	T1/M7	T1/M8	T1/M8	T1/M8	T1/M9	T1/M10	T1/M11		T2	T2	T2
7031.3	MTCY164		T1	T1	T1					T2			T1																				T1	T1	T1
	MTCI364								M1		M2				M3	T1	T1																		
	MTV012										M1				M2		T1		T2																
	MTCY07A7																							M1											
	MTCY10G2																							M1					M2						
7031.4	MTCY98		D	T1		T2	T2		T3	T4	T3	T3				T5	T5																	T6	T7
	MTCY98																																		
7031.5	MTV016/MTMOAIS	D	T1	T1	T1	T2	T2	T3					T1				T4											T5	T6	T7			T1	T1	T1
7031.6	MTV043		T1	T1	T1								T1														T2						T1	T1	T1
7031.7	MTCY130		T1	D	D								D													D		D					T1	T1	T1
	MTCY130									D					D	D	D																		
	not found		D			D					D			D				D	D	D		D	D					D		D		D	D	D	D
7031.8	MTU242907		D	D	D	T1	T2		T1	T1	D	T1	D	D	D	D	D	D	D				T2	T2		T1						T1	D	T3	T3
7031.9	MTV018/MTCY39															M1	M1	M1	M1															T1	T1
	MTCY28					D	D		T1	T1	T1	D		D	D	D	D															D	D		
7031.10	MTCY28												M1																						
7031.11	MTV014					T1	T1			T2																			T2	T3	T3		T4	T5	T5
7031.12	MTC25C11																														T1			T2	T2
7031.13	MTCY98		T1	T2		T3	T3		T5/M3	T4/M2	T5/M3	T5/M3		M1	M1	T6/M4	T6/M4		M1										M1			T7/M5		T8/M6	T3/M7
	MTCY98																													D					

Ancestral domains (non-mutated) are depicted by blank spaces, although these domains hybridize to their respective probes. D - deletion (unknown size). M - mutational event/s (number indicates different mutational events within a specific chromosomal domain). T - mutation by IS*6110* insertion/s (number indicates the different insertion points within a specific chromosomal domain). T/M - mutation caused by both IS*6110* insertion and chromosomal mutation in the domain adjacent to the insertion point (number indicates different events). Identical mutational events, detected by different probes, were only recorded once.

identified in different isolates (Table 3.5). This demonstrates that preferential integration regions are common in the chromosome of *M. tuberculosis*.

To identify the approximate position of an IS*6110* element within each of the chromosomal domains analyzed, the RFLP fragment lengths (minus the IS*6110* domain) were aligned with the complementary *Pvu*II fragments in H37Rv (Cole *et al.*, 1998), RvD2 (Gordon *et al.*, 1999a) and MTMOAIS (Fang *et al.*, 1999a). The accuracy of this technique was calculated by comparing the ISL RFLP fragments lengths (non-mutated) and the complementary *M. tuberculosis* or *M. bovis* sequence-derived *Pvu*II fragment lengths. The standard deviation between the RFLP and sequence fragment lengths was found to be 46 bp over the range 1000-8000 bp. Mapping of the position of each insertion in relation to the annotated ORFs in H37Rv (Cole *et al.*, 1998), RvD2 (Gordon *et al.*, 1999a) and MTMOAIS (Fang *et al.*, 1999a), showed the disruption of 28 ORFs by an IS*6110* element.

Alignment of the ISL probe fragment sequences with the *M. tuberculosis* and *M. bovis* BCG genome sequence enabled the identification of chromosomal deletions (identified by the absence of a hybridization signal). Of the 14 deletions identified, 13 overlapped with 17 predicted ORFs (compare Tables 3.1 and 3.4).

In order to gain an understanding of the mutational mechanisms leading to band shifts (excluding IS*6110* insertion), the *M. tuberculosis* H37Rv sequence data were analyzed to identify sequence features which, when mutated, could explain the change in fragment length. From this analysis, the following mutational mechanisms could be deduced: First, band shifts could result from either mutation or DNA modification of the terminal *Pvu*II restriction site, such that the resulting fragment increased in length to include the domain up to the following

Table 3.5. Preferential IS*6110* integration regions identified in the chromosomes of clinical isolates of *M. tuberculosis*.

Preferential integration region	Cosmid	N° insertions
Rv3323c-Rv3324	MTV016	4
Intergenic (Rv3324c-*moaA*)	MTMOAIS	3
Intergenic (Rv0001-Rv0002) (Kurepina *et al.*, 1998)	MTV029	2
Intergenic (Rv3345c-Rv3347c)	MTV004	2
Rv0797	MTCI429B	8
Rv1755c region 1 (Sampson *et al.*, 1999)	MTCY28	5
Rv1755c region 2 (Sampson *et al.*, 1999)	MTCY28	3
Rv1758 (Sampson *et al.*, 1999)	MTU242907	3
Rv1777	MTC25C11	2
Rv1917c	MTCY180	2
Rv2351c-Rv2352c	MTCY98	4
Intergenic-Rv2353c	MTCY98	3
Rv2353c	MTCY98	8
Rv2818c-Rv2819c	MTCY16B7	2
Rv3125c	MTCY164	5
Rv3183	MTV014	4
Rv3327 (Fang and Forbes, 1997)	MTV016	13
RvD1-ORF2-Rv2024c	MBO18650	2

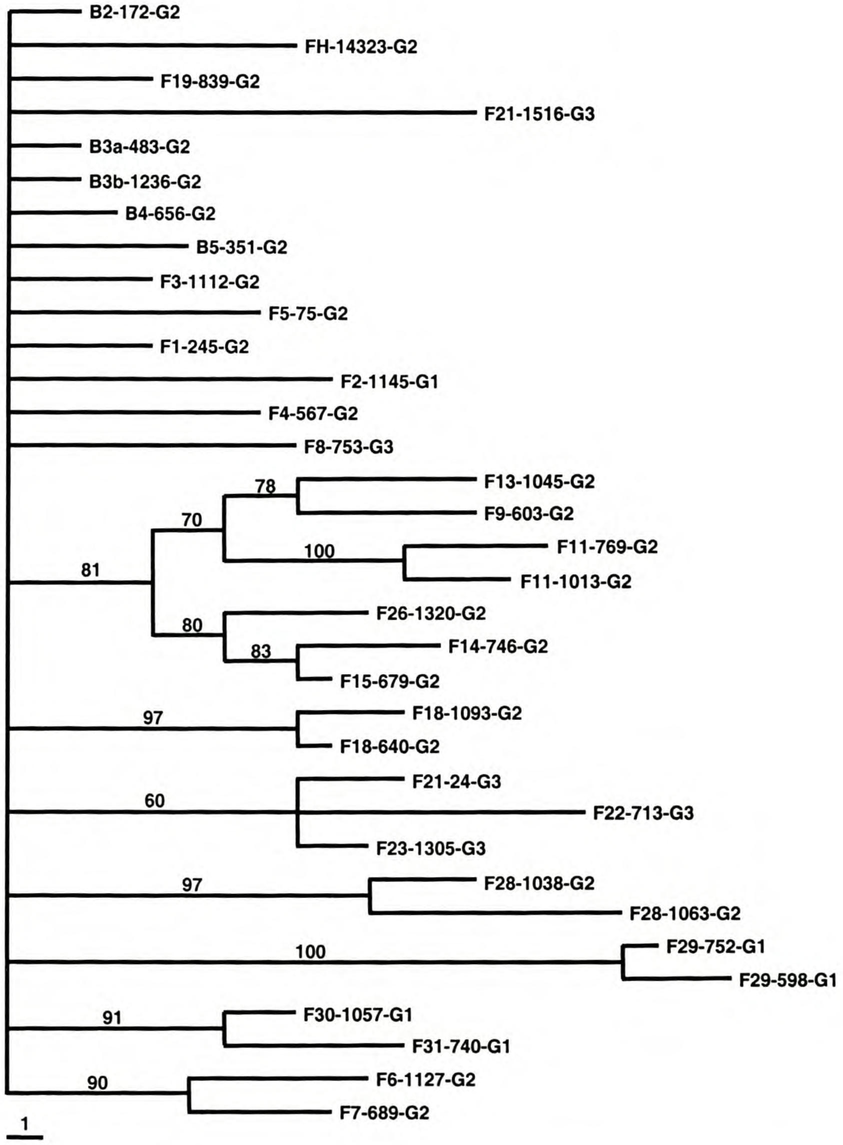
*Pvu*II site (Fig. 3.2C, compare isolate 1145 with 567). To determine whether inhibition of *Pvu*II restriction was due to DNA methylation, the DNA sequences were analyzed to identify the previously described DNA methylase recognition sequence (cagctgggagc) (van Soolingen *et al.*, 1996). This recognition sequence was not identified in any of the cosmids analyzed. Second, the identification of repeat sequences, of which the consensus repeat length corresponded to the change in fragment length, as seen in Rv1753c, suggested mutation via expansion or contraction of the repeat elements. Third, mutations leading to a decrease in fragment length, and which could not be ascribed to the deletion of repeat elements, were assumed to have occurred by mutation leading to the creation of an additional *Pvu*II restriction site, or by small deletions. Lastly, additional mutational mechanisms were also observed, however, these could not be explained from the sequence data but could include insertions (see Fig. 3.2C, isolates 598, 753, 1013 and 1305).

To establish a possible relationship between IS*6110* insertion and chromosomal mutation, the number of mutational events (including chromosomal mutation and deletion) and the number of IS*6110* insertions within each isolate were analyzed. The resulting analysis showed a weak although significant correlation ($r = 0.36$, $p = < 0.05$), suggesting that IS*6110* transposition events may predispose the surrounding chromosome to further mutation events, or that chromosomal domains which show a high chromosomal mutation rate tolerate IS*6110* insertion.

3.3.4 Phylogenetic reconstruction

To reconstruct the evolutionary history of the mutational events observed, the RFLP data were subjected to parsimony analysis using the heuristic algorithm [PAUP* 4.0; phylogenetic analysis using parsimony (*and other methods)] (Lawrence *et al.*, 1989). A confidence interval for the topology of the tree was established using the bootstrapping method, whereby only nodes occurring in > 50% of the trees were assumed to be significant (Felsenstein, 1985). The resulting consensus tree may be considered to be robust, showing a consistency index of 0.662 (Fig. 3.3). Interestingly, the topology of the tree does not demonstrate the sequential acquisition of mutations in the *katG* and *gyrA* genes used to classify the pathogenic groups G1, G2 and G3 (Sreevatsan *et al.*, 1997) (Fig. 3.3). The phylogenetic tree shows that the low copy number strains arrange close to the hypothetical ancestral strain, demonstrating that the chromosomal domains analyzed in these strains have remained stable over their evolutionary history (Fig. 3.3). In contrast, most of the high copy number strains partitioned into independently evolving lineages, each characterized by varying degrees of evolution within the chromosomal domains analyzed (Fig. 3.3). Comparison between the total number of mutations identified in each strain family and the frequency of the strain family isolates identified within the study community (Warren *et* al.,

Figure 3.3. Phylogenetic reconstruction using the bootstrapping method in association with the heuristic algorithm PAUP* 4.0. A consensus tree was generated from 10000 bootstrapped trees using the majority rule formula, and only nodes occurring in > 50% of the trees were considered to be significant (internal labels). Branch lengths are proportional to the number of evolutionary events (scale bar equals one evolutionary event), and all branches with a zero length were collapsed. The tree length was 222 and showed a consistency index of 0.662 and a homoplasy index of 0.338. The isolate number at the terminus of each branch is preceded by the strain family classification number (Warren *et al.*, 1999) (F: high copy number strain; B: low copy number strains) and followed by the pathogenic strain classification according to Sreevatsan *et al.* (1997).



1999) showed a moderately significant correlation (r = 0.48, p = < 0.01) (Table 3.6). This suggests that mutation may influence the overall fitness of strains within these strain family groupings, or that strain proliferation has increased the chance for these strains to acquire mutations.

Analysis of the mutational events occurring at each node shows both convergent and divergent evolution. The identification of convergent IS*6110* insertion events supports the concept of preferential integration regions, although the frequency of insertion into identical positions within a defined chromosomal domain was low (homoplasy index 0.25). Convergent chromosomal mutations were found to occur at a similar frequency (homoplasy index 0.255). In contrast, the frequency of chromosomal deletions, occurring as convergent evolutionary events, was significantly higher (homoplasy index 0.674). This demonstrates that chromosomal domains encompassing the regions of homology with the ISL probes and possible adjacent regions have been deleted by independent evolutionary events.

Table 3.6. Accumulation of mutations in different strain family representatives.

Strain frequency[a]	Strain family	Accumulated mutations
1	F3	4
1	F5	7
2	F8	8
6	F1	4
7	B3b	2
7	F31	11
7	F19	4
7	F22	12
8	B3a	2
11	F13	13
11	F6	10
12	F15	9
12	F23	8
12	F26	10
15	F30	8
17	F7	9
23	F18	10
24	B4	3
26	B2	2
26	F4	7
27	B5	5
27	F9	13
28	F21	11
31	F14	12
47	F2	9
88	F28	15
159	F29	19
208	F11	15

[a] Updated from Warren *et al.* 1999.

3.4 Discussion

Sreevatsan *et al.* (1997) examined the sequences of 26 genes in *M. tuberculosis* strains. Their results showed very little change and extreme conservatism in the genome of *M. tuberculosis* according to these genes. However, it is well known that isolates of *M. tuberculosis* can be distinguished on the basis of differences in their genotype. In order to reconcile these apparently contradictory findings, and to establish the importance of various ORFs in the genome of *M. tuberculosis*, it is important that other regions of the genome be examined. We have therefore examined flanking regions of IS*6110* elements in the genome of clinical isolates of *M. tuberculosis*.

In this study, mapping of these chromosomal domains clearly demonstrates that these domains are polymorphic. Therefore it is likely that these domains evolve substantially faster than the 26 genes described by Sreevatsan *et al.* (1997). Mutation was found to be predominantly driven by insertion of an IS*6110* element, supporting the notion of Sreevatsan *et al.* (1997), that transposition may play an important evolutionary role in *M. tuberculosis*. In addition, it is evident that chromosomal mutation, including point mutation, expansion or contraction of tandem repeat elements and larger chromosomal deletions contribute significantly to the evolutionary process.

This study has not established the exact mechanism of the chromosomal mutation events, however, comparison between the change in fragment length and the H37Rv genome sequence has allowed possible mechanisms to be suggested. These include mutations which will result in the loss or gain of *Pvu*II recognition sequences, a gain or loss of tandem repeats, the deletion of regions of the chromosome which may be small or large, or further mechanisms which are currently unknown.

Mapping of the insertion points on the chromosome demonstrates the presence of numerous preferential integration regions, where a preferential integration site has been defined as a domain of < 500 bp where different IS*6110* insertions have been identified in more than one *M. tuberculosis* isolate. In addition to the previously described preferential integration regions (Fang and Forbes, 1997; Kurepina *et al.*, 1998; Sampson *et al.*, 1999) a further 13 preferential integration regions were identified in this study, demonstrating that preferential integration regions are common in the *M. tuberculosis* genome. This result is in concordance with the conclusions derived from computer-based analysis of the IS*6110* RFLP banding patterns (McHugh and Gillespie, 1998), although the methods used to derive these conclusions differs vastly. Our study showed that IS*6110* insertion into a preferential integration region does not necessarily yield bands of identical electrophoretic mobility. Furthermore, it is evident from this study that chromosomal domains flanking the insertion element may undergo evolution, thereby contributing to IS*6110* RFLP diversity. This has direct implications for strain comparisons used in epidemiological studies, as fragment size rather than the IS*6110* chromosomal position forms the basis of the cluster analysis.

The relatively high number of mutational events occurring within the chromosomal domains analyzed contrasts significantly to previous results demonstrating genome stability (Sreevatsan *et al.*, 1997). Therefore, it may be hypothesized that the observed increase in mutation rate may be a direct result of the disruptive effect of IS*6110* insertion on the function of the surrounding DNA. Comparison between the number of IS*6110* insertions and the number of chromosomal mutations in each isolate showed a weak although significant correlation. As a causal relationship was not demonstrated, it could not be concluded whether IS*6110* insertion predisposed the surrounding chromosomal domains to an increased mutation rate or whether chromosomal mutation was an indicator of the tolerance of IS*6110*-driven mutation.

Two recent studies have used comparative hybridization to analyze the genomes of *M. tuberculosis*, *M. bovis* and *M. bovis* BCG (Gordon *et al.*, 1999a; Behr *et al.*, 1999). The results showed that 16 regions had been deleted from the *M. bovis* BCG chromosome, in different lineages, during *in vitro* culturing (Behr *et al.*, 1999). The authors hypothesize that these deletions have arisen due to progressive adaptation to laboratory conditions, leading to further attenuation as measured by the efficacy of vaccination (Behr *et al.*, 1999). Mapping of these domains in *M. tuberculosis* showed a coincidental deletion in the RD6 domain in some clinical isolates (Gordon *et al.*, 1999a). In addition, the authors identified two chromosomal domains (RvD1 and RvD2) which were present in *M. bovis* BCG but deleted from the chromosome of H37Rv and some clinical isolates (Gordon *et al.*, 1999a). Comparison of these results with the results presented in this study confirmed that deletions in the RvD2 region are common in clinical isolates of *M. tuberculosis*. A further 13 deletions were also identified in clinical isolates using the ISL probe set, two of which corresponded to neighbouring regions within the RD5 deletion (Gordon *et al.*, 1999a) and RD7 deletion (Behr *et al.*, 1999), and have not been previously described in *M. tuberculosis*. The mechanism whereby such deletions occur is as yet unclear, although it has been suggested that homologous recombination between two IS*6110* elements may lead to the deletion of the chromosomal domain situated between the two insertion elements (Fang *et al.*, 1999a).

Phylogenetic analysis of the *M. bovis* BCG genome demonstrates that the chromosome has continued to evolve (Behr *et al.*, 1999). However, it is unclear whether the mutations observed in clinical isolates of *M. tuberculosis* represent an ongoing evolutionary process or a past evolutionary bottleneck leading to mutations being inherited by descent. To differentiate between these two scenarios, the ISL RFLP data was subjected to parsimony analysis. The resulting phylogenetic reconstruction predicts that a number of evolutionary lineages have evolved independently from a common ancestral strain. The low-copy-number strains appear to have evolved independently of the high-copy-number strains, demonstrating that strain evolution is not necessarily a process of the sequential acquisition of additional copies of IS*6110* elements. This result supports a previous finding which showed that chromosomal domains disrupted by IS*6110* insertion in low-copy-number strains were only rarely disrupted in high-copy-number strains (Fomukong *et al.*, 1998). Interestingly, this phylogenetic reconstruction does not demonstrate the sequential evolutionary path of the pathogenic strain groups suggested by Sreevatsan *et al.* (1997). This implies that the mutations in the *katG* and *gyrA* genes used to classify the pathogenic groups may represent distant evolutionary events which occurred prior to the mutations detected by the ISL probes.

Analysis of the mutational events occurring within each lineage demonstrates both divergent and convergent evolution. The evolution of the IS*6110* banding pattern would be expected to be divergent, given the presumed random nature of replicative transposition. However, the presence of preferential integration sites could explain the convergent evolution seen using the ISL probes. In contrast, chromosomal deletions, detected by the absence of a hybridization signal, showed a high degree of convergent evolution. This may be due to the inability of the hybridization technique to determine the extent of each deletion. However, this demonstrates that the chromosomal domains homologous to the ISL probes have been deleted as independent evolutionary events. Together, this data demonstrates that the genome of clinical isolates of *M. tuberculosis* continues to evolve.

From the branch lengths of the phylogenetic tree it is evident that the number of evolutionary events occurring within the chromosomal domains analyzed differs considerably for different lineages. Evaluation of the IS*6110* database of clinical isolates collected from a defined community clearly demonstrates the over-representation of three strain family groupings (Warren *et al.*, 1999). These strains are positioned on separate evolutionary lineages, each showing a high number of mutational events. Comparison between the number of mutations identified in each strain family grouping and the frequency of these family groupings within the study community demonstrated a moderately significant correlation. This result suggests that either acquisition of mutations could enhance strain "fitness" or that a high bacterial population number has increased the chance for a strain to acquire mutations.

Comparative genomics in combination with phenotypic analysis could enable the identification of mutational events conferring an advantage to strains of *M. tuberculosis*. This in turn will provide an understanding of pathogenic mechanisms.

CHAPTER 4

IS*6110*-MEDIATED DELETION POLYMORPHISM IN THE DR REGION OF CLINICAL ISOLATES OF *MYCOBACTERIUM TUBERCULOSIS*

NOTE: The results presented in the following chapter will be submitted to *The Journal of Clinical Microbiology* as: "**IS*6110*-mediated deletion polymorphism in the DR region of clinical isolates of *Mycobacterium tuberculosis*.** Sampson, S.L., Warren, R.M., Richardson, M., Victor, T.C., Jordaan, A.M., Van Der Spuy, G.D., Van Helden, P.D."

(The style of the text and numbering of sections has been altered to conform with the style of this thesis. Literature cited in the text takes the form of author name and year of publication as opposed to the number format specified by The Journal of Clinical Microbiology. *All references are compiled into a single list at the end of the thesis for ease of reference.)*

4.1 Introduction

The recent surge in availability of mycobacterial genome sequence information (Cole *et al*., 1998; Pym and Brosch, 2000; Cole *et al*., 2001) has added great impetus to the field of tuberculosis research. Comparative genomics can now be exploited to identify intra- and interspecies variation, which could ultimately allow the identification and functional analysis of genes essential for virulence, pathogenesis and persistence, and thereby facilitate targeted drug design (Brosch *et al*., 1998; Pym and Brosch, 2000). In addition, investigation of the mechanisms whereby laboratory strains have become attenuated may contribute to the rational design of a more effective vaccine (Behr *et al*., 1999; Gordon *et al*., 1999a). Identification of genetic differences between vaccine strains, environmental mycobacteria and clinical isolates is also important for the development of diagnostic tools which have no cross reactivity to the vaccine strain or environmental mycobacteria. Finally, comparative genomics provides a powerful tool to define evolutionary relationships between mycobacterial species and strains.

One of the first studies to investigate genomic differences between *Mycobacterium bovis* BCG and *Mycobacterium tuberculosis* H37Rv utilized subtractive hybridization to demonstrate the deletion of 3 regions (RD1, RD2 and RD3) from *M. bovis* BCG relative to *M. tuberculosis* H37Rv (Mahairas *et al*., 1996) (Table 4.1). Hybridization of *M. tuberculosis* H37Rv bacterial artificial chromosome (BAC) arrays with *M. bovis* BCG (Pasteur) DNA confirmed these deletions and identified a further 7 deletions (RD4-RD10) (Gordon *et al*., 1999a) (Table 4.1). An alternative approach combining Southern blotting, sequence analysis and PCR, confirmed a 12.7 kb deletion from *M. bovis* relative to *M. tuberculosis*, which corresponds to RD7 (Zumárraga *et al*., 1999) (Table 4.1).

Table 4.1. Summary of regions deleted from *M. bovis* and *M. bovis* BCG relative to *M. tuberculosis*

Name	Strains where missing	Number of ORFs	ORFs deleted/ disrupted	Size (kb)	Authors
RD1	All BCG	9	Rv3871-Rv3879c	9.5	Mahairas *et al.*, 1996 Gordon *et al.*, 1999a Behr *et al.*, 1999
RD2	some BCG[a]	11	Rv1978-Rv1988	10.7	Mahairas *et al.*, 1996 Gordon *et al.*, 1999a Behr *et al.*, 1999
RD3	*M. bovis,* some *M. tuberculosis*	14	Rv1573-Rv1586c (prophage)	9.3	Mahairas *et al.*, 1996 Gordon *et al.*, 1999a Behr *et al.*, 1999
RD4	*M. bovis*, some BCG[b], *M. tuberculosis*[c]	11	Rv1506-Rv1516c	12.7	Brosch *et al.*, 1998 Gordon *et al.*, 1999a Behr *et al.*, 1999 (=RD6)[d]
RD5	*M. bovis*, all BCG, *M. microti*, *M. tuberculosis*[c]	8	Rv2346c-Rv2353c	9.0	Leão *et al.*, 1995 Gordon *et al.*, 1999a Behr *et al.*, 1999 (=RD7)[d]
RD6	*M. bovis*, all BCG, some *M. tuberculosis*[c]	5	Rv3425-Rv3429	4.9	Gordon *et al.*, 1999a Behr *et al.*, 1999 (=RD11)[d]
RD7	*M. bovis*, all BCG, some *M. tuberculosis*[c], *M. microti*	15	Rv1963c-Rv1977	12.7	Gordon *et al.*, 1999a Behr *et al.*, 1999 (=RD15)[d] Zumárraga *et al.*, 1999
RD8	*M. bovis*, all BCG, *M. microti*, some *M. tuberculosis*[c]	7	Rv3617-Rv3623	5.9	Gordon *et al.*, 1999a Behr *et al.*, 1999 (=RD9)[d]
RD9	*M. bovis*, all BCG, *M. microti*, *M. africanum*, some *M. tuberculosis*[c]	4	Rv2072- Rv2075	2.0	Gordon *et al.*, 1999a Behr *et al.*, 1999 (=RD12)[d]
RD10	*M. bovis*, all BCG, *M. microti*, *M. africanum*, some *M. tuberculosis*[c]	3	Rv0221-Rv0223	1.9	Gordon *et al.*, 1999a Behr *et al.*, 1999 (=RD4)[d]
RD11	*M. bovis*, all BCG	16	Rv2645-Rv2660	11	Behr *et al.*, 1999 (=RD13)[d]
RD12	*M. bovis*, all BCG	5	Rv3117-Rv3121	2.8	Behr *et al.*, 1999 (=RD5)[d]
RD13	*M. bovis*, all BCG	3	Rv1255c-Rv1257	3.0	Behr *et al.*, 1999 (=RD10)[d]
RD14	BCG-Pasteur	8	Rv1766-Rv1773	9.1	Behr *et al.*, 1999 (=RD14)[d] Ho *et al.*, 2000
RD15	BCG-Frappier, BCG-Connaught	4	Rv0309-Rv0312	3.0	Behr *et al.*, 1999 (=RD8)[d]
RD16	BCG-Moreau	6	Rv3400-Rv3405c	7.6	Behr *et al.*, 1999 (=RD16)[d]
RD17	*M. bovis*	1	Rv1563c	0.8	Gordon *et al.*, 2001
NRD17	BCG-Pasteur	1	Rv3479	0.7	Salamon *et al.*, 2000
NRD18	BCG-Pasteur	2	Rv1190-Rv1191	1.5	Salamon *et al.*, 2000
NRD25	BCG-Pasteur	3	Rv3737-Rv3739c	1.2	Salamon *et al.*, 2000

a. BCG originating from *M. bovis* BCG (Pasteur) after 1925: BCG-Danish, BCG-Prague, BCG-Glaxo, BCG-Frappier, BCG-Connaught, BCG-Phipps, BCG-Tice, BCG-Pasteur

b. BCG-Pasteur, BCG-Glaxo, BCG-Denmark

c. Denotes some *M. tuberculosis* clinical isolates

d. Alternative naming by Behr *et al.* (1999)

DNA microarray technology has been exploited to compare the genomes of *M. tuberculosis* H37Rv, *M. bovis* and a collection of *M. bovis* BCG daughter strains (Connaught; Danish; Pasteur 1173; Frappier; Glaxo; Japan; Moreau; Phipps; Prague; Sweden; Tice) (Behr *et al.*, 1999). PCR products representing ORFs in the *M. tuberculosis* H37Rv genome were hybridized with *M. bovis* and *M. bovis* BCG DNA. The results confirmed the deletions RD1 to RD10, and identified a further 6 deletions (Table 4.1). Three of these were found in *M. bovis* and *M. bovis* BCG (RD11 to RD13) while the remaining three deletions were restricted to specific *M. bovis* BCG daughter strains (RD14 to RD16) (Table 4.1). A subsequent study employed a more sensitive technique, where computational analysis of hybridization signals to high-density oligonucleotide arrays was applied to identify a further 3 deletions from *M. bovis* BCG (Pasteur) relative to *M. tuberculosis* H37Rv (NRD17, NRD18, NRD25) (Salamon *et al.*, 2000). This technology has recently been applied to analysis of deletions in 19 *M. tuberculosis* clinical isolates, identifying 25 deletion regions involving 93 ORFs (Kato-Maeda *et al.*, 2001). Although this technique represents a significant technological advance in the field of comparative genomics, it is relatively inaccessible to many molecular biology labs at present.

Deletions from *M. tuberculosis* H37Rv relative to *M. bovis* and *M. bovis* BCG have also been identified. Direct comparison of BAC clones allowed the identification of 2 deletions (RvD1 and RvD2) (Table 4.2) from *M. tuberculosis* H37Rv relative to *M. bovis.* (Gordon *et al.*, 1999a). This work was extended by sequence analysis and PCR screening to identify a further three deletions (RvD3 to RvD5) (Table 4.2) from *M. tuberculosis* H37Rv relative to *M. bovis* (Brosch *et al.*, 1999). The deletions RvD2, RvD3 (Ho *et al.*, 2000), and RvD5 (Fang *et al.*, 1999a) have also been detected in *M. tuberculosis* clinical isolates.

Table 4.2. Summary of regions deleted from *M. tuberculosis* H37Rv relative to *M. bovis* BCG

Name	ORFs deleted/disrupted	Size (kb)	Authors
RvD1	RvD1-ORF1; RvD1-ORF2; Rv2024c	5.0	Gordon *et al.*, 1999a
RvD2	Rv1755 (plcD)[a]; IS*6110*; RvD2-ORF1; RvD2-ORF2; RvD2-ORF3; Rv1758	5.0	Gordon *et al.*, 1999a Ho *et al.*, 2000
RvD3	IS*6110*; RvD3-ORF1	1	Brosch *et al.*, 1999 Ho *et al.*, 2000
RvD4	IS*6110*; Rv2353c (PPE)[b]	0.8	Brosch *et al.*, 1999
RvD5	IS*6110*; RvD5-ORF1; RvD5-ORF2; RvD5-ORF3; RvD5-ORF4	1.1	Fang *et al.*, 1999a Brosch *et al.*, 1999

a. *plcD* interrupted by IS*6110* in H37Rv.
b. PPE interrupted by IS*6110* in H37Rv.

The importance of transposable elements as agents of genome evolution has been established both for eucaryotes (Fedoroff, 2001) and procaryotes (Blot, 1994; Naas *et al.*, 1994). It has been suggested that insertion sequence-mediated deletion events are also an important mechanism driving mycobacterial genome variation (Sreevatsan *et al.*, 1997; Fang *et al.,* 1999a; Brosch *et al.*, 1999). Homologous recombination between directly repeated IS*6110* elements has been proposed as a likely mechanism for genomic deletions in clinical isolates (Fang *et al.*, 1999a). On insertion, the IS*6110* element generates 3-4 bp duplications of the sequence immediately flanking the point of insertion (Dale, 1995), and the absence of these 3-4 bp direct repeats is interpreted to reflect homologous recombination events between two IS*6110* elements (Fang *et al.*, 1999a; Brosch *et al.*, 1999). This knowledge was used in conjunction with *in silico* analysis of sequences flanking the 16 IS*6110* elements in the *M. tuberculosis* H37Rv genome to identify the deletions RvD3, RvD4 and RvD5 (Brosch *et al.*, 1999). Similarly, analysis of sequences immediately flanking IS*6110* elements in a 20 kb variable region of the chromosome provides further examples of the absence of flanking direct repeats, again implying IS*6110*-mediated deletion events (Ho *et al.*, 2000). However, to our knowledge, no study has yet investigated confirmed predecessor strains (prior to the deletion event) to substantiate this hypothesis.

Thus far, the phenotypic impact of the described deletions remains uncertain, with much speculation based on the proposed function of deleted and disrupted genes (Mahairas *et al.*, 1996; Behr *et al.*,1999; Brosch *et al.*, 1999; Gordon *et al.*, 1999a; Zumárraga *et al.*, 1999). Intriguingly, increasing amounts of genomic deletion have been associated with a reduced likelihood for the strain to cause pulmonary cavitation (Kato-Maeda *et al.*, 2001). Also, it has been suggested that there is a correlation between the progressive accumulation of deletions and BCG attenuation (Behr *et al.*, 1999). However, the genetic basis for these phenomena is unknown. It is evident that assessing the biological significance of deletion events based on the function of deleted genes is not straightforward. Due to the limited sensitivity of available technologies which are unable to easily detect small deletions (<350 bp, Salamon *et al.*, 2000) and point mutations, it is possible that critical mutational events will be missed by current approaches, further complicating the elucidation of phenotypic differences.

However, comparative genomics has proved valuable in terms of elucidating evolutionary questions. For instance, it has challenged the previous assumption that *M. tuberculosis* evolved in a linear fashion from *M. bovis*. Deletions have been observed both from the *M. tuberculosis* genome relative to *M. bovis*, and vice versa. The bi-directionality of the events suggest that their evolutionary history involves a common ancestor (Gordon *et al.*, 1999a), from which both species subsequently evolved. Deletions observed in various BCG daughter strains were used to infer a BCG phylogeny, which correlated with the historical record for these strains (Behr *et al.*, 1999).

To date, attention has focused mainly on comparisons of the mycobacterial strains for which the complete genome sequence is available, for example *M. tuberculosis*, *M. bovis* and *M. leprae*. Clinical isolates of *M. tuberculosis* have been relatively neglected by the field of comparative genomics. However, some investigators have examined deletion polymorphism in clinical isolates, specifically focusing on the direct repeat (DR) region of the genome (Fang *et al*., 1998; Van Embden *et al*., 2000). This locus consists of numerous copies of a 36 base pair direct repeat, separated by variable spacer sequences of 35-41 base pairs in length (Hermans *et al*., 1991). (Each direct repeat unit with its associated spacer is referred to as a direct variable repeat (DVR) (Groenen *et al*., 1993).) The DR region has been described as a "hotspot" for the insertion of an IS*6110* element in members of the *M. tuberculosis* complex (Hermans *et al*., 1991). It is speculated that this could reflect true preferential insertion or reduced frequency of excision of the element once inserted into this region (Hermans *et al*., 1991). Alternatively, this region may represent the ancestral IS*6110* integration site in the *M. tuberculosis* chromosome, with the current chromosomal arrangement of the various copies of IS*6110* reflecting outward migration from this region (Dale, 1995; Philipp *et al*., 1996).

The DR region exhibits polymorphism in *M. tuberculosis* , and this has been exploited for strain typing using a novel PCR-based fingerprinting method known as spoligotyping (Goyal *et al.,* 1997; Kamerbeek *et al.,* 1997), which is dependent on the presence or absence of spacer sequences between the direct repeats. The current understanding is that homologous recombination between adjacent or spatially distant DR elements is responsible for the observed variability (Hermans *et al*., 1991; Groenen *et al*., 1993; Fang *et al*., 1998; Van Embden *et al*., 2000). In addition, mutational events mediated by IS*6110* also contribute to diversification of this region. These events can include transposition and recombination leading to deletion (Groenen *et al*., 1993; Fang *et al*., 1998; Van Embden *et al*., 2000).

Investigation of genetic differences in the DR region in a set of closely related clinical isolates of *M. tuberculosis* has been used to elucidate an evolutionary history (Fang *et al.*, 1998).

This study addresses the phenomenon of "deletion polymorphism" in the DR region of clinical isolates of *M. tuberculosis*. Five discrete deletion polymorphisms in the DR region were investigated. Sequencing of the deletion junctions, and comparison to putative predecessor strains (as defined by genotyping) have been utilized to elucidate the mechanisms leading to genome deletions. The results support the involvement of the insertion sequence IS*6110* in the observed deletions, but suggest that this occurs either via unidentified intermediate strains or by mechanisms other than homologous recombination between directly repeated copies of the element. Finally, the results obtained suggest that some caution should be exercised in the interpretation of spoligotyping results.

4.2 Methods

4.2.1 Study setting and strain selection

An ongoing molecular epidemiology study of tuberculosis in a high incidence community in the Western Cape, South Africa, has produced a database of more than 2000 *Mycobacterium tuberculosis* isolates from over 800 patients (Warren *et al.*, 1996a; Warren *et al.*, 1999; M. Richardson, Dept. Medical Biochemistry, University of Stellenbosch, personal communication). These isolates have been typed by IS*6110* DNA fingerprinting according to an internationally standardized protocol (Van Embden *et al.*, 1993). Five isolates (392, 397, 704, 780, 973; Fig. 4.1) were selected for further analysis on the basis of direct repeat (DR) region spoligotypes which indicated deletion of the 5' portion of the DR region. An additional three isolates (Fig. 4.1) representing putative predecessors of 392, 397 and 704 (1985, 176 and 1227 respectively) were investigated to establish the mechanisms whereby the observed deletions occurred. The putative predecessors were identified by either (i) GelCompar comparisons of IS-3' and IS-5' DNA fingerprints (1985, 176) or (ii) phylogenetic analysis of a group of closely related strains (1227) (Warren *et al.*, 2001). The isolates investigated represent examples of the 3 dominant strain family groupings within the study community (Warren *et al.*, 1999), where a family of strains is defined as isolates which are > 65% related in terms of IS*6110* DNA fingerprinting, as determined by GelCompar (Applied Maths, Kortrijk, Belgium).

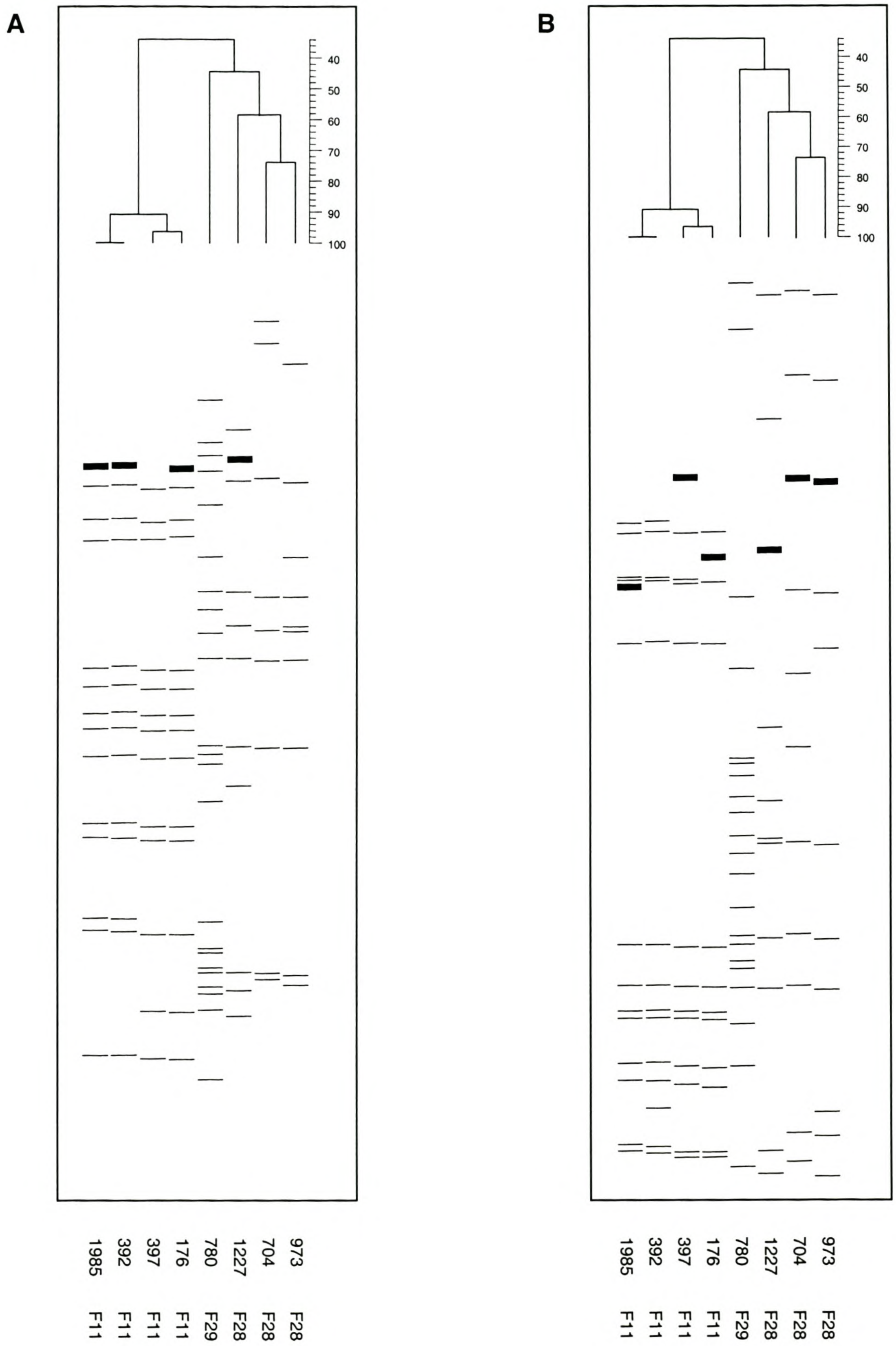
4.2.2 DNA Hybridization

*Pvu*II-digested genomic DNA was electrophoretically fractionated together with 8 ng DNA Molecular Weight Marker X (Roche) in 0.8% agarose, as previously described (Warren *et al.*, 1996a). The DNA was Southern transferred onto a Hybond N+ membrane, and fixed by heating at 80°C for 2 hours. The membranes were sequentially hybridized with ECLlabelled IS-3' (Van Embden *et al.*, 1993) (Fig. 4.1A), IS-5' (Warren *et al.*, 2000) (Fig. 4.1B) and Marker X (Roche), as recommended by the manufacturer (Amersham). To identify DR-containing *Pvu*II fragments, membranes were hybridized with ^{32}P-labelled DRr probe (Hermans *et al.*, 1991), as previously described (Warren *et al.*, 1996a). DNA fingerprints were visualized by autoradiography. Between each hybridization step, the membranes were incubated in 0.4 M NaOH at 50°C for 45 min to denature all DNA hybrids. The membranes were then neutralized in 0.2 M Tris-HCl pH 8.0, 0.1x SSC and 0.1x SDS at 50°C for 15 min prior to re-hybridization. To facilitate the alignment of autoradiographs, membranes were spotted with *M. tuberculosis* genomic DNA and Marker X prior to hybridization.

4.2.3 Spoligotyping

DNA polymorphism in the DR locus was detected by spoligotyping according to a standardized protocol (Molhuizen *et al.*, 1999). Briefly, the method relies on the presence or absence of spacer sequences between the characteristic 35–41 bp direct repeats. A commercially available membrane (Isogen Biosciences BV, Maadsen, The Netherlands), with covalently linked parallel rows of 43 synthetic oligonucleotides representing the unique spacer sequences between direct repeat units in the DR region, was used. The oligonucleotides DRa (5'-biotinylated) and DRb (Table 4.3), complementary to either side of the DR unit, were utilized to amplify the intervening spacer regions from clinical isolates of *M. tuberculosis*. The PCR cycling conditions were as follows: denature at 95°C for 3 min; then cycle 30 times at 95°C for 1 min; 55°C for 1 min; 72°C for 1 min; followed by extension at 72°C for 5 min. The amplified DNA was hybridized to the membrane-bound spacer oligonucleotides and detected with the ECL (Amersham) detection system to generate a characteristic spoligotype pattern for each isolate.

Figure 4.1. GelCompar representation of IS*6110* DNA fingerprints of strains utilized for DR region investigation. (A) IS-3' DNA fingerprints, (B) IS-5' DNA fingerprints. Heavy lines indicate *Pvu*II fragments which co-hybridize with DR bands.



4.2.4 DNA Manipulation

All PCR amplification reactions were performed with the Expand Long Range PCR system (Roche). Reaction mixtures contained 1 μg DNA template, 50 pmol of each primer (Table 4.3, Fig. 4.3A), 1x Reaction Buffer 3, and 5 U enzyme. Cycling conditions included a 2 min 94°C denaturation step, followed by 35 cycles of 94°C for 30 s, 30 s at the specified annealing temperature, and 68°C for the specified extension time, followed by a 10 minute extension step at 68°C (see Table 4.4 for specified times and temperatures). PCR products were electrophoretically fractionated in 0.8% agarose (1x TBE, pH 8.3), then visualized under UV after ethidium bromide staining. PCR products were purified with the Wizard PCR purification kit (Promega), and subsequently cloned using the pGEM-T Easy vector system (Promega). All cloning steps were performed according to standard protocols (Sambrook *et al.*, 1989). Plasmids were isolated using the Wizard DNA Purification system (Promega), then restriction mapped and/or sequenced. All DNA sequencing reactions were performed on an ABI3100 automated sequencer.

Table 4.3. Oligonucleotides utilized in this study

Name	Sequence (5' to 3')	Application
DRa	(5′ biotin)ggt ttt ggg tct gac gac	Spoligotyping
DRb	ccg aga ggg gac gga aac	Spoligotyping
16B7del	cct tgc tgt ccc gcc aat ac	Amplification of DR region and flanking sequence
1919del	gcc gaa gtc acg gca gac tg	Amplification of DR region and flanking sequence
16B7delRb	tgc aga aga agc tgg cga ag	Sequencing of cloned regions
IS-5'	ggt acc tcc tcg atg aac cac	Sequencing and PCR amplification from 5' terminal of IS*6110*
IS-3'	ttc aac cat cgc cgc ctc tac c	Sequencing and PCR amplification from 3' terminal of IS*6110*
16B7delF41	tga tcg acg cga acc tgt c	Sequencing of cloned regions
16B7delfF47	gct gcg gat gtg gtg ctg g	Sequencing of cloned regions
16B7delF63	tga tag aag ccg gaa agc tcc	Sequencing of cloned regions

Table 4.4. PCR products

Isolate name	Primer pair	Product size (kb)	Annealing temp. (°C)	Extension time (min)	GenBank Accession numbers
H37Rv (cosmid MTCY16B7)	16B7del / 1919del	9.8	65	12	Z81331
1227	16B7del / IS-3'	3.0	62	6	AF390041 AF390042
	IS-5' / IS-5'	3.6	62	6	AF390043 AF390044
	1919del / IS-3'	1.6	62	6	AF390045 AF390046
704	16B7del / 1919del	5.5	65	12	AF390047 AF390048 AF390049 AF390050
973	16B7del / 1919del	5.5	65	12	AF390051 AF390052 AF390053 AF390054
176	16B7del / IS-3'	3.8	62	6	AF390058 AF390059
	IS-5' / IS-5'	2.8	62	6	AF390060 AF390061
	1919del / IS-3'	1.6	62	6	AF390062 AF390063
397	16B7del / IS-3'	3.4	62	10	AF390064 AF390065
	1919del / IS-5'	1.5	62	10	AF390066 AF390067 AF390068
1985	16B7del / IS-5'	6.0	62	6	AF411182 AF411183
	1919del / IS-3'	1.6	62	6	AF390056 AF390057
392	16B7del / IS-5'	3.6	62	10	AF390069 AF390070
	1919del / IS-3'	1.6	62	10	AF390071
780	16B7del / 1919del	3.2	65	12	AF390039 AF390040

4.2.5 Sequence analysis

The DNA sequence of the *M. tuberculosis* MTCY16B7 cosmid was downloaded from the National Center for Biotechnology Information (NCBI) website (http://www.ncbi. nlm.nih.gov). The BLASTN algorithm was used to identify the positions of IS*6110* insertions, and to localize the deletion junctions. The precise points of insertion into the DR region were confirmed by comparison to the published MTCY16B7 sequence, and recently reported sequence data (Beggs *et al*., 2000; Van Embden *et al*., 2000). DVRs were numbered according to Van Embden *et al*. (2000). For comparative purposes, the original numbering of DVRs (Kamerbeek *et al*., 1997), and the rational nomenclature proposed by Dale *et al*. (2001) are also shown (Fig. 4.2 and Table 4.5, respectively).

4.3 Results

4.3.1 Spoligotype analysis of the DR region

Spoligotyping of the strains 392, 397, 704, 973 and 780 (Fig. 4.2) demonstrated deletion of the 5' portion of the DR region in these strains, in contrast to the strains 176, 1985 and 1227 which contained both 3' and 5' DVR sequences (Fig. 4.2). The genotypically closely related strains 392 and 397 (as defined by IS-3' and IS-5' DNA fingerprinting) shared identical spoligotype patterns (which were also identical to that of 704), while the closely related strains 704 and 973 differed by the presence of a hybridizing signal for one DVR (direct repeat unit plus associated spacer). The strain 780 demonstrated a distinct spoligotype pattern, with deletion of the 5' portion of the DR region. For comparative purposes, Fig. 4.2 illustrates nomenclature according to Van Embden *et al.* (2000), aligned with original nomenclature by Kamerbeek *et al.* (1997).

Figure 4.2. Spoligotype patterns of strains utilized for DR region investigation. Each row represents the spoligotype pattern of one isolate. Each square represents one DVR, numbered according to (A) Van Embden *et al.* (2000) and (B) Kamerbeek *et al.* (1997). Filled squares indicate positively hybridizing DVRs.

	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	■	■	■	■	■	■	■	H37Rv
	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	□	□	□	□	■	■	■	■	■	■	■	973
	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■	□	□	□	□	■	■	■	■	■	■	■	704
	■	■	■	■	■	■	■	■	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	■	■	■	■	■	■	■	1227
	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	780
	■	■	■	■	■	■	■	■	□	□	□	■	■	■	■	■	■	■	■	■	□	□	□	□	■	■	■	■	■	■	■	■	□	□	□	□	■	■	■	■	■	■	■	176
	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■	□	□	□	□	■	■	■	■	■	■	■	397
	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■	□	□	□	□	■	■	■	■	■	■	■	392
	■	■	■	■	■	■	■	■	□	□	□	■	■	■	■	■	■	■	■	■	□	□	□	□	■	■	■	■	■	■	■	■	□	□	□	□	■	■	■	■	■	■	■	1985
A	2	3	4	12	13	14	15	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	46	47	51	52	53	62	63	64	65	
B	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	

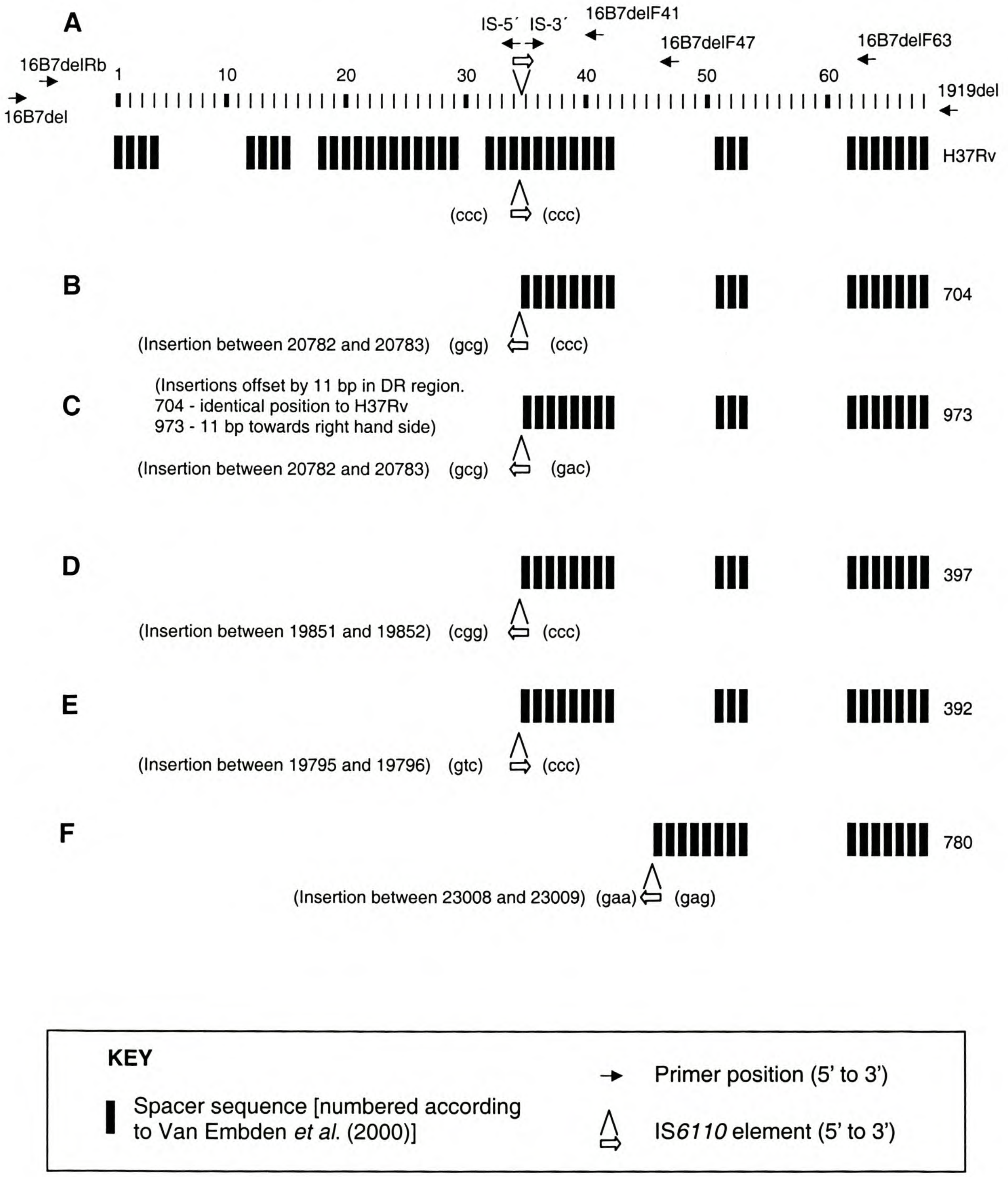
Table 4.5 Rational nomenclature (binary, octal and hexadecimal assignments). Spoligotype data shown in Fig. 4.2 is presented in the binary, octal and hexadecimal formats recently proposed by Dale *et al.* (2001) to facilitate inter-laboratory spoligotype comparisons.

Strains	Assignment	System
H37Rv	1111111111111111111001111111111100001111111	Binary
	777777477760771	**Octal**
	7F-7F-7C-7F	Hexadecimal
973	0000000000000000000000000111111100001111111	Binary
	000000003760771	**Octal**
	00-00-00-07-F0-7F	Hexadecimal
704,	0000000000000000000000001111111100001111111	Binary
392,	**000000007760771**	**Octal**
397	00-00-00-0F-F0-7F	Hexadecimal
1227	1111111100111111111111111111111100001111111	Binary
	776377777760771	**Octal**
	7F-4F-7F-7F-F0-7F	Hexadecimal
780	0000000000000000000000000000000000111111111	Binary
	000000000003771	**Octal**
	00-00-00-00-03-7F	Hexadecimal
176,	1111111100011111111100001111111100001111111	Binary
1985	**776177607760771**	**Octal**
	7F-47-7E-0F-F0-7F	Hexadecimal

4.3.2 PCR Amplification and cloning of the DR deletion regions

The primers 16B7del and 1919del were utilized to amplify the region between nucleotides 14093 and 23843 (cosmid MTCY16B7 numbering) which encompassed the DR region and deletion junctions (Fig. 4.3A). This region was amplified from the strains 704, 973 and 780 (Table 4.4). For the remaining strains, appropriate combinations of the primers 16B7del, IS-3' , IS-5' and 1919del were utilized to amplify either side of the DR region and deletion junctions (Table 4.4, Fig. 4.3A). All amplification products were subsequently cloned and sequenced.

Figure 4.3. Arrangement of DVRs and IS*6110* insertions in the DR region of clinical isolates relative to *M. tuberculosis* H37Rv. The position and orientation of the IS*6110* insertion, arrangement of DVRs and flanking 3 bp direct repeats (in brackets) are shown for (A) *M. tuberculosis* H37Rv, (B) 704, (C) 973, (D) 397, (E) 392 and (F) 780. The positions of oligonucleotides utilized in this study are shown relative to *M. tuberculosis* H37Rv, (positions of DR flanking sequence primers not to scale).


A
IS-5´ IS-3´
16B7delF41
16B7delF47
16B7delF63
16B7delRb
1
10
20
30
40
50
60
1919del
16B7del
H37Rv
(ccc)
(ccc)
B
704
(Insertion between 20782 and 20783)
(gcg)
(ccc)
C
(Insertions offset by 11 bp in DR region.
704 - identical position to H37Rv
973 - 11 bp towards right hand side)
973
(Insertion between 20782 and 20783)
(gcg)
(gac)
D
397
(Insertion between 19851 and 19852)
(cgg)
(ccc)
E
392
(Insertion between 19795 and 19796)
(gtc)
(ccc)
F
780
(Insertion between 23008 and 23009)
(gaa)
(gag)
KEY
Spacer sequence [numbered according to Van Embden et al. (2000)]
Primer position (5' to 3')
IS6110 element (5' to 3')

4.3.3 Sequence analysis

For reference purposes, sequence data is expressed relative to the *Mycobacterium tuberculosis* H37Rv cosmid MTCY16B7 sequence, and DVRs (direct repeats plus associated spacer sequences) are numbered as recently proposed by Van Embden *et al.* (2000) (See Fig. 4.2 and Table 4.5 for alternative nomenclature). Sequence encompassing the DVRs is referred to as the direct repeat (DR) region, and sequence outside of this is referred to as DR flanking sequence.

Analysis of the presence/absence and arrangement of DVRs in 392, 397, 704 and 973 demonstrated that these strains all had the same DVRs, and in the same order, as the 3' side of the *M. tuberculosis* H37Rv DR region (Fig. 4.3A – 4.3E). The arrangement and sequence of DVRs in the 780 DR region was identical to that found in strains recently described by other investigators (Beggs *et al.*, 2000; Van Embden *et al.* 2000) (Fig. 4.3F), and contains 5 unique DVRs not found in *M. tuberculosis* H37Rv.

Analysis of sequence from the closely related strains 704 and 973 demonstrates that the point of integration in the DR flanking region is identical in the two strains and that 2129 bp of DR flanking sequence is deleted (Fig. 4.4B). The orientation of the DR IS*6110* elements is the same in the two strains (opposite to that in *M. tuberculosis* H37Rv), but the point of integration within the DR region differs. In 704, the IS*6110* element is in an identical position to the *M. tuberculosis* H37Rv DR region insertion, while in 973, the insertion point is offset by 11 bp, towards the 3' end of the DR region. It is generally believed that the DR region evolves either by (a) IS*6110*-mediated recombination, or (b) homologous recombination or slipped strand mispairing leading to loss of discrete DVR units (Groenen *et al.*, 1993; Fang *et al.*, 1999a; Van Embden *et al.*, 2000). Therefore, the offset position is suggestive of an additional, discrete transposition into the DR region, associated with a subsequent recombination-mediated deletion event. It is interesting to note that the IS*6110* element disrupts the spoligotyping primer DRb site in the DR unit flanking spacer 35. Under the PCR conditions employed, amplification does not proceed through the disrupting IS*6110* element, resulting in the absence of hybridization signal for this DVR.

To investigate the possible deletion mechanism in 704, a putative predecessor strain was identified with the aid of a multi-probe phylogenetic analysis of a closely related group of strains (Warren *et al.*, in press). This data identified 1227 as an ancestral strain to 704 and 973 (see Appendix A for phylogenetic data), with a DR flanking region IS*6110* insertion in an identical position, but with an intact DR region. Sequence data demonstrated that the DR region IS*6110* insertion in this strain was identical to that of *M. tuberculosis* H37Rv (Fig. 4.4B). The DR region insertion was therefore in an identical position to the insertion in 704, although in the opposite orientation. The flanking region insertion in 1227 occurred in an identical position and orientation to that of 704, suggesting a common evolutionary origin for the two strains. Partial sequencing, spoligotyping and PCR product size analysis demonstrated that the intervening sequence between the DR insertion and the DR flanking insertion was intact in 1227. This data suggests that 704 arose from 1227 (or a similar closely related strain) via an IS*6110*-mediated deletion event. Elsewhere, IS*6110*-mediated deletion events have been suggested to occur via homologous recombination between directly repeated copies of the insertion sequence (Fang *et al.*, 1999a). In contrast, results presented here demonstrate that the ancestral strain has inversely orientated IS*6110* elements in the DR region and DR flanking region. This is suggestive of an alternative recombination mechanism or an unidentified intermediate strain.

Analysis of the strains 392 and 397, which are genotypically closely related (Fig. 4.1), demonstrates that the IS*6110* insertion occurred in identical positions within the DR region in the two strains (Fig. 4.4C and 4.4D). This corresponds to the position of the *M. tuberculosis* H37Rv DR region IS*6110* insertion (Fig. 4.4A). However, the point of insertion into the DR flanking sequence is offset by 56 bp in the two strains (Fig 4.4C and 4.4D). This represents deletions of 1198 bp and 1142 bp of 5' DR-flanking sequence from 392 and 397 respectively. A further point of difference between 392 and 397 is the opposing orientations of the IS*6110* elements in the two strains. In both strains, the element is in an identical position to the *M. tuberculosis* H37Rv DR region insertion (Hermans *et al*., 1991), however, in 397, the element lies in the opposite orientation (Fig. 4.4C and 4.4D, respectively). It is of note that the identical spoligotypes for the two strains (Fig. 4.2) do not reflect this structural difference.

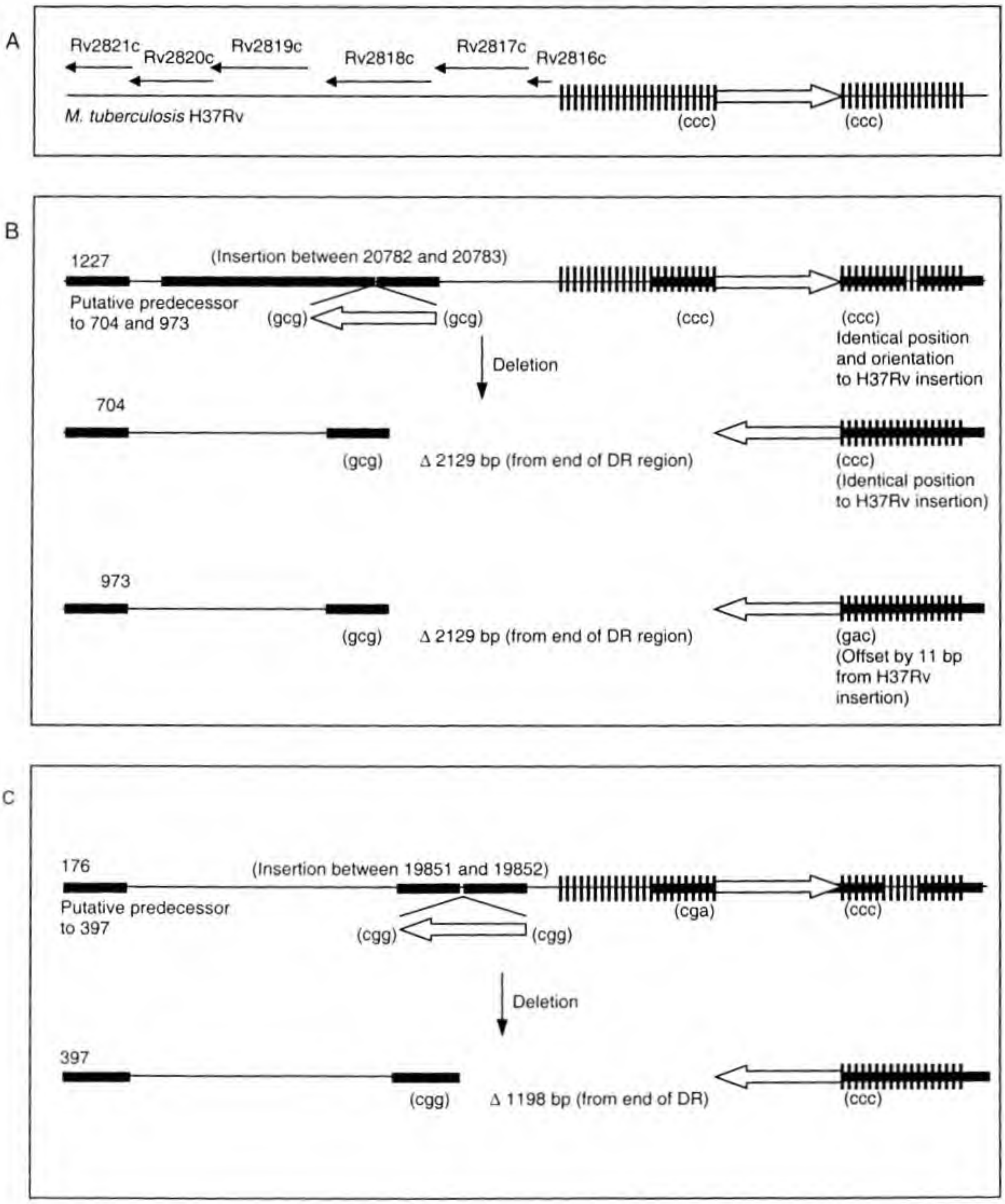
To investigate the possible mechanisms giving rise to the observed structure of the DR region in 392 and 397, putative predecessor strains were examined (Fig. 4.4C and 4.4D). GelCompar analysis of IS-3' and IS-5' DNA fingerprints was utilized in conjunction with spoligotype pattern data to identify strains most closely related to 392 and 397, but with an intact DR region. Using this approach, the isolate 176, with a DR flanking insertion in an identical position to that of 397, but with an intact intervening sequence and DR region, was identified (Fig. 4.4C). Sequence data from the two strains shows that the DR region insertion in the predecessor strain is in an identical position and orientation to that in *M. tuberculosis* H37Rv, with the DR flanking region insertion in an opposite orientation. The overall structure of the DR region and flanking sequence in 176 is therefore very similar to that observed in 1227, the predecessor to 704 (compare Fig. 4.4B and 4.4C).

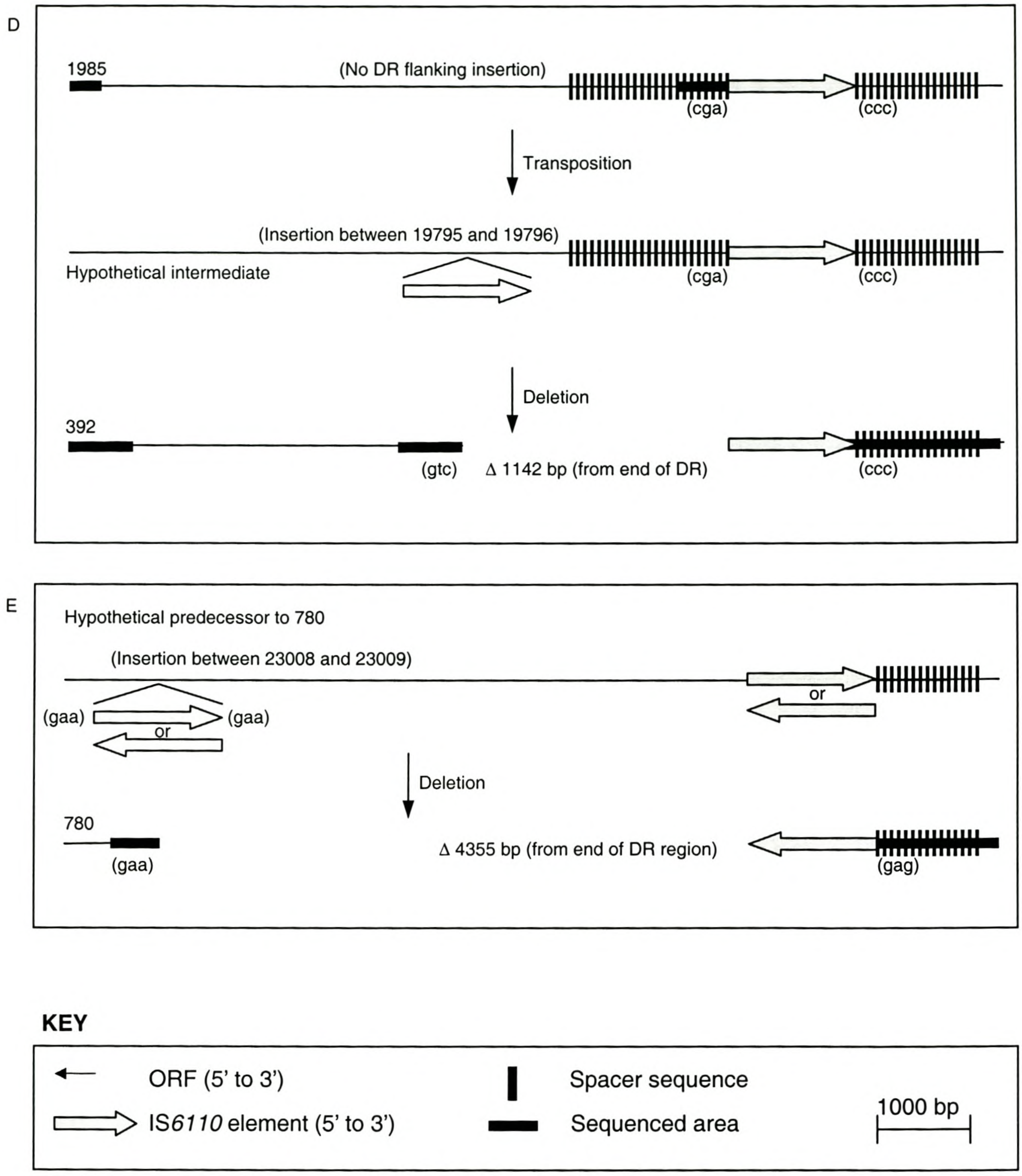
A search of the complete DNA fingerprint database (>2000 isolates) failed to identify an immediate predecessor strain for 392. The most closely related strain that was identified, namely 1985, contained an identical (in terms of position and orientation) DR region insertion, but no DR flanking region insertion. It is proposed that 392 evolved from 1985 (or a closely related strain) via an intermediate with an insertion between 19795 bp and 19796 bp (Fig. 4.4D). Interestingly, sequence data from 176 and 1985 suggests that these strains have undergone a prior IS*6110*-mediated deletion event in the DR region, as evidenced by the absence of DVRs 31 - 34 on the spoligotype pattern and the lack of 3 bp direct repeats flanking the insertion element.

The strain 780 exhibits the largest deletion of sequence flanking the DR region (4355 bp) in these isolates, and is identical in structure to that recently described by other investigators (Beggs *et al.*, 2000; Van Embden *et al.* 2000) (Fig. 4.4E). Once again, a database search failed to identify a predecessor strain, and therefore the relative orientation of IS*6110* elements which may have mediated the deletion event is unknown.

In total, four different points of IS*6110* insertion were identified in the sequence flanking the 5' side of the DR region (Fig. 4.4). Three discrete points of integration were identified within the DR region (Fig. 4.4). Interestingly, two of these occurred within slightly offset positions within DVR 35, the originally described "hotspot" for IS*6110* integration (Hermans *et al.*, 1991). Deletions of DR flanking sequence ranging in size from 1142 bp to 4355 bp were identified.

Figure 4.4. Deletion polymorphism in the DR region and flanking sequence of clinical isolates. (A) *M. tuberculosis* H37Rv, (B) putative predecessor 1227, and descendants 704 and 973; (C) putative predecessor 176 and descendant 397; (D) putative predecessor 1985, hypothetical intermediate and descendant 392; (E) Hypothetical predecessor and descendant 780. Three basepair duplications which arise due to transposition indicated in brackets. Note that the DR region insertions in the strains 176 and 1985 lack 3 bp duplications, which is thought to indicate a prior IS*6110*-mediated deletion event.


A
Rv2821c
Rv2820c
Rv2819c
Rv2818c
Rv2817c
Rv2816c
M. tuberculosis H37Rv
(ccc)
(ccc)
B
1227
(Insertion between 20782 and 20783)
Putative predecessor to 704 and 973
(gcg)
(gcg)
(ccc)
(ccc) Identical position and orientation to H37Rv insertion
Deletion
704
(gcg)
Δ 2129 bp (from end of DR region)
(ccc) (Identical position to H37Rv insertion)
973
(gcg)
Δ 2129 bp (from end of DR region)
(gac) (Offset by 11 bp from H37Rv insertion)
C
176
(Insertion between 19851 and 19852)
Putative predecessor to 397
(cgg)
(cgg)
(cga)
(ccc)
Deletion
397
(cgg)
Δ 1198 bp (from end of DR)
(ccc)


D
1985
(No DR flanking insertion)
(cga)
(ccc)
Transposition
(Insertion between 19795 and 19796)
Hypothetical intermediate
(cga)
(ccc)
Deletion
392
(gtc)
Δ 1142 bp (from end of DR)
(ccc)
E
Hypothetical predecessor to 780
(Insertion between 23008 and 23009)
or
(gaa)
(gaa)
or
Deletion
780
(gaa)
Δ 4355 bp (from end of DR region)
(gag)
KEY
ORF (5' to 3')
IS*6110* element (5' to 3')
Spacer sequence
Sequenced area
1000 bp

4.4 Discussion

Analyses of sequence data from *M. tuberculosis* clinical isolates (Fang *et al.*, 1999a; Ho *et al.*, 2000) and *in silico* sequence comparisons (Brosch *et al.*, 1999) have suggested that the insertion sequence IS*6110* may promote chromosomal deletion events mediated by homologous recombination. The absence of short direct repeats flanking the IS*6110* element has been cited as evidence of such events. However, studies to date have not provided substantial evidence in the form of sequence data from predecessor strains, and little attention has been focused on clinical isolates of *M. tuberculosis*. To address these shortcomings, this study investigates examples of deletion polymorphism in the DR region of clinical isolates of *M. tuberculosis*. Strains were selected for investigation on the basis of spoligotyping data which demonstrated the absence of the 5' portion of the DR region. PCR amplification and sequence comparisons were used to map the deletion junctions. Where possible, putative predecessor strains were analyzed to investigate mechanisms whereby deletion events occurred.

Analysis of the structure of the DR region in the strains 704 and 397, and comparison to putative predecessor strains (1227 and 176, respectively) demonstrates that in contrast to previous studies (Fang *et al.*, 1999a; Brosch *et al.*, 1999), the putative predecessors of 704 and 397 both contain DR flanking region insertions in inverse orientation to those in the DR region itself. Although the sequence data demonstrates deletion of one insertion element and the intervening sequence from the predecessor, conventional understanding dictates that homologous recombination between directly repeated sequences results in deletion events, whereas homologous recombination between inverted repeats leads to inversion of the intervening sequence (Kleckner *et al.*, 1979; Petes and Hill, 1988; Roth, 1996). Homologous recombination is often RecA-mediated, and usually requires sequence homologies of 50 bp or more (Lloyd and Low, 1996). Rarely, homologous recombination can occur between shorter regions of homology (>20 bp), and in such cases, is often RecA independent (Lloyd and Low, 1996). It is difficult to reconcile the inversely orientated IS*6110* elements with deletions by a classical homologous recombination pathway. It is also unlikely that the inverted repeats of the IS*6110* elements act as sites promoting homologous recombination, as these are only 28 bp in length. Furthermore, the repeats are imperfect and demonstrate a 3 bp difference, and inspection of the sequence data does not reveal rearrangement of any of the inverted repeats.

The data does not however, rule out homologous recombination as a possible deletion mechanism, as it is possible to speculate on an alternative scenario (Fig. 4.5A) where a second IS*6110* element integrates into the DR region, immediately flanking, but in opposite orientation to the existing insertion. The presence of an IS*6110* element in the same orientation as the DR flanking region insertion could then mediate deletion via a conventional homologous recombination pathway (Fig. 4.5A). If the DR locus is indeed a true preferential integration locus, then this is a plausible hypothesis. In this regard, it is interesting to note that sequence data suggests the occurrence of a second discrete insertion in the 973 isolate within the same DVR (DVR 35). Furthermore, Groenen *et al*. (1993) reported a second DR region insertion into a duplication of DVR 35, which might suggest that some inherent sequence characteristic promotes frequent insertion into this particular site (Groenen *et al*., 1993). In addition, multiple independent insertions into identical sites within the *ipl* locus have been observed (Fang and Forbes, 1997; Fang *et al*., 1999b; K. Forbes, personal communication), sometimes in opposing orientations, which is likely to be a general characteristic of preferential integration loci.

An alternative mechanism is that of RecA-independent, spontaneous deletions mediated by short sequence homologies (Albertini *et al*., 1982; Lloyd and Low, 1996). In this model, short sequence homologies of as little as 5-8 bp can mediate large deletions (700-1000 bp) by a slipped mispairing mechanism. It is possible to apply this model to the strains investigated here, by considering the terminal part of the inverted repeats (excluding the 3 bp difference) as the substrate for this type of recombination event (Fig. 4.5B). Slipped mispairing between the two right hand inverted repeats of the IS*6110* elements could then lead to deletion of one element and the intervening sequence.

Other possible deletion mechanisms were discussed by Fang *et al.* (1999a), and included site-specific recombination and transpositionally mediated mechanisms. Site-specific recombination involves recombination between predetermined loci (Hallet and Sherratt, 1997) but as with homologous recombination, site-specific recombination between inverse repeats will lead to inversion of the intervening sequence (Nash, 1996), and is therefore unlikely to be occurring here. The transpositionally-mediated mechanisms include deletion of target DNA on IS integration, IS excision with associated deletion and deletion of sequence flanking the donor IS copy on duplicative transposition (Craig, 1996). However, these mechanisms are all considered to be unlikely, as two independent examples of closely related predecessor strains with a second IS*6110* copy in an identical position to the descendants have been identified. It seems implausible that this would occur by chance following a transpositionally-mediated recombination event. It is evident that numerous mechanisms can be invoked to explain the observed deletion events. While homologous recombination via an unidentified intermediate strain may be the most likely explanation, in the absence of confirmed immediate predecessor strains, and knowledge of IS*6110* transposition mechanisms, this remains unproven.

Figure 4.5 Proposed models for deletion mechanisms. (A) Transposition followed by homologous recombination between directly repeated IS*6110* elements. 1: IS*6110* transposition in opposite orientation to, and immediately flanking, existing DR region insertion; 2: Head to head conformation of two IS*6110* elements within DR region, possibly an unstable intermediate; 3: Homologous recombination between external, directly-repeated IS*6110* elements; 4: Resulting deletion mutant, with loss of intervening sequence and two IS*6110* elements.

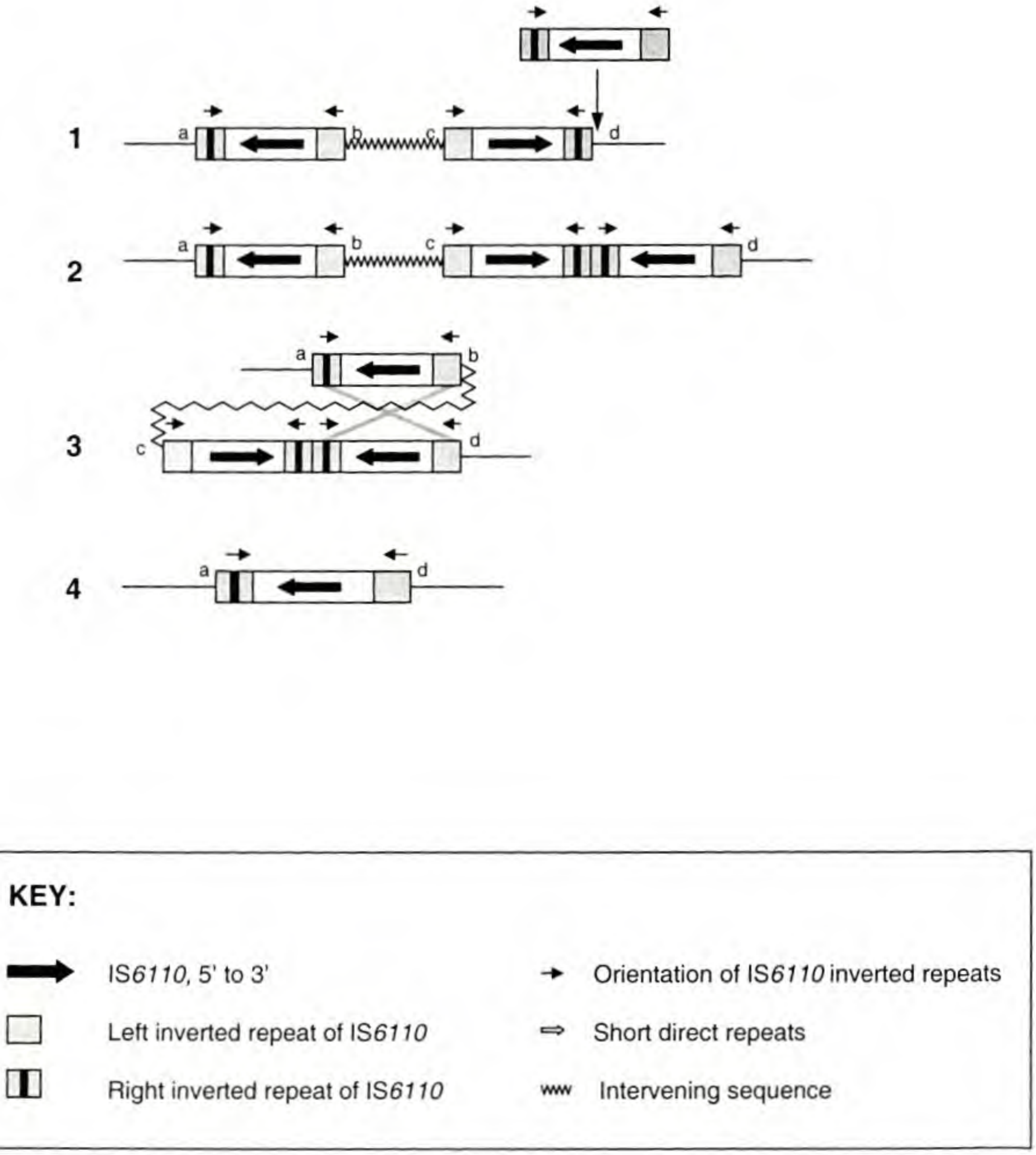
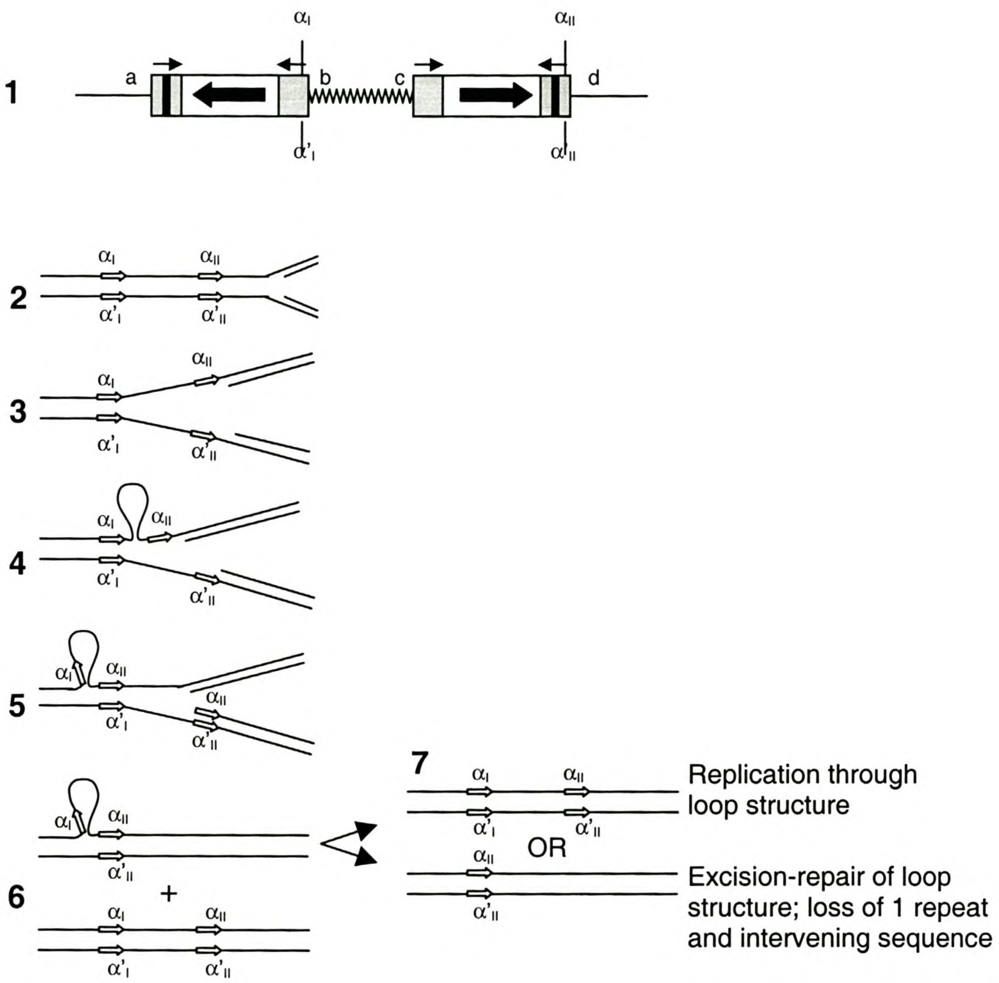

1
a
b
c
d
2
a
b
c
d
3
a
b
c
d
4
a
d
KEY:
IS*6110*, 5' to 3'
Left inverted repeat of IS*6110*
Right inverted repeat of IS*6110*
Orientation of IS*6110* inverted repeats
Short direct repeats
Intervening sequence

Figure 4.5 Proposed models for deletion mechanisms (cont.). (B) Spontaneous deletion mediated by slipped mispairing between short regions of homology within inverted repeats of IS*6110*. 1: Possible sites of short sequence homology within terminal repeats of IS*6110*. 2-7: slipped mispairing occurs during DNA replication, leading to spontaneous deletion, as proposed by Albertini *et al*. (1982). Figure adapted from Albertini *et al*, 1982.


1
a
b
c
d
αI
αII
α'I
α'II
2
3
4
5
6
+
7
OR
Replication through loop structure
Excision-repair of loop structure; loss of 1 repeat and intervening sequence

Interestingly, in the strain 392, the point of IS*6110* integration into the DR region is in an identical position but the element is in opposite orientation to that of the closely related 397. It is proposed that the opposing orientation of the remaining IS*6110* element is the result of two distinct recombination events. This is supported by sequence analysis of the DR flanking region, which revealed slightly offset IS*6110* integration points in the two strains, indicative of discrete IS*6110* transposition events. In the case of 392, deletion may be mediated by homologous recombination between directly repeated IS*6110* elements. However, in the absence of a defined predecessor strain, this remains speculative.

Analysis of the strain 780 revealed the largest deletion relative to *M. tuberculosis* H37Rv, identical to the deletion recently identified by Van Embden *et al*. (2000). In this case, the DR IS*6110* insertion occurs into a newly identified DVR (Beggs *et al*., 2000; Van Embden *et al*., 2000), and is not identical to the *M. tuberculosis* H37Rv insertion. Without the support of data from predecessor strain types, it is difficult to speculate on the mechanism whereby this deletion occurred. A point of interest however, is that this deletion appears to be strain-family specific, and has been applied as a marker to genotype multidrug-resistant strains resembling "Strain W" in the Ravensmead-Uitsig study community (South Africa) (Van Rie *et al*., 1999a). Strain W, first identified in several large multidrug-resistant tuberculosis outbreaks in New York in the 1990's (Coronado *et al*., 1993; Frieden *et al*., 1993; Friedman *et al*., 1995; Bifani *et al*., 1996; Moss *et al*., 1997), is related to the drug-susceptible "Beijing" strain group which is predominant in Asian countries (Van Soolingen *et al*., 1995; Huh *et al*., 1995; Lin *et al*., 1996; Palittapongarnpim *et al*., 1997). It is hypothesized that the Beijing-type strains are ancestors of the Strain W-type strains, and that the relative genetic homogeneity of these strain groups reflects relatively recent clonal expansion (Van Soolingen *et al*., 1995; Kurepina *et al*., 1998; Bifani *et al*., 1999). The failure to identify a predecessor strain to 780 with an intact DR region may therefore reflect an evolutionary event which occurred prior to the global dissemination of the strain family group, and its introduction into the Ravensmead-Uitsig study community.

This study has identified a number of discrete IS*6110* integration points within an 8 kb region encompassing part of the DR locus (n = 3) and its flanking sequence (n = 4) in the 5 strains analysed. Other investigators have reported 3 additional integration points within the DR region (Groenen *et al.*, 1993; Van Embden *et al.*, 2000; Filliol *et al.*, 2000; Benjamin *et al.*, 2001). Therefore, the DR region and its flanking sequence represents a further example of a preferential insertion locus. The preferential integration of IS*6110* apparently causes the region to be prone to IS*6110*-mediated deletion events. A similar scenario was suggested for a 20 kb hypervariable region recently described by Ho *et al.* (2000). Factors influencing the frequency of recombination events within a preferential integration site are as yet unknown, but it is expected that proximity and relative orientation of IS*6110* elements, as well as the sequence of the surrounding chromosomal region will play a role. For example, it is tempting to speculate that the failure to identify a predecessor strain for 392 may indicate that such a predecessor is an unstable intermediate either due to favorable orientation and/or close proximity of IS*6110* elements. Alternatively, this could simply reflect a rare strain type, and this remains to be investigated.

Further investigation will be necessary to establish whether deletion-associated hypervariability is a feature of other IS*6110* preferential integration regions that have been identified (Fang *et al.*, 1997; Kurepina *et al.*, 1998; Sampson *et al.*, 1999; Warren *et al.*, 2000; Sampson *et al.*, 2001). Preliminary analysis (not shown) demonstrates that the majority (11/12) of deletions identified in a previous study (Warren *et al.*, 2000) correspond to preferential integration regions as identified by mapping of IS*6110* insertion sites to the *M. tuberculosis* H37Rv genome (Sampson *et al.*, 1999; Warren *et al.*, 2000; Sampson *et al.*, 2001). Together, these findings support the role of IS*6110* as a mediator of genome diversity in *M. tuberculosis*. In the absence of horizontal gene transfer (Cole *et al*., 1998) and limited sequence diversity (Sreevatsan *et al*., 1997; Musser *et al*., 2000) that is observed in the pathogen, this may represent an important mechanism for strain adaptation.

Results presented here highlight two important cautions for the interpretation of spoligotyping patterns, and the method's widespread application for strain typing. Firstly, the absence of a signal for a specific DVR does not necessarily imply the absence of that DVR. This can occur as a result of disruption of the spoligotyping oligonucleotide priming site as exemplified by 973. This results in the absence of a PCR product encompassing the flanking spacer sequence, and in turn leads to the absence of hybridizing signal for the spacer in question. This study demonstrates absence of hybridizing signal for DVR 35 despite its presence in 973. Similar results have been reported for DVRs 41 and 19 in other clinical isolates (Filliol *et al*., 2000; Benjamin *et al*., 2001), where IS*6110* insertion was shown to have disrupted the priming sites immediately adjacent to the spacer sequences for these DVRs. Secondly, contrary to a recent report which suggested that "fortuitous convergence" of spoligotype patterns is unlikely (Sola *et al*., 2001), it is evident that identical spoligotype patterns may evolve in parallel. This is demonstrated by comparison of the strains 392, 397 and 704, which fall within 2 unrelated family groupings. These strains share identical spoligotypes, but sequence data suggests three substantially different origins for these patterns. These considerations are important both for epidemiological interpretations and evolutionary studies.

This study underscores the importance of investigating the molecular basis for IS*6110* DNA fingerprint and spoligotype diversity. Understanding of the mechanisms and frequency of IS*6110*-mediated genome rearrangements is essential for the accurate interpretation of

IS*6110*-derived DNA fingerprint patterns. Results presented here strongly support the occurrence of preferential IS*6110* integration into the DR region and its flanking sequence. This implies that IS*6110* DNA fingerprint patterns and spoligotype patterns may evolve in concert. Therefore, the contribution of IS*6110* transposition events and rearrangements to spoligotype diversity should be taken into account in the context of both epidemiological and evolutionary studies.

CHAPTER 5

THE PPE GENE FAMILY: EXTENT AND MECHANISMS OF VARIATION

The results presented in the following chapter will be submitted to *The Journal of Bacteriology* as: **The PPE gene family of *Mycobacterium tuberculosis*: extent and mechanisms of variation in clinical isolates**. Sampson, S.L., Warren, R.M., Richardson, M., Van Der Spuy, G.D., Van Helden, P.D.

(The style of the text and numbering of sections has been altered to conform with the style of this thesis. Literature cited in the text takes the form of author name and year of publication as opposed to the number format specified by The Journal of Bacteriology. *All cited literature is compiled into a single list at the end of the thesis for ease of reference.)*

5.1 Introduction

Analysis of the complete genome sequence of *Mycobacterium tuberculosis* H37Rv highlighted the existence of two large gene families, namely the PE and PPE gene families (Cole *et al.*, 1998). These glycine-rich families comprise about 10% of the genome and are named after the Pro-Glu and Pro-Pro-Glu motifs at the predicted N-termini of their encoded proteins (Cole *et al.*, 1998). The function of the proteins encoded by the PE and PPE gene families has not been established, although numerous hypotheses exist. It has been suggested that they may be of immunological importance, either by inhibition of antigen processing in host cells, or by providing a source of antigenic variation in an otherwise genetically homogenous bacterium (Cole *et al.*, 1998).

The PE gene family contains about 100 members, grouped according to a relatively conserved 110 amino acid (aa) N-terminal region (Cole *et al.*, 1998). The predicted proteins vary considerably in length, attaining up to 1400 aa. The gene family can be divided into a number of subgroups, the largest of which is the PE-PGRS (polymorphic GC-rich sequence) subgroup. The PE-PGRS genes are characterized by stretches of the PGRS consensus repeat CGGCGGCAA. These encode multiple tandem repeats of glycine-rich motifs (often Gly-Gly-Ala or Gly-Gly-Asn).

The PPE family contains 68 members that have in common a relatively conserved 180 aa N-terminal region (Cole *et al.*, 1998), and range in size from 77 aa to 3300 aa. As for the PE gene family, the PPE family can also be classified into a number of subgroups, one of which is the PPE-MPTR (major polymorphic tandem repeat) subgroup. The MPTR core consensus repeat is GCCGGTGTTG, separated by 5 bp spacers (Cole and Barrell, 1998) and encodes multiple copies of the motif Asn-X-Gly-X-Gly-Asn-X-Gly. A second group of PPE genes is a class characterized by a conserved motif (Gly-X-X-Ser-Val-Pro-X-X-Trp) at around position 350 in the amino acid sequence (Cole *et al*., 1998).

There is accumulating evidence that members of the two gene families are expressed during host infection. For instance, Espitia *et al.* (1999) demonstrated recognition of a recombinant PE protein by patient sera. Indirect evidence for expression of the PPE genes was provided by a study of a serine-rich antigen from *M. leprae* (Vega-Lopez *et al*., 1993), which was recognized by serum from both leprosy and tuberclosis patients. The protein was later recognized as one of the PPE family. Serological expression cloning has demonstrated that the PPE genes Rv1196 and Rv0915c encode potent T cell antigens, which induce protection in a mouse model of tuberculosis (Dillon *et al*., 1999; Skeiky *et al*., 2000). Further support for *in vivo* PPE gene expression is provided by *in situ* hybridization of human tissue samples with PPE gene probes (see Chapter 6).

Recent studies have provided indications that members of both gene families may play an important role during infection. Firstly, it was demonstrated that two PE-PGRS genes expressed in *Mycobacterium marinum* were essential for replication in macrophages and granuloma persistence, implicating the PE-PGRS proteins in virulence (Ramakrishnan *et al*., 2000). Secondly, it has been shown that a transposon mutant of the PPE gene Rv3018c is attenuated for growth in macrophages, implicating the gene as a possible virulence factor (Camacho *et al*., 1999).

It has been speculated that the PE and PPE gene families may provide the *M. tuberculosis* pathogen with the ability to vary antigens (Cole *et al*., 1998; Cole, 1999). Utilization of the PGRS (Ross *et al*., 1992) and MPTR (Hermans *et al*., 1992) repeats (found within the PE and PPE gene families respectively) as DNA fingerprinting probes, demonstrated polymorphism within domains containing these repeats, although to a lesser extent with MPTR-containing domains. PE and PPE gene polymorphisms were also identified by comparison between the genome sequences of *M. tuberculosis* H37Rv and *M. bovis* BCG (Cole *et al*., 1998).

A subset of the PPE-MPTR subgroup contains, in addition to the short MPTRs, other repetitive features, consisting of longer (69–78 bp) tandem repeats. Polymorphism associated with these repeat regions was first recognized during investigation of the region upstream of the *katG* gene (Zhang and Young, 1994; Goyal *et al.*, 1994b). Sequence homology comparisons demonstrate that this region corresponds to 75 bp tandem repeats within the PPE-MPTR gene, Rv1917c (Sampson *et al*., unpublished observation). A recent study also identified extensive polymorphism associated with 69 bp tandem repeats in the Rv1917c gene in *M. bovis* (O' Brien *et al*., 2000). These results are interesting in light of the finding that the *M. tuberculosis* genome is relatively conserved at the nucleotide level (Sreevatsan *et al*., 1997), which has lead to the hypothesis that the identification of polymorphic genes may be an indication of responsiveness to host immune pressure (Sreevatsan *et al*., 1997).

Emerging data clearly justifies further exploration of the PPE gene family. Therefore, in this study, a systematic investigation of polymorphisms associated with the gene family was undertaken. Firstly, phylogenetic analysis was applied to identify different PPE gene groupings. These results were utilized in the selection of representative genes for further investigation. Probes complementary to representatives of the various subgroups were applied to gain an understanding of polymorphisms associated with the PPE gene family in clinical isolates of *M. tuberculosis*. Polymorphic variants of PPE genes were identified in clinical isolates and examined to establish the mechanism responsible for these variants. The stability of PPE genes was assessed by both hybridization- and PCR-based approaches.

5.2 Methods

5.2.1 Study setting and strain selection

M. tuberculosis isolates obtained from an ongoing molecular epidemiology study of tuberculosis in a high incidence community in the Western Cape, South Africa have been typed by IS*6110* DNA fingerprinting according to an internationally standardized protocol (Van Embden *et al*, 1993; Warren *et al*., 1996a; Warren *et al*., 1999). Strains that are >65% related according to IS-3' DNA fingerprint patterns are routinely classified into strain family groupings (Warren *et al*., 1999), and typical representatives of various strain families were selected for further analysis.

Eighty-six isolates (plus the reference strain Mt14323, (Van Embden *et al*., 1993)) were analyzed by hybridization with the probes complementary to the 5' termini of the PPE genes Rv1787, Rv1917c, Rv2123, Rv3018c and Rv3429 (see section 5.2.4). Strain groups analyzed included: (i) a group of 18 randomly selected unrelated isolates representing 12 strain families, (ii) 51 representatives of the three dominant high IS*6110* copy number strain family groupings identified in the study community (families 11, 28 and 29; with an average of 14, 10 and 19 IS*6110* copies respectively; Warren *et al*., 1999), and (iii) 17 low copy number strains (IS*6110* copy number = 4). An additional 16 pairs of isolates representing relapse cases were hybridized with all 5 probes to assess PPE gene stability (Van Rie *et al*., 1999b). A further 280 isolates were hybridized with only the Rv1787 and Rv1917c probes. These included representatives of various strain family groupings identified within the study community (n = 194), as well as isolates cultured from multiple sites of infection from post-mortem material (n = 86) (Du Plessis *et al*., 2001).

One-hundred and thirty eight clinical isolates were investigated by PCR analysis of tandem repeat domains. These included 60 isolates representing strain family grouping 11; 18

isolates from strain family grouping 28; 41 isolates with 4 copies of IS*6110*; and an additional 19 isolates representing a range of strain family groupings. This set of 138 isolates included 11 paired first and last isolates from patients with persistent disease, and therefore represented 127 patients.

5.2.2 Sequence analysis

Sixty-eight PPE DNA and protein sequences from *M. tuberculosis* H37Rv were downloaded from the Tuberculist website (http://www.genolist.pasteur.fr/Tuberculist/). Ten of the predicted PPE proteins did not contain the characteristic N-terminal "PPE" motif. In 5 of these cases (Rv0305c, Rv3425, Rv3426, Rv3429, Rv3892c), one or other of the proline residues within the PPE motif was substituted, but the remainder of the protein sequence showed significant homology to other PPE proteins. These were included in further analyses. In the other five cases, the upstream region was disrupted either by IS*6110* insertion or apparent frameshift mutations leading to gene disruption and loss of the PPE motif from the proteins encoded by these genes. These five examples (Rv0304c Rv0354c, Rv2353c, Rv3021c, Rv3738c) were excluded from further analyses. *M. tuberculosis* CDC1551 PPE sequences were obtained from the TIGR website (http://www.tigr.org/tdb/CMR/gmt/htmls/SplashPage.html). PPE gene and pseudogene sequences from *M. leprae* (Cole *et al*., 2001) were obtained from the Leproma website (http://www.genolist.pasteur.fr/Leproma/).

Amino acid and nucleotide sequence repeats were identified using DNAMAN software (Version 4.1, Lynnon BioSoft). Dot matrix comparisons and the direct repeat finder feature of DNAMAN were used to identify and localize repeat regions in nucleotide and amino acid sequences. Protein hydrophobicity and hydrophilicity plots were also performed using DNAMAN software. *Pvu*II recognition sequences were identified using DNAMAN, and their positions on the *M. tuberculosis* H37Rv chromosome were used to calculate the lengths of PPE-containing *Pvu*II restriction fragments. Sequence homology comparisons were performed using the BLASTN, BLASTP and BLASTX algorithms at the NCBI (http://www.ncbi.nlm.nih.gov) and Tuberculist websites.

The PPE genes were analyzed for the presence of signal peptide motifs with the SignalP program (Version 1.1) (http://www.cbs.dtu.dk/services/SignalP/; Nielsen *et al.*, 1997), which uses a neural network method trained on a gram-positive data set. SignalP scores were determined for the first 70 aa of each PPE protein, using the default cut-off parameters of the program. Possible transmembrane helices within the PPE proteins were identified with the TMHMM (Version 1.0) program (http:www.cbs.dtu.dk/services/ TMHMM-1.0; Sonnhammer *et al.*, 1998). This program uses a hidden Markov model, trained on a dataset of 160 proteins (eucaryotic and procaryotic). The cut-off value of the program was 0.8, but weak predictions of between 0.5 and 0.8 were also included. To identify conserved protein motifs, the Prosite database was scanned with the PrositeScan program (http://www.isrec.isb-sib.ch/software/PSTSCAN_form.html).

5.2.3 Phylogenetic analysis

The first 180 aa of sixty-three PPE protein sequences were aligned using ClustalW software accessed via the Pôle Bio-Informatique Lyonaise website (http://pbil.ibcp.fr/cgi-bin/align_clustalw). Five protein sequences which did not contain the characteristic N-terminal "PPE" motif (Rv0304c Rv0354c, Rv2353c, Rv3021c, Rv3738c) and could not be reliably aligned were excluded from the analysis. The alignment data was subjected to phylogenetic analysis using the neighbor joining algorithm PAUP* 4.0 (phylogenetic analysis using parsimony [*and other methods]), Sinauer Associates (Lawrence *et al.*, 1989). The PPE gene Rv3873 was selected as the outgroup, based on an unrelated study of ESAT-6 gene clusters (Gey van Pittius *et al.*, 2001). The ESAT-6 gene clusters contain PPE genes, and analysis of these regions suggested that Rv3873 was present within the most evolutionarily ancient of the regions investigated.

A confidence interval for the internal topology of the tree was established using the bootstrapping method, whereby only nodes occurring in > 50% of the trees were assumed to be significant (Felsenstein, 1985). A consensus tree was generated using the majority rule formula. All branches with a zero branch length were collapsed.

5.2.4 DNA Hybridization

Genomic DNA was digested with *Pvu*II and electrophoretically fractionated together with 8 ng DNA Molecular Weight Marker X (Roche) in 0.8% agarose, as previously described (Warren *et al.*, 1996a). The DNA was Southern transferred onto a Hybond N+ membrane, and fixed by heating at 80°C for 2 hours. The membranes were sequentially hybridized with ECL-labeled IS-3' (Van Embden *et al.*, 1993), IS-5' (Warren *et al.*, 2000) and Marker X as recommended by the manufacturer (Amersham).

PPE 5' terminal probes were generated by separate PCR amplification of regions from the genes Rv1787, Rv2123, Rv1917c, Rv3018c and Rv3429, with primers and conditions as detailed in Table 5.1. Probes were radio-labeled with [α^{32}P] dCTP using the "Prime-It" Random labeling kit (Stratagene), according to the manufacturer's instructions. Hybridization was performed in formamide hybridization buffer (50% formamide, 5X SSPE, 0.1% SDS, 5X Denhardts solution, 0.1 mg of denatured herring sperm DNA per ml) at 42°C for 16h. The membranes were washed at moderate stringency (2X SSC, 0.1% SDS; 55°C), and the resulting hybridization patterns were visualized by autoradiography. Between each hybridization step, the membranes were stripped by incubation in 0.4 M NaOH at 50 °C for

45 min to denature all DNA hybrids. The membranes were then neutralized in 0.2 M Tris-HCl pH 8.0, 0.1x SSC and 0.1x SDS at 50 °C for 15 min.

5.2.5 DNA manipulation

All PCR amplification reactions were performed with the Qiagen HotStar Taq system (Qiagen), according to the manufacturer's instructions. Reaction mixtures contained 50-100 ng chromosomal DNA template, 50 pmol of each primer (Tables 5.1 and 5.2), 1x Reaction Buffer, 2 mM $MgCl_2$ and 1 U enzyme. Cycling conditions included a 15 min incubation at 95°C, followed by 35 cycles of 94°C for 30 s, 30 s at the specified annealing temperature, and 72°C for 1 min, followed by a 10 minute extension step at 68°C (Tables 5.1 and 5.2). PCR products were electrophoretically fractionated in 2% agarose (1x TBE, pH 8.3), then visualized under UV after ethidium bromide staining. PCR products were purified with the Wizard PCR purification kit (Promega), and subsequently cloned using the pGEM-T Easy vector system (Promega). All cloning steps were performed according to standard protocols (Sambrook *et al.*, 1989). Plasmid DNA was isolated using the Wizard DNA Purification system (Promega), then restriction mapped and/or sequenced. All DNA sequencing reactions were performed on an ABI automated sequencer.

Table 5.1 PCR primers used to amplify PPE 5' termini.

Name	Sequence (5'-3')	Tm	Application
ppe-15	TGG ACT TCG GGG CGT TAC	58°C	*Amplification of 499 bp 5' terminal region from Rv1787 and Rv1790*
ppe-16	AAC GGA ATC AAC CGC GAC	56°C	
ppe-17	TTC AAC TCC GTG ACG TCG	56°C	*Amplification of 471 bp 5' terminal region from Rv1917c*
ppe-18	CAG CAC ACC CTT GGA ACT G	60°C	
2123F	ATG TGG TTC GCA GTT CCG C	60°C	*Amplification of 227 bp 5' terminal region from Rv2123*
2123R	GTT AGC CAA TAC CGG AAC GG	62°C	
3018cF	ATT CGG CGC TGC TAA GTG C	60°C	*Amplification of 160 bp 5' terminal region from Rv3018c and Rv3021c*
3018cR	AAC TCA GCA CTG GGA CCC TG	64°C	
3425F	CAT CCA ATG ATA CCA GCG GAG	64°C	*Amplification of 148 bp 5' terminal region from Rv3429*
3429R	GCT CGC CGA GCC TGT CGG	64°C	

Table 5.2 PCR primers used to amplify PPE tandem repeat regions

Name	Sequence (5'-3')	Tm	Application
ppe-1	CGC ACC GGA ATT GAA GAA G	58°C	*Amplification of 75 bp tandem repeats from Rv1753c*
ppe-2	CAT TGA CCG GCC CTA TTG TC	62°C	
ppe-5	ACC TGA TCT GAC TCT GCC GC	64°C	*Amplification of 78 bp tandem repeats from Rv1753c*
ppe-6	ACT TCC GGA ATC TGC AAT GG	60°C	
ppe-8	CAA GTT CAG GGG GGA TCC	58°C	*Amplification of 2 adjacent sets of 69 bp tandem repeats from Rv1917c*
ppe-9	ACT GAG CGT CGA AGT GAA TG	60°C	
ppe-11	GTG ACA GTG AGT GGT CAA ATC G	66°C	*Amplification of 75 bp tandem repeats from Rv1917c*
ppe-12	GTT CCA GAA GCC AGA TCC G	60°C	
ppe-13	CTT CCG TCT CTG GAA ATA CCC	64°C	*Amplification of 78 bp tandem repeats from Rv1918c*
ppe-14	AGC CGC CTA TAC TTA TTT GGG	62°C	

5.3 Results

5.3.1 Phylogenetic analysis of the PPE gene family

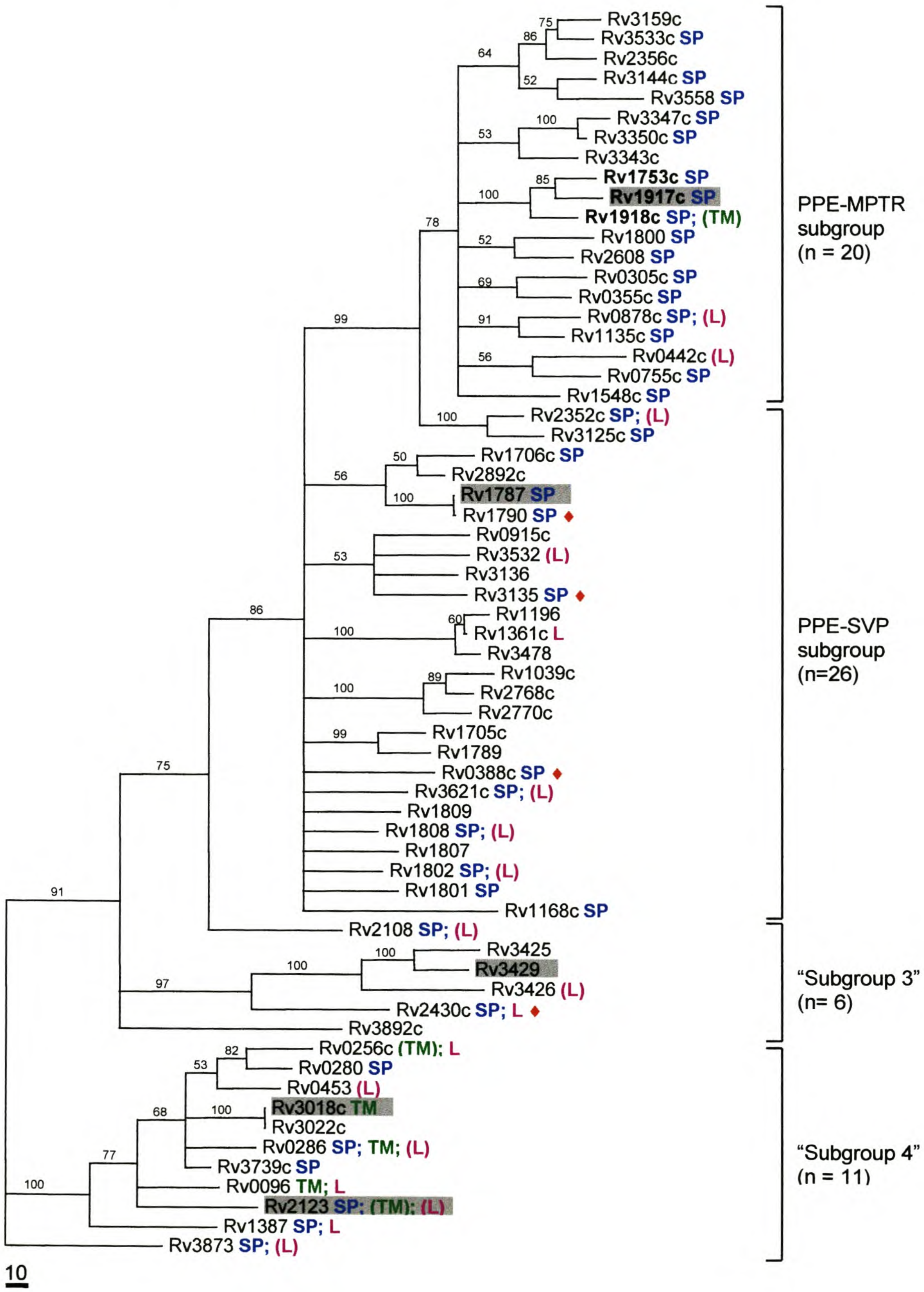
The evolutionary history of the PPE gene family was reconstructed by genetic distance analysis using the neighbor joining algorithm PAUP* 4.0 (Lawrence *et al.*, 1989). Analysis of evolutionary relationships between the PPE proteins reveals at least 4 distinct subgroups (Fig. 5.1). The two largest of these, as previously described by Cole *et al.* (1998) are (i) the PPE-MPTR subgroup (n = 20) and (ii) the subgroup characterized by the amino acid motif GXXSVPXXW centered around position 350 in the amino acid sequence (n = 26), hereafter referred to as the "PPE-SVP" subgroup (Fig. 5.1). At least 2 smaller subgroups, designated subgroups 3 and 4, are also evident, with 6 and 11 members respectively (Fig. 5.1). The phylogenetic tree suggests that the PPE-MPTR subgroup has evolved from the PPE-SVP subgroup. This is supported by *in silico* analysis of sequence data which demonstrates the occurrence of single MPTRs within isolated members of the PPE-SVP group (Rv1361c, Rv3135 and Rv3136). Interestingly, PPE gene or pseudogene homologues identified within *M. leprae* (Cole *et al.*, 2001) are distributed throughout the various subgroups (Fig. 5.1), suggesting that the differentiation into distinct subgroups represents distant evolutionary events.

To determine whether the phylogenetically defined subgroups shared any common characterisitics, the programs SignalP, TMHMM and PrositeScan were used to identify signal peptides, transmembrane helices and conserved protein motifs respectively. Interestingly, the majority (16/20) of the PPE-MPTR subgroup members have a putative signal peptidase cleavage site (Fig 5.1), while there are proportionally fewer predicted signal peptides in the other 3 subgroups (12/26; 2/6; 6/11 for the PPE-SVP subgroup, subgroup 3 and subgroup 4, respectively). Five members of subgroup 4 have predicted transmembrane regions. Three of these proteins have values above the default 0.8 cut-off of the program, while the values of

Figure 5.1. **Phylogenetic reconstruction of evolutionary relationships between 180 aa N terminal domains of PPE proteins.** The bootstrapping method in association with the neighbour joining algorithm PAUP* 4.0 was used to generate a consensus tree from 1000 bootstrapped trees using the majority rule formula. Only nodes occurring in > 50% of the trees were considered to be significant (internal labels). Branch lengths are proportional to the number of evolutionary events (scale bar equals ten evolutionary events), and all branches with a zero length were collapsed. The tree length was 2946 and showed a consistency index of 0.432 and a homoplasy index of 0.568. The PPE Rv number is given at the terminus of each branch.

Four distinct subgroups are indicated. These include the PPE-SVP subgroup (characterized by a conserved motif (Gly-X-X-Ser-Val-Pro-X-X-Gly) at around amino acid position 350), and the PPE-MPTR subgroup (characterized by stretches of the 15 bp MPTR [major polymorphic tandem repeat, Hermans *et al*., 1992]). The three PPE-MPTR proteins in bold type (Rv1753c, Rv1917c and Rv1918c) indicate those containing long tandem repeat regions. Proteins highlighted in gray indicate the products of genes selected for design of DNA hybridization probes. The letters "SP" denote proteins with predicted signal peptide motifs. "TM" indicates those proteins with predicted transmembrane helices (score >0.8), with weak predictions (between 0.5 and 0.8) indicated by brackets. Leprae gene and pseudogene homologs are indicated by "L" and "(L)", respectively. Diamonds signify genes investigated by Musser *et al*. (2000).

Figure 5.1 Phylogenetic reconstruction of evolutionary relationships between 180 aa N terminal domains of PPE proteins. (See opposite for legend)



the remaining two lie between 0.5 and 0.8. Only one other PPE protein has predicted transmembrane helices, namely Rv1918c. This protein falls into the PPE-MPTR subgroup and its TMHMM values are weakly predictive (between 0.5 and 0.8) of transmembrane helices. PrositeScan identified a procaryotic membrane lipoprotein attachment site (PS00013) within the PPE-SVP protein Rv1706, which also has a predicted signal peptide motif. Motifs corresponding to gram-positive cocci surface proteins anchoring hexapeptides (PS00343) were identified in two PPE proteins, Rv1387 and Rv1808, which fall within subgroup 4 and the PPE-SVP subgroup respectively. Both of these proteins also contain signal peptide motifs.

5.3.2 Assessment of PPE gene family polymorphism by hybridization with 5' terminal DNA probes

To gain an overview of the stability of PPE-containing genomic domains, a hybridization-based approach was utilized. For this purpose, the genes Rv1917c, Rv1787, Rv2123, Rv3018c and Rv3429 were selected to represent various PPE subgroupings as defined by the phylogenetic tree (gray-shaded text on Fig. 5.1). Probes complementary to the 5' terminus of each gene were generated by PCR amplification using the primers detailed in Table 5.1. One-hundred and eighteen clinical isolates representing 12 strain families, with an IS*6110* copy number ranging from 3 to 22 were investigated (see Fig. 5.2A as an example). *Pvu*II-digested genomic DNA from these strains and the reference strain Mt14323 (Van Embden *et al.*, 1993) was sequentially hybridized with the five 5' terminal probes (Figs. 5.2B – 5.2F).

The Rv1917c and Rv1787 probes hybridized to the largest number of fragments, approximately 7 and 12, respectively (Fig. 5.2B and 5.2C). Both the Rv1917c and Rv1787 probes demonstrate one or two strongly hybridizing bands, along with a number of weaker intensity bands (Fig. 5.2B and 5.2C). The strongly hybridizing bands represent hybridization to the gene (or closely related homologue thereof) against which the probe was designed, while the lower intensity bands are thought to represent hybridization to less closely related PPE genes (Fig. 5.2B and 5.2C). This is supported by (i) *in silico* DNA sequence homology comparisons, and (ii) the *Pvu*II restriction fragment lengths predicted by analysis of the *M. tuberculosis* H37Rv genome sequence, which correlate with the sizes of the strongly hybridizing bands. Rv1917c and its neighboring gene, Rv1918c, share 69% nucleotide (nt) homology in the 5' probe region, and are predicted to lie on the same 12.0 kb *Pvu*II fragment. The gene Rv1753c (73% nt homology to Rv1917c in the probe region) is expected to lie on a 4.4 kb *Pvu*II fragment. Strongly hybridizing fragments of these approximate sizes are detected with the Rv1917c probe (Fig 5.2B). Rv1787 and the closely related Rv1790 (100% nt homology in probe region) are predicted to lie on 1.1 kb and 1.9 kb *Pvu*II fragments, respectively, and fragments of these sizes are detected with the Rv1787 5' terminal probe (Fig. 5.2C).

In contrast to the Rv1917c and Rv1787 probes, the probe complementary to the 5' region of Rv2123, hybridized to only one *Pvu*II fragment corresponding to its predicted size of 3.6 kb in the majority (78/86) of isolates (Fig. 5.2D). The Rv3018c 5' terminal probe hybridized to 2 *Pvu*II fragments in the majority (74/86) of isolates (Fig. 5.2E). These correspond to a *Pvu*II fragment containing Rv3018c, with a predicted size of 2.4 kb, and a fragment containing its closely related homologue, Rv3022c (100% nt homology in the 5' probe region), with a predicted size of 4.0 kb. Similarly, the Rv3429 probe hybridizes to two fragments in the majority (64/86) of cases. These correspond to fragments containing Rv3429 and its closely related homologue, Rv3425 (83% nt homology in 5' region), with predicted sizes of 2.9 kb and 4.0 kb, respectively. The limited number of hybridizing fragments observed for the Rv2123, Rv3018c and Rv3429 probes correlates

Figure 5.2 Hybridization analysis of clinical isolates. 18 randomly selected clinical isolates plus the reference strain Mt14323 were sequentially hybridized with IS-3', IS-5' (not shown) and five PPE 5' terminal probes. (A) IS-3', (B) Rv1917c, (C) Rv1787, (D) Rv2123, (E) Rv3018c, (F) Rv3429. Lanes (1) Mt14323; (2) 780; (3) 785; (4) 790; (5) 791; (6) 798; (7) 800; (8) 813; (9) 816; (10) 817; (11) 818; (12) 820; (13) 821; (14) 822; (15) 823; (16) 826; (17) 828; (18) 829; (19) 832. The domains complementary to the PPE gene probes and their homologues are indicated on the right of (B) to (E) (according to predicted *Pvu*II fragment sizes).

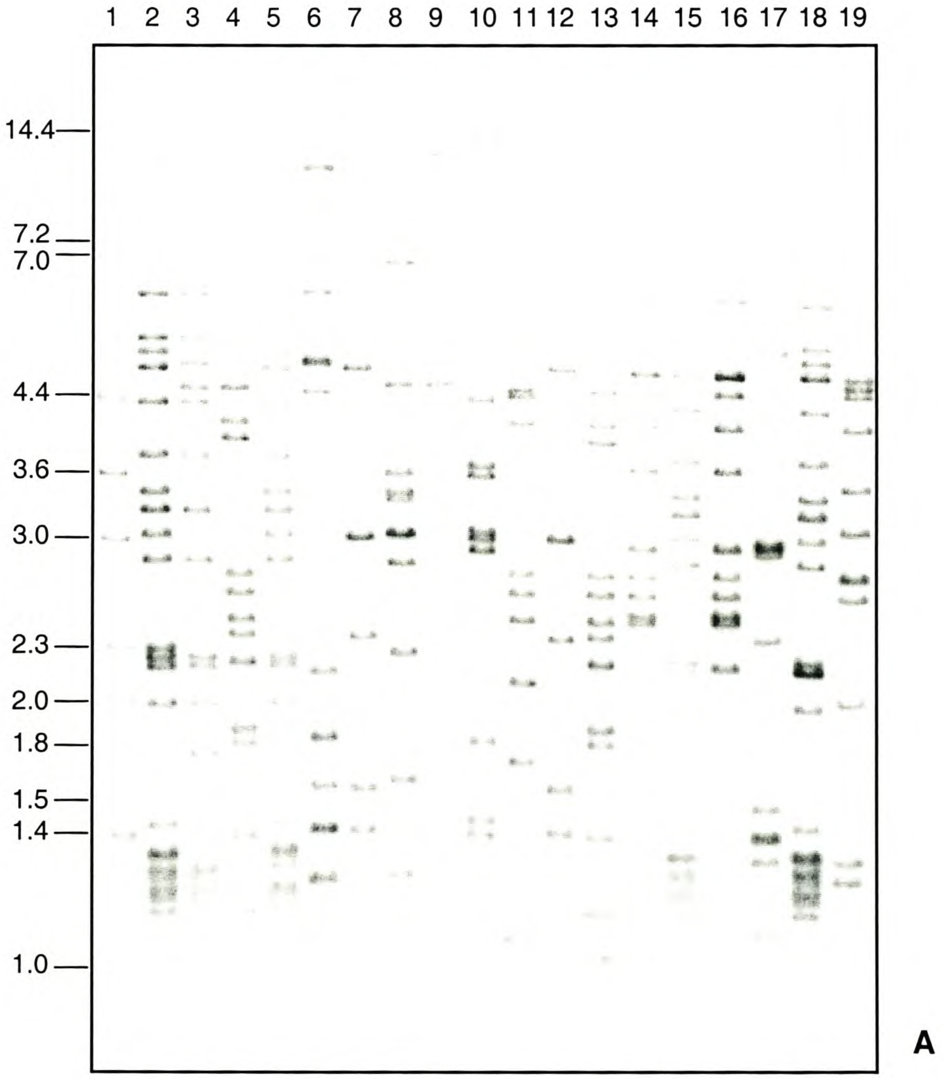
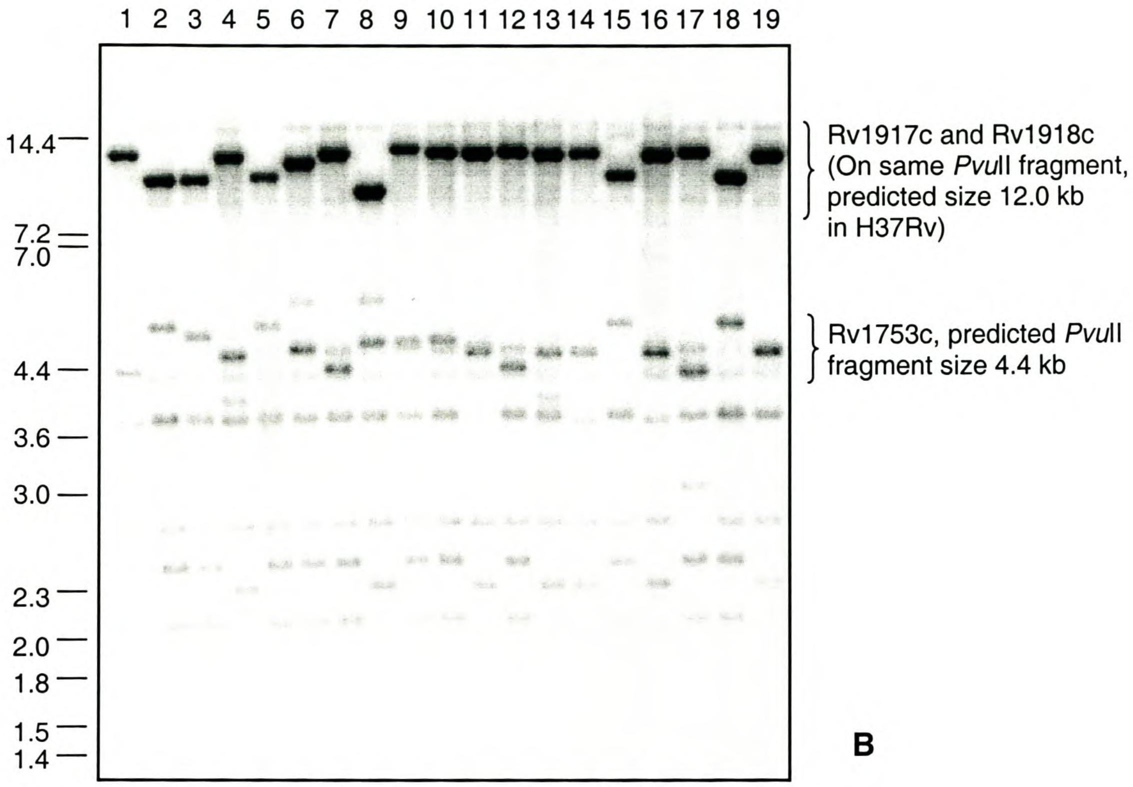


Figure 5.2 Hybridization analysis of clinical isolates (cont).


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
14.4
7.2
7.0
4.4
3.6
3.0
2.3
2.0
1.8
1.5
1.4
Rv1917c and Rv1918c (On same PvuII fragment, predicted size 12.0 kb in H37Rv)
Rv1753c, predicted PvuII fragment size 4.4 kb
B

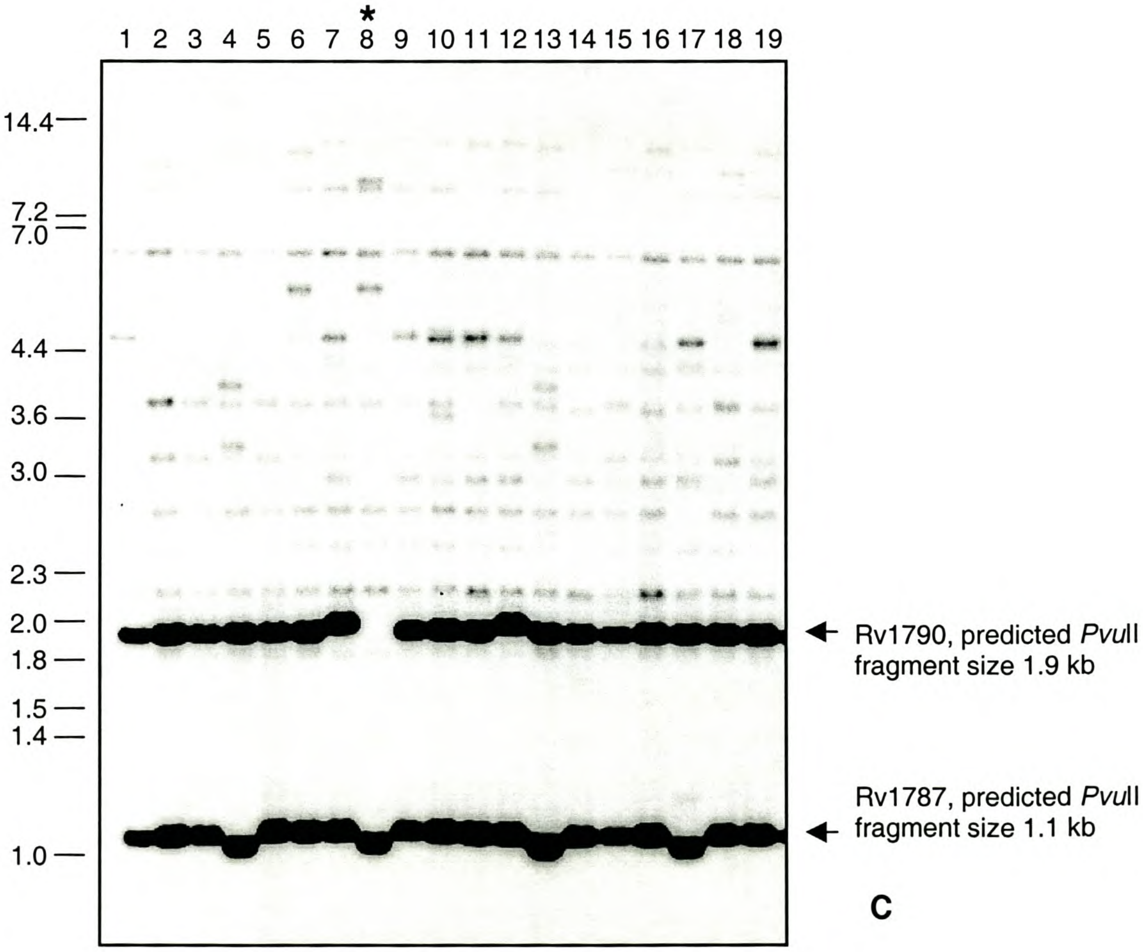
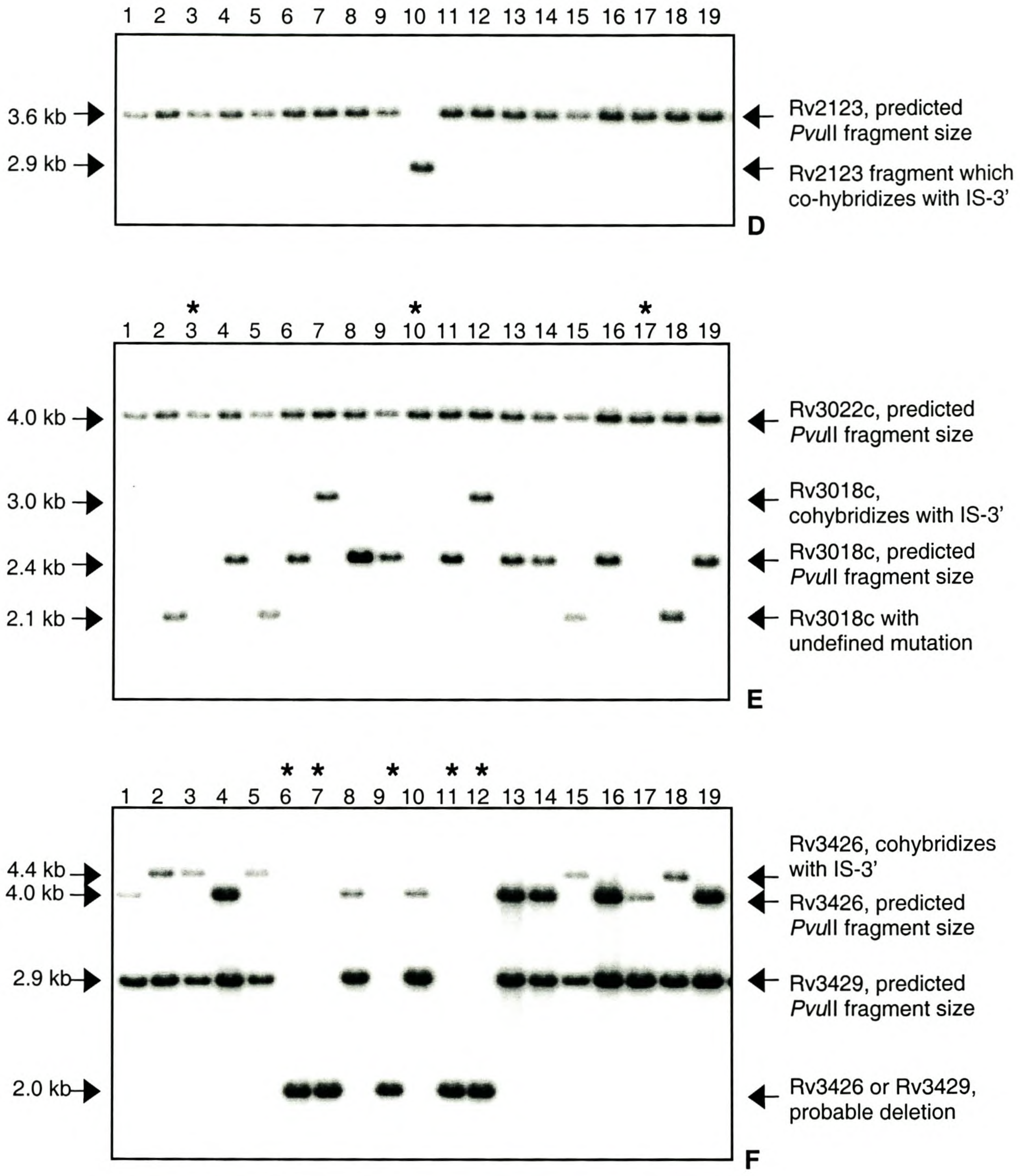

*
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
14.4
7.2
7.0
4.4
3.6
3.0
2.3
2.0
1.8
1.5
1.4
1.0
Rv1790, predicted PvuII fragment size 1.9 kb
Rv1787, predicted PvuII fragment size 1.1 kb
C

Figure 5.2 Hybridization analysis of clinical isolates (cont).


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
3.6 kb
2.9 kb
Rv2123, predicted *Pvu*II fragment size
Rv2123 fragment which co-hybridizes with IS-3'
D
* * *
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
4.0 kb
3.0 kb
2.4 kb
2.1 kb
Rv3022c, predicted *Pvu*II fragment size
Rv3018c, cohybridizes with IS-3'
Rv3018c, predicted *Pvu*II fragment size
Rv3018c with undefined mutation
E
* * * * *
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
4.4 kb
4.0 kb
2.9 kb
2.0 kb
Rv3426, cohybridizes with IS-3'
Rv3426, predicted *Pvu*II fragment size
Rv3429, predicted *Pvu*II fragment size
Rv3426 or Rv3429, probable deletion
F

* indicates putative deletion events

with the smaller size of the phylogenetic subgroups (subgroups 3 and 4) into which these genes are classified.

Analysis of the hybridization results demonstrates that while some hybridizing fragments are conserved in terms of size, others are highly variable. The Rv1917c probe, for example, hybridizes strongly to two hypervariable fragments (Fig. 5.2B). *Pvu*II restriction fragment analysis and DNA sequence homology comparisons suggest that these correspond to a fragment containing the Rv1917c/Rv1918c PPE-MPTR genes and a fragment containing Rv1753c, also a member of the PPE-MPTR subgroup. These genes all contain MPTR and longer tandem repeat regions, which could potentially give rise to the observed hypervariability (see below). However, the subgroups represented by the other four probes do not contain tandem repeat regions, therefore other mechanisms must also be responsible for the observed variation. For example, superimposition of the PPE 5' terminal hybridization patterns on IS-3' and IS-5' DNA fingerprints (not shown) reveals that some of the band shifts correlate with IS-3' (or IS-5') co-hybridization (see Tables 5.3 and 5.4). Although the IS/PPE co-hybridization could be co-incidental, previous studies have identified IS*6110* insertions within PPE genes (Cole *et al.*, 1998, Sampson *et al.*, 1999)). This supports the hypothesis that disruption of the hybridizing band by IS*6110* insertion with a subsequent change in electrophoretic mobility has occurred. An additional mechanism giving rise to RFLP is the deletion of PPE-containing domains (Figs. 5.2C, 5.2E and 5.2F; lanes marked by asterisk). Single nucleotide polymorphisms leading to the loss or gain of *PvuII* recognition sites, and other as yet undefined mechanisms may account for further variation that is observed.

To compare the stability of PPE-containing domains within different strain types, four strain family groupings (defined by >65% relatedness according to GelCompar analysis of IS-3' DNA fingerprint patterns) were analyzed with the five PPE 5' terminal probes. The strain family groupings were low copy number isolates (IS*6110* copy number = 4), and 3 groups of high copy number isolates (strain families 11, 28 and 29) representing the dominant strain family groupings within the study community (Warren *et al.*, 1999). Only the strongly hybridizing fragments detected with the Rv1787 probe were included in the analysis. Due to the complexity of the banding patterns, analysis of the Rv1917c hybridization results is not reported here, although domains detected with the Rv1917c probe were analyzed in more detail by PCR and sequencing, as described below. The hybridization results are summarized in Tables 5.3 and 5.4.

The hybridization patterns are conserved within the IS-3'-defined clusters, and a limited degree of variation was observed within the strain families (Table 5.3). In contrast, inter-family comparisons demonstrate genetic variability in the PPE-containing domains. These results may reflect relatively recent clonal expansion of strain clusters, which has limited the opportunities for the PPE domains to undergo genetic variation. The Rv3018c probe generates the most variable hybridization patterns, and the Rv2123 probe the most stable. As with the random selection of isolates, the Rv1787 and Rv1917c probes hybridized strongly to 1 or 2 bands, as well as to other lower intensity bands (not shown).

Table 5.3 PPE Hybridization variants. (Rv1917c results not included here; domains complementary to this probe were analyzed in more detail by a PCR-based approach.)

Family	Cluster number	Number of Samples	Rv1787	Rv2123	Rv3018c	Rv3429
Random isolates (12 strain families)	Clusters 1-16	18	A = 12 B = 3 C = 2 D = 1	A = 17 B = 1	A = 9 B = 4 C = 3 D = 2	A = 8 B = 5 C = 5
4B	Cluster1	7	A = 7	A = 7	D = 7	C = 7
	Cluster2	10	A = 5 D = 5	A = 3 C = 7	D = 10	C = 10
F29	Clusters 1 & 2	3	A = 3	A = 3	C = 3	B = 3
	Clusters 3 – 5	12	A = 12	A = 12	B = 12	B = 12
F28	Clusters 1 – 3	6	D = 6	A = 6	E = 6	A = 6
	Cluster 4	2	D = 2	A = 2	F = 2	A = 2
	Clusters 5 – 12	10	D = 10	A = 10	A = 10	A = 10
F11	Clusters 1 – 8	18	ND	A = 18	A = 18	A = 18

Table 5.4 Distribution of PPE Hybridization variants

Rv1787	Rv2123	Rv3018c	Rv3429

A = 57% (39/68)
B = 4% (3/68)
C = 3%(2/68)
D = 35% (24/68)
	A = 91% (78/86)
	B = 1% (1/86)
	C = 8% (7/86)
		A = 43% (37/86)
		B = 19% (16/86)
		C = 7% (6/86)
		D = 22% (19/86)
		E = 7% (6/86)
		F = 2% (2/86)
			A = 51% (44/86)
			B = 23% (20/86)
			C = 26%(22/86)

KEY:

ND = Not determined.

Rv1787
A = Rv1787 and Rv1790 fragments of predicted size (approx. 0.9 kb and 1.8 kb, respectively)
B = Rv1787 decrease (approx. 0.8 kb); Rv1790 predicted size
C = Rv1790 increase (approx. 1.9 kb); Rv1787 predicted size
D = deletion of Rv1790; decrease of Rv1787 (approx. 0.8 kb)

Rv2123
A = Rv2123 fragment of predicted size
B = Rv2123 decrease (approx. 3.0 kb), IS-3' co-hybridization
C = Rv2123 decrease (approx. 2.6 kb)

Rv3018c
A = Rv3018 and Rv3022 fragments of predicted size (approx. 2.4 kb and 4.0 kb, respectively)
B = Rv3018 decrease (approx. 2.0 kb); Rv3022 predicted size)
C = Rv3018 deleted; Rv3022 predicted size
D = Rv3018 increase, IS-3' / Rv3018 co-hybridization; Rv3022 predicted size
E = Rv3018 deleted; Rv3022 decrease (approx. 3.8 kb)
F = Rv3018 increase (approx. 2.7 kb), IS-3' and IS-5' co-hybridization; Rv3022 predicted size

Rv3429
A = Rv3429 and Rv3426 predicted size (approx. 4.0 kb and 2.9 kb respectively)
B = Rv3426 increase (approx. 4.4 kb), IS-3' co-hybridization; Rv3429 predicted size
C= Deletion, remaining fragment approx. 2.1 kb

5.3.3 Identification of tandem repeat regions

To investigate the molecular basis for the extensive polymorphism detected with the Rv1917c probe, *in silico* analysis of PPE gene and protein sequence was performed using DNAMAN software (Version 4.1, Lynnon Biosoft). DotMatrix comparisons and visual inspection confirmed the presence of MPTR repeats in all of the members of the PPE-MPTR subgroup defined by phylogenetic analysis (see Fig. 5.3A for example). In addition to MPTR repeat regions, regions consisting of longer tandem repeats (LTRs) were also identified in 3 of the PPE-MPTR genes (Rv1753c, Rv1917c and Rv1918c). No LTR regions were identified in any other PPE genes. The longer tandem repeat regions were localized using the direct repeat finding feature of DNAMAN and visual inspection of the gene and protein sequences (Fig. 5.3 and Fig. 5.4). Hydrophobicity and hydrophilicity plots were performed using DNAMAN (see Chapter 7, Fig. 7.2 for example). The LTR regions corresponded to hydrophobic stretches, whereas the MPTR repeat regions tended to be less hydrophobic.

5.3.4 PCR amplification of tandem repeat regions

The long tandem repeat regions identified by analysis of the Rv1753c, Rv1917c and Rv1918c PPE gene sequences suggested that tandem repeat expansion / contraction could be responsible for the extensive polymorphism detected by the Rv1917c probe (Fig. 5.2B). This hypothesis was investigated by PCR analysis of tandem repeat domains identified within Rv1753c, Rv1917c, and Rv1918c. Five PCR primer pairs were designed to amplify 6 blocks of tandem repeats from the 3 genes (Fig. 5.4). The regions amplified consisted of 78 bp and 75 bp repeats from Rv1753c; two regions of non-identical 69 bp repeats and a region of 75 bp repeats from Rv1917c; and a region of 78 bp repeats from Rv1918c (Fig. 5.4). LTR regions within the genomes of 127 clinical isolates, representing 10 strain family groupings (>65% related, as defined by GelCompar), and the reference strain *M. tuberculosis* H37Rv were assayed with the 5 primer pairs.

Figure 5.3 PPE sequence analysis. PPE sequence analysis was performed using DNAMAN software. The results for the PPE-MPTR gene, Rv1917c, are shown as an example. (A) A DotMatrix plot of the gene sequence against itself was used to identify repeat regions, which are highlighted by circles. (B) The direct repeat finding algorithm and visual analysis were combined to localize the repeat regions identified by DotMatrix comparisons. The repeat regions in Rv1917c are illustrated, aligned with the DotMatrix plot.

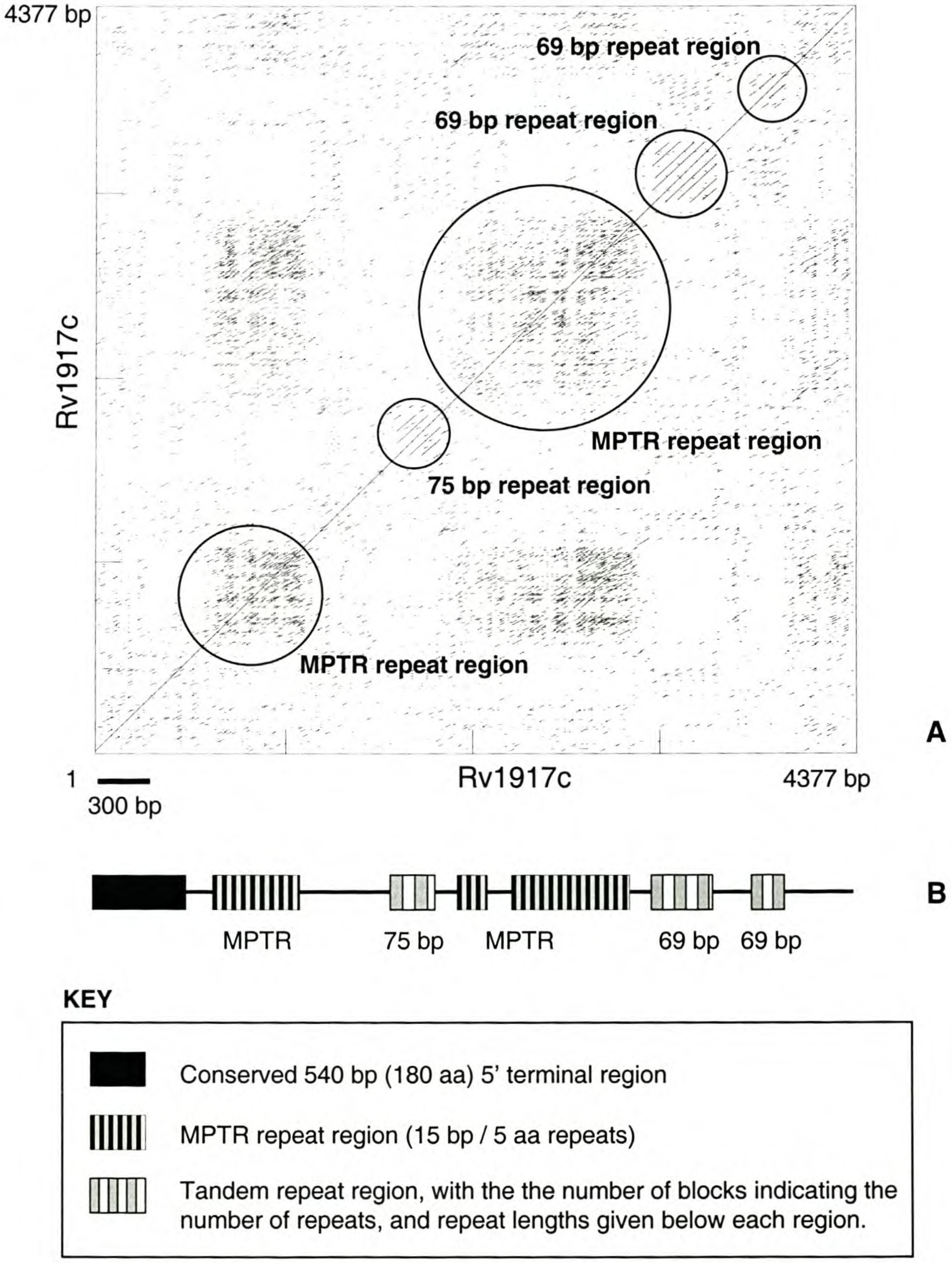
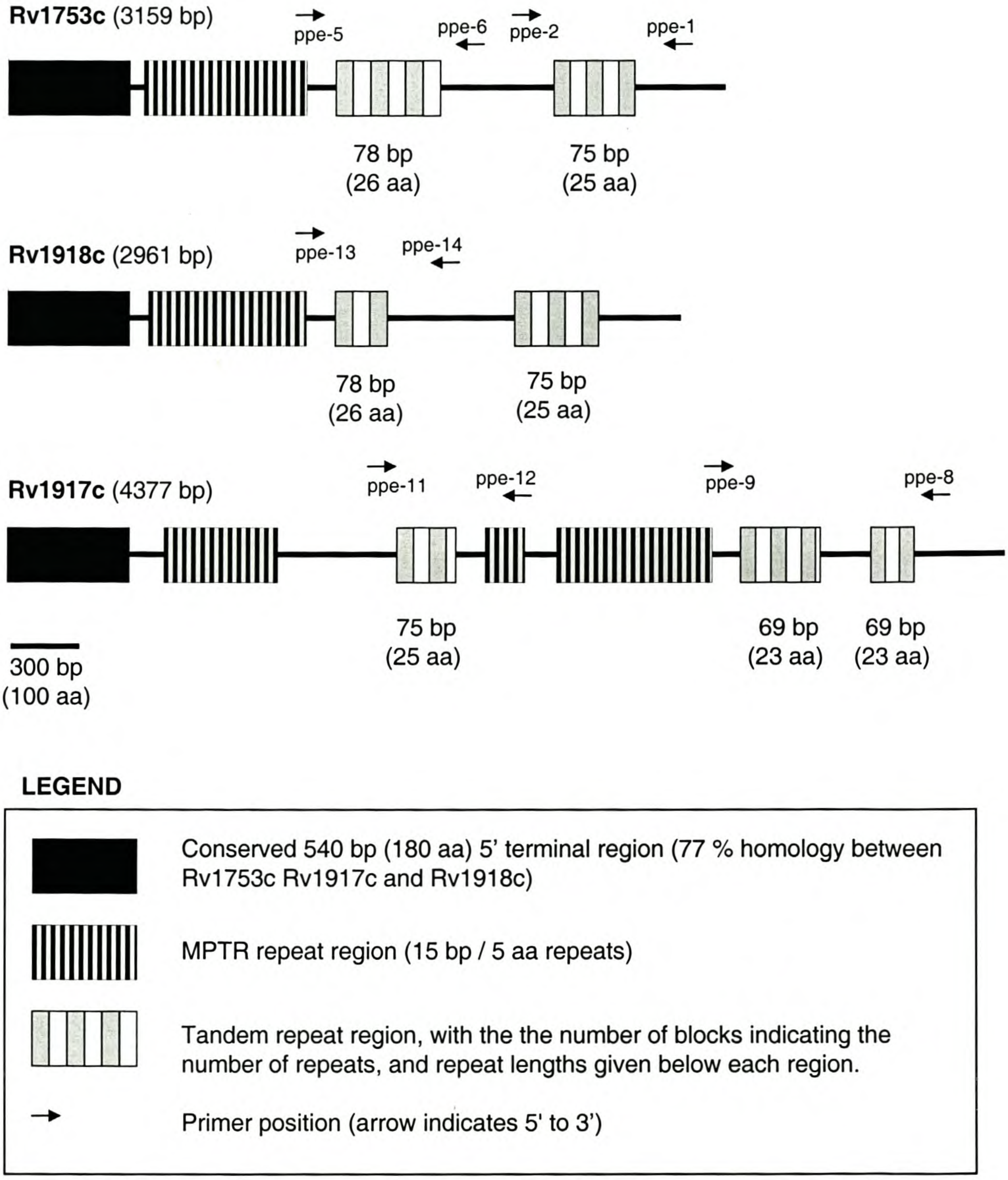


Figure 5.4 PPE Rv1753c, Rv1918c and Rv1917c tandem repeat regions. Note that different tandem repeat blocks represent different consensus repeats, although the LTR regions in Rv1753c and Rv1918c share similar consensus sequences. The positions of primers used to amplify the different repeat regions are shown.


Rv1753c (3159 bp)
ppe-5
ppe-6
ppe-2
ppe-1
78 bp (26 aa)
75 bp (25 aa)
Rv1918c (2961 bp)
ppe-13
ppe-14
78 bp (26 aa)
75 bp (25 aa)
Rv1917c (4377 bp)
ppe-11
ppe-12
ppe-9
ppe-8
75 bp (25 aa)
69 bp (23 aa)
69 bp (23 aa)
300 bp (100 aa)
LEGEND
Conserved 540 bp (180 aa) 5' terminal region (77 % homology between Rv1753c Rv1917c and Rv1918c)
MPTR repeat region (15 bp / 5 aa repeats)
Tandem repeat region, with the the number of blocks indicating the number of repeats, and repeat lengths given below each region.
Primer position (arrow indicates 5' to 3')

The PCR results (see Fig. 5.6 for example of results obtained) were interpreted under the assumption that the different PCR product sizes reflected different numbers of tandem repeats, and are therefore described in terms of the numbers of tandem repeats predicted from the PCR product size (summarized in Tables 5.5 and 5.6). Due to the limited sample number analyzed for each strain family grouping, it is difficult to draw conclusions regarding differences in the degree of inter- and intra-strain family variation. However, when the results for all the isolates investigated are viewed together (Table 5.6), it is evident that the number of variants obtained differed considerably for the different tandem repeat regions. For example, the 78 bp repeat region in Rv1753c demonstrates a total of 10 variants, while the 78 bp repeat region in Rv1918c showed only 2 variants.

Table 5.5 Tandem repeat variation: strain family comparisons. PCR variants analyzed according to strain family groupings (IS-3' DNA fingerprints >65% related according to GelCompar); clusters are groups of identical strains (according to IS-3' DNA fingerprints) within the strain family groups. The distribution of tandem repeat variants is given for each primer pair, and the corresponding Rv number is shown. (see Fig. 5.4 for primer positions)

Strain family	Cluster classification	Number of strains	Rv1753c		Rv1917c		Rv1918c
			1/2	5/6	8/9	11/12	13/14
4B	Cluster 1	33	**6**: 8/33 **8**: 7/33 **9**: 18/33	**6**: 1/33 **7**: 25/33 **8**: 7/33	**8**: 2/33 **9**: 24/33 **10**: 1/33 **11**: 5/33 **12**: 1/33	**3**: 33/33	**3**: 33/33
	Cluster 2	8	**6**: 7/8 **9** : 1/8	**7**: 1/8 **8**: 7/8	**9**: 1/8 **11**: 7/8	**2**: 2/8 **3**: 6/8	**3**: 8/8
F11	Cluster 1	16	**5**: 16/16	**5**: 16/16	**8**: 3/16 **9**: 13/16	**2**: 16/16	**3**: 16/16
	Non-clustered	40	**4**: 1/40 **5**: 39/40	**2**: 1/40 **5**: 39/40	**8**: 9/40 **9**: 29/40 **10**: 1/40 **IS**[a]: 1/40	**1**: 1/40 **2**: 39/40	**3**: 40/40
F28	Non-clustered	16	**5**: 4/16 **6**: 11/16 **6**[a]: 1/16	**9**: 1/16 **10**: 15/16	**9**: 4/16 **10**: 8/16 **11**: 2/16 **IS**[b]: 2/16	**3**: 9/16 **4**: 7/16	**2**: 1/16 **3**: 15/16

a: PCR product size is between that predicted for 5 and 6 repeats, suggesting the occurrence of a small deletion / insertion

b: PCR product > 1.5 kb due to IS*6110* insertion into tandem repeat region

Table 5.6 Distribution of PCR variants. The distribution of tandem repeat variants is given for each primer pair, and the corresponding Rv number is shown. (see Fig. 5.4 for primer positions). (These results include samples in Table 5.5, plus an additional 14 randomly selected isolates representing different strain family groupings.)

Rv1753c		Rv1917c		Rv1918c
1/2	**5/6**	**8/9**	**11/12**	**13/14**
4: 0.8% (1/127)	**2**: 0.8% (1/127)	**7**: 0.8% (1/127)	**1**: 0.8% (1/127)	**3**: 99.2% (126/127)
5: 49.6% (63/127)	**3**: 47.2% (60/127)	**8**: 12.6% (16/127)	**2**: 48.8% (62/127)	**2**: 0.8% (1/127)
6: 22.8% (29/127)	**4**: 0.8% (1/127)	**9**: 59.1% (75/127)	**3**: 40.9% (52/127)	
6[a]: 0.8% (1/127)	**5**: 0.8% (1/127)	**10**: 11.0% (14/127)	**4**: 8.7% (11/127)	
8: 9.4% (12/127)	**6**: 1.6% (2/127)	**11**: 11.8% (15/127)	**5**: 0.8% (1/127)	
9: 15.7% (20/127)	**7**: 22.0% (28/127)	**12**: 0.8% (1/127)		
IS[b]: 0.8% (1/127)	**8**: 11.0% (14/127)	**15**: 0.8% (1/127)		
	9: 2.4% (3/127)	**IS**[b]: 3.1% (4/127)		
	10: 12.6% (16/127)			
	11: 0.8% (1/127)			

a: PCR product size is between that predicted for 5 and 6 repeats, suggesting the occurrence of a small deletion / insertion

b: PCR product > 1.5 kb due to IS*6110* insertion into tandem repeat region

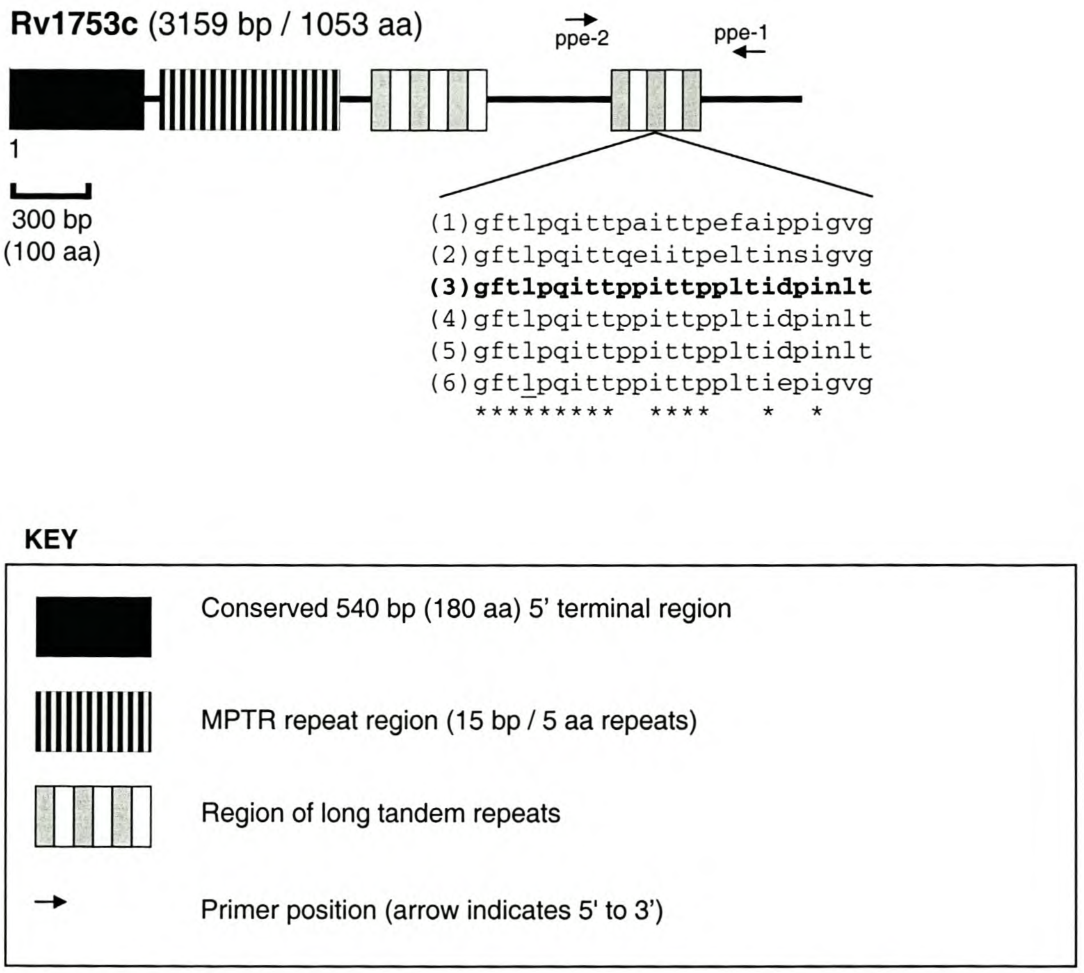
5.3.5 Sequencing of PCR variants

To prove the hypothesis that the observed variation in PCR product size was due to variable numbers of tandem repeats, 9 PCR variants were cloned and sequenced. These included 3 variants obtained from the 75 bp repeat region of Rv1753c (GenBank Accession numbers: AY029798, AY029799 and AY029800; Fig. 5.5) and 6 variants of the two 69 bp repeat regions from Rv1917c (GenBank Accession numbers: AF082287, AF082288, AF082289, AF082290, AF082291, AH007018, see Chapter 7, Figure 7.3).

Analysis of the sequence data demonstrated that expansion / contraction of tandem repeat regions was responsible for the observed PCR variants. Slipped-strand mispairing or homologous recombination are two possible mechanisms whereby this might occur (Levinson and Gutman, 1987; Lloyd and Low, 1996), although this remains to be elucidated. It should be noted that, within a single tandem repeat region, the sequence of the repeat units are not identical, and therefore variation can occur both in terms of numbers and sequence of tandem repeats. However, there does appear to be selection against amino acid changes in the repeat regions. For example, comparison of ClustalW nucleotide and amino acid sequence alignments of the sequence obtained from the 69 bp repeat regions in Rv1917c from 6 clinical isolates demonstrated that the 7 single nucleotide polymorphisms (relative to *M. tuberculosis* H37Rv) only resulted in one amino acid change. This contrasts with the results of Musser *et al.* (2000), who found that 6 of 7 nucleotide changes in 5 antigen-encoding genes resulted in amino acid changes.

To further analyze repeat region variation, sequence data from *M. tuberculosis* H37Rv, *M. tuberculosis* CDC1551 and *M. bovis* was compared. This revealed very little MPTR repeat variation, with only one example of MPTR repeat expansion/contraction identified in Rv0355c. This result was not unexpected, as DNA hybridization with the MPTR probe demonstrates very little restriction fragment length polymorphism in *M. tuberculosis* complex strains (Hermans *et al.*, 1992). However, the LTR regions in Rv1753c and Rv1917c are a source of variation between *M. tuberculosis* H37Rv, *M. tuberculosis* CDC1551 and *M. bovis*, and examples of loss or gain of discrete repeat blocks were identified in these 3 strains. Rv1918c does not exhibit any variation in either MPTR or LTR regions between the 3 strains.

Figure 5.5 Analysis of sequence variants of tandem repeat regions from Rv1753c. A region of the PPE gene Rv1753c containing 75 bp tandem repeats region was amplified from 3 clinical isolates (299, 366 and 727) with primer pair ppe-1/ppe-2. The amplification products were cloned and sequenced, and sequence data was compared to *M. tuberculosis* H37Rv and CDC1551. The amplification product from 299 demonstrated the presence of an additional repeat unit (3), which was identical to repeat units 4 and 5 in *M. tuberculosis* H37Rv. The amplification product from 727 showed a single L → R amino acid change in repeat unit 6, at underlined position.

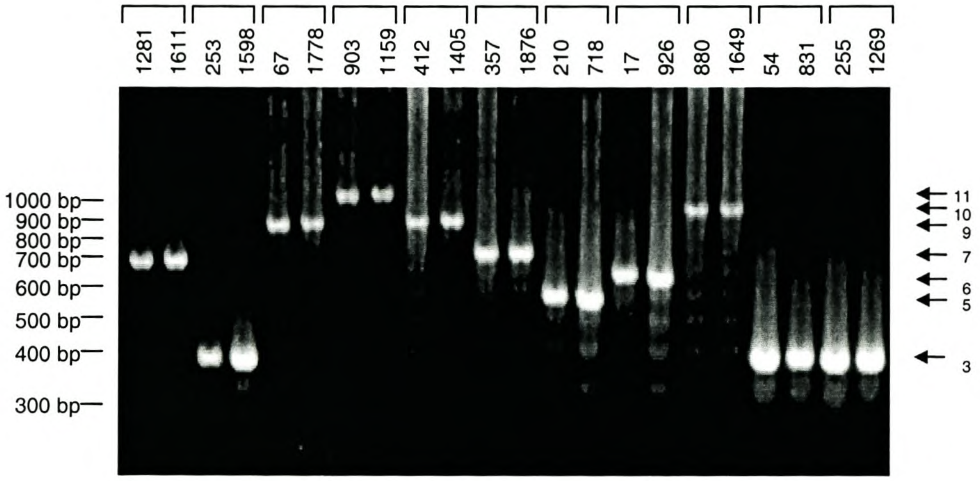

Rv1753c (3159 bp / 1053 aa)
ppe-2
ppe-1
1
300 bp
(100 aa)
(1)gftlpqittpaittpefaippigvg
(2)gftlpqittqeiitpeltinsigvg
(3)gftlpqittppittppltidpinlt
(4)gftlpqittppittppltidpinlt
(5)gftlpqittppittppltidpinlt
(6)gft_l_pqittppittppltiepigvg
********* ****   *   *
KEY
Conserved 540 bp (180 aa) 5’ terminal region
MPTR repeat region (15 bp / 5 aa repeats)
Region of long tandem repeats
Primer position (arrow indicates 5' to 3’)

5.3.6 Stability of PPE-containing domains

PPE gene stability was assessed by both hybridization and PCR analysis. To assess gene stability in isolates recovered from patients with disseminated infection, 86 isolates recovered from multiple sites of infection in 16 post-mortem cases (Du Plessis *et al.,* 2001) were hybridized with the Rv1787 and Rv1917c probes (results not shown). All hybridizing fragments (weak and strong intensity) were analyzed. The only changes in hybridization patterns that were detected could be attributed to a mixed infection (1 patient) or to an IS*6110* transposition event (1 isolate). No other polymorphisms were detected within single patients. In a further experiment, isolates from 16 patients who relapsed after curative treatment of post-primary tuberculosis (Van Rie *et al*., 1999b) were hybridized with the five 5' terminal probes. Of these, 4 pairs of isolates demonstrated identical IS-3' fingerprints before and after relapse. No change was detected with any of the probes in these four pairs of isolates.

To determine whether the long tandem repeat regions in Rv1753c, Rv1917c and Rv1918c were stable over a prolonged period of infection, first and last isolates (with identical IS-3' fingerprints) obtained from 11 patients with persistent disease over periods ranging from 12 to 91 months were investigated. These isolates represented 10 IS-3'-defined clusters within 8 strain family groups. Isolates from six of the patients were drug sensitive, while isolates from 5 patients were resistant to one or more antibiotics. The five primer pairs listed in Table 5.2 were used to amplify 6 tandem repeat regions from Rv1753c, Rv1917c and Rv1918c. No change was detected in any of the tandem repeat regions investigated (see Fig. 5.6 for example).

Figure 5.6 Stability of 78 bp LTR region in PPE-MPTR gene Rv1753c. The region containing 78 bp tandem repeats in Rv1753c was PCR amplified from a selection of *M. tuberculosis* clinical isolates using the primers ppe-5 and ppe-6. PCR products were fractionated in 2% agarose (1x TBE, pH 8.0), and visualized under UV after ethidium bromide staining. The samples shown are paired first and last isolates from single patients with persistent disease over time periods ranging from 12 to 91 months. The number of tandem repeats represented by each size variant is indicated.



5.4 Discussion

The *Mycobacterium tuberculosis* genome demonstrates a relatively low level of silent nucleotide substitutions in structural genes, and is thought to be relatively young in evolutionary terms (Sreevatsan *et al.*, 1997). Based on this finding, the authors suggested that the identification of highly polymorphic genes could indicate possible targets of the host immune response (Sreevatsan *et al.*, 1997). This hypothesis was recently investigated further by analysis of putative antigenic proteins from *M. tuberculosis*, which demonstrated an unusually low level of amino acid diversity (Musser *et al.*, 2000). These results contrast with numerous reports which have demonstrated extensive polymorphism associated with the transposable element IS*6110*, the polymorphic GC-rich sequences (PGRS) and, to a lesser degree, the major polymorphic tandem repeat (MPTR) (Van Soolingen *et al.*, 1993; Ross *et al.*, 1992; Hermans *et al.*, 1992). The potentially important role of IS*6110* as an agent of genome evolution has been acknowledged (Sreevatsan *et al.*, 1997; Fang *et al.*, 1999a; Sampson *et al.*, 1999; Beggs *et al.*, 2000), but the role of variation associated with PGRS- and MPTR-containing domains remains largely unaddressed. These repeats are found within members of the PE and PPE gene families respectively (Cole *et al.*, 1998). The gene families account for approximately 10% of the *M. tuberculosis* genome, and sequence comparisons have demonstrated that they are polymorphic (Cole *et al.*, 1998), leading to the suggestion that they may provide the pathogen with a source of antigenic variation. To investigate this hypothesis, this study addresses the question of PPE gene stability.

A previous report described a number of distinct subgroups within the PPE gene family (Cole *et al.*, 1998), but provided little detail regarding the size of, and relationships between, these subgroups. This result was therefore expanded by performing a phylogenetic analysis of evolutionary relationships among PPE proteins in *M. tuberculosis* H37Rv (see Fig. 5.1). The results confirm the subdivision of the PPE gene family into a number of subgroups, but also suggest a possible evolutionary history for the gene family. The majority of PPE proteins fall into two subgroups, namely the PPE-SVP and PPE-MPTR subgroups. As previously described (Cole *et al.*, 1998), these are characterized by a distinct motif centered around position 350 in the amino acid sequence, and by extensive MPTR repeats, respectively. The topology of the phylogenetic tree suggests that the PPE-MPTR subgroup has evolved from the PPE-SVP subgroup. This is supported by the presence of isolated MPTR repeats within selected members of the PPE-SVP subgroup, suggesting the existence of a common progenitor gene from which the PPE-MPTR subgroup expanded. The remainder of the PPE proteins fall into two smaller subgroups, which appear to be more ancient in evolutionary terms. The topology of the phylogenetic tree suggests an evolutionary progression, with the PPE-SVP subgroup, and subsequently the PPE-MPTR subgroup, arising from these smaller subgroups. It is possible that the phylogenetic subgroups may reflect relatively recent clonal expansion from common progenitor genes. However, since the much-reduced *M. leprae* genome (Cole *et al.*, 2001) contains PPE gene and pseudogene homologues from all of the subgroups, this seems unlikely. Alternatively, the subgroups identified may represent functionally specialized subdivisions, with conservation of particular sequence features. Bioinformatic analysis suggests possible common characteristics associated with the different subgroups, although this remains to be experimentally confirmed.

Analysis of the PPE gene family should take into account the possibility that the subgroups may reflect functionally distinct classes of proteins. A recent study which analyzed 4 different PPE genes only identified polymorphisms associated with 1 of these genes (Musser *et al.*, 2000). However, closer analysis reveals that 3 of the 4 genes examined by these investigators fell within the PPE-SVP subgroup (see Fig. 5.1), which may have lead to an under-estimation of PPE gene polymorphism. To avoid such a bias, this study utilized the results of the phylogenetic analysis to select PPE genes representing different subgroups of the gene family for further investigation. To gain an understanding of overall stability of the gene family, a hybridization-based approach was utilized. Analysis of the hybridization results suggests that PPE-containing domains are relatively stable within clusters of strains, which may reflect the recent evolutionary history of these clusters. However, there is extensive variation between strain families, particularly associated with the PPE-MPTR subgroup. Deletions, single nucleotide polymorphisms and IS*6110* insertions were all found to contribute to variation of PPE-containing domains. This confirms previous studies that have identified PPE gene knockouts resulting from IS*6110* transposition (Sampson *et al.*, 1999) and deletion of regions of the genome containing PPE genes (Behr *et al.*, 1999; Gordon *et al.*, 1999a; Warren *et al.*, 2000).

In addition to variation driven by IS*6110*-mediated gene disruption and deletion events, an additional level of PPE gene variation is provided by the presence of long tandem repeats (LTRs) within selected members of the PPE-MPTR subgroup. DNA sequence analysis identified stretches of long tandem repeats within three closely related members of this subgroup, namely Rv1753c, Rv1917c and Rv1918c. Previous reports had demonstrated polymorphism associated with tandem repeat regions in Rv1917c (Zhang and Young, 1994; Goyal *et al.*, 1994b; O'Brien *et al.*, 2000). This study confirmed this result and further demonstrated that expansion/contraction of tandem repeat domains was responsible for the observed polymorphisms in all 3 genes.

To determine whether different strain types demonstrate differing degrees of tandem repeat variation, closely related strain family groups were compared. Similar numbers of LTR variants were observed within the three strain family groups analyzed, although greater sample numbers will be required to calculate whether the mutation rates are similar within the different strain families. However, it is clear that the different repeat regions demonstrate considerably different levels of variation. The degree of variation associated with the different repeat regions may be related to the mechanism whereby the variation arises, or alternatively may reflect different selective pressure on the 3 genes. Currently, the mechanism whereby PPE tandem repeat regions undergo expansion/contraction is unknown, although homologous recombination is one possibility (Lloyd and Low, 1996). This has been invoked as a potential mechanism giving rise to antigenic variation associated with tandem repeat regions in antigenic proteins in other pathogens (Hollingshead *et al*., 1987).

The functional significance of PPE gene polymorphisms requires further investigation. The tolerance of PPE gene disruption by deletion and IS*6110* insertion might be interpreted to reflect functional redundancy among members of the gene family. The hybridization results suggest that in the event of a knockout, at least one intact copy of closely related PPE genes is maintained. This implies that PPE gene knockouts would not necessarily demonstrate an altered phenotype, due to complementation. However, not all PPE genes have closely related homologues, as evidenced by the hybridization patterns observed with the Rv2123 probe. Incidentally, this probe yields highly conserved hybridization patterns, which may indicate that there is selective pressure to maintain an intact copy of the gene. These results suggest that PPE gene disruption could lead to subtle phenotypic differences which may be deleterious to the organism. Functional analysis of members of the gene family will be necessary to examine this possibility, although careful consideration should be given to the selection of candidate genes to be studied.

The phenotypic implications of the tandem repeat variation also remains to be elucidated. However, it is intriguing to draw parallels with tandem repeat region variation in antigenic surface structures of other pathogens (Hollingshead *et al*., 1987; Zheng *et al*., 1995; Gravekamp *et al*., 1998; Lysnyansky *et al*., 1999; Labandeira-Rey *et al*., 2001). It has recently been demonstrated that the PPE gene Rv1917c is cell-wall associated (Sampson *et al.*, 2001; see Chapter 7). Although it remains to be determined whether the tandem repeat domains are surface-exposed, these results support the hypothesis that selected PPE proteins may interact with the host immune system and provide the pathogen with a source of antigenic variation.

Although antigenic variation may be expected to manifest at the gene sequence level as hypervariability of PPE-containing domains within single patients, this was not detected in this study. The stability of tandem repeat regions was investigated by analysis of PCR variants obtained from patients persistently infected with the same strain. Despite the identification of extensive inter-patient strain variation, no change was observed in tandem repeat regions in any of the isolates obtained from a single patient. Hybridization results provide further evidence that PPE genes remain stable within single patients. However, this does not rule out the PPE protein family as a potential source of antigenic variation, as this phenomenon may also be mediated at the transcriptional and translational level.

The results described here provide evidence that there is extensive polymorphism associated with specific members of the PPE gene family. Tandem repeat regions and other types of PPE gene variation may provide a "bonus" in terms of diversity in an otherwise stable genome. This might be important in terms of the bacterial population as a whole, rather than at the level of a single strain. Should this prove to be the case, it may provide a possible mechanism for *M. tuberculosis* to cause exogenous re-infection, a phenomenon that has been reported by a number of investigators (Chaves *et al.*, 1999; Van Rie *et al.*, 1999b; Bandera *et al.*, 2001; Caminero *et al.*, 2001; Du Plessis *et al.*, 2001). This also holds important implications for vaccine development. The results presented here validate investigation of the

PPE gene family as a potential source of antigenic variation. Functional analysis, particularly in terms of *in vivo* gene and protein expression profiles, is required.

CHAPTER 6

ANALYSIS OF *IN VITRO* AND *IN VIVO* PPE GENE EXPRESSION

NOTE: The results presented in the following chapter will be submitted to *Tuberculosis* as: **Differential expression of the PPE gene family of *Mycobacterium tuberculosis***. Sampson, S.L., Warren, R.M., Moses, L., Fenhalls, G., Stevens, L., Van Helden, P.D.

(The style of the text and numbering of sections has been altered to conform with the style of this thesis. Literature cited in the text takes the form of author name and year of publication as opposed to the number format specified by Tuberculosis. *All cited literature is compiled into a single list at the end of the thesis for ease of reference.)*

6.1 Introduction

The PE and PPE gene families constitute approximately 10% of the *Mycobacterium tuberculosis* genome. Their maintenance in the genome suggests that they fulfill a potentially important function, and much speculation has centered around this (Cole *et al.*, 1998; Cole and Barrell, 1998; Tekaia *et al.*, 1999). It has been suggested that the proteins encoded by the two gene families may inhibit antigen processing or alternatively, provide the pathogen with a source of antigenic variation (Cole *et al.*, 1998). Support for the latter hypothesis is provided by *in silico* sequence comparisons between *M. tuberculosis* H37Rv and *Mycobacterium bovis* which demonstrate PE and PPE sequence variation (Cole *et al.*, 1998; Gordon *et al.*, 2001). Furthermore, analysis of a subset of PPE genes has revealed extensive variation in clinical isolates of *M. tuberculosis* (see Chapter 5).

It has been reported that the *M. tuberculosis* genome demonstrates very little sequence variation in structural genes (Sreevatsan *et al.*, 1997). Therefore, the finding that selected members of the PPE gene family demonstrate extensive polymorphism in clinical isolates of *M. tuberculosis* is potentially significant, as it may reflect the influence of host selective pressure on these genes (Sreevatsan *et al.*, 1997). However, it is premature to draw such conclusions without knowledge of whether the genes are in fact expressed during acrive disease, and evidence for this is very limited. The first proof of *in vitro* expression of PPE genes was provided by a study which utilized differential display PCR to demonstrate the up-regulation of the PPE gene Rv0755c in H37Ra relative to H37Rv (Rivera-Marrero *et al.*, 1998). Similarly, the PPE gene Rv2770c was shown to be up-regulated in *M. tuberculosis* H37Rv vs H37Ra (Rindi *et al.*, 1999). Recently, the up-regulation of the PPE gene Rv2123 under low iron conditions *in vitro* was reported, leading to speculation that the gene product may be involved in iron acquisition via siderophore uptake (Rodriguez *et al.*, 1999).

Early evidence for *in vivo* expression of the PPE genes was provided by a study of a serine-rich antigen (sra) from *Mycobacterium leprae* (Vega-Lopez *et al.*, 1993), which was recognized by serum from both leprosy and tuberculosis patients. Sequence homology comparisons demonstrate that this corresponds to the *M. tuberculosis* PPE gene Rv2108 (S. Sampson, unpublished observation). More recently, it was shown that a mutant strain of *M. tuberculosis* with a transposon-generated knockout of the PPE gene Rv3018c, is attenuated for growth in lungs of mice (Camacho *et al.*, 1999). This has been suggested to imply a role for this gene in early infection (Camacho *et al.*, 1999), although *in vivo* expression of the gene has yet to be confirmed, and the relevance of this finding remains to be explored in the context of clinical isolates. Intriguingly, T-cell expression cloning has demonstrated that the products encoded by the PPE genes Rv1196 and Rv0915 invoke a T cell response (Dillon *et al.*, 1999; Skeiky *et al.*, 2000). The recognition of these proteins by patient sera (Dillon *et al.*, 1999; Skeiky *et al.*, 2000) is indicative of *in vivo* expression.

Emerging data suggests that members of the PPE gene family may fulfil a role in disease pathogenesis (Rodriguez *et al.*, 1999; Camacho *et al.*, 1999; Dillon *et al.*, 1999; Skeiky *et al.*, 2000), and it is therefore likely that they will be expressed during host infection. A recent study has demonstrated a role for *Mycobacterium marinum* homologues of the PE gene family in macrophage replication and persistence in frog granulomas (Ramakrishnan *et al.*, 2000). While it is tempting to extrapolate these results to the PPE gene family, it is essential to first establish whether the genes are expressed, particularly within the granuloma, an important site of host-pathogen interaction.

The granuloma is a characteristic feature of tuberculosis disease, and its formation is preceded by infection of the host by inhalation of aerosol-borne bacilli (Wells, 1955; Riley *et al.*, 1959). The bacilli are phagocytosed by alveolar macrophages, which possess numerous bacterial killing mechanisms (Flesch and Kaufman, 1991; Chan *et al.*, 1992; O'Brien *et al.*, 1996; Placido *et al.*, 1997). However, *M. tuberculosis* is able to subvert many of these mechanisms, and the bacteria frequently survive and replicate within the macrophage (Gordon *et al.*, 1980; Chan *et al.*, 1989; Andersen *et al.*, 1991; McNeil and Brennan, 1991; Sturgill-Koszycki *et al.*, 1994). Alternatively, the infected macrophages may proceed to dendritic cells which take up the mycobacteria or components thereof (Henderson *et al.*, 1997), and migrate to draining lymph nodes (Saunders and Cooper, 2000; Cutler *et al.*, 2001). In either case, the mycobacterial antigens can be complexed with MHC Class II molecules on the surface of the antigen presenting cell (APC), and presented to naïve T cells to initiate the cell-mediated immune response (Barnes *et al.*, 1994; Ivanyi and Thole, 1994; Boom, 1996).

It is well-established that the control of infection is largely dictated by the T-cell response, particularly by those of the CD4$^+$ subgroup, otherwise known as T helper (T_h) cells (Mossman and Coffman, 1989; Barnes *et al.*, 1994; Zhang *et al.*, 1995; Romagnani, 1997; Wangoo *et al.*, 2001). Successful containment of initial infection is largely dependent on IFN-γ secretion by T_h type I (T_h1) cells (Flynn *et al.*, 1993; Cooper *et al.*, 1993; Jouanguy *et al.*, 1996; Condos *et al.*, 1998), although a complex and finely regulated network of other cytokines is also involved, including IL-12 (Altare *et al.*, 1998; Flynn and Bloom, 1996) and TNF-α (Flynn *et al.*, 1995; Bekker *et al.*, 2000; Mohan *et al.*, 2001).

Long-term disease outcome is dependent on adequate granuloma formation, a process initiated by macrophage infection and continued by the orchestrated recruitment of various immune cells to the site of infection. Neutrophil influx to the site of infection occurs early after *M. tuberculosis* infection (Pedrosa *et al.*, 2000), and is closely followed by the recruitment of monocytes which mature into macrophages (Adams, 1974). Secretion of various chemokines and cytokines by these cells serves to attract and activate other immune cells, among them T_h ($CD4^+$), T_c (cytotoxic T lymphocytes, $CD8^+$) and B lymphocytes (Friedland, 1994; Fenhalls *et al.*, 2000; Gonzalez-Juarrero *et al.*, 2001). This focused accumulation of cells matures into an organized granuloma structure, which classically contains a central region of epitheloid macrophages (Canetti, 1955; Dannenberg and Rook, 1994; Saunders and Cooper, 2000), some of which may fuse to form multinucleated giant cells (Papadimitriou and Van Bruggen, 1986; Hernandez-Pando *et al.*, 2000a). Some of the macrophages and/or giant cells may contain mycobacteria (Dannenberg and Rook, 1994; Fenhalls *et al.*, submitted for publication). The central region is surrounded by a dense margin of lymphocytes and monocytes (Canetti, 1955; Dannenberg and Rook, 1994; Kobzik and Schoen, 1994; Saunders and Cooper, 2000). This lymphocyte cuff consists predominantly of $CD4^+$ lymphocytes, with some $CD8^+$ cells situated around the periphery, and B cells are also observed (Dannenberg and Rook, 1994; Fenhalls *et al.*, 2000; Gonzalez-Juarrero *et al.*, 2001). The mature granuloma represents an organized structure in which a network of immune cells are situated in close proximity to one another, and therefore better able to interact and fulfil their respective functions (Saunders and Cooper, 2000). The granuloma also serves to confine the potentially toxic accumulation of cytokines, which could otherwise damage host tissue (Saunders and Cooper, 2000). In addition, the granuloma environment is unfavourable to bacterial growth, and it acts as a physical barrier which prevents bacterial dissemination (North and Izzo, 1993; Doenhoff, 1997; Emile *et al.*, 1997; Saunders *et al.*, 1999).

The granuloma is by no means a static structure, and after some time, macrophages in the central region may start to degenerate, a process known as caseating necrosis (Canetti, 1955; Dannenberg and Rook, 1994). This process has been associated with apoptosis (programmed cell death) of macrophages (Cree *et al.*, 1987). The necrotic region is thought to be unfavourable for bacterial growth, and few acid-fast bacilli are detected within the solid caseous focus (Canetti, 1955; Dannenberg and Rook, 1994), although bacteria can persist here for many years (Robertson, 1993). The necrotic granuloma may eventually resolve by fibrin deposition, sclerosis and calcification to form a lesion with no detectable bacilli (Canetti, 1955; Dannenberg and Rook, 1994). Alternatively, under certain circumstances, the necrotic region may start to liquefy, creating an ideal environment for bacterial proliferation (Canetti, 1955; Dannenberg and Rook, 1994). Eventually, rupture of the granuloma can lead to the disruption of lung architecture and the formation of pulmonary cavities characteristic of advanced disease in humans. Expulsion of the granuloma contents with its high bacterial load into nearby airways can result in the transmission of infection, as well as reseeding and renewal of the infection cycle in the original host (Canetti, 1955; Dannenberg and Rook, 1994).

An extensive, albeit complex, body of data regarding the host response to tuberculosis infection is starting to emerge, yet little is known about bacterial gene expression within the granuloma. Even fundamental understanding of the state of bacterial growth within the granuloma remains controversial. It has long been suggested that the granuloma represents an oxygen-poor and potentially toxic environment, not conducive to bacterial growth, wherein mycobacteria reside in a latent or dormant state (Dannenberg, 1982), similar to that observed in oxygen-depleted stationary cultures (Wayne, 1976; Wayne, 1994). An alternative viewpoint is that the infection should be viewed as "persistent and chronic" (D. Russell, quoted in Butler, 2000), with mycobacteria actively metabolizing and undergoing limited growth, but with their numbers held in check by the host immune system. Either way, although a handful of recent studies have identified mycobacterial mutants which are impaired in their ability to survive within the granuloma (McKinney *et al.*, 2000; Glickman *et al.*, 2000; Ramakrishnan *et al.*, 2000), very little is yet known about *M. tuberculosis* gene expression in the context of the human granuloma.

This study addresses the question of PPE gene expression within the human granuloma. As a preliminary screen to determine whether PPE genes are actively transcribed *in vitro*, total RNA from liquid cultures was analyzed by RT-PCR. *In vivo* PPE gene expression was subsequently analyzed by RNA:RNA *in situ* hybridization of human lymph node biopsy sections.

6.2 Methods

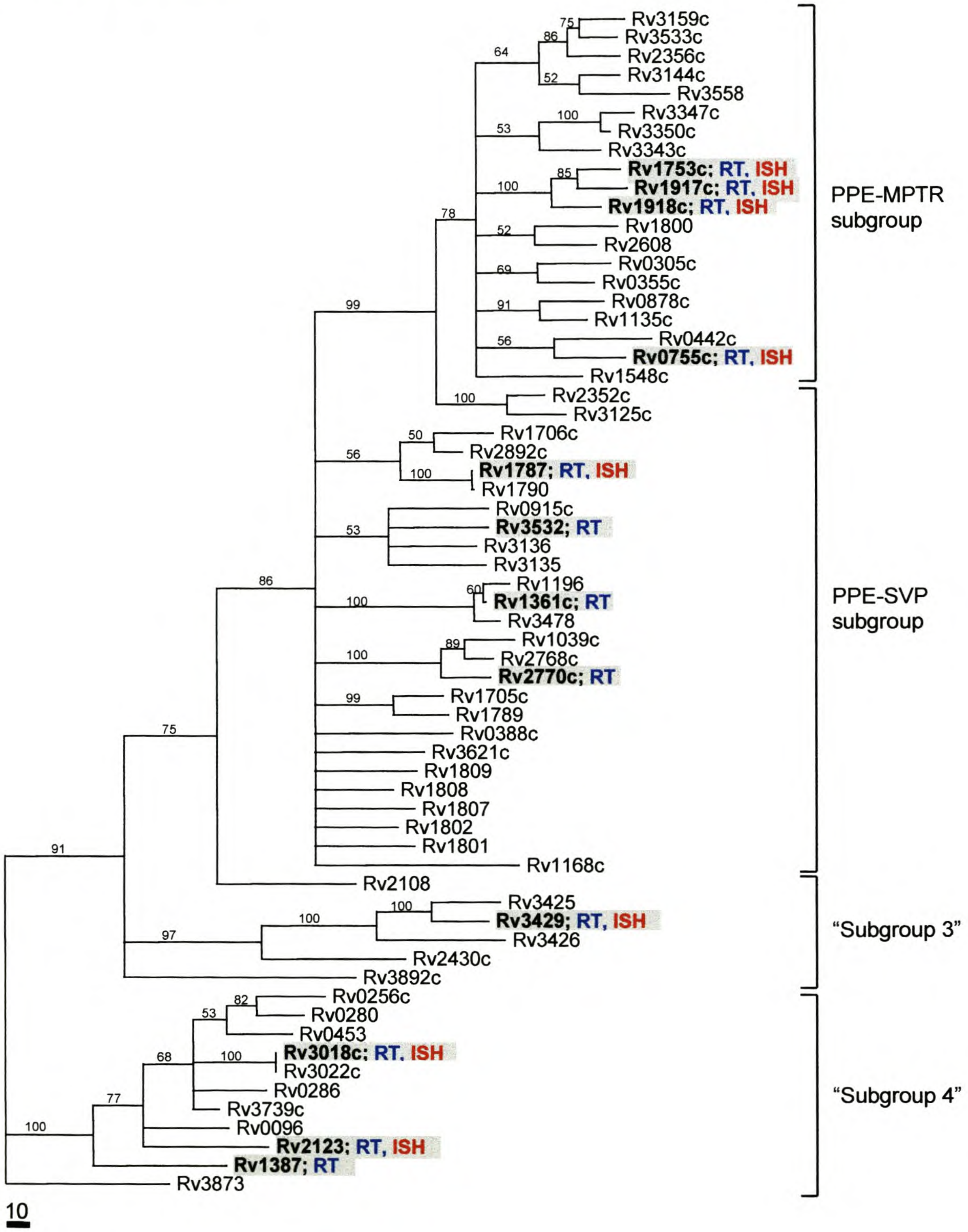
6.2.1 Bacterial strains and culture conditions

Escherichia coli XL-1 Blue cells were electroporated according to a standard method (Sambrook *et al.*, 1989), and plated onto LB-agar containing 50 μg/ml ampicillin. Liquid cultures of *E. coli* were grown in LB medium containing 50 μg/ml ampicillin. Stirred cultures of *Mycobacterium tuberculosis* H37Rv were incubated in 50 ml screw-cap tubes containing 10 ml of Middlebrooks 7H9 medium supplemented with 10% ADC enrichment and 0.05% Tween-80 (MADCTW; Allen, 1999). Mycobacterial growth phase was ascertained by comparison to a pre-determined growth curve, generated under identical culture conditions by plotting OD_{600} against time point over a period of 28 days (data not shown).

6.2.2 RNA manipulation

A previously constructed PPE phylogenetic tree (see Chapter 5), was used to select 12 PPE genes for analysis by RT-PCR (Fig. 6.1). The selection was such that each subgroup defined by the tree was represented. Total RNA was extracted from mid-log, late log and stationary phase (7, 14 and 21 day) liquid cultures of *M. tuberculosis* H37Rv. Cells were pelleted by centrifugation at 4,000 rpm for 5 minutes and the cell pellet was resuspended in Trizol (Gibco BRL). Resuspended cells were added to FastPrep Blue tubes (Bio101) containing silica beads, and ribolysed at speed 6.5 for 45 s. The tubes were centrifuged at 13,000 rpm for 1 minute to pellet cell debris and beads. The Trizol layer was removed and chloroform-extracted. After isopropanol precipitation, RNA was resuspended in nuclease-free H_2O (Promega) with 10 mM DTT and 1U/μl RNasin (Promega). DNaseI buffer (400 mM Tris-HCl (pH 8.0), 100 mM $MgSO_4$, 10 mM $CaCl_2$) and DNaseI (0.5 U/μl) (Roche) were added, and incubated at 37°C for 30 minutes. The reaction was stopped by addition of 12.5

Figure 6.1. Selection of PPE genes for RT-PCR and RNA:RNA *in situ* hybridization analysis. PPE gene family subgroupings previously defined by phylogenetic analysis (Chapter 5) were used to select examples of each subgroup for further investigation. "RT" denotes genes analyzed by RT-PCR, while "ISH" indicates those investigated by *in situ* RNA:RNA hybridization.



mM EDTA and the RNA was subsequently purified using the QIAGEN RNeasy mini-kit, as described by the manufacturer. A second round of DNaseI-treatment and column purification was performed. RNA integrity was determined after electrophoresis in 1% agarose (1× TBE, pH 8.0) by visualization under UV after ethidium bromide staining. RNA concentration was determined spectrophotometrically at OD_{260}.

Twelve pairs of primers were utilized to assay for the presence of PPE mRNA in *in vitro* cultures (see Table 6.1). The primer pair Rv3018F/Rv3018R can potentially amplify mRNA from both the Rv3018c and Rv3022c genes, due to 100% sequence homology at the priming sites in the two genes. Similarly, the primer pair ppe-15/Rv1787R can amplify both Rv1787 and Rv1790. In all other cases, primers were designed to be specific for the particular gene, particularly at the 3' end of the primer. Two primer pairs were designed for the PPE gene Rv1917c, to amplify products from the 5' and 3' regions of the gene respectively.

RT-PCR amplification was performed using the Titan One-Tube RT-PCR kit (Roche), with the oligonucleotide pairs listed in Table 6.1, together with 1 μg total RNA as template. Each preparation of RNA was tested by amplification with primers specific for the *hspX* gene, which is known to be expressed under *in vitro* conditions (Hu and Coates, 1999). Controls included RNaseA (Roche) pre-treated samples, DNaseI (Roche) pre-treated samples, PCR with heat-inactivated reverse transcriptase, and a water blank (see Chapter 7, Fig. 7.4B). To ensure that positive signal was not due to PCR amplicon contamination, a nuclease-free water control (with no added template) was included in all subsequent experiments. RT-PCR products were fractionated by electrophoresis in 3% MetaPhor agarose (BioWhittaker Molecular Applications) (1× TBE, pH 8.0). To confirm that the correct region had been

Table 6.1 Oligonucleotides utilized for RT-PCR and generation of *in situ* hybridization riboprobes. "RT" indicates oligonucleotides utilized for RT-PCR, while "ISH" indicates oligonucleotides utilized for the generation of riboprobes.

Name	Sequence (5'-3')	Tm*	Application
hsp394	CGT CGA CAT TAT GGT CCG C	60°C	*Amplification of 206 bp product from Rv2031c (hspX) control* (**RT**)
hsp599	TTG GCT TCC CTT CCG AAA C	58°C	
Rv0355cF	ATG GAC TGG CCG ACG AAT TG	60°C	*Amplification of 138 bp product from Rv0355c* (**RT**)
Rv0355cR	CGC TCA ACC AGC CCA GAT AG	64°C	
Rv0755cF	TTG CCT CCG GAG ACC AAT TC	62°C	*Amplification of 214 bp product from Rv0755c* (**RT**)
Rv0755cR	TCA ACC AGC TCG CAT ACG G	60°C	
0755ISH_F	ACT CGG GCC AAT CGG TAA C	56°C	*Amplification of 238 bp product from Rv0755c* (**ISH**)
0755ISH_R	AAG CCC GAC ACG AAA TCG	60°C	
Rv1361cF	CGC TGG TGG CCG CCG CG	64°C	*Amplification of 114 bp product from Rv1361c* (**RT**)
Rv1361cR	ACG AAC CTA TCC ACG ATC CCG	66°C	
Rv1387F	GTT GGG CCG ATG TTG ATC TC	62°C	*Amplification of 165 bp product from Rv1387* (**RT**)
Rv1387R	CCA CGC CAG AAA CGG CAT G	62°C	
1753ISH_F	CGG TGG CTT TAG TCT ACC TGC	66°C	*Amplification of 279 bp product from Rv1753c* (**ISH**)
1753ISH_R	CCG GTC AAT GTG TAT GGG TG	62°C	
ppe-15	TGG ACT TCG GGG CGT TAC	58°C	*Amplification of 166 bp product from Rv1787 and Rv1790* (**RT**)
Rv1787R	GGC GCA CCG GTC AGC TC	60°C	
ppe-15	as above	58°C	*Amplification of 499 bp 5' terminal region from Rv1787 and Rv1790* (**ISH**, *preliminary screen*)
ppe-16	AAC GGA ATC AAC CGC GAC	56°C	
ppe-17	TTC AAC TCC GTG ACG TCG	56°C	*Amplification of 471 bp 5' terminal region from Rv1917c* (**ISH**; *preliminary screen*)
ppe-18	CAG CAC ACC CTT GGA ACT G	60°C	
Rv1917cF	AAC TCG GCC CTC ATA TTC GG	62°C	*Amplification of 204 bp product from Rv1917c* (**RT**)
Rv1917cR	CGC GGC AAG CCA TCC TAG G	64°C	
1917ISH_F	CGT CCC TAT TTC CCG CAT C	60°C	*Amplification of 220 bp product from Rv1917c* (**RT**, **ISH**)
1917ISH_R	GCT CAG CGG AAT ATT CAA ACC	62°C	
1918ISH_F	TAG GCG GCT TTA GCA CTC C	60°C	*Amplification of 235 bp product from Rv1918c* (**RT**, **ISH**)
1918ISH_R	TTG ATT TGA CCT CCA CCC AC	60°C	
Rv2123F	ATG TGG TTC GCA GTT CCG C	60°C	*Amplification of 227 bp product from Rv2123* (**RT**, **ISH**)
Rv2123R	GTT AGC CAA TAC CGG AAC GG	62°C	
Rv2770cF	CAC TCG ACC AGA AAC GGA AC	62°C	*Amplification of 144 bp product from Rv2770c* (**RT**)
Rv2770cR	TGC TCA CCT CGA CAG CTA TG	62°C	
Rv3018F	ATT CGG CGC TGC TAA GTG C	60°C	*Amplification of 160 bp product from Rv3018c and Rv3022c* (**RT**, **ISH**)
Rv3018R	AAC TCA GCA CTG GGA CCC TG	64°C	
Rv3429F	CAT CCA ATG ATA CCA GCG GAG	64°C	*Amplification of 148 bp product from Rv3429* (**RT**, **ISH**)
Rv3429R	GCT CGC CGA GCC TGT CGG	64°C	
Rv3532F	TGT TCA TGG ATT TCG CGA TG	58°C	*Amplification of 195 bp product from Rv3532* (**RT**)
Rv3532R	TCA CAG ACG ATG GAC CCA GC	64°C	
T7	GCG ATT AAG TTG GGT AAC GCC	64°C	*Generation of riboprobes for ISH from pGEM-T Easy vector*
SP6	CAC TTT ATG CTT CCG GCT CG	62°C	

*Tm = 4(G + C) + 2(A +T)

amplified, amplification products were cloned into the pGEM-T Easy vector (Promega) and sequenced on an ABI automated sequencer at the University of Stellenbosch Core Sequencing Facility.

6.2.3 Tissue specimens

Lymph node tissue was obtained from 2 HIV-negative adults (18 year-old male and 49 year-old female) with no previous history of tuberculosis, undergoing lymph node biopsies at Tygerberg hospital, Cape Town, South Africa. Diagnosis of TB was confirmed by Ziehl-Neelsen staining, and bacterial culture. Informed consent was obtained from each patient, and the study was approved by the University of Stellenbosch Ethical Review Committee. Tissue was formalin fixed, embedded in paraffin, then cut into 5 µm sections using a microtome (Department of Anatomical Pathology, Faculty of Health Sciences, University of Stellenbosch, Tygerberg, South Africa; Fenhalls *et al.*, 1999). Consecutive sections were applied to RNase-free aminopropyl-triethoxysilane coated slides (5 µg/ml; Sigma-Aldrich).

6.2.4 Preparation of riboprobes

Eight genes were selected as targets for *in situ* RNA:RNA hybridization analysis based on their positions within a previously constructed PPE phylogenetic tree (see Chapter 5; Fig. 6.1). Products were amplified from 100 ng *M. tuberculosis* H37Rv genomic DNA using Qiagen HotStar *Taq* DNA polymerase with the oligonucleotide primer pairs listed in Table 6.1, according to the manufacturer's instructions. Amplification products were cloned using the pGEM-T Easy vector system (Promega), and purified with the Wizard DNA purification kit (Promega). Plasmids were sequenced to confirm that the correct products had been amplified and cloned, and to determine the orientation of the insert DNA. The insert DNA was amplified using Qiagen HotStar *Taq* DNA polymerase with the T7 and SP6 primers (see

Table 6.1), then purified with the Concert Rapid PCR Purification System (Gibco BRL, Switzerland), as described by the manufacturer.

Antisense and sense biotin-labelled probes were generated by transcription from the purified PCR templates with the T7 or SP6 RNA polymerases (Gibco BRL, Switzerland), depending on insert orientation. The following were mixed together in RNase-free tubes: 1 μg DNA, 1x T7 or SP6 transcript buffer, 1 mM of each dATP, dGTP, dUTP and Biotin-14-dCTP, 0.5 mM dCTP, 5 mM DTT, 120 U RNase OUT Inhibitor (Gibco BRL), 100 U T7 polymerase or 52.5 U SP6 polymerase and RNAse-free H_2O to a final volume of 50 μl. The reactions were incubated at 37°C for 2 hours. The riboprobes were precipitated with ethanol, washed in 70% ethanol, then dried in a Speed Vac Concentrator (Savant, USA) for approximately 5 minutes. The riboprobes were resuspended in 100 μl RNAse-free H_2O, incubated at 37°C for 30 minutes, then stored at -20°C until further use. Riboprobe concentrations were between 0.1 and 0.5 μg/ml.

To confirm biotinylation of the riboprobes, 5–10 μl of each probe was spotted onto Hybond N+ membranes (Amersham, UK), then detected with streptavidin-conjugated alkaline phosphatase in conjunction with the NBT/BCIP (nitroblue-tetrazolium / 5-bromo-4-chloro-3-indolyl-phosphate) substrate (Gibco BRL, Switzerland).

6.2.5 RNA:RNA *In situ* Hybridization

RNA:RNA *in situ* hybridization (ISH) was performed as described elsewhere (Fenhalls *et al.*, 1999). Briefly, the paraffin-embedded lymph node tissue sections were deparaffinized in two changes of xylene for 10 minutes each. The sections were rehydrated in graded ethanols and diethyl pyrocarbonate-treated water, then incubated in phosphate-buffered saline (PBS). Proteinase K (1 μg/ml) treatment was performed in prewarmed (37°C)

10 mM Tris-HCl (pH 7.5), 5 mM EDTA for 45 min at 37°C in a humidified chamber. The sections were washed with PBS, then refixed in 0.4% paraformaldehyde, and washed with RNAse-free H_2O. Acetylatation was carried out in a 1:400 (vol/vol) solution of triethanolamine-acetic anhydride with stirring for 10 min. The slides were rinsed in PBS, then dehydrated in graded ethanols, and air-dried.

RNA:RNA hybridization was carried out using the *In Situ* Hybridization and Detection System (Gibco BRL). Briefly, 5μg/ml of the antisense or sense biotin-labelled riboprobe was added to the hybridization mixture (20 μl of 20% dextran sulphate solution and 20 μl of 2x hybridization buffer). The probe mixture was added to the sections, and covered with a coverslip. No denaturation was performed prior to hybridization. The sections were incubated for 16 – 18 hours at 50°C in a humidified chamber, then washed twice in 2x SSC for 15 minutes at room temperature.

For signal detection, the sections were covered with 200 μl Blocking Solution (Gibco ISH Kit), and incubated in a humid chamber at room temperature for 20 minutes. The Blocking Solution was removed from by touching absorbent paper to the edge of each slide. Each section was covered with 100 μl of working conjugate solution (10 μl of streptavidin-alkaline phosphatase conjugate mixed with 90 μl of conjugate dilution buffer (Gibco ISH Kit)) and incubated in a humid chamber at room temperature for 15 minutes. The slides were washed twice in Buffer A (100 mM Tris-HCl, pH 7.5; 150 mM NaCl) for 15 minutes each at room temperature, then incubated in prewarmed alkaline-substrate buffer, Buffer B (100 mM Tris-HCl, pH 9.5; 50 mM $MgCl_2$; 100mM NaCl), at 37°C for 5 minutes. Three hundred microlitres of NBT and 249 μl BCIP were added to 75 ml Buffer B prewarmed at 37°C. Levamisole (200 μg/ml) (Southern Cross Biotechnologies) was added to the alkaline-substrate solution to inhibit endogenous phosphatase activity. The slides were incubated in the

NBT/BCIP solution at 37°C until the desired level of signal was achieved (10 min to 2 hours), then rinsed briefly in distilled water.

To allow the visualization of cells negative for *M. tuberculosis* mRNA, slides were counterstained by immersing in a methyl green solution (DAKO, USA) for 5 minutes, then rinsing with 95% EtOH to remove excess reagent. Sections were dehydrated with graded alcohols, and finally incubated twice in xylene for 2 minutes. The slides were then mounted with Dako Faramount.

6.2.6 Photography and assessment of slides

Images were captured using a Zeiss (Germany) Axioskop 2 microscope fitted with a Sony 3CCV video camera, then saved using Zeiss Axiovision software. ISH is an empirical staining technique which cannot be accurately quantitated. Granulomas were therefore scored as either negative (0) or positive (1) for the presence of *M. tuberculosis* mRNA. The number of lymph node granulomas positive for *M. tuberculosis* mRNA was counted on 2.5x magnification, by two observers.

6.3 Results

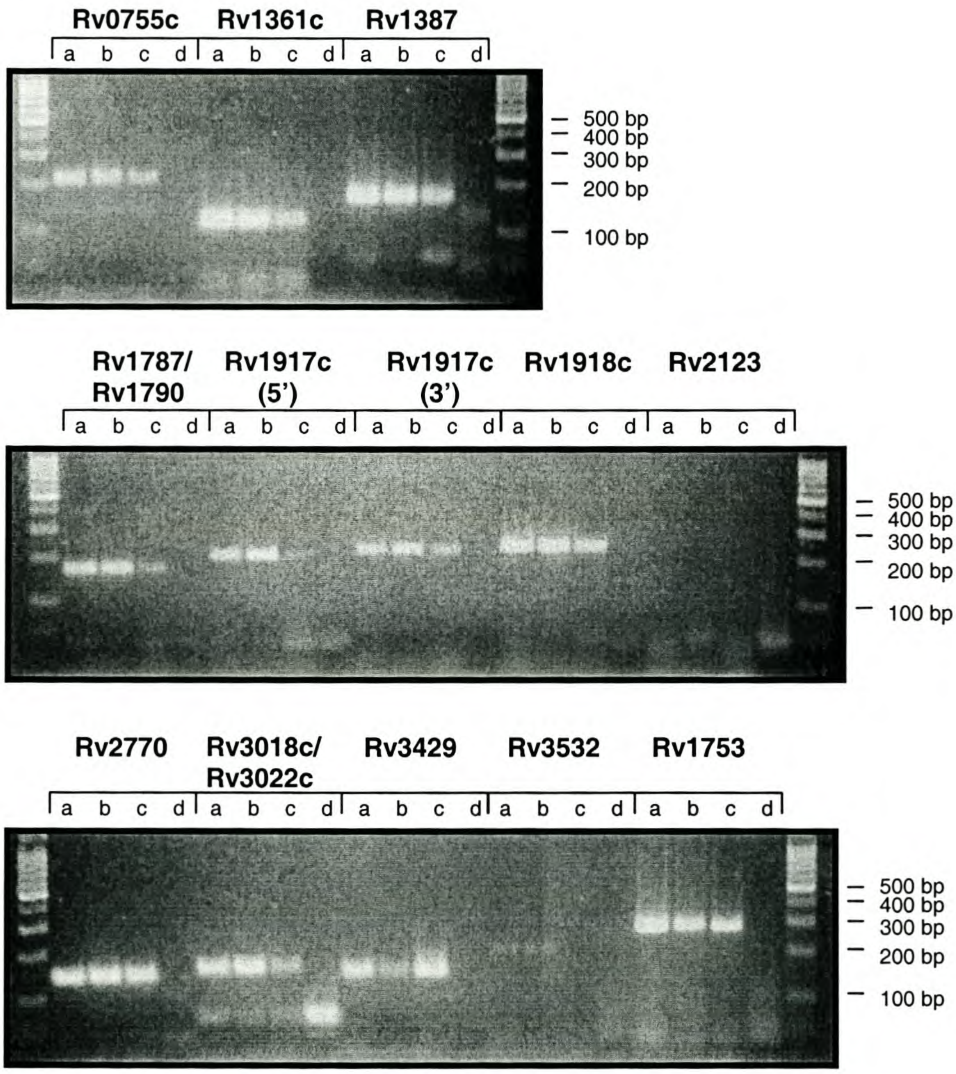
6.3.1 Analysis of *in vitro* expression

RT-PCR was utilized to assay for the presence of PPE mRNA in mid-log phase, late log and stationary phase (7, 14 and 21 day) liquid cultures of *Mycobacterium tuberculosis* H37Rv. Twelve primer pairs were used, and PPE mRNA was detected for all genes analyzed, although with varying signal intensity (Fig. 6.2). The identity of the amplified products was confirmed by DNA sequencing (data not shown). The results confirm that at least 12 PPE genes are transcriptionally active under *in vitro* conditions in liquid cultures of *M. tuberculosis* H37Rv.

RT-PCR signal was compared between different genes at different time points, using a fixed amount of total RNA template. While it is acknowledged that this methodology is only semi-quantitative, the results demonstrate inter-gene and temporal variation in signal intensity. For example, the gene Rv2770c consistently demonstrated the strongest signal intensity, and Rv2123 the lowest. Comparison of the signal obtained for Rv3018c and Rv3429 shows peak expression for Rv3018c at day 14, while Rv3429 demonstrates its lowest expression at this time point (Fig. 6.2). To confirm that the fluctuations in signal were not due to variable PCR efficiency, the experiments were repeated at least twice, using different preparations of total RNA template. In addition, the amplification of Rv1917c was performed using two different primer pairs situated at the 5' or 3' terminal regions of the gene respectively. This produced very similar results, confirming the validity of the observations.

Under the conditions utilized for the inter-gene comparison shown in Fig. 6.2, the Rv2123 gene gave no detectable signal. However, adding increased amounts of template to the reaction yielded detectable signal (not shown). The low signal level obtained for this gene

Figure 6.2. RT-PCR detection of *in vitro* PPE gene expression. Total RNA was isolated from (a) mid-log, (b) late-log and (c) stationary phase cultures of *M. tuberculosis* (7, 14 and 21 day cultures, respectively). PPE mRNA was detected by RT-PCR performed with primers complementary to the PPE genes as shown. A water blank was included concurrently for each primer pair (d). The samples were electrophoretically fractionated in 3% MetaPhor agarose (1X TBE, pH 8.0), with a 100 bp DNA ladder (Promega) run in the outer lanes of each gel. The gels were stained with ethidium bromide, and the products were visualized under UV.



did not appear to be due to a sub-optimal PCR reaction, as standard PCR using *M. tuberculosis* H37Rv genomic DNA as template produced comparable signal intensity for Rv2123 and other genes assayed under identical conditions. It appears therefore, that the gene is expressed, albeit at low levels under *in vitro* growth conditions. This is consistent with the finding of Rodriguez *et al*. (1999), who ascribed their failure to map the transcription start site of Rv2123 to very low *in vitro* mRNA levels.

6.3.2 Analysis of *in vivo* expression

To determine whether PPE genes are expressed within the granuloma micro-environment during host infection, expression profiles were investigated by RNA:RNA *in situ* hybridization (ISH). As a preliminary screen, biotin-labeled antisense riboprobes complimentary to the 5' region of the PPE genes Rv1917c and Rv1787 (representatives of the two largest PPE subgroups) were utilized for RNA:RNA ISH (see Table 6.1 for primer pairs used to generate the probes). Previous results from DNA hybridization to genomic DNA had demonstrated that DNA probes complementary to these regions hybridized to multiple loci (Sampson *et al*., unpublished data; see Chapter 5), due to sequence conservation within the 5' terminal of the gene family. Therefore, during *in situ* hybridization these probes would potentially cross-hybridize to mRNA from more than one PPE gene, maximizing the chance of obtaining a positive signal. The preliminary screen with the two probes demonstrated positive signal in lymph node tissue sections from the 49 year-old female patient (data not shown), validating further investigation of the gene family with this methodology.

PPE gene expression was dissected further by RNA:RNA ISH of lymph node biopsy sections (from a single patient) containing 15 discrete granulomas (Fig. 6.3). Seven biotin-labelled probes representing genes from 3 previously defined phylogenetic subgroups of the PPE gene family were utilized (see Chapter 5 and Fig. 6.1). DNA sequence homology

Figure 6.3. Schematic representation of the granulomas within an adult lymph node biopsy section. A lymph node biopsy specimen from an 18 year old, HIV-negative male TB patient was embedded in paraffin wax, mounted and analyzed by RNA:RNA *in situ* hybridization. The spatial arrangement of 15 granulomas, identified on 2.5x magnification, is indicated below. Non-necrotic granulomas are shaded, while those exhibiting caseous necrosis are unshaded.

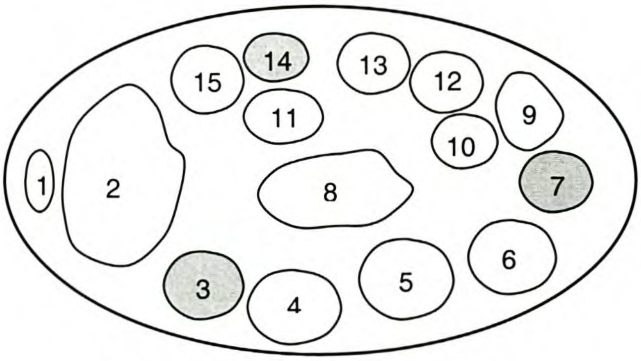


Table 6.2 Summary of RNA:RNA *in situ* hybridization results. The lymph node biopsy section represented above was screened with 7 biotin-labelled PPE-specific riboprobes. Granulomas are numbered according to Fig. 6.2, and were scored as necrotic or non-necrotic. The seven probes and their phylogenetic subgroup are indicated. Granulomas were scored as negative (0) or positive (1) for blue signal.

Granuloma number	Caseous necrosis	Rv0755c (MPTR)	Rv 1753c (MPTR)	Rv 1917c (MPTR)	Rv 1918c (MPTR)	Rv 2123 (Subgroup 4)	Rv 3018c (Subgroup 4)	Rv 3429 (Subgroup 3)
1	Y	1	1	0	0	1	0	0
2	Y	1	1	0	0	1	0	0
3	N	1	1	1	0	1	0	0
4	Y	1	1	1	0	1	0	0
5	Y	1	1	1	0	1	0	0
6	Y	1	1	1	0	1	0	0
7	N	1	1	0	0	1	0	0
8	Y	1	1	0	0	1	0	0
9	Y	1	1	0	0	1	0	0
10	Y	1	1	1	0	1	0	0
11	Y	1	1	1	0	1	0	0
12	Y	1	1	1	0	1	0	0
13	Y	1	1	1	0	1	0	0
14	N	1	1	1	0	1	0	0
15	Y	1	1	0	0	1	0	0

comparisons and DNA hybridization (data not shown) confirmed that these probes were specific for either one (Rv0755c, Rv1753c, Rv1917c, Rv1918c and Rv2123c) or two (Rv3018c and Rv3429) PPE genes (see Chapter 5, Fig 5.2 E and F). The ISH results are summarized in Table 6.2, and illustrated in Figs. 6.4 to 6.6.

Three of the probes (Rv1918c, Rv3018c and Rv3429) demonstrated no detectable signal (Fig. 6.4E, G, H) in any of the granulomas, despite mRNA from these genes having been detected by RT-PCR in an *in vitro* system (Fig. 6.2). In contrast, mRNA from three genes (Rv0755c, Rv1753c and Rv2123c) was detected in all granulomas analyzed (Fig. 6.4B, C, F). Interestingly, the PPE gene Rv1917c demonstrated variable expression, with no detectable signal in 6 granulomas, and positive staining in the other 9 granulomas (Table 6.2; see Fig. 6.4D, 6.5A and 6.5B). The presence or absence of signal did not relate to the presence or absence of necrosis.

Positive signal was predominantly localized to the lymphocyte cuff of non-necrotic and necrotic granulomas. Positively staining cells with the morphology of macrophages were evident (see Fig. 6.6 B, C). A limited amount of signal was observed in the central region of non-necrotic granulomas, but no signal was observed in regions of caseous necrosis (see Figs. 6.4 B, C, D and F).

To confirm the specificity of hybridization, the negative controls included hybridization with the sense probe, which will not hybridize to target mRNA (Fig. 6.4A). To detect non-specific binding of streptavidin to the sections, hybridization was also carried out with no biotinylated probe added (not shown, but identical results to Fig. 6.4A). The results of these controls demonstrate that the positive signal was dependent on specific binding of the biotinylated antisense probes to the target mRNA, and was not an artefact of the methodology. The specificity of the staining was further demonstrated by the differential staining observed, particularly at higher magnifications. (See for example Fig. 6.6B, where negative (green staining) and positive (blue staining) cells are easily distinguished.)

Figure 6.4. ***In situ*** **hybridization of sections though granulomas 3 and 4.** Consecutive tissue sections were probed with (A) sense probe (negative control), and the following antisense riboprobes: (B) Rv0755c, (C) Rv1753c, (D) Rv1917c, (E) Rv1918c, (F) Rv2123, (G) Rv3018c and (H) Rv3429. The region shown is representative of results across the lymph node biopsy. A non-necrotic granuloma is shown on the left of the picture (granuloma 3; Gr. 3), while the granuloma on the right of the picture (granuloma 4; Gr. 4) exhibits caseous necrosis (CN) in the central region. The negative control (hybridization with sense probe), as well as hybridization with the Rv1918c, Rv3018c and Rv3429 riboprobes produced no detectable signal (A, E, G and H). The remaining probes produced positive (blue) signal, which was predominantly concentrated around the lymphocyte cuff (LC) of the granulomas (B, C, and F), although signal obtained with the Rv1917c probe was predominantly associated with the region of granuloma 3 indicated by arrows. Some staining was observed in the central regions of the non-necrotic granulomas shown in B, C and F, but no staining of the central region of caseous necrosis was observed for any of the probes. Folding of tissue sections during the hybridization process is visible as dark lines running across the sections, seen in C-G. (2.5× magnification)

Figure 6.4 ***In situ*** **hybridization of sections through granulomas 3 and 4** (legend opposite)

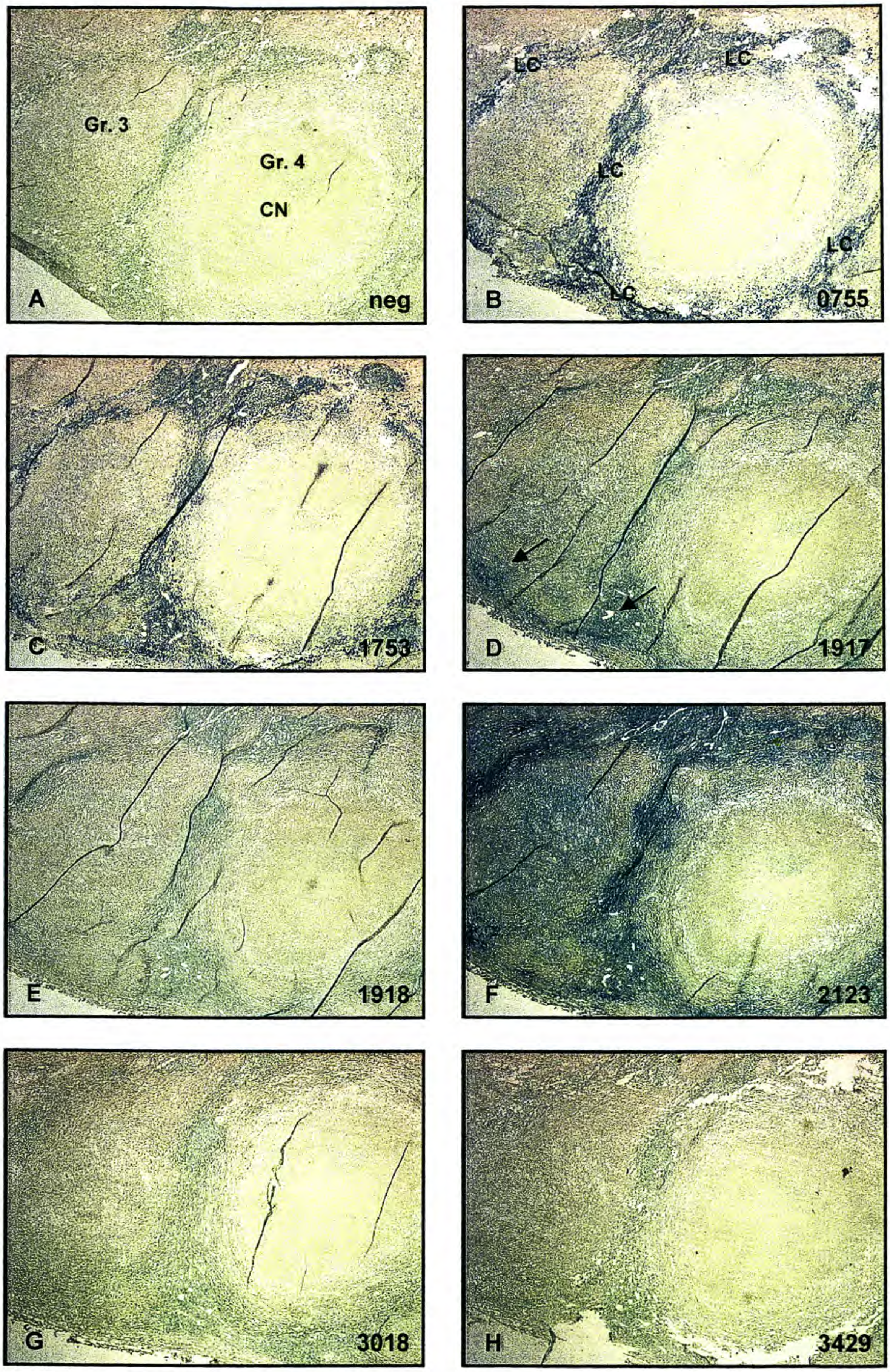


Figure 6.5. *In situ* hybridization of sections though granulomas 7 and 9 with Rv1917c riboprobe. The tissue section was probed with the Rv1917c riboprobe, which demonstrates variable expression in granulomas from a single host. (A) Granuloma 9 demonstrates staining around the periphery and within the granuloma centre, while (B) the immediately adjacent granuloma 7 shows no positive signal. The granuloma shown in (A) does not show signs of necrosis, while the granuloma in (B) exhibits caseous necrosis in the central region (2.5× magnification).

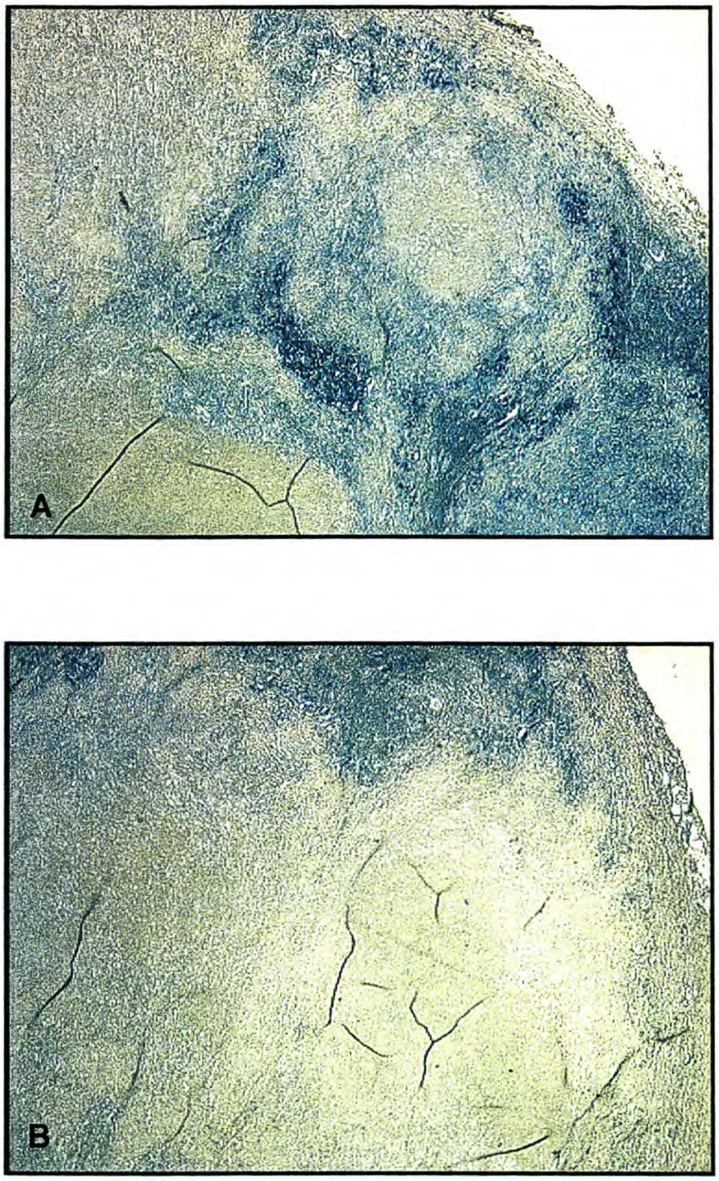
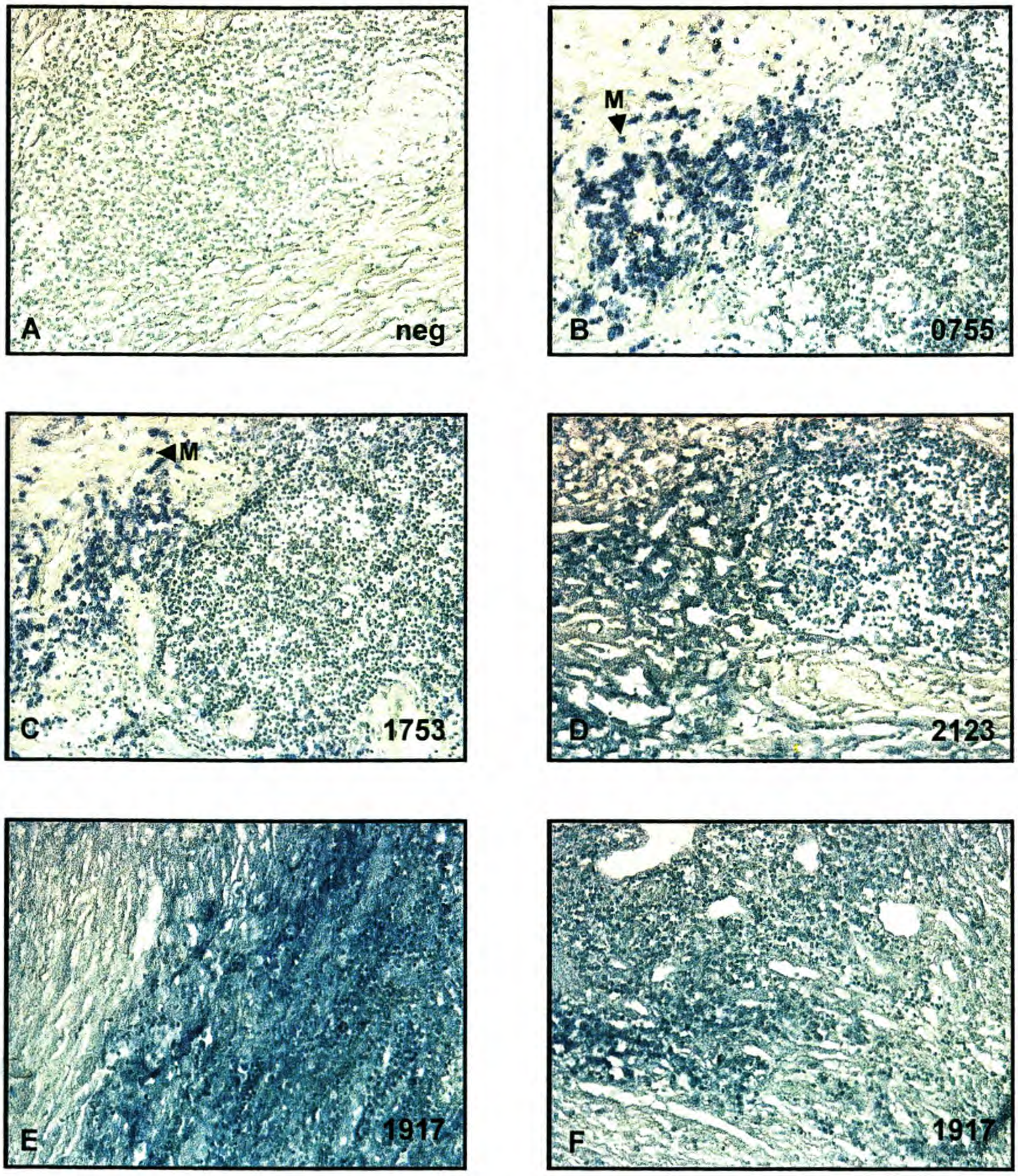


Figure 6.6. Higher magnification view of *in situ* hybridization results. Tissue sections through the peripheral region above granulomas 3 and 4 were probed with (A) sense probe (negative control), and the following riboprobes: (B) Rv0755c, (C) Rv1753c, and (D) Rv2123. Staining of tissue sections adjacent to granulomas 7 and 9 with the Rv1917c riboprobe are shown in (E) and (F), respectively. (B) and (C) demonstrate similar staining patterns, with a clear contrast between negatively stained cells (green) and positively stained cells (blue). The Rv1917c riboprobe produce very intense staining in (E), and less intense staining in (F), with a clear distinction between positively and negatively stained cells. M indicates positively staining cells with the morphology of macrophages. (20x magnification).



6.4 Discussion

The PE and PPE gene families comprise 10% of the *M. tuberculosis* genome, and it has been speculated that they may be of antigenic significance (Cole *et al*., 1998). A growing body of evidence suggests that the PPE gene family plays a potentially important role during host infection (Camacho *et al*., 1999; Dillon *et al*., 1999; Skeiky *et al*., 2000). However, there is very limited data regarding either *in vitro* or *in vivo* expression profiles of members of the gene family. To our knowledge, all PPE expression data published to data have been determined as a result of whole-genome screens, either at the mRNA (Rivera-Marrero *et al*., 1998; Rindi *et al*., 1999) or protein level (Vega-Lopez *et al*., 1993; Dillon *et al*., 1999; Skeiky *et al*., 2000). While these studies have uncovered valuable data regarding the PPE gene family, many subtleties of gene expression patterns may be missed by such approaches. This study addresses this shortcoming by analysis of *in vitro* and *in vivo* expression profiles of selected members of the PPE gene family.

In vitro PPE gene expression was demonstrated by RT-PCR. Interestingly, the results suggest qualitatively variable expression levels across the spectrum of genes analyzed. The temporal differences in gene expression may indicate differential regulation associated with entry into different growth phases. Although the apparent differential expression of the PPE genes requires more rigorous analysis, the results clearly demonstrate that at least 12 members of the gene family are actively transcribed *in vitro*.

A more physiologically relevant approach was provided by RNA:RNA *in situ* hybridization (ISH). This allowed the localization of gene expression patterns within the human granuloma, the site of interaction between the host immune system and the mycobacteria. Analysis of granulomatous lymph node tissue allowed the observation of differential expression of PPE genes *in vivo*. Three of the probes demonstrated no detectable signal, while mRNA from three genes was detected in all granulomas analyzed. Interestingly, the PPE gene Rv1917c showed variable expression between the different granulomas.

The absence of detectable *in vivo* signal for Rv1918c, Rv3018c and Rv3429 (in contrast with *in vitro* findings) does not necessarily exclude these genes as potentially important elements of bacterial pathogenesis. For example, a transposon mutant of the PPE gene Rv3018c is attenuated for growth in mouse lungs, suggesting a possible role during early infection (Camacho *et al.*, 1999), which could imply that the gene is not required for long-term survival within the granuloma. The absence of detectable ISH signal may therefore be attributable to the established nature of the disease within the patient analyzed. Alternatively, the gene may exhibit inter-patient or inter-strain variation in expression patterns, although this remains to be determined.

The results described here highlight the importance of analyzing gene expression within host tissue, as opposed to *in vitro* culture systems. For example, signal for Rv2123 was detected in all the granulomas analyzed, in contrast to the very weak signal obtained with RT-PCR for this gene. Intriguingly, Rv2123 was recently shown to be upregulated under low iron conditions, leading to the suggestion that it may encode a protein involved in siderophore uptake (Rodriguez *et al.*, 1999), an essential component in iron acquisition, and therefore of mycobacterial survival within the human host (De Voss *et al.*, 2000; Olakanmi *et al.*, 2000; Lounis *et al*, 2001). Whether this relates to the observation presented here of Rv2123 expression within human granulomas, and whether the gene is similarly expressed in all patients is as yet unknown.

In contrast to the consistent signal observed in 6 of the PPE genes analyzed, whether positive or negative, the PPE gene Rv1917c demonstrates a range of signal intensities (from no detectable signal to very intense staining) within a single tissue sample, and even within adjacent granulomas on the same tissue section. RT-PCR also revealed fluctuations in Rv1917c expression levels over a time course, with peak expression during late log phase growth, and lowest expression levels during stationary phase, although it is unknown how this relates to the situation within the host. The differential expression of Rv1917c is particularly intriguing in the light of recent findings that the Rv1917c-encoded protein is cell-wall associated (Chapter 7) and that the gene is highly variable in clinical isolates of *M. tuberculosis* (Chapter 5). These features are reminiscent of cell wall antigens in other pathogens, and may imply a similar role for the Rv1917c protein. The differential expression patterns may therefore represent an immune avoidance strategy, akin to "phase variation" observed in other pathogens (Serkin and Seifert, 1998; Park *et al.*, 2000; White-Ziegler *et al.*, 2000; Zhang and Wise, 2001), an area which deserves further attention. The results presented here provide a possible explanation for reports of the variable human serum responses to PPE proteins (Vega-Lopez *et al.*, 1993; Dillon *et al.*, 1999; Skeiky *et al.*, 2000).

This study demonstrates that even closely related PPE genes may be differentially expressed. For example, the Rv1753c, Rv1917c and Rv1918c proteins demonstrate 79% sequence homology within their 180 aa N-terminal regions, but show substantially different expression patterns within the host. Rv1753c mRNA is detected in all the granulomas analyzed, while Rv1918c mRNA is not detectable in any granulomas, and signal for Rv1917c is present in some granulomas, but absent in others. This may reflect functional redundancy among these closely related genes. Alternatively, the gene products may fulfil different functions, and therefore be differentially expressed.

Interestingly, the Rv1917c and Rv1918c genes lie adjacent to one another with 340 bp of intervening sequence, which might suggest that the two genes reside within an operon structure. Similar arrangements have been observed for numerous other PE and PPE genes.

However, this study clearly demonstrates that the two genes are differentially expressed, which does not support an operon structure for this locus.

This study provides the first evidence of PPE gene expression at the site of infection within the human host. PPE mRNA is predominantly associated with the peripheral region of the granuloma, which is classically a dense, lymphocyte-rich margin, with interspersed macrophages and monocytes (Canetti, 1955; Dannenberg and Rook, 1994; Saunders and Cooper, 2000). Two PPE proteins have recently been demonstrated to be potent T cell antigens (Dillon *et al.*, 1999; Skeiky *et al.*, 2000), and the close juxtaposition of antigen presenting cells (macrophages) with T cells in this region could facilitate the presentation of these and other PPE-derived antigens to the host immune system. A limited amount of signal is evident in the central region of non-necrotic granulomas, which would be expected to consist primarily of macrophages, some of which would be infected with *M. tuberculosis* bacilli (Dannenberg and Rook, 1994). No detectable signal for PPE mRNA is observed in the regions of caseous necrosis, which are commonly thought to be hostile to bacterial growth (Dannenberg and Rook, 1994). These findings have yet to be correlated with host cell types, localized cytokine profiles, and the state of bacterial growth. Elucidation of these parameters will be essential to clarifying the interplay between the host and pathogen, as well as understanding PPE gene function.

While the results presented here are consistent with the hypothesis that the PPE gene family may provide the bacterium with a source of antigenic variation (Cole *et al.*, 1998), they highlight the possibility that this is not only mediated by gene-level sequence variation. It is interesting to speculate that these findings may provide a partial explanation for the poor protection against exogenous re-infection, a phenomenon that has recently been observed by a number of investigators (Chaves *et al.*, 1999; Van Rie *et al.*, 1999b; Bandera *et al.*, 2001;

Caminero *et al.*, 2001; Du Plessis *et al.*, 2001). Finally, the differential expression patterns may represent functional specialization of selected members of the gene family, with associated differences in expression profiles.

NOTE: Various people at the University of Stellenbosch Medical School are thanked for their invaluable contributions to the in situ *hybridization study, namely Dr. Junita Bezuidenhout, Anatomical Pathology (Preparation of sections); Ms. Lorrainne Moses (in situ hybridization, analysis and scoring of sections); Dr. Gael Fenhalls (analysis and scoring of sections); Ms. Liesl Stevens (photography).*

CHAPTER 7

EXPRESSION, CHARACTERIZATION AND SUBCELLULAR LOCALIZATION OF THE PPE GENE RV1917C

NOTE: The results presented in the following chapter were published in full as "**Expression, characterization and subcellular localization of the *Mycobacterium tuberculosis* PPE gene Rv1917c.** Sampson, S.L., Lukey, P., Warren, R.M., Van Helden, P.D., Richardson, M. Everett, M.J. *Tuberculosis* (2001) 81:305-17"

(The style of the text and numbering of sections has been formatted to conform with the style of this thesis. Literature cited in the text takes the form of author name and year of publication as opposed to the number format specified by Tuberculosis*. All cited literature is compiled into a separate list at the end of the thesis for ease of reference. Figure 7.1 does not appear in the manuscript, but is included here to illustrate the cloning procedure. No other changes have been made to the body of the text.)*

7.1 Introduction

Sequencing of the genome of *Mycobacterium tuberculosis*, the causative organism of tuberculosis (TB), revealed two large glycine-rich gene families which together account for 10% of the genome (Cole *et al.*, 1998). These families have been termed PE and PPE after the characteristic Pro-Glu and Pro-Pro-Glu motifs near the N-termini of their encoded proteins. The PPE family contains 68 members which have a relatively conserved 180 amino-acid N-terminal region in common, followed by varying C-terminals, some of which contain stretches of polymorphic highly repetitive sequences (Cole *et al.*, 1998). An early report describing the characterization of a novel repetitive DNA sequence in *Mycobacterium bovis*, speculated that these repeats occurred within a family of genes encoding functionally related proteins, which were likely to be cell wall associated (Doran *et al.*, 1992). These repeats have been named the major polymorphic tandem repeats (MPTRs) (Hermans *et al.*, 1992), and it is now known that they occur within a subgroup of the PPE gene family (Cole *et al.*, 1998).

The function of the PPE proteins has not been established, although numerous hypotheses exist. One suggestion is that they provide a potential source of antigenic variation in an otherwise genetically homogenous bacterium (Cole *et al.*, 1998). In support of this hypothesis, PPE gene polymorphism has been shown to be a source of variation between *M. tuberculosis* H37Rv and *M. bovis* BCG (Cole *et al.*, 1998; Gordon *et al.*, 2001). This is potentially significant as the *M. tuberculosis* genome demonstrates limited sequence diversity, and could indicate that such genes are responsive to host immune pressures (Sreevatsan *et al.*, 1997). A further hypothesis is that PPE gene products may function as storage proteins for the relatively rare amino acid asparagine (Cole, 1999). A recent report demonstrated that the PPE gene Rv2123 is upregulated under low iron conditions, leading to the suggestion that the gene product is involved in iron acquisition via siderophore uptake (Rodriguez *et al.*, 1999). It has further been shown that a transposon mutant of the PPE gene Rv3018c was attenuated for growth in macrophages (Camacho *et al*., 1999), suggesting a role in *in vivo* survival of the bacterium.

In vitro evidence for expression of PPE genes was provided by a study which demonstrated that the PPE gene Rv0755c, a member of the MPTR subgroup, is upregulated in the avirulent strain H37Ra relative to the virulent strain H37Rv, as determined on the basis of differential display PCR (Rivera-Marrero *et al*., 1998). Similar methodology has been utilized to show the upregulation of the PPE gene Rv2770c in H37Rv compared to H37Ra (Rindi *et al*., 1999). Evidence for *in vivo* expression of PPE proteins is suggested by two studies which demonstrated that the PPE proteins encoded by Rv1196 (Dillon *et al*., 1999) and Rv0915c (Skeiky *et al*., 2000) are potent T cell antigens.

Taken together, the available data suggest an important role for the PPE gene family in host-pathogen interactions; however, data regarding PPE gene products is very limited. We have therefore investigated one member of the *M. tuberculosis* PPE gene family, Rv1917c. This gene is a member of the MPTR subgroup of PPE genes and is of interest in that it has been shown to be highly polymorphic between clinical isolates of *M. tuberculosis* (Zhang and Young, 1994; Chapter 5). A recent investigation also demonstrated Rv1917c gene polymorphism in *M. bovis* isolates (O'Brien *et al*., 2000). In this study, we have investigated the molecular basis for PPE gene polymorphism, and have shown that the PPE gene Rv1917c is expressed in liquid cultures of *Mycobacterium tuberculosis* H37Rv. In addition, in order to facilitate detection of the protein product, we have cloned and expressed Rv1917c as a C-terminal green fluorescent protein (GFP) fusion in *Mycobacterium smegmatis* and *Mycobacterium bovis* BCG, and have characterized the recombinant protein in terms of its subcellular location and glycosylation status.

7.2 Methods

7.2.1 Bacterial strains, plasmids and materials

Clinical *M. tuberculosis* isolates were collected as part of an ongoing molecular epidemiology study of TB in the Western Cape, South Africa. Genomic DNA was extracted and typed according to an internationally standardized protocol (Van Embden *et al.*, 1993). Further bacterial strains and plasmids utilized in this study are detailed in Table 1. All materials were purchased from SIGMA, unless otherwise stated.

Table 7.1 Bacterial strains and plasmids used in this study. (Plasmids carrying ampicillin, hygromycin and kanamycin resistance markers are indicated by *amp*^{R}, *hyg*^{R} and *kan*^{R}, respectively.)

Bacterial strain / plasmid	Description	Source
E. coli XL-1 Blue	Competent cells, cloning host	Stratagene
M. smegmatis mc^{2}155	Mycobacterial host strain	(Snapper *et al.*, 1990)
M. bovis BCG (Pasteur)	Mycobacterial host strain	Glaxo SmithKline
M. tuberculosis H37Rv	*M. tuberculosis* reference strain	Gift from V. Mizrahi (University of the Witwatersrand, South Africa)
pEGFP-N1	Commercial vector containing *gfp* orf (*kan*^{R})	Clonetech
pACE-1	acetamide-inducible promoter in shuttle vector (*hyg*^{R})	(Parish *et al.*, 1997)
pATB10	shuttle vector (*hyg*^{R})	This study
pGEM T-Easy	TA cloning vector (*amp*^{R})	Promega
pBS4G4	*M. tuberculosis* library clone with near full-length Rv1917c orf (*amp*^{R})	Glaxo SmithKline
pATB45	acetamide-inducible promoter in shuttle vector, with additional cloning sites (*hyg*^{R})	This study
pATB46	full-length Rv1917c orf (*amp*^{R})	This study
pATB47	full-length Rv1917c orf, 3'-terminally fused to gfp (*kan*^{R})	This study
pATB48	full-length Rv1917c orf, 3'-terminally fused to *gfp* in shuttle vector, *ace* promoter (*hyg*^{R})	This study
pATB49	full-length Rv1917c orf in shuttle vector, *ace* promoter (*hyg*^{R})	This study
pATB50	*gfp* in shuttle vector, *ace* promoter (*hyg*^{R})	Glaxo SmithKline
pATB51	*gfp* in shuttle vector, *hsp60* promoter (*hyg*^{R})	Glaxo SmithKline
pATB52	Rv1917c-*gfp* in shuttle vector, *hsp60* promoter (*hyg*^{R})	This study

7.2.2 DNA manipulation

Tandem repeat regions in Rv1917c were amplified from *M. tuberculosis* H37Rv and clinical isolate genomic DNA with Qiagen HotStar *Taq* DNA polymerase using the primer pairs ppe-8 (5'-CAAGTTCAGGGGGATCC-3') and ppe-9 (5'-ACTGAGCGTCGAAGT GAATG-3'); or ppe-11 (5'-GTGACAGTGAGTGGTCAAATCG-3') and ppe-12 (5'-GTTCC AGAAGCCAGATCCG-3') (see Fig. 7.2). Amplification products were cloned using the pGEM-T Easy vector system (Promega), and purified with the Wizard DNA purification system (Promega) prior to sequencing.

7.2.3 Construction of plasmids

Construction of the mycobacterial/*E.coli* shuttle vector pATB45 and subsequently the plasmids pATB48 and pATB49, containing the Rv1917c-*gfp* and Rv1917c orfs respectively, was performed according to standard protocols (Sambrook *et al.*, 1989) (Figure 7.1; all constructs are detailed in Table 1). Plasmid pATB45 was constructed by cloning the entire acetamidase promoter region from pACE-1 into pATB10 which contains the *hsp60* ribosome binding consensus sequence followed by a translational start. Insertion of genes into the single *Nde*I site in pATB45 allows expression under the control of the acetamidase promoter.

The *M. tuberculosis* library clone pBS4G4 contains the near full-length Rv1917c orf, but lacks the 5' terminal region of the gene. The 5' terminal sequence was therefore amplified from H37Rv genomic DNA with the primers ppeA (5'-GAGCTCCATATGAA TTTTTCAACATTGCCAC-3') and ppeB (5'-GACGTCACGGAGTTGAAAGA-3'), using *Taq* DNA polymerase (Promega) with the addition of 10% DMSO. The primers incorporated *Sac*I and *Aat*II sites at either end of the amplification product to facilitate later cloning steps. The amplification product was cloned into the PCR cloning vector pGEM-T Easy (Promega) and the *Sac*I-*Aat*II fragment was later excised and inserted into pBS4G4 to generate the full-length Rv1917c orf (in pATB46). The Rv1917c orf was excised from pATB46 and inserted into pEGFP-N1 (Clonetech) to generate a 3'-terminal translational fusion with the EGFP gene, termed pATB47. Subsequent steps involved insertion of Rv1917c-*gfp* (from pATB47) and Rv1917c (from pATB46) into the *Nde*I site of pATB45 to create pATB48 and pATB49 respectively.

Plasmids pATB50, containing the *gfp* orf immediately downstream of the inducible acetamidase promoter; and pATB51, containing the *gfp* orf downstream of the constitutive *hsp60* promoter were obtained from F. Cook (Glaxo SmithKline). Plasmid pATB52 was constructed by replacing the *gfp* orf in pATB51 with the Rv1917c-*gfp* orf from pATB48, such that Rv1917c-*gfp* is placed downstream of the *hsp60* promoter.

To confirm the integrity of plasmids transformed into mycobacterial cells, plasmids were transferred to *E. coli* by electroduction (Baulard *et al*., 1992), isolated using standard procedures (Sambrook *et al*., 1989), and restriction mapped and/or sequenced. All DNA sequencing reactions were performed on an ABI automated sequencer.

Figure 7.1 Scheme for cloning of pATB48, pATB49 and pATB50. The shuttle vector pATB45, containing the acetamide-inducible promoter, was constructed from pACE-1 and pATB10 (A). The 5'-terminal region of Rv1917c was PCR amplified from H37Rv genomic DNA (B) and ligated into a TA cloning vector. The 5' terminal region was excised from the pGEM-T Easy vector and ligated into pBS4G4 (* indicates Rv1917c with absent 5' terminal), to generate pATB46 (C), containing the full-length Rv1917c orf. This was excised from pATB46 and cloned into pATB45, to generate pATB49 (D) containing the full-length Rv1917c orf downstream of the acetamide-inducible promoter. The full-length Rv1917c orf was excised with *Pvu*II and ligated into pEGFP-N1 to construct pATB47 (E), containing the full-length Rv1917c orf fused to the *gfp* orf at its 3'-terminus. The Rv1917c-*gfp* fusion was excised from pATB47 and ligated into pATB45, to generate pATB48 (F), a 3'-terminal Rv1917-*gfp* translational fusion, downstream of the acetamide-inducible promoter. The construct pATB50 (G), contains the *gfp* orf downstream of the acetamide-inducible promoter.

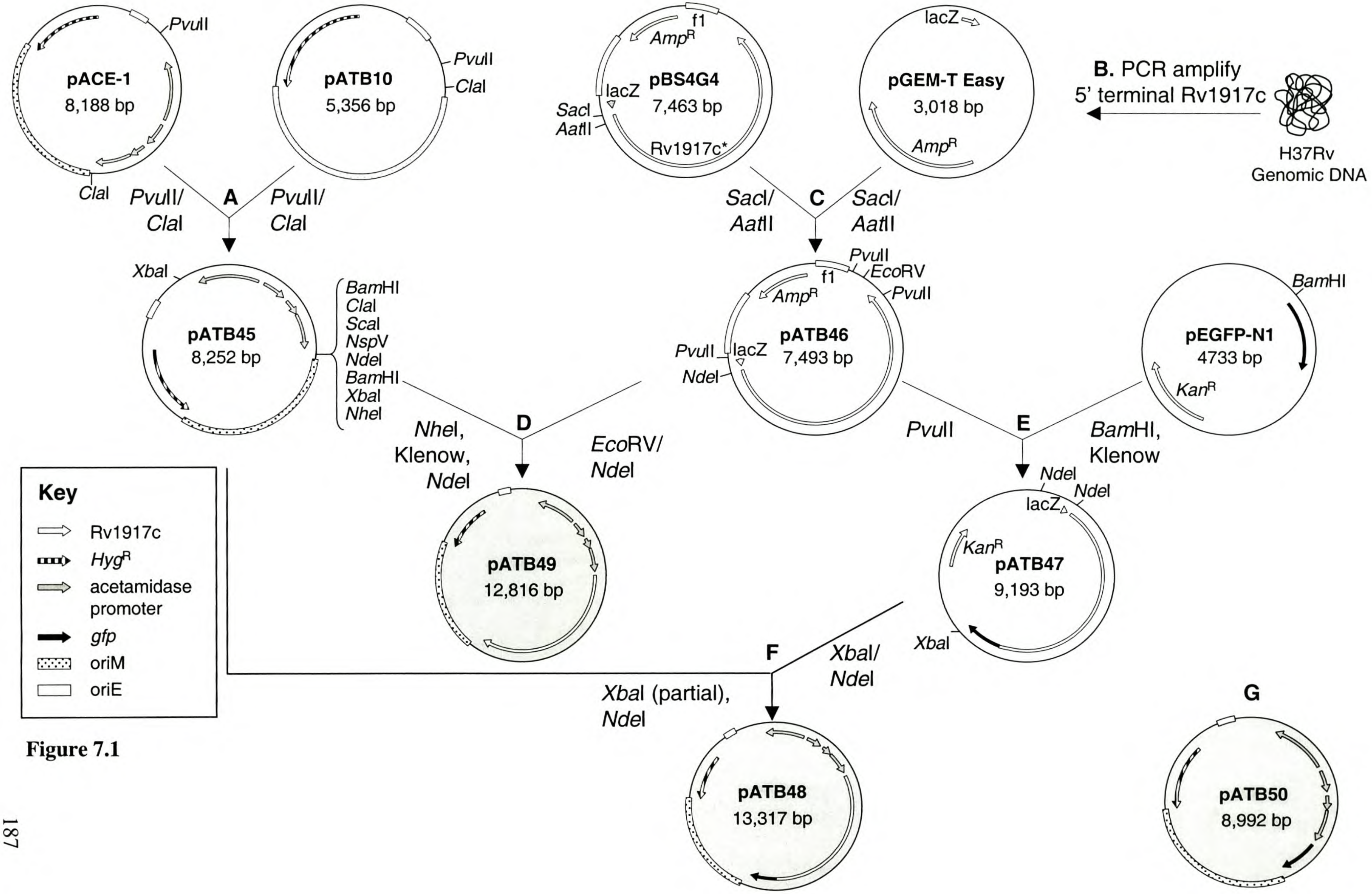

pACE-1
8,188 bp
PvuII
ClaI
pATB10
5,356 bp
PvuII
ClaI
PvuII/
ClaI
A
PvuII/
ClaI
pBS4G4
7,463 bp
f1
AmpR
lacZ
SacI
AatII
Rv1917c*
pGEM-T Easy
3,018 bp
lacZ
AmpR
B. PCR amplify
5' terminal Rv1917c
H37Rv
Genomic DNA
SacI/
AatII
C
SacI/
AatII
pATB45
8,252 bp
XbaI
BamHI
ClaI
ScaI
NspV
NdeI
BamHI
XbaI
NheI
pATB46
7,493 bp
PvuII
EcoRV
PvuII
f1
AmpR
PvuII
lacZ
NdeI
pEGFP-N1
4733 bp
BamHI
KanR
NheI,
Klenow,
NdeI
D
EcoRV/
NdeI
PvuII
E
BamHI,
Klenow
pATB49
12,816 bp
pATB47
9,193 bp
NdeI
NdeI
lacZ
KanR
XbaI
Key
Rv1917c
HygR
acetamidase
promoter
gfp
oriM
oriE
F
XbaI/
NdeI
XbaI (partial),
NdeI
pATB48
13,317 bp
G
pATB50
8,992 bp

Figure 7.1

7.2.4 RNA manipulation

Total RNA was extracted from mid-log phase, late log and stationary phase (7, 14 day and 21 day) liquid cultures of *M. tuberculosis* H37Rv. Cells were pelleted by centrifugation at 4,000 rpm for 5 minutes and the cell pellet was resuspended in Trizol (GibcoBRL). Resuspended cells were added to FastPrep Blue tubes (Bio101) containing silica beads, and ribolysed at speed 6.5 for 45 s. The tubes were centrifuged at 13,000 rpm for 1 minute to pellet cell debris and beads. The Trizol layer was removed and chloroform-extracted. After isopropanol precipitation, RNA was resuspended in nuclease-free H_2O (Promega) with 10 mM DTT and 1U/μl RNasin (Promega). DNaseI treatment was performed as recommended by the manufacturer (Roche). RNA was subsequently purified using the Qiagen RNeasy Mini kit, and stored in aliquots at -20°C. The integrity of the RNA was determined by electrophoresis in 1% agarose (1X TBE, pH 8.0).

RT-PCR was performed using the Titan One-Tube RT-PCR kit (Roche). The oligonucleotides Rv1917cF (5'-AACTCGGCCCTCATATTCGG-3') and Rv1917cR (5'-CGCGGCAAGCCATCCTAGG-3') were used to specifically amplify a 204 bp 5'-terminal product from Rv1917c (see Fig. 7.2B). In addition, the primers 1917ISHF (5'-CGTCCTATTTCCCGCATC-3') and 1917ISHR (5'-GCTCAGCGGAATATTCAAACC-3') were used to specifically amplify a 220 bp 3' terminal product from Rv1917c (see Fig. 7.2B). Controls included RNaseA-treated samples, DNaseI-treated samples, PCR with heat-inactivated reverse transcriptase, and a water blank. RT-PCR products were electrophoretically fractionated in 2% agarose (1X TBE, pH 8.0), and visualized under UV after ethidium bromide staining. Controls were also Southern blotted, and hybridized with a ^{32}P-labelled probe complementary to the 5' terminal of Rv1917c (nucleotides 133 to 603). To confirm that the correct regions had been amplified, the amplification products were cloned and sequenced.

7.2.5 Bacterial transformations and culture conditions

E. coli XL-1 Blue competent cells (Stratagene) were transformed by heat shock, according to a standard method (Sambrook *et al.*, 1989). Transformed cells were plated onto LB-agar containing 100 µg/ml ampicillin, 50 µg/ml kanamycin or 100 µg/ml hygromycin. Competent *M. smegmatis* mc^2155 and *M. bovis* BCG (Pasteur) were transformed by electroporation (BioRad GenePulser) (Parish and Stoker, 1999b). Transformed mycobacterial cells were plated onto Middlebrooks 7H11 agar supplemented with 0.2% glycerol, 10% OADC enrichment and 50 µg/ml hygromycin. Liquid cultures were grown in LB medium (*E. coli*), M63 medium (Triccas *et al.*, 1998) (*M. smegmatis* mc^2155) or Middlebrooks 7H9 medium supplemented with 0.2% glycerol, 10% ADC enrichment and 0.05% Tween-80 (MADCTW) (*M. bovis* BCG), with appropriate antibiotics. To induce expression of recombinant genes from the acetamidase promoter (Parish *et al.*, 1997; Triccas *et al.*, 1998), transformed cells were cultured with or without 0.2% acetamide (Triccas *et al.*, 1998) in M63 medium (*M. smegmatis* mc^2155), or in MADCTW (*M. bovis* BCG). To ascertain optimal growth/induction conditions for *M. bovis* BCG, alternative growth media were assessed; these were Hopwood (Hopwood *et al.*, 1985), adapted Hopwood (glucose replaced with succinate and L-asparagine replaced with $(NH_4)_2SO_4$) (Hopwood *et al.*, 1985), Proskauer and Beck (Allen, 1999) and M63. Stirred cultures of *M. tuberculosis* H37Rv were incubated in 28 ml screw-cap vials containing 10 ml of Middlebrooks 7H9 medium supplemented with 10% ADC enrichment and 0.05% Tween-80.

7.2.6 Protein extraction and cell fractionation

To monitor the expression of recombinant genes, crude protein extracts were prepared as follows: *M. smegmatis* cells were lysed using the B-PER Bacterial Protein Extraction Reagent (Pierce). The resulting supernatant and pellet fractions were stored at -20°C until further use. *M. bovis* BCG cells were lysed using the Hybaid Recovery Ribolyser kit (blue).

Cells pelleted from 50 ml mid-log phase culture were washed twice in 1 ml phosphate buffered saline (PBS) with 1% Tween-20. A volume of blue Hybaid beads equal to the pellet size was added to the washed pellet. Three-hundred microlitres of SDS lysis buffer (0.3% SDS, 200 mM DTT, 28 mM Tris-HCl, 22 mM Tris-base, 1mM PMSF) or Triton X-114 lysis buffer (1% Triton X-114 in PBS) was added and mixed. The mixture was ribolysed for 45s, speed 6.5, then boiled for 5 minutes. The sample was cooled to room temperature, then centrifuged (13,000 rpm, 10 minutes). The supernatant was removed, the insoluble fraction was resuspended in 300 µl lysis buffer, and both fractions were stored at -20°C until further use.

A more refined cell fractionation was performed on *M. smegmatis* using the detergent Triton X-114 (Radolf *et al.*, 1988; Parish *et al.*, 1999). *M. smegmatis* cells were pelleted, and the culture supernatant was retained and concentrated 50X using Centricon Plus-20 tubes (Amicon). The cell pellet was resuspended in ice-cold PBS containing 1 mM PMSF, and sonicated (10x 15 s bursts, Sanyo Soniprep 150). The sonicate was centrifuged at 13,000 rpm for 30 minutes at 4°C. The resulting supernatant was centrifuged at 100,000 xg for 2 hours (4°C), to separate membrane (pellet) and cytosolic (supernatant) fractions. The pellet was washed twice in ice-cold PBS (100,000 xg, 1 hour, 4°C), then resuspended in an ice-cold 1% Triton X-114 solution in PBS with 1mM PMSF. The pellets were vigorously vortexed, and stored at 4°C overnight. Pellets were vortexed again, then centrifuged at 13,000 rpm for 30 minutes at 4°C. The supernatant was then phase-partitioned and back-extracted 4 times as described elsewhere (Radolf *et al.*, 1988).

Protein fractions were analyzed for expression of recombinant proteins by SDS-PAGE and immunoblotting with 1:1000 mouse monoclonal anti-GFP (Roche), followed by 1:5000 goat anti-mouse-horseradish-peroxidase (HRP) conjugate for detection.

7.2.7 Immunoprecipitation

Mycobacterial lysates (prepared by sonication or ribolysis, as described above) were centrifuged at 13,000 rpm for 30 mins at 4°C to pellet unbroken cells, prior to immunoprecipitating the Rv1917c-GFP and GFP proteins using the Protein G Immunoprecipitation kit (Roche) in conjunction with affinity-purified rabbit polyclonal anti-GFP (R. Wilson, Glaxo SmithKline).

7.2.8 Glycan detection

Glycan detection was performed on crude protein extracts and immunoprecipitated proteins. For glycan staining with the DIG Glycan/Protein Double labelling kit (Roche), membranes were stored dry until use, then pre-wetted in methanol prior to following the protocol recommended by the manufacturer. For ConA detection of glycoproteins, membranes were blocked overnight in PBS-T (PBS with 0.1% Tween 20), then probed with ConA-HRP, as described elsewhere (Herrman *et al.*, 1996).

7.2.9 Anti-GFP binding to whole cells and FACS analysis

M. smegmatis cells were pelleted from 3 ml mid-log phase culture by centrifugation at 13,000 rpm for 10 min, then washed twice in PBS with 0.05% Tween-20. Cells were resuspended in 100 μl wash buffer, prior to incubating with affinity-purified rabbit polyclonal anti-GFP or pre-immune serum at 1:1000, 1:10,000 and 1:100,000 dilutions at room temperature (RT) for 1 hour with shaking. Cells were washed twice, then incubated with goat anti-rabbit IgG R-Phycoerythrin conjugate (GAR-PE) at 1:40 dilution in 100 μl wash buffer, for 1 hour at RT with shaking. Cells were washed twice, then resuspended in 50 μl wash buffer. Controls included cells with no antibody, and cells with GAR-PE only. The resuspended cells were fixed by the addition of 4% paraformaldehyde, then mixed with PBS, after which the samples were analyzed by flow cytometry (Coulter Epics XL-MCS), with a cut-off at 50,000 events.

7.2.10 Microscopy

Bacterial cells expressing Rv1917c-GFP and GFP were visualized using fluorescence microscopy. Fluorescence was detected and imaged at 510 nm emission (excitation at 488 nm) using a 63x objective lens on a Leica TCS NT confocal microscope (Leica UK Ltd).

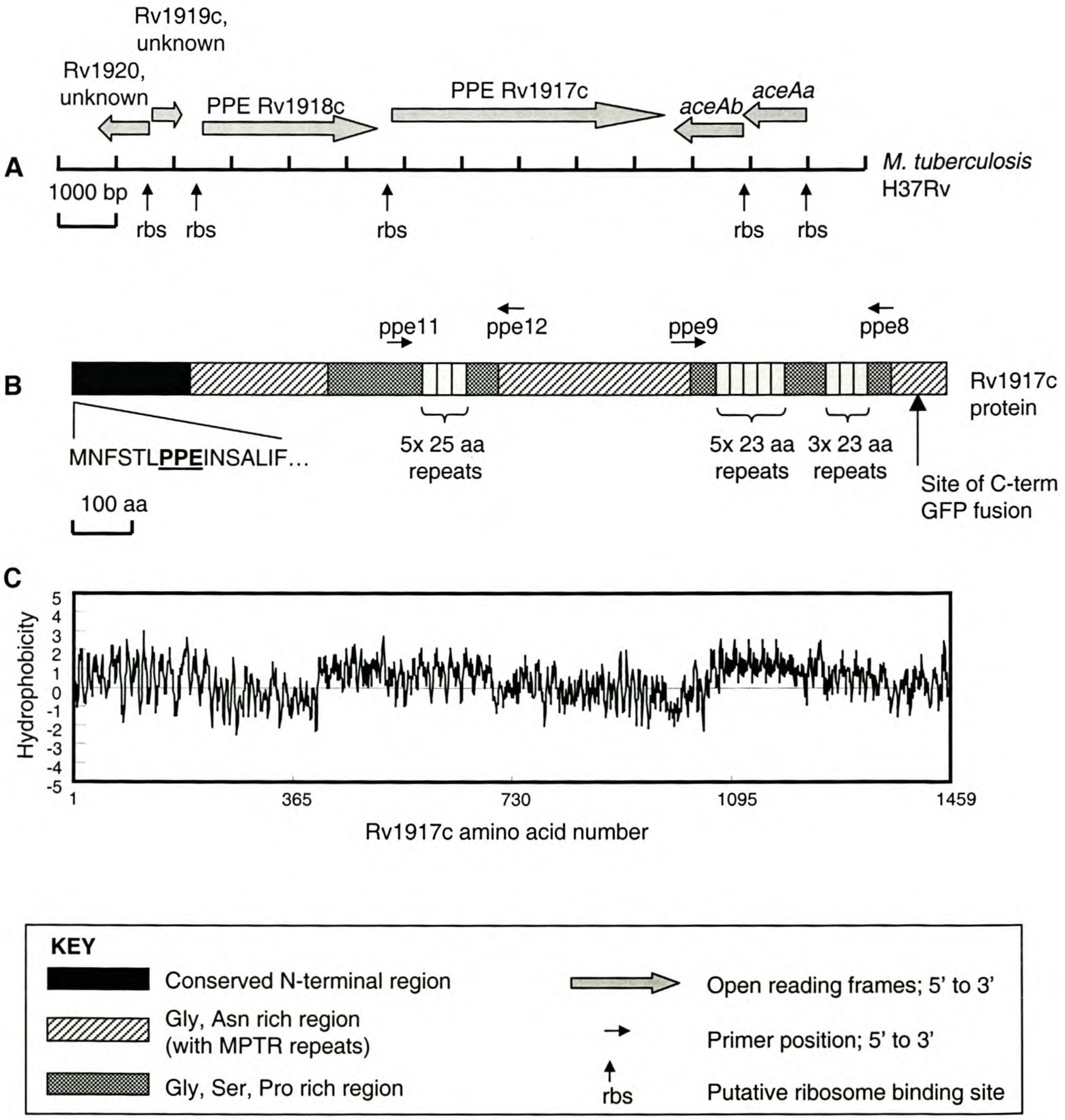
7.3 Results

7.3.1 Bioinfomatic analysis

Analysis of the genomic environment of the Rv1917c gene (http://genolist.pasteur.fr/ TubercuList) reveals that it is situated downstream of Rv1918c, another member of the PPE gene family (Fig. 7.2A). Rv1917c is situated downstream of, and in opposite orientation to the putative isocitrate lyase genes *aceAa* and *aceAb* (Fig. 7.2A). Rv1917c and Rv1918c lie in the same orientation, and may therefore be co-transcribed, although this remains to be experimentally determined. Rv1917c and Rv1918c encode putative proteins of 143 kDa and 98 kDa respectively, both with a relatively conserved 180 aa N terminal region which contains the characteristic PPE motif at positions 7-9. Rv1917c and Rv1918c are members of the PPE-MPTR subgroup of PPE genes, of which there are 22 in *M. tuberculosis* H37Rv, and contain stretches of 15 bp MPTR repeats, encoding Gly- and Asn-rich regions. Both genes also encode Gly-Ser-Pro rich regions, which contain long tandem repeats, ranging from 69 to 78 bp in length, a feature shared with one other PPE gene, Rv1753. Rv1917c encodes 3 blocks of tandem repeats, of 75 bp, 69 bp and 69 bp, all with different consensus repeats (Figs. 7.2B and 7.3). A Hopp-Woods hydrophobicity plot (Fig. 7.2C) demonstrates that the long tandem repeat regions are hydrophobic.

In light of the hypothesis that PE and PPE proteins may represent a source of antigenic diversity in *M. tuberculosis* (Cole *et al*., 1998), and evidence that glycosylation may modulate the antigenic properties of mycobacterial proteins (Romain *et al*., 1999), it was of interest to determine the glycosylation status of Rv1917c. Analyses of several PE and PPE genes using the NetOGlyc program (Hansen *et al*., 1995) predicted many putative O-glycosylation sites in these proteins. In Rv1917c, 13 high-scoring residues were identified (not shown). The number of transmembrane helices predicted for Rv1917c was different for each of five programs used (HMMTOP, DAS, TMPRED, SOSUI, TMHMM), and ranged from 0 to 19.

Figure 7.2. Bioinformatic analysis of the Rv1917c gene and its encoded product. (A) Genomic environment of Rv1917c gene. (B) Schematic representation of the Rv1917c protein showing the mosaic structure typical of many PPE proteins. The positions of tandem repeat blocks of 25 aa, 23 aa and 23 aa are indicated. The relative positions of the 5' and 3' RT-PCR products and primers used to amplify polymorphic tandem repeat regions are shown. (C) Hopp-Woods hydrophobicity plot of Rv1917c protein, aligned with protein structure in (B).



7.3.2 Analysis of Rv1917c polymorphism

Previous studies (Zhang and Young, 1994; O'Brien *et al.*, 2000) and results from this laboratory (see Chapter 5) had shown that the PPE gene Rv1917c demonstrated tandem repeat polymorphism, and the molecular basis for this was investigated. PCR amplification of the Rv1917c tandem repeat regions from 184 clinical isolates demonstrated eight variants for the ppe-8/ppe-9 primer pair (2 blocks of 69 bp tandem repeats), and five variants for the ppe-11/ppe-12 primer pair (75 bp tandem repeats) (results not shown). The regular spacing of size variants suggested that differing numbers of tandem repeats were responsible for the observed polymorphism (see Fig. 7.3A). This was investigated further by cloning and sequencing 3 ppe-8/ppe-9 variants (Fig. 7.3A) from the strains 336, 500 and 747 (> 70% related as defined by IS*6110* DNA fingerprinting). Comparison of this sequence data to that of *M. tuberculosis* H37Rv and CDC1551 confirmed that differences in the numbers of tandem repeats were responsible for the observed variants (GenBank accession numbers: AF082288, AF082289 and AF082291) (Fig. 7.3B; see also Chapter 5).

7.3.3 Detection of PPE mRNA

To determine whether Rv1917c is expressed in *M. tuberculosis*, cultures of *M. tuberculosis* H37Rv were grown and analyzed for expression of Rv1917c message. RT-PCR analysis reproducibly confirmed the presence of Rv1917c mRNA in mid-log phase liquid cultures (day 7), as well as in late log and stationary phase cultures (day 14 and 21, respectively) (Fig. 7.4A, lanes 1-3). The control reactions were hybridized with a Rv1917c 5' probe, which confirmed that the signal was due to mRNA, and that there was no background signal due to genomic DNA or amplicon contamination (Fig. 7.4B). Although there is some nucleotide sequence homology between the 5' regions of PPE genes, it was still possible to design primers in this region to ensure specific amplification of Rv1917c. Nonetheless, a

Figure 7.3. Analysis of PCR variants. (A) Two blocks of 69 bp tandem repeats near the C-terminal of Rv1917c were amplified from *M. tuberculosis* clinical isolates to produce 3 PCR variants, which were fractionated in 2% agarose (1X TBE, pH 8.0), and visualized under UV after ethidium bromide staining. Strain numbers as follows: Lanes: 1, 336; 2, 500; 3, 747 (811 bp PCR product obtained for H37Rv not shown). (B) Schematic representation of nucleotide sequence variants relative to *M. tuberculosis* H37Rv and CDC1551. The consensus amino acid sequence for the two sets of repeats is given. Repeats within the two regions are similar, but non-identical, and the different shading of blocks represents nucleotide sequence variation. The intervening sequence is identical in all isolates sequenced.

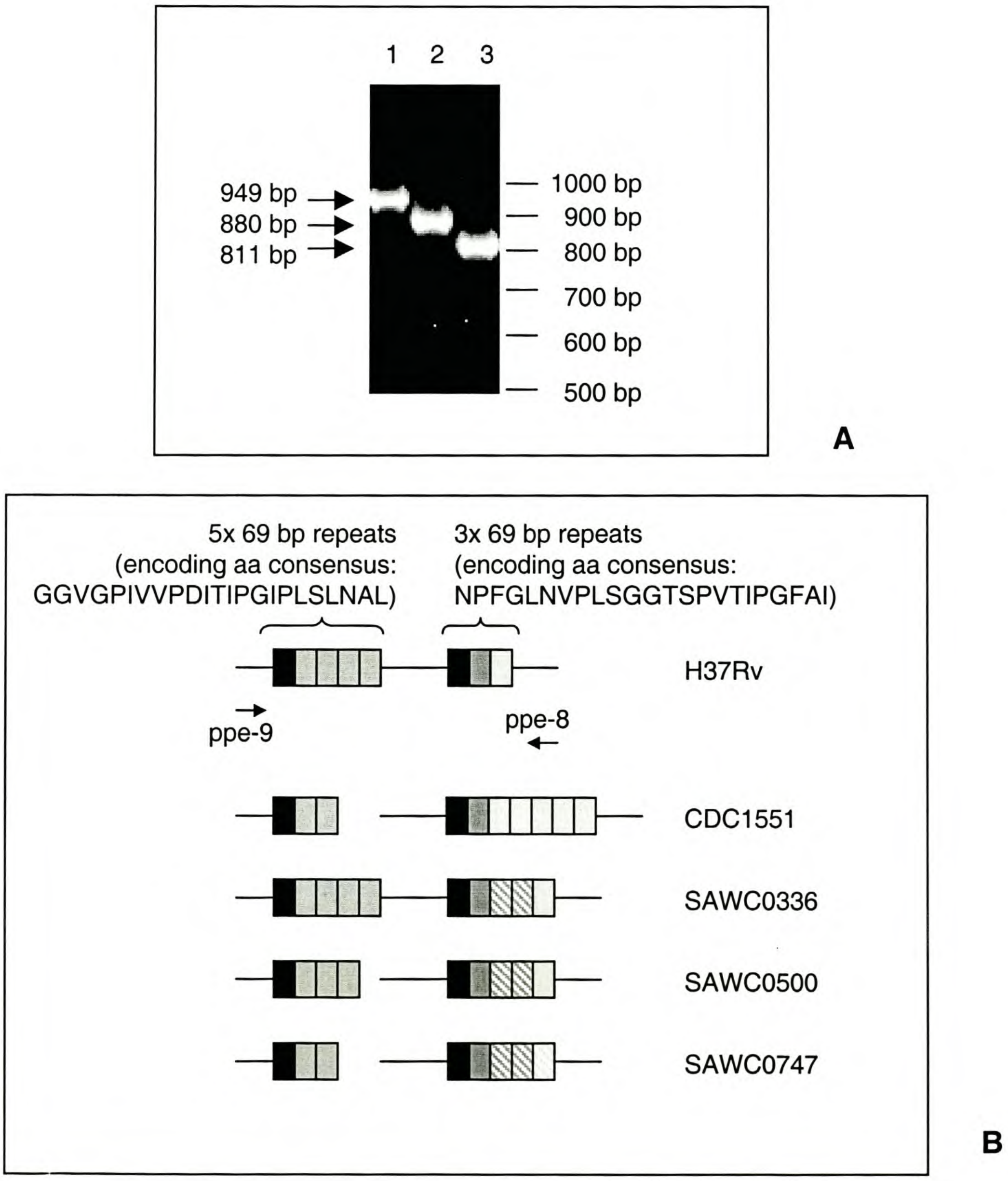
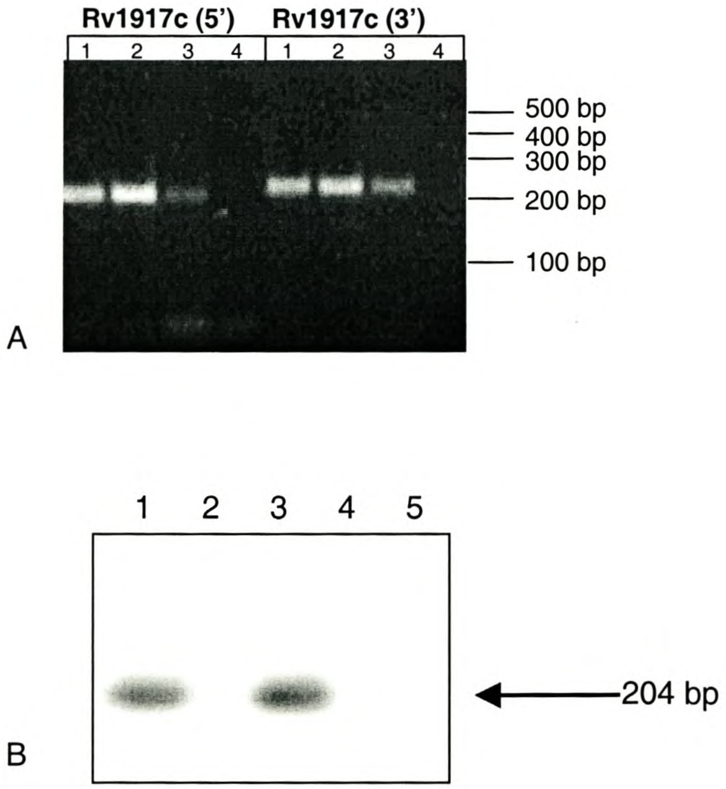


Figure 7.4. RT-PCR detection of Rv1917c mRNA in liquid cultures of H37Rv. RT-PCR products were amplified from either the 5' or 3' regions of Rv1917c from *M. tuberculosis* H37Rv RNA extracted from mid-log, late log and stationary phase cultures. (A) Products were fractionated in 2% agarose (1X TBE, pH 8.0), and visualized under UV. Lanes: 1, mid-log; 2, late log; 3, stationary phase; 4, water blank. (B) Control reactions using 5' terminal primers on total RNA extracted from mid-log phase cultures. Products were detected by Southern hybridization with a ^{32}P-labelled probe complementary to the 5' terminal of Rv1917c. Lanes: 1, RT-PCR; 2, Heat-inactivated reverse transcriptase; 3, DNaseI-treated RNA; 4, RNaseA-treated RNA; 5, water blank.



second set of primers complementary to the unique 3' sequence of the gene were also utilized. Furthermore, sequencing of the RT-PCR products confirmed the specificity of amplification.

7.3.4 Expression of Rv1917c-GFP and GFP in *M. smegmatis* and *M. bovis* BCG

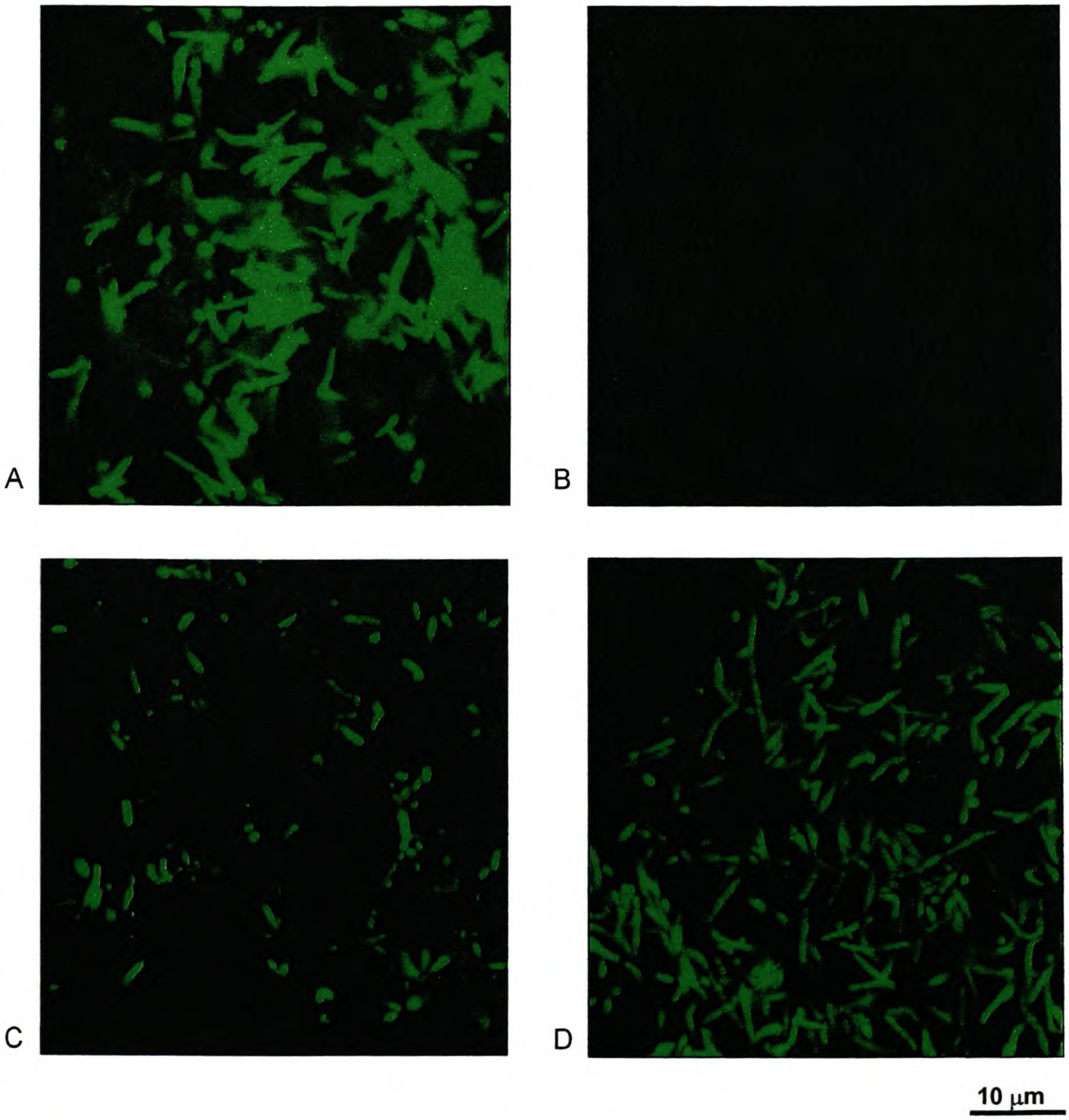
In the absence of specific antibodies to Rv1917c, a fusion protein strategy was adopted, whereby expression and localization of recombinant protein could be determined using antibodies specific to the fusion moiety. GFP was chosen as the fusion moiety as its intrinsic fluorescence additionally facilitated direct visualization of expressed proteins and enabled dual labelling experiments as described later in the Results section.

To visualize the expression of recombinant proteins, the plasmids pATB48 (*ace* promoter, Rv1917c-*gfp*), pATB50 (*ace* promoter, *gfp*), pATB51 (*hsp60* promoter, *gfp*) and pATB52 (*hsp60* promoter, Rv1917c-*gfp*) were introduced into *M. smegmatis* mc^{2}155 and *M. bovis* BCG (Pasteur) and the transformed cultures were analyzed for expression of GFP and GFP fusion products by fluorescence confocal microscopy. High levels of GFP expression were observed from the *hsp60* promoter (pATB51) in both *M. smegmatis* (Fig. 7.5A) and *M. bovis* BCG (not shown). However, when Rv1917c-*gfp* was introduced under the control of the same strong *hsp60* promoter (pATB52) only small colonies were recovered and these failed to grow in liquid media (not shown). *M. smegmatis* pATB52 transformants picked directly from plates demonstrated limited fluorescence when visualized by microscopy. Few colonies demonstrated fluorescence, and where fluorescence was detected, this was weak and unevenly distributed throughout the cells (not shown). This suggested that high-level expression of Rv1917c-GFP was poorly tolerated in the host cells. In order to check that rearrangement of the plasmid had not occurred, pATB52 was transferred from *M. smegmatis* into *E. coli* by electroduction (Baulard *et al.*, 1992) and demonstrated to be correct by sequencing and restriction map analysis.

To overcome the problems experienced with high-level expression of the recombinant proteins, Rv1917c-*gfp* was introduced under the control of the inducible acetamidase promoter (pATB48) which had no obvious effect on growth of either *M. smegmatis* or *M. bovis* BCG, in the presence or absence of acetamide. No fluorescence was observed in cultures of *M. smegmatis* carrying pATB49 (Rv1917c) (Fig. 7.5B). In contrast, expression of Rv1917c-GFP (pATB48) and GFP (pATB50) in *M. smegmatis* was induced by the addition of 0.2% acetamide to the M63 growth medium, and detected by fluorescence microscopy (Fig. 7.5C and 7.5D, respectively). Fluorescence was also detected by flow cytometric analysis, using pATB49 (Rv1917c) transformed cells as a non-fluorescent baseline (see later in text and Fig. 7.9). It was evident from both the fluorescence microscopy and flow cytometric analysis that pATB50 (*gfp*) transformed cells demonstrated higher levels of induction (both in terms of intensity of fluorescence and proportion of cells fluorescing) than pATB48 (Rv1917c-*gfp*) transformed cells (compare Fig. 7.5C and 7.5D)).

M. bovis BCG does not grow in M63 minimal medium, therefore four alternative culture media were investigated (see Materials and Methods). Induction was found to be optimal in the adapted Hopwood medium, although growth was poor. The best combination of growth and induction was obtained using MADCTW + 0.2% acetamide. As with *M. smegmatis,* induction of expression in *M. bovis* BCG was more efficient for pATB50 (*gfp*) than for pATB48 (Rv1917c-*gfp*) (results not shown).

Figure 7.5. ***M. smegmatis* expressing Rv1917c-GFP and GFP products.** (A) *M. smegmatis* cells transformed with pATB51 (*hsp60-gfp* construct) exhibit a high level of fluorescence, as visualized by confocal fluorescence microscopy, consistent with constitutive expression of GFP. (B) *M. smegmatis* cells transformed with pATB49 (*ace*-Rv1917c construct) provide a non-flourescent negative control after induction with 0.2% acetamide. (C) pATB48 (*ace*-Rv1917c-*gfp* construct) and (D) pATB50 (*ace-gfp* construct) after induction with 0.2% acetamide. Bacteria in panels C and D exhibit fluorescence demonstrating the expression of the Rv1917c-GFP and GFP proteins, respectively.



7.3.5 Detection of Rv1917c, Rv1917c-GFP and GFP in crude protein extracts

In order to confirm expression of recombinant proteins, crude extracts were prepared and proteins visualized on silver stained SDS-PAGE gels and western blots. A number of different lysis procedures were investigated, namely addition of B-PER protein extraction reagent (Pierce), disruption using the ribolyser and sonication. The B-PER reagent produced partial cell lysis when mixed with *M. smegmatis* pellets. These were separated into soluble and insoluble fractions by centrifugation. SDS-PAGE electrophoresis of the *M. smegmatis* extracts demonstrated the presence of protein bands of the predicted sizes for Rv1917c-GFP (166 kDa) and Rv1917c (143 kDa) in the insoluble fractions in cells transformed with pATB48 and pATB49, respectively (Fig. 7.6A, lanes 1 and 3)). Immunodetection with anti-GFP antibody confirmed the presence of the Rv1917c-GFP protein (Fig. 7.6B, lane 1) and indicated expression of GFP protein (28 kDa) from cells transformed with pATB50 (Fig. 7.6B, lanes 5 and 6). Some of the GFP protein was found in the B-PER-soluble fraction, as expected for a cytosolic protein, but much was associated with the insoluble fraction. While this could reflect the presence of inclusion bodies, it should be noted that the B-PER lysis reagent was developed for use with *E. coli* cells, which are much more easily lysed than mycobacterial cells. Therefore, this is probably predominantly due to the presence of unlysed cells. In contrast, the Rv1917c-GFP and Rv1917c proteins were exclusively associated with the B-PER-insoluble fraction. A number of lower molecular weight bands were also evident below the Rv1917c-GFP band in Fig. 7.6B. These co-immunoprecipitated with this product and are likely to be degradation products of the full-length fusion protein retaining the C-terminal GFP epitope.

Both B-PER extraction and sonication were found to be ineffective at lysing *M. bovis* BCG. Efficiency of lysis was determined by visual inspection of lysates as well as visualization of protein lysates on silver stained SDS-PAGE gels (not shown). Improved

Figure 7.6. Crude protein extracts from *M. smegmatis*. Proteins were extracted from *M. smegmatis* expressing Rv1917c-GFP (pATB48), Rv1917c (pATB49) and GFP (pATB50) using the B-PER reagent, and the insoluble (pellet) and soluble (supernatant) fractions were fractionated on 4-12% NuPage Bis-Tris SDS-PAGE gels (NOVEX). (A) Silver stained gel; (B) Western blot of proteins transferred onto Immobilon-P and probed with mouse monoclonal anti-GFP (at 1:1000, and goat anti-mouse-HRP at 1:5000). Lanes: 1, pATB48, IS; 2, pATB48, S; 3, pATB49, IS; 4, pATB49, S; 5, pATB50, IS; 6, pATB50, S; where IS = insoluble fraction, and S = soluble fraction.

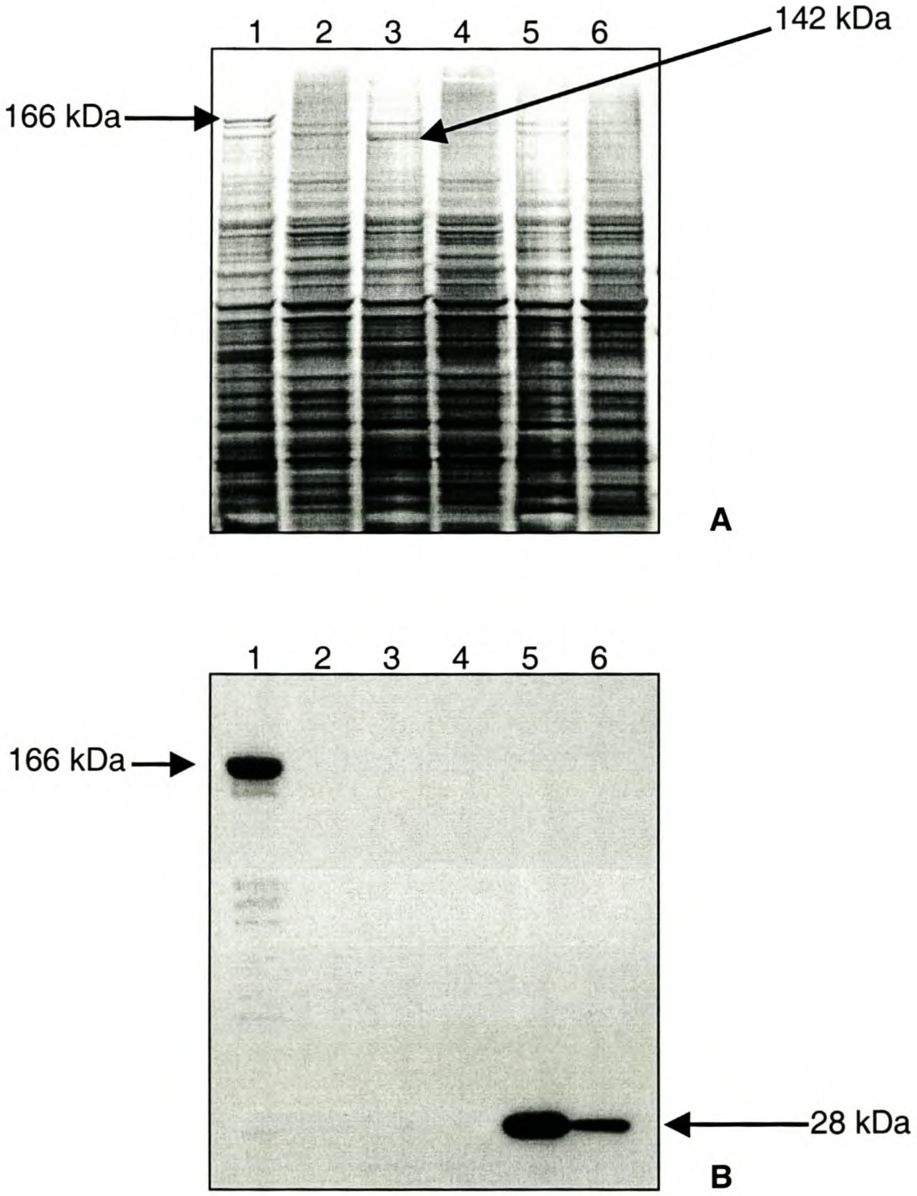
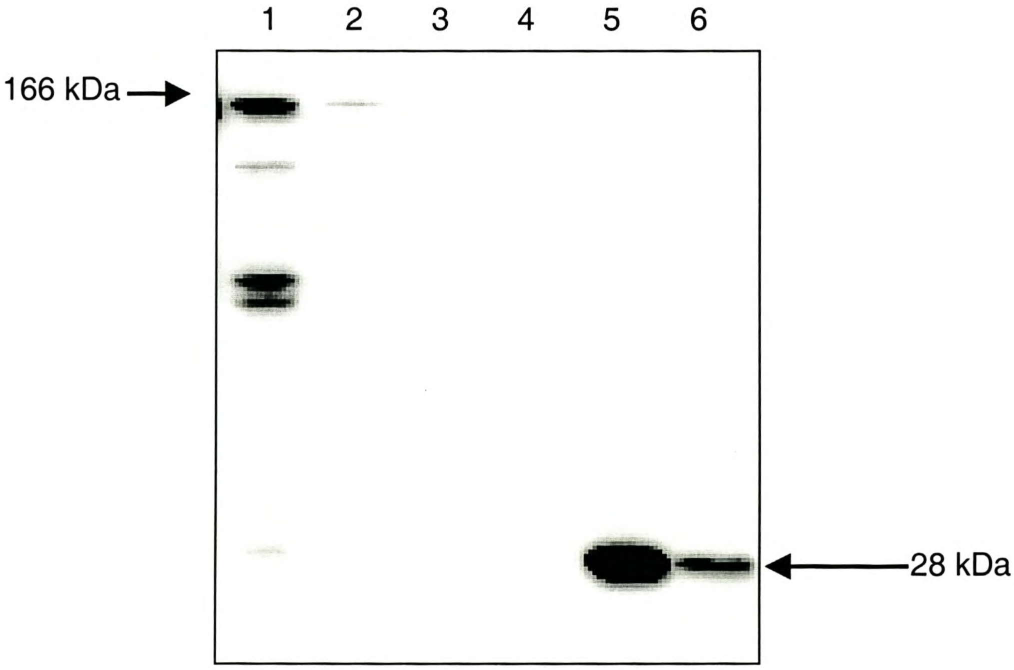


Figure 7.7. Crude protein extracts from *M. bovis* BCG. Proteins extracted from *M. bovis* BCG by ribolysis in SDS lysis buffer were fractionated on a 4-12% NuPage Bis-Tris SDS-PAGE gel (NOVEX). Proteins were transferred onto Immobilon-P and immunoblotted with mouse monoclonal anti-GFP (at 1:1000, and goat anti-mouse-HRP at 1:5000). Lanes: 1, pATB48, S; 2, pATB48, IS; 3, pATB49, S; 4, pATB49, IS; 5, pATB50, S; 6, pATB50, IS; where S = soluble fraction, and IS = insoluble fraction.



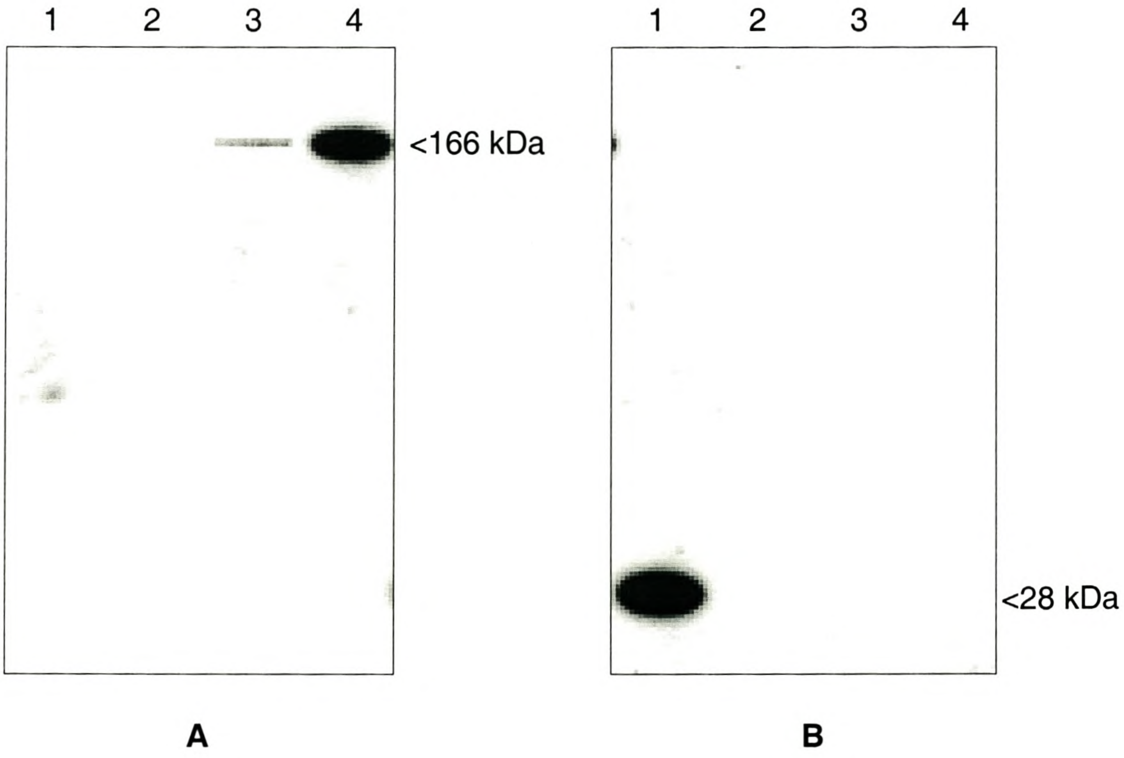
results were obtained from this organism by ribolysis in SDS lysis buffer (Fig. 7.7), as measured by the release of soluble protein. Both GFP and Rv1917c-GFP protein were found in the supernatant fraction indicating solubility in the SDS lysis buffer. As with *M. smegmatis,* immunodetection identified lower molecular weight derivatives of the Rv1917c-GFP protein. The fragment patterns obtained with *M. smegmatis* and *M. bovis* BCG were reproducible between similar extracts, but varied between strains and the lysis buffer employed.

7.3.6 Cellular location of Rv1917c-GFP

To gain insight into the cellular location of the recombinant protein, it was necessary to perform a more refined cell fractionation on *M. smegmatis* cells expressing the Rv1917c-GFP and GFP products. In this procedure, cell envelope material was solubilized in Triton X-114 at 0°C and insoluble material removed by centrifugation. The supernatant is warmed to 37°C at which temperature it partitions into a lower detergent phase containing the most hydrophobic proteins and an upper aqueous phase containing less hydrophobic proteins (Radolf *et al.*, 1988). The Rv1917c-GFP fusion product was mostly insoluble even in Triton X-114 (Fig. 7.8A); however, the small proportion which was soluble partitioned into the detergent phase, suggesting that the protein is strongly hydrophobic. In contrast, GFP fractionated exclusively with the cytosolic fraction (Fig. 7.8B). No Rv1917c-GFP or GFP was detected in the culture supernatant (not shown).

Visual inspection of lysates and visualization of silver-stained lysates on SDS-PAGE gels suggested that lysis of *M. bovis* BCG by ribolysing in Triton X-114 buffer was very inefficient, therefore this method was not suitable for fractionating *M. bovis* BCG extracts.

Figure 7.8. Cell fractionation with Triton X-114. Proteins extracted from *M. smegmatis* expressing (A) Rv1917c-GFP (pATB48) and (B) GFP (pATB50) were fractionated using the detergent Triton X-114, and separated on 4-12% NuPage Bis-Tris SDS-PAGE gels. Proteins were detected with mouse monoclonal anti-GFP (at 1:1000, and goat anti-mouse-HRP at 1:5000). Lanes: 1, cytosolic fraction; 2, aqueous phase from Triton X-114 extraction (hydrophilic); 3, detergent phase from Triton X-114 extraction (hydrophobic); 4, insoluble material.



7.3.7 Glycosylation status of Rv1917c-GFP isolated from *M. smegmatis* and *M. bovis* BCG

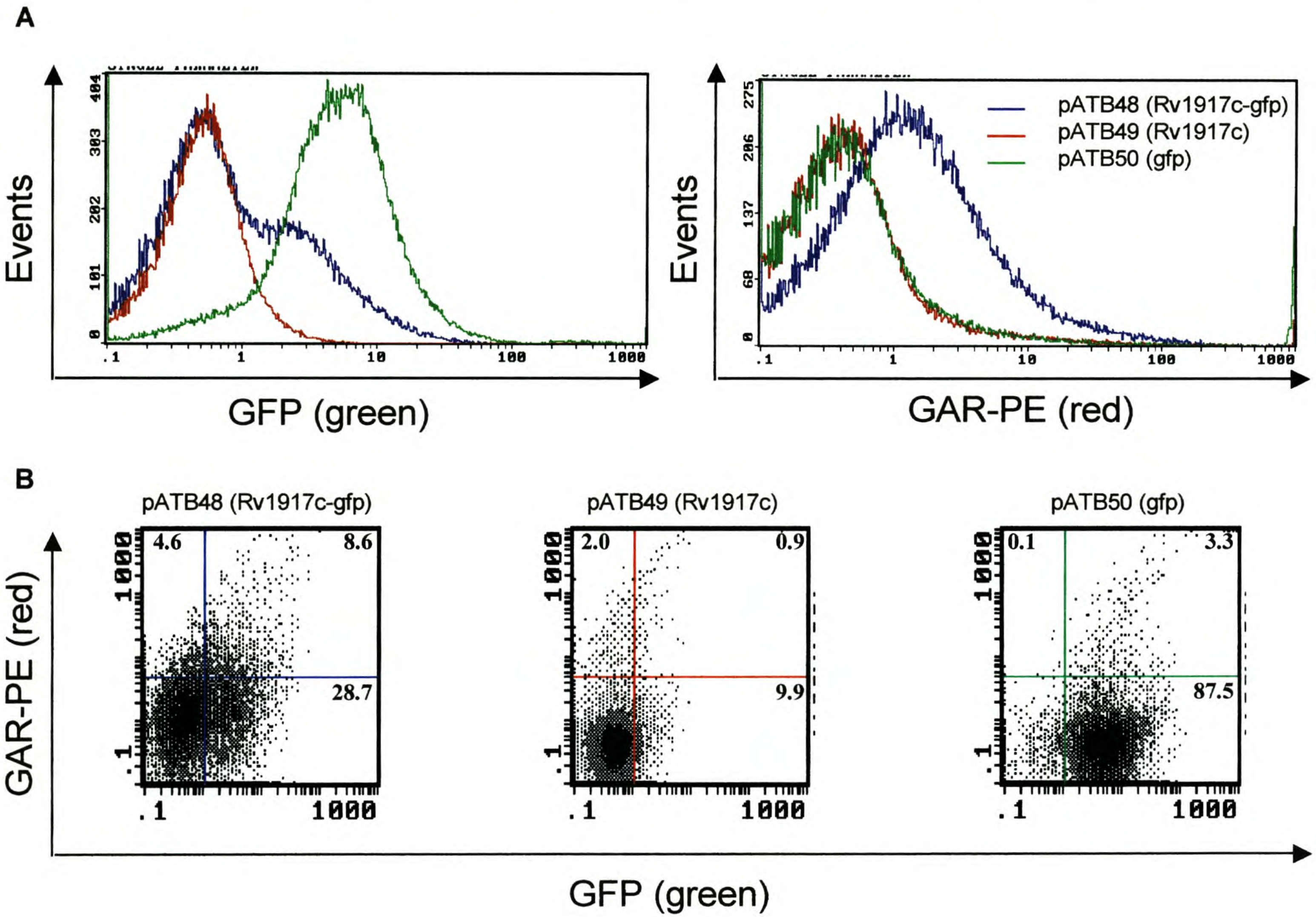
Glycosylation has been shown to modulate the antigenic properties of mycobacterial proteins (Romain *et al*., 1999), and PPE gene products have been speculated to be of immunological importance (Cole *et al*., 1998). It was therefore of interest to determine whether the recombinant Rv1917c protein was glycosylated. Two different methods were used to detect glycans on recombinant Rv1917c-GFP protein expressed in both *M. smegmatis* and *M. bovis* BCG, namely staining with the Roche Glycan/Protein Double labelling kit and detection with the ConA lectin (Herrman *et al*., 1996). Both methods failed to demonstrate glycosylation of the recombinant proteins, although a number of other host-derived glycoproteins were detected (not shown). To ensure that the negative result was not due to insufficient protein or interference from cell envelope lipids, the staining experiments were repeated on recombinant protein purified by immunoprecipitation. Despite significant purification of the recombinant protein, positive glycan staining was not detected (not shown).

7.3.8 Anti-GFP binding to whole cells expressing Rv1917c, Rv1917c-GFP and GFP

Cell fractionation data suggested that the recombinant protein may be associated with the hydrophobic cell wall. To determine whether the recombinant Rv1917c protein was exposed on the surface of mycobacterial cells, whole cell antibody binding experiments were perfomed. Cells expressing Rv1917c-GFP specifically and reproducibly bound polyclonal anti-GFP, as evidenced by flow cytometric analysis (Fig. 7.9). In this experiment the level of GFP expression was quantified by measuring green fluorescence, and the proportion of cells expressing GFP on the cell surface quantified by labelling with anti-GFP antibody followed by GAR-PE detection (red fluorescence). GFP was expressed in 37.3% of *M. smegmatis* containing pATB48 (Rv1917c-*gfp*) and 8.6% of bacilli also expressed GFP on the surface. In contrast, GFP was expressed in 90.8% of bacilli containing pATB50 (*gfp*), however surface expression was equivalent to that of the control containing pATB49 (Rv1917c). In a subsequent experiment very similar results were obtained, with 15% of pATB48-containing bacilli demonstrating surface expression of GFP (data not shown).

Figure 7.9. Flow cytometric analysis of antibody binding to intact *M. smegmatis* cells. (A) One parameter histograms of green fluorescence (GFP) and red fluorescence (GAR-PE) are shown. *M. smegmatis* containing pATB49 (Rv1917c) is the negative control (red line), pATB48 (Rv1917c-GFP fusion) has detectable levels of GFP in a subpopulation of bacilli (blue line), whilst pATB50 (GFP alone) has the majority of the bacilli expressing higher levels of GFP (green line). Analysis of GFP on the surface is shown by labelling intact bacilli with a rabbit polyclonal antibody to GFP and subsequent detection by goat anti-rabbit immunoglobulin conjugated to phycoerythrin (GAR-PE). GFP is not present on the surface of *M. smegmatis* containing pATB49 or pATB50 (red and green lines overlay) but is present on bacilli containing pATB48 (blue line has a significant shift to the right). (B) Two parameter histograms of the same data are shown. The negative control (pATB49) has background levels of green and red fluorescence. GFP is expressed in 37.3% of *M. smegmatis* containing pATB48 and 8.6% of bacilli also express GFP on the surface. GFP is expressed in 90.8% of bacilli containing pATB50, however surface expression is equivalent to that of the control (pATB49).

Figure 7.9 Flow cytometric analysis of antibody binding to intact *M. smegmatis* cells (legend opposite)



7.4 Discussion

The PPE gene family constitutes several percent of the *M. tuberculosis* genome, and the putative protein products have been speculated to be of immunological importance (Cole *et al.*, 1998). Emerging data supports a potentially important role for this gene family (Rodriguez *et al.*, 1999; Camacho *et al.*, 1999; Dillon *et al.*, 1999; Skeiky *et al.*, 2000), however, knowledge of its basic biology is limited. In this study we have demonstrated that the polymorphic Rv1917c PPE gene is expressed in liquid cultures of *M. tuberculosis* H37Rv and that this gene encodes a protein of the predicted size when cloned and expressed in *M. smegmatis* or *M. bovis* BCG. Furthermore, we have presented evidence that the recombinant Rv1917c protein is potentially cell wall associated and at least partially exposed on the cell surface.

Analysis of the nucleotide sequence of Rv1917c demonstrated that in addition to stretches of short (15 bp) major polymorphic tandem repeats (MPTRs), the gene also contains 3 stretches of longer (75 bp, 69 bp and 69 bp) tandem repeats. Previous investigation had suggested that the repeat regions were polymorphic (see Chapter 5; Zhang and Young, 1994; O'Brien *et al.*, 2000). In this study, we demonstrated that expansion / contraction of the tandem repeat regions is responsible for the observed polymorphisms. While the functional implication of this finding is unclear, it is interesting to note the occurrence of similar tandem repeat variation in surface antigens of other pathogens (Jones *et al.*, 1988; Anderson *et al.*, 1990; Lysnyansky *et al.*, 1999; Lachenauer *et al.*, 2000). The observed polymorphisms may add support to the hypothesis that PPE proteins provide the pathogen with a source of antigenic variation, although this requires further investigation.

It has long been suggested that PPE proteins may be cell-envelope associated (Doran *et al.*, 1992) yet little evidence is available to support this claim. Bioinformatic analyses of

PPE genes is of limited value and is not conclusively predictive of transmembrane regions. For instance, analysis of Rv1917c using five different transmembrane prediction programs gave five different answers. This may indicate that the protein resides elsewhere than in a classical membrane bilayer. To investigate this hypothesis, physical analysis of the location of the expressed Rv1917c and Rv1917c-GFP fusion proteins in *M. smegmatis* was undertaken using crude B-PER cell fractionation, and also more refined fractionation with the detergent Triton X-114. Neither the Rv1917c nor the Rv1917c-GFP fusion products were detected in the B-PER supernatant fraction. In contrast, the GFP product was highly soluble in that fraction. This result is corroborated by Triton X-114 fractionation, in which the GFP product was detected only in the cytosolic fraction, whilst the recombinant Rv1917c-GFP product was associated with the hydrophobic fraction. Despite prolonged incubation and vigorous mixing, only a small amount of the fusion product was extractable in Triton X-114, and this partitioned with the detergent phase, indicating that the protein is strongly hydrophobic. It is unclear why the majority of the protein is insoluble in Triton X-114. It may be that with this particular expression system much of the protein is expressed in inclusion bodies. Alternatively, the protein may be integrally associated with the waxy mycolic acid-containing outer membrane characteristic of mycobacteria and hence resistant to detergent extraction.

Surface exposure of antigenic proteins is a common theme amongst many pathogens (Jones *et al*., 1988; Anderson *et al*., 1990; Lysnyansky *et al*., 1999; Lachenauer *et al*., 2000). To explore the possibility that the recombinant Rv1917c-GFP protein may be surface-exposed, whole cell antibody binding experiments were performed. Binding of polyclonal anti-GFP antibodies to intact *M. smegmatis* cells expressing Rv1917c, Rv1917c-GFP and GFP demonstrates specific binding of anti-GFP to cells expressing the Rv1917c-GFP product. This indicates that at least the GFP moiety of the recombinant protein is exposed on the mycobacterial cell surface and supports the hypothesis that Rv1917c is a cell wall associated protein. Only a relatively small proportion of cells expressing the Rv1917c-GFP fusion demonstrated anti-GFP binding which may indicate masking of GFP epitopes by cell surface structures or the large PPE peptide itself. This may also reflect the expression of a proportion of the fusion protein as inclusion bodies.

The Rv1917c, Rv1917c-GFP and GFP products were also successfully expressed in *M. bovis* BCG though further analyses were hampered by low yields of protein obtained after cell disruption. The value of an inducible expression system is highlighted by the difficulties experienced in trying to express the Rv1917c-*gfp* construct using the constitutive *hsp60* promoter. It is likely that the acetamide inducible system will be useful for expressing other recombinant genes that are similarly poorly tolerated when overexpressed.

Glycosylation has been suggested to modulate the antigenic properties of mycobacterial proteins (Romain *et al.*, 1999). As PPE genes are suspected to be of immunological importance (Cole *et al.*, 1998; Dillon *et al.*, 1999; Skeiky *et al.*, 2000), it was therefore of interest to determine the glycosylation status of the Rv1917c protein. Although numerous putative O-glycosylation sites were identified within the Rv1917c protein, it was necessary to test these predictions in laboratory experiments. Two different methods were used to look for glycosylation of the Rv1917c-GFP fusion product expressed in *M. smegmatis* and *M. bovis* BCG. In each case the results obtained were negative, even when the protein was purified by immunoprecipitation. It is acknowledged that the negative results obtained do not rule out the possibility that the native Rv1917c protein is glycosylated when expressed in *M. tuberculosis*.

In conclusion, these results demonstrate that the Rv1917c PPE gene is polymorphic with respect to variable numbers of tandem repeats and is expressed in *M. tuberculosis*.

Expression of a recombinant Rv1917c protein in *M. smegmatis* suggests that it is likely to be both cell wall associated and surface exposed, although this has yet to be confirmed for the native protein in *M. tuberculosis*. These findings are consistent with the hypothesis that Rv1917c, and perhaps PPE proteins more generally, interact with the host immune system. Continued investigation of these intriguing genes is warranted.

CHAPTER 8

DISCUSSION, RECOMMENDATIONS AND CONCLUSIONS

8.1 *Mycobacterium tuberculosis* demonstrates extensive genetic variation

This study has contributed to a growing body of knowledge which demonstrates that genetic diversity in *Mycobacterium tuberculosis* can arise by a number of mechanisms. These include IS*6110* transposition, which can lead to disruption of genes and possibly also regulatory regions (Sampson *et al.*, 1999; Warren *et al.*, 2000; Beggs *et al.*, 2000; Sampson *et al.*, 2001). There is also evidence supporting IS*6110*-mediated genomic deletions via recombination between adjacent copies of the element (Fang *et al.*, 1999a; Chapter 4). These events are reflected in the results of various Southern hybridization and PCR-based strain-typing techniques (which exploit the presence of repetitive DNA within the *M. tuberculosis* genome), which have demonstrated a large number of strain types. The most widely-used strain-typing method, which utilizes the transposable element IS*6110*, has demonstrated extensive genetic diversity among clinical isolates of *M. tuberculosis* (Hermans *et al.*, 1990; Thierry *et al.*, 1990b; Cave *et al.*, 1991; Van Soolingen *et al.*, 1991; Mazurek *et al.*, 1991; Chevrel-Dellagi *et al.*, 1993; Yang *et al.*, 1994; Goyal *et al.*, 1994a; Huh *et al.*, 1995; Warren *et al.*, 1999; O'Brien *et al.*, 1997; Yang *et al.*, 1998).

Genetic variation can also be mediated by IS-independent mechanisms. For example, the presence of other classes of repetitive DNA can give rise to genetic variability, by promoting expansion and contraction of repeat regions via slipped-strand mispairing (SSM) and/or homologous recombination (Levinson and Gutman, 1987; Lloyd and Low, 1996). This type of variability has been reported in *M. tuberculosis* in numerous studies which have identified extensive diversity associated with the direct repeat (DR), polymorphic GC-rich sequence (PGRS), major polymorphic tandem repeat (MPTR), variable number of tandem repeat (VNTR) and multiple interspersed tandem repeat (MIRU) loci (Goyal *et al.*, 1997; Ross *et al.*, 1992; Hermans *et al.*, 1992; Filliol *et al.*, 2000b; Supply *et al.*, 1997). The work reported in this study further demonstrates that a number of mechanisms, including IS*6110* transpositions, deletions, point mutations and long tandem repeat variation can give rise to the extensive genetic diversity associated with the PPE gene family (Chapter 5).

These results contrast with early opinions regarding *M. tuberculosis* strain variation. For instance, the low discriminatory power of traditional *Mycobacterium tuberculosis* strain typing methods, such as phage typing (Rado *et al.*, 1975; Jones and Greenberg, 1978), serotyping (Jones and Kubica, 1968), biotyping (Román and Sicilia, 1984) and antibiotic susceptibility profiling (Gruft *et al.*, 1984), gave rise to the opinion that there was little heterogeneity amongst *M. tuberculosis* strains. This view has, to some extent, been perpetuated even with the advent of more sensitive molecular techniques (Frothingham *et al.*, 1994; Kapur *et al.*, 1994). A frequently-quoted study utilized DNA sequencing to demonstrate very limited sequence diversity in structural genes (Sreevatsan *et al.*, 1997). Similar results were obtained from a subsequent analysis of presumed antigenic proteins (Musser *et al.*, 2000). As the spontaneous mutation frequency for *M. tuberculosis* is similar to that observed in other bacteria (David and Newman, 1971), it has been hypothesized that the limited allelic diversity observed in *M. tuberculosis* indicates that the organism is relatively young in evolutionary terms (Kapur *et al.*, 1994; Sreevatsan *et al.*, 1997; Musser *et al.*, 2000).

There is therefore a striking contrast between the presumed conservation of the genome on the one hand, and extensive DNA fingerprint diversity on the other. While it may be true that the *M. tuberculosis* genome is relatively homogeneous at the single nucleotide level (particularly in ORFs), it is misleading to interpret this as an accurate reflection of the full extent of genetic variation. Emerging data (Chapters 2 to 5; Sampson *et al.*, 1999; Warren *et al.*, 2000) supports the view that the *M. tuberculosis* genome is a dynamic entity, with the capacity to generate genetic diversity by several mechanisms, besides point mutations. If we accept that the *M. tuberculosis* genome is relatively conserved at the single nucleotide level (Sreevatsan *et al.*, 1997; Musser *et al.*, 2000), then variation associated with repetitive DNA (either mobile genetic elements or simple repetitive DNA) represents an important mechanism for bacterial evolution and adaptation to changing environments. This has implications for vaccine development and drug design strategies, which should take into account the capacity of the pathogen to vary its genetic structure. This study explores the relationship between evolutionary mechanisms and the impact of genomic change on the phenotype of clinical isolates.

8.2 IS*6110* is an agent of *M. tuberculosis* genome evolution

8.2.1 IS*6110*: Chromosomal distribution and evolutionary mechanisms

At the inception of this study, little was known about the chromosomal location of copies of the IS*6110* element in *M. tuberculosis*. Results from this study have clearly demonstrated that the majority of IS*6110* insertions investigated, occurred within predicted coding regions (Chapters 2-4; Sampson *et al.*, 1999; Warren *et al.*, 2000; Sampson *et al.*, 2001). This finding is based on a substantially larger number of insertion sites than that of an earlier report (Philipp *et al.*, 1996), and is likely to be a more accurate reflection of the capacity of the element to disrupt coding regions. This underscores the role of IS*6110* transposition as a potentially important evolutionary mechanism for the pathogen, and may explain why the element has been maintained throughout the evolutionary history of the pathogen.

The collation of IS*6110* insertion site data has also permitted observations regarding the chromosomal distribution of the element. To some extent, the results support earlier findings that certain regions of the genome have few or no IS*6110* insertions (Philipp *et al.*, 1996). Specifically, the chromosomal domain flanking the origin of replication is devoid of insertions (Philipp *et al.*, 1996), which may suggest that this region fulfils an essential function. A similar distribution bias was observed in a study of all insertion sequences (in addition to IS*6110*) found in *M. tuberculosis* H37Rv (Gordon *et al.*, 1999b), and was also revealed in a transposon mutagenesis study using the transposon Tn*5367* (Bardarov *et al.*, 1997). However, continued surveillance and collection of IS*6110* insertion site data from clinical isolates of *M. tuberculosis* is advocated, as this may yet identify insertions within this region.

In contrast to the absence of insertions in certain chromosomal regions, the mapping of IS*6110* integration sites reveals that there are a number of IS*6110* preferential integration loci within the genome (Chapters 2 - 4; Sampson *et al.*, 1999; Warren *et al.*, 2000; Sampson *et al.*, 2001). The phenomenon of preferential integration, defined here as multiple independent insertions (within a few hundred base pairs of each other, or into identical positions) into particular areas of the genome, was originally reported by Fang *et al.* (1997), and has since been observed by other investigators (Kurepina *et al.*, 1998; Ho *et al.*, 2000), and ourselves (Sampson *et al.*, 1999; Warren *et al.*, 2000; Sampson *et al.*, 2001). It is unknown whether the presence of preferential integration loci merely reflects tolerance of insertion into these areas, or whether this represents an adaptive strategy on the part of the pathogen. For example, it may represent a selective advantage for the organism to maintain multiple copies of the element, which can promote genetic variation, and thereby enhance its ability to adapt to changing environmental conditions. Clearly, this is not the only adaptive mechanism available to the pathogen, as evidenced by the continued proliferation of strains with no or low IS*6110* copy number (Yuen *et al.*, 1993). However, low copy number and high copy number strains are believed to have arisen as separate lineages (Fomukong *et al.*, 1998), which might suggest that they have evolved distinct mechanisms of adaptation and evolution.

Interestingly, the findings presented here suggest a possible trend with regards to preferential integration loci, namely that these may be "deletion hotspots", i.e. regions of the genome that are prone to IS*6110*-mediated deletion events. Recently, a group of investigators described the occurrence of "deletion hypervariability" within a 20 kb region of the genome (Ho *et al.*, 2000), which was previously identified as a preferential integration region (Chapter 2; Sampson *et al.*, 1999). In this study, detailed analysis of the DR region of the chromosome and its flanking sequence provides another example of a deletion-prone preferential integration site (Chapter 4). Mapping of chromosomal deletions identified in this study

(Chapter 3) demonstrates that the majority of these are associated with preferential integration regions. Furthermore, superimposition of the RvD regions, (deleted from *M. tuberculosis* H37Rv relative to *M. bovis* (Gordon *et al*., 1999a)), demonstrates that these are all associated with preferential integration loci (Fig. 2.3).

It is therefore evident that in addition to genome variation mediated directly by transposition events, the IS*6110* element can also contribute significantly to genome variation by promoting recombination events leading to deletion of genomic domains. Numerous other *M. tuberculosis* insertion sequences have been identified (Collins and Stephens, 1991; Fang *et al*., 1999b; Gordon *et al*., 1999b), which may further contribute to overall genome diversification mediated by transposition and recombination events. This is consistent with a study of genome evolution in *E. coli*, where genome variation was predominantly ascribed to transpositions, deletions, and other types of chromosomal rearrangement, frequently associated with insertion sequences, rather than point mutations (Papadopoulos *et al*., 1999).

8.2.2 Phenotypic aspects of IS*6110*-driven genome diversity

While a number of examples of phenotypic change mediated by insertion sequences have been described in other organisms (Collins and Gutman, 1992; Podglajen *et al*., 1994; Hammerschmidt *et al*., 1996; Hubner and Hendrickson, 1997; Hall *et al*., 1998), such an effect remains to be shown for IS*6110*. The potential phenotypic implications of IS*6110*-mediated gene disruption are discussed in detail in chapter 2. Briefly, factors such as gene function and genetic redundancy are expected to influence the impact of a particular IS*6110* transposition event or IS*6110*-mediated deletion on strain phenotype.

An added level of complexity is introduced if one considers a recent study, which demonstrated that despite disruption of the *M. tuberculosis ctpD* gene by IS*6110*, mRNA for that gene is still detectable by RT-PCR (Beggs *et al*., 2000). This may suggest that IS*6110* is able to provide a promoter for gene transcription (Beggs *et al*., 2000). This finding has two important implications. Firstly, an IS*6110*-mediated "gene knockout" may still encode a functional product, depending on the relative position of the insertion sequence within the gene. Alternatively, an intergenic insertion may activate a previously silent gene. These possibilities paint an increasingly complex picture, and the precise impact of specific insertions remains to be empirically determined.

This study has clearly demonstrated the capacity of IS*6110* to create "natural gene knockouts" and "natural deletion mutants" with potential phenotypic consequences for the organism. If one views the study population from which these mutants were isolated as a "field experiment", it might be argued that the IS*6110*-associated mutations found are not appreciably disadvantageous to the organism, as the strains were all isolated from patients with active disease. However, this viewpoint does not take into account the potentially subtle and cumulative effects of these mutations. In this regard, it is interesting to note the findings of a recent study, which utilized high density oligonucleotide arrays to map deletions in clinical isolates of *M. tuberculosis* (Kato-Maeda *et al*., 2001). The results were related with radiographical findings, and a correlation between increased deletion and reduced pulmonary cavitation was identified (Kato-Maeda *et al*., 2001). In a similar vein, progressive attenuation of BCG strains was correlated with accumulation of deletions identified by microarray technology (Behr *et al*., 1999). While neither of these studies demonstrate a causal relationship, the approaches suggest a relevant and feasible design for future studies of the consequences of defined mutations for strain fitness.

8.2.3 Implications for molecular epidemiology

One indicator of strain fitness is the capability of the organism to be transmitted and cause disease. Accurate measurement of disease transmission requires reliable tools to trace the movement of particular strains through well-defined study communities. IS*6110* DNA fingerprinting is the most commonly used method to measure disease transmission, but some concerns regarding the accurate interpretation of this data exist. In this regard, the findings of this study hold potentially important implications for the accurate interpretation of IS*6110* DNA fingerprint data, and for the molecular epidemiology of *M. tuberculosis*.

The demonstration that preferential integration regions are common (Chapters 2-4) raises the issue of convergent evolution of IS*6110* DNA fingerprints. It has been suggested elsewhere that the presence of IS*6110* preferential integration loci might lead to an over-estimation of clustering (and therefore an over-estimation of transmission) as defined by IS*6110* DNA fingerprinting (McHugh and Gillespie, 1998; Gillespie *et al.*, 2000). However, as reported in this study, insertion points within a defined preferential integration region are often sufficiently offset to be detectable as unique insertions within the resolution of standard fingerprinting techniques (Chapters 2-4). Furthermore, the chromosomal domains flanking IS*6110* frequently undergo mutational events (Chapters 2-4), which can lead to changes in restriction fragment lengths, thereby countering this effect. Therefore, it is unlikely that two high IS*6110* copy number strains with different origins will convergently evolve to have the same IS*6110* banding pattern.

An important question that has received some attention in the recent literature is that of the rate of change of IS*6110* DNA fingerprints. Despite the innate ability of IS*6110* to transpose (Fomukong and Dale, 1993; Wall *et al.*, 1999), it is considered to be sufficiently stable for use as a molecular probe. Early studies included *in vitro* (Van Soolingen *et al.*, 1991) and *in vivo* (Hermans *et al.*, 1990) serial passage experiments, as well as studies of isolates from relapsed patients, persistent emitters, multiple sites of infection and drug resistant isolates (Van Soolingen *et al.*, 1991; Otal *et al.*, 1991; Godfrey-Fausset *et al.*, 1993; Chevrel-Dellagi *et al.*, 1993; Cave *et al.*, 1994). In conjunction, these results were assumed to indicate sufficient stability of the element for accurate application as a molecular probe, and strains with IS*6110* DNA fingerprints with one or two band changes are therefore not classified as part of a chain of recent transmission. However, conflicting results have been presented, which suggest a greater rate of change than previously assumed (Yeh *et al.*, 1998; Alito *et al.*, 1999), and these authors advise the inclusion of strains with one or two band differences in chains of recent transmission.

The controversy surrounding the stability of IS*6110* RFLPs has lead to attempts to define the “molecular clock” of IS*6110* (De Boer *et al.*, 1999). Based on fingerprint changes observed in serial isolates, the authors suggest a half-life for IS*6110* RFLPs of 3.2 years (De Boer *et al.*, 1999). However, this result should be treated with some caution, and cannot be universally applied. Firstly, it is based on a specialized sub-population of tuberculosis patients. Secondly, as Niemann and co-workers have pointed out, assumptions of stability based on examination of serial isolates from a single patient are not necessarily valid for patient to patient transmission (Niemann *et al.*, 2000). Recently it was demonstrated that states of low oxygen tension might promote IS*6110* transposition (Ghanekar *et al.*, 1999). In physiological terms, this might relate to the granuloma micro-environment which may vary between patients and with disease progression. Hence, the "molecular clock" may run faster in some patients than in others. Thirdly, the results from a low incidence community cannot necessarily be extrapolated to another geographical area, particularly high incidence communities. As reported in this study (Chapter 3; Warren *et al.*, 2000), there is a moderately significant correlation between strain family frequency and increased number of mutations, which might indicate that certain strains in high incidence communities are more prone to genetic variation. Finally, the use of the term "half-life" is problematical, as it suggests that an entire database of fingerprints will have a half-life of 3.2 years. This is most unlikely, as strains commonly identified at the onset of our collection (in 1993) are still commonly isolated from new patients today, almost 8 years later (Warren *et al.*, 1996a; Warren *et al.*, 1999).

In addition to patient-related factors, there are also strain-specific factors which may influence *M. tuberculosis* genetic variation. In agreement with results presented in this study (Chapter 3; Warren *et al.*, 2000), it was recently conceded that IS*6110* stability could vary from strain to strain (Niemann *et al.*, 2000). This is supported by recent results from this laboratory, which demonstrate variation in rates of change in 2 clonal groups within a single strain family (Warren *et al.*, 2001). A possible explanation for strain to strain variation in IS*6110* transposition rates is provided by a study which suggests that IS*6110* transposition could be stimulated by external promoters flanking the element (Wall *et al.*, 1999). The very low variation observed among isolates with low IS*6110* copy numbers could therefore be a function of these strains being "frozen" in terms of IS*6110* transposition capability (Wall *et al.*, 1999). Conversely, this may also imply that once transposition into an "active" locus has occurred, a particular strain may be more prone to variation due to increased transposition (Wall *et al.*, 1999). The relative arrangement of IS*6110* elements could also promote deletion hypervariability, as discussed above (and in Chapter 4). These mechanisms may give rise to highly mutable strains, better able to adapt in response to environmental challenges. Such strains have not yet been identified, but may well exist, as found in other bacterial species (Sniegowski *et al.*, 1997).

The rate and biological significance of IS*6110*-associated genome diversification, and the implications for molecular epidemiology are concepts that are narrowly intertwined. To progress towards an understanding of the biological significance of strain variation, it is essential for future studies to incorporate the valuable data to be gained from well-defined study communities. In practical terms, it is envisaged that the correlation of genetic variants with epidemiological and clinical parameters will contribute to an understanding of the functional implications of strain diversity.

8.3 The PPE gene family and strain variation

In addition to the contribution of insertion sequences to the genetic heterogeneity of clinical isolates of *M. tuberculosis* and other pathogens, there are alternative mechanisms whereby strain variation may arise. Genetic variation associated with repetitive DNA and multi-gene families in other bacteria frequently has functional consequences, in terms of antigenic properties and pathogenic characteristics (Robertson and Meyer, 1992; Dybvig, 1993; Moxon *et al.*, 1994; Van Belkum *et al.*, 1998; Metzgar and Wills, 2000). Examples include expansion or contraction of repeat domains which can alter the expression and antigenic properties of surface molecules, and the programmed switching of expression of large, multi-gene families which provides an effective means of immune evasion (Robertson and Meyer, 1992; Dybvig, 1993).

To initiate work in this area in *M. tuberculosis*, this study has focused on variation associated with the PPE gene family of *M. tuberculosis*. This large family of genes occupies approximately 4% of the coding capacity of the genome, and it has been suggested that the encoded proteins may provide the pathogen with a source of antigenic variation (Cole *et al.*, 1998). This study provides support for this hypothesis, as it demonstrates that there is extensive sequence variation associated with representative members of the gene family. To assess the potential significance of this variation, various aspects of the gene family have been investigated.

Intuitively, the maintenance of large numbers of PE and PPE genes within the genome of members of the *M. tuberculosis* family suggests that these gene families fulfil an important function. Alternatively, it might be argued that this is simply a reflection of recent expansion from common ancestral genes, with no functional significance. However, PE and PPE gene homologues have been identified in other mycobacteria, including *Mycobacterium leprae* and the distantly related *Mycobacterium smegmatis* (Poulet and Cole, 1995; Hermans *et al.*, 1992; Gey van Pittius, in press) It is therefore likely that the progenitor of the gene family arose in the distant evolutionary past, prior to the divergence of these species, which may imply that there has been sufficient time for different members of this gene family to have evolved specialized functional roles. Intriguingly, analysis of PPE genes in the *M. leprae* genome, which has undergone substantial reductive evolution (Cole *et al.*, 2001), identifies members of each of 4 phylogenetically defined subgroups (Chapter 5). Although the majority of these are pseudogenes, 5 intact PPE genes are maintained within the much-reduced genome of this mycobacterium, hinting at a possible functional significance. Although this could simply reflect ongoing genome downsizing, the 5 intact PPE genes may be associated with the specialized niche occupied by *M. leprae*. Further investigation will be necessary to fully understand the evolutionary history and importance of the gene family.

Recent literature reflects an inclination to view members of the PPE gene family as a homogeneous group, and to extrapolate observations from a limited subset of the gene family to all its members (Musser *et al.*, 2000). However, it may be incorrect to assume that members of the gene family all fulfil identical or even similar functions. A phylogenetic tree demonstrates that the gene family is composed of a number of distinct subgroups, with unique sequence characteristics (Chapter 5). While the biological significance of the different groupings has yet to be established, they may point to functional specialization of the PPE subgroups. The products of various subgroups within the gene family may therefore fulfill diverse functional roles. In support of this hypothesis, this study demonstrates noteworthy differences in terms of sequence characteristics, the extent of genetic variation and expression profiles (both *in vitro* and *in vivo*) among different members of the gene family (Chapters 5 and 6).

This study utilized a hybridization-based approach to gain a broad overview of genetic variation among different members PPE gene family. DNA hybridization probes were based on members of the different PPE subgroups (as defined by the phylogenetic analysis), and this approach has highlighted different degrees of genetic variation associated with different subgroups. Some genes demonstrate relative conservation across all strains analyzed. This may indicate that there is selective pressure to maintain intact copies of particular genes, associated with their specific functional niche. In this regard, it is interesting to note a recent report which suggested that the Rv2123 gene (subgroup 4) may encode a siderophore involved in iron uptake, based on its up-regulation under low iron conditions (Rodriguez *et al.*, 1999). If this hypothesis is proved correct, it might suggest that it is advantageous for the bacteria to maintain a functionally intact copy of the gene, and intriguingly this study demonstrates that this gene shows little variation in clinical isolates (Chapter 5).

In contrast to the relative stability observed in some PPE genes (Chapter 5, Musser *et al.*, 2000), other members of the gene family demonstrate substantial variation. This variation can arise by a number of mechanisms, including point mutation, deletion, IS*6110* transposition and expansion/contraction of repetitive domains (Chapters 2, 3 and 5). A recurring theme in many pathogens is hypervariability associated with genes encoding surface-exposed, or other proteins that interact directly with the host immune system (Jones *et al.*, 1988; Anderson *et al.*, 1990; Lysnyansky *et al.*, 1999; Lachenauer *et al.*, 2000). These highly mutable "contingency loci" are regions that allow the pathogen to rapidly adapt to unfavorable environmental conditions by facilitating the generation of antigenic diversity (Moxon *et al.*, 1994). The observed variation in *M. tuberculosis* PPE genes may therefore represent a mechanism whereby the pathogen is able to adapt to, and subvert host immune responses.

An alternative explanation for the observed variation is that it is merely due to tolerance of genetic diversification as a consequence of genetic redundancy, and that it reflects a lack of selective pressure to maintain intact copies of these genes rather than an adaptive mechanism. However, in this respect, it is interesting to note the finding that a transposon mutant of the PPE gene Rv3018c is attenuated for growth in a murine model (Camacho *et al.*, 1999). Although the possibility of functional redundancy among the many members of the PPE gene family cannot be ruled out, this finding cautions the assumption of redundancy based on nucleotide sequence alone.

An intriguing form of PPE gene variation is that associated with tandem repeats (Chapter 5, Zhang and Young, 1994; O'Brien *et al.*, 2000). Studies of other pathogens have described numerous examples of antigenic proteins which contain tandem repeat motifs, and many of these studies further demonstrate that tandem repeat variation can directly impact on antigenic and opsonic properties of surface-exposed molecules (Jones *et al.*, 1988; Zheng *et al.*, 1995; Gravekamp *et al.*, 1998). A recent study of *M. tuberculosis* demonstrated that 3 serologically immunodominant proteins that are expressed early in infection (Singh *et al.*, 2001) contain tandem repetitive domains, which might suggest a role for the repeat domains in eliciting an immune response. One of these proteins is a member of the PE-PGRS subgroup of the PE family, and there is recent evidence to suggest that it is specifically the PGRS repeat regions of PE-PGRS proteins that elicit an antibody response (Delogu and Brennan, 2001). Intriguingly, PGRS domains are highly variable in clinical isolates of *M. tuberculosis*, which is consistent with the hypothesis that these genes may be involved in antigenic variation.

The PGRS repeats of the PE gene family and the MPTR repeats of the PPE gene family are speculated to have a common origin, on the basis of sequence similarity (Poulet and Cole, 1995). However, in contrast to the observed instability of PGRS-containing domains, the MPTR repeats that are associated with the PPE gene family demonstrate little genetic variation in clinical isolates of *M. tuberculosis* (Hermans *et al.*, 1992). This may suggest different functional roles for the two types of repeats. PGRS repeats are known to encode antigenic epitopes (Delogu and Brennan, 2001), therefore their variation may be driven by host immune responses. The limited variation of the MPTR repeats may indicate that the domains encoded by these repeats are not "seen" by the host immune system. Alternatively, their conservation may be necessary for their functional role, and this may represent a selective advantage for the organism.

An interesting feature of 3 specialized members of the PPE-MPTR subgroup (Rv1753c, Rv1917c and Rv1918c) is the presence of regions of long tandem repeats (LTRs), which demonstrate extensive strain-to-strain variation. This variation is mediated by differences in numbers of LTRs, a feature reminiscent of surface-exposed antigens in a number of other pathogens (Jones *et al.*, 1988; Anderson *et al.*, 1990; Lysnyansky *et al.*, 1999; Lachenauer *et al.*, 2000). Intriguingly, investigation of one of the hypervariable genes described above demonstrates that the gene product is cell-wall associated, and surface-exposed. The observed tandem repeat variation may therefore provide the pathogen with a means to alter surface epitopes, and thereby represent an important immune evasion strategy.

It is important to note that, as with other pathogens, sequence-level variation is not necessarily the only adaptive strategy available to *M. tuberculosis*, and variation may also be mediated at the transcriptional and/or translational level (Weiser *et al.*, 1989; Van der Woude *et al.*, 1996; Biegel Carson *et al.*, 2000; Park *et al.*, 2000; Zhang and Wise, 2001). Results from this study demonstrate variable expression of PPE mRNA both *in vitro* and *in vivo*, supporting the possibility that PPE variation may also be mediated at the transcriptional level. Variable expression of one particular PPE gene in human granulomas suggests that PPE genes may be differentially regulated during host infection. Should the proteins encoded by these particular PPE genes prove to be recognized by the host immune system, differential regulation of their expression may provide a further mechanism for immune system evasion. While these results have yet to be confirmed at the protein level and within different hosts, recent studies provide indirect evidence that other members of the PPE gene family demonstrate variable expression *in vivo* (Dillon *et al.*, 1999; Skeiky *et al.*, 2000).

Many questions regarding the PPE gene family remain unresolved. However, results presented in this thesis, in conjunction with newly emerging data, suggest that further investigation of the gene family is warranted. The ability to invoke a T cell response (Dillon *et al.*, 1999; Skeiky *et al.*, 2000), and possible cell-wall association of one or more of its members (Chapter 7) supports a potentially important role in host-pathogen interaction. The combination of sequence diversity and variable expression may provide the pathogen with a vast antigenic repertoire and suggests possible mechanisms for immune evasion. This could have important implications both for vaccine design strategies and tuberculosis control programs. Biological evaluation of the gene family may reveal an important role in disease pathogenesis, but further investigation must not lose sight of the possibility that the different PPE subgroups are functionally distinct.

8.4 Recommendations

- Establish internet-based genome map of collated IS*6110* data, to provide a valuable and accessible reference tool for a number of applications.
- Focus future investigation of IS*6110*-mediated deletion mechanisms on confirmed predecessor / deletion descendant isolates.
- A long-term goal will be to obtain an accurate measure of the IS*6110* "molecular clock" in different strain families, by using a multi-probe approach. Ultimately, this can be included as a component of epidemiological studies
- Correlate clinical and molecular epidemiological parameters with specific genetic mutational events, for example deletions.
- Establish an evolutionary history for the PPE gene family. (A study combining phylogenetic- and hybridization-based approaches to establish the distribution, evolutionary history and relative stability of PPE genes in different mycobacterial species, has been initiated.)
- Correlate PPE gene expression with host cytokine profiles.
- Investigate the antigenic properties of PPE proteins.
- Elucidate factors controlling PPE gene expression.
- Investigate the subcellular location of other representatives of the PPE gene family.

8.4 Conclusions

The results described here have substantially contributed to the understanding of genetic diversification in *Mycobacterium tuberculosis*. The study has combined a clinical base with genomic approaches to challenge the paradigm of genetic homogeneity. A number of molecular mechanisms were identified whereby the pathogen can alter its genetic structure. It was shown that the mobile genetic element IS*6110* plays a significant role in this regard, and the polymorphic PPE gene family was demonstrated to contribute an additional level of genetic diversity. The observed genetic variation may impact on strain phenotype, which could partially explain the observation of different frequencies of strain families within the study community. In addition, these findings have potential implications for future vaccine development initiatives. In conclusion, this study has emphasized the value of applying data from a clinical context to formulate and address relevant biological and epidemiological questions.

REFERENCES

A

Abou-Zeid, C., Garbe, T., Lathigra, R., Wiker, H.G., Harboe, M., Rook, G.A., Young, D.B. (1991) Genetic and immunological analysis of *Mycobacterium tuberculosis* fibronectin-binding proteins. *Infect Immun.* 59:2712-8

Adams, D. (1974) The structure of mononuclear phagocytes differentiating *in vivo*: I. Sequential fine and histologic studies of the effect of bacillus Calmette-Guérin (BCG). *Am J Pathol*. 76:17-48

Albertini, A.M., Hofer, M., Calos, M.P., Miller, J.H. (1982) On the Formation of Spontaneous Deletions: The Importance of Short Sequence Homologies in the Generation of Large Deletions. *Cell*. 29:319-328

Alito, A., Morcillo, N., Scipioni, S., Dolmann, A., Romano, M.I., Cataldi, A., Van Soolingen, D. (1999) The IS*6110* Restriction Fragment Length Polymorphism in Particular Multidrug-Resistant *Mycobacterium tuberculosis* Strains May Evolve Too Fast for Reliable Use in Outbreak Investigation. *J Clin Microbiol.* 37:788-791

Alland, D., Kalkut, G.E., Moss, A.R., McAdam, R.A., Hahn, J.A., Bosworth, W., Drucker, E., Bloom, B.R. (1994) Transmission of Tuberculosis in New York City: An analysis by DNA Fingerprinting and Conventional Epidemiologic Methods. *N Engl J Med*. 330:1710-1716

Allen, B.W. (1999) Mycobacteria. In *Methods in Mol Microbiol, Vol 101: Mycobacteria Protocols*, pp. 15-29. Edited by T. Parish and N. Stoker. Humana Press Inc

Altare, F.A., Durandy, A., Lammas, J.F., Emile, S., Lamhamedi, S., Le Deist, F., Drysdale, P., Jouanguy, E., Doffinger, R., Bernaudin, Jeppsson, O., Gollob, J.A., Meinl, E., Segal, A.W., Fischer, A., Kumararatne, D., Casanova, J.L. (1998) Impairment of mycobacterial immunity in human interleukin-12 receptor deficiency. *Science*. 280:1432-1435

Andersen, P., Askgaard, D., Ljungqvist, L., Bennedsen, J., Heron, I. (1991) Proteins released from *Mycobacterium tuberculosis* during growth. *Infect Immun.* 59:1905-1910

Anderson, B.E., McDonald, G.A., Jones, D.C., Regnery, R.L. (1990) A protective antigen of *Rickettsia rickettsii* has tandemly repeated, near-identical sequences. *Infect Immun.* 58:2760-2769

Azad, A.K., Sirakova, T.D., Rogers, L.M., Kolattukudy, P.E. (1996) Targeted replacement of the mycocerosic acid synthase gene in M*ycobacterium bovis* BCG produces a mutant that lacks mycosides. *Proc Natl Acad Sci USA*. 93:4787-92.

B

Balasubramanian, V., Pavelka, M.S. Jr, Bardarov, S.S., Martin, J., Weisbrod, T.R., McAdam, R.A., Bloom, B.R., Jacobs, W.R. Jr. (1996) Allelic exchange in *Mycobacterium tuberculosis* with long linear recombination substrates. *J Bacteriol*. 178:273-9.

Bandera, A., Gori, A., Catozzi, L., Esposti, A.D., Marchetti, G., Molteni, C., Ferrario, G., Codecasa, L., Penati, V., Matteelli, A., Franzetti, F. (2001) Molecular Epidemiology Study of Exogenous Reinfeciton in an Area with a Low Incidence of Tuberculosis. *J Clin Microbiol*. 39:2213-2218

Bardarov, S., Kriakov, J., Carriere, C., Yu, S., Vaamonde, C., McAdam, R.A., Bloom, B.R., Hatfull, G.F., Jacobs, W.R. (1997) Conditionally replicating mycobacteriophages: a system for transposon delivery to *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA*. 94:10961-10966

Barnes, P.F., Bloch, A.B., Davidson, P.T., Snider, D.E. (1991) Tuberculosis in patients with human immunodeficiency virus infection. *N Engl J Med*. 324:1644-1650

Barnes, P.F., Modlin, R.L., Ellner, J.J. (1994) T-Cell Responses and Cytokines. In *Tuberculosis: Pathogenesis, Protection and Control*. pp. 13-23. Edited by B. Bloom, ASM Press, Washington, DC.

Barnes, P.F., Yang, Z., Preston-Martin, S., Pogoda, J.M., Jones, B.E., Otaya, M., Eisenach, K.D., Knowles, L., Harvey, S., Cave, M.D. (1997) Patterns of tuberculosis transmission in Central Los Angeles. *JAMA*. 278:1159-1163

Bates, J.H., Stead, W.W. (1993) The history of tuberculosis as a global epidemic. *Med Clin North Am.* 77:1205-1217

Baulard, A., Jourdan, C., Mercenier, A. & Locht, C. (1992) Rapid mycobacterial plasmid analysis by electroduction between *Mycobacterium* spp and *Escherichia coli*. *Nucleic Acids Res.* 20:4105

Beggs, M.L., Eisenach, K.D., Cave, M.D. (2000) Mapping of IS*6110* insertion sites in two epidemic strains of *Mycobacterium tuberculosis*. *J Clin Microbiol.* 38:2923-2928.

Behr, M.A., Small, P.M. (1999) A historical and molecular phylogeny of BCG strains. *Vaccine.* 17:915-922

Behr, M.A., Wilson, M.A., Gill, W.P., Salamon, H., Schoolnik, G.K., Rane, S., Small, P.M. (1999) Comparative Genomics of BCG Vaccines by Whole-Genome DNA Microarray. *Science. 284:1520-1523*

Bekker, L.-G., Moreira, A.L., Bergtold, A., Freeman, S., Ryffel, B., Kaplan, G. (2000) Immunopathologic effects of tumor necrosis factor alpha in murine mycobacterial infection are dose dependent. *Infect Immun.* 68:6954-6961

Benjamin, W.H., Lok, K.H., Harris, R., Brook, N., Bond, L., Mulcahy, D., Robinson, N., Pruitt, V., Kirkpatrick, D.P., Kimerling, M.E., Dunlap, N.E. (2001) Identification of a Contaminating *Mycobacterium tuberculosis* Strain with a Transposition of an IS*6110* Element Resulting in an Altered Spoligotype. *J Clin Microbiol.* 39:1092-1096

Biegel Carson, S.D., Stone, B., Beucher, M., Fu, J., Sparling, P.F. (2000) Phase variation of the gonococcal siderophore receptor FetA. *Mol Microbiol* 36:585-593

Bifani, P.J., Plikaytis, B.B., Kapur, V., Stockbauer, K., Pan, X., Lutfey, M.L., Moghazeh, S.L., Eisner, W., Daniel, T.M., Kaplan, M.H., Crawford, J.T., Musser, J.M., Kreiswirth, B.N. (1996) Origin and Interstate Spread of a New York City Multidrug-Resistant *Mycobacterium tuberculosis* Clone Family. *JAMA.* 275:452-457

Bifani, P.J., Mathema, B., Liu, Z., Moghazeh, S.L., Shopsin, B., Tempalski, B., Driscoll, J., Frothingham, R., Musser, J.M., Alcabes, P., Kreiswirth, B.N. (1999) Identification of a W Variant Outbreak of *Mycobacterium tuberculosis* via Population-Based Molecular Epidemiology. *JAMA.* 282:2321-2327

Bishai, W. (2000) Lipid lunch for persistent pathogen. *Nature.* 406:683-685

Blot, B. (1994) Transposable elements and adaptation of host bacteria. *Genetica.* 93:5-12

Boom, W.H. (1996) The role of T-cell subsets in *Mycobacterium tuberculosis* infection. *Infect Agents Dis.* 5:73-81

Bothamley, G.H., Catty, D., Clifton-Hadley, R., Griffin, F., Hewinson, G., Pollock, J. (1999) Immunodiagnosis of mycobacterial infection. pp. 180-197. In *Mycobacteria*: *Molecular Biology and Virulence,* pp. 287-306. Edited by C. Ratledge and J. Dale. Blackwell Science Ltd.

Braden, C.R. (1997) Current Concepts in *Mycobacterium tuberculosis* DNA Fingerprinting. *Infect Dis Clin Pract* . 6:89-95

Braden, C.R., Templeton, G.L., Cave, M.D. Valway, S., Onorato, I.M., Castro, K.G., Moers, D., Yang, Z., Stead, W.W., Bates, J.H. (1997) Interpretation of restriction fragment length polymorphism analysis *Mycobacterium tuberculosis* isolates from a state with a large rural population. *J Infect Dis.* 175:1446-1452

Brosch, R., Gordon, S.V., Billault, A., Garnier, T., Eiglmeier, K., Soravito, C., Barrell, B.G., Cole, S.T. (1998) Use of a *Mycobacterium tuberculosis* H37Rv Bacterial Artificial Chromosome Library for Genome Mapping, Sequencing, and Comparative Genomics. *Infect Immun.* 66:2221-2229

Brosch, R., Philipp, W.J., Stavropoulos, E., Colston, M.J., Cole, S.T., Gordon, S.V. (1999) Genomic Analysis Reveals Variation between *Mycobacterium tuberculosis* H37Rv and the Attenuated *M. tuberculosis* H37Ra Strain. *Infect Immun.* 67:5768-5774

Brudney, K., Dobkin, J. (1991) Resurgent tuberculosis in New York City. *Am Rev Respir Dis.* 121:313-316

Butler, D. (2000) New fronts in an old war. *Nature.* 406:670-672

C

Camacho, L.R., Ensergieux, D., Perez, E., Gicquel, B., Guilhot, C. (1999) Identification of a virulence gene cluster of *Mycobacterium tuberculosis* by signature-tagged transposon mutagenesis. *Mol Microbiol.* 34:257-267

Camarena, L. Poggio, S., Campos, A., Bastarrachea, F., Osorio, A. (1998) An IS*4* Insertion at the *glnA* Control Region of *Escherichia coli* Creates a New Promoter by Providing the –35 Region of Its 3'-End. *Plasmid.* 39:41-47

Caminero, J., Pena, M.J., Campos-Herrero, M.I., Rodríguez, J.C., Afonso, O., Martin, C., Pavón, J.M., Torres, M.J., Burgos, M., Cabrera, P., Small, P.M., Enarson, D.A. (2001) Exogenous Reinfeciton with Tuberculosis on a European Island with a Moderate Incidence of Disease. *Am J Respir Crit Care Med.* 163:717-720

Campbell, A. (1981) Evolutionary Significance of Accessory DNA Elements in Bacteria. *Annu Rev Microbiol.* 35:55-83

Canetti, G. (1955) The tubercle bacillus in the pulmonary lesion of man: histobacteriology and its bearing on the thearapy of pulmonary tuberculosis. *Springer Publishing Company, Inc. New York, NY*

Cave, M.D., Eisenach, K.D., McDermott, P.F., Bates, J.H., Crawford, J.T. (1991) IS*6110*: Conservation of Sequence in the *Mycobacterium tuberculosis* complex and its utilization in DNA fingerprinting. *Molecular and Cellular Probes*. 5:73-80

Cave, M.D., Eisenach, K.D., Templeton, G., Salfinger, M., Mazurek, G., Bates, J.H., Crawford, J.T. (1994) Stability of DNA Fingerprint Pattern Produced with IS*6110* in Strains of *Mycobacterium tuberculosis*. *J Clin Microbiol.* 32:262-266

Chan, J., Fujiwara, T., Brennan, P. McNeil, M., Turco, S.J., Sibille, J.C., Snapper, M., Aisen, P., Bloom, B.R. (1989) Microbial glycolipids: posible virulence factors that scavenge oxygen radicals. *Proc Natl Acad Sci USA*. 86:2453-2457

Chan, J., Xing, Y., Magliozzo, R.S., Bloom, B.R. (1992) Killing of virulent *Mycobacterium tuberculosis* by reactive nitrogen intermediates produced by activatged murine macrophages. *J Exp Med*. 175:1111-1122

Chan, J., Kaufman, S.H.E. (1994) Immune mechanisms of protection. In *Tuberculosis: Pathogenesis, Protection and Control.* pp. 389-416. Edited by B. Bloom, ASM Press, Washington, DC

Chandler, M., Fayet, O. (1993) Translational frameshifting in the control of transposition in bacteria. *Mol Microbiol*. 7:497-503

Charlier, D., Piette, J., Glansdorff, N. (1982) IS3 can function as a mobile promotor in *E. coli*. *Nucleic Acids Res.* 10:5935-5948

Chaulet, P., Boulahbal, F., Grosset, J. (1995) Surveillance of drug resistance for tuberculosis control: why and how? *Tuber Lung Dis.* 76:487-492

Chaves, F., Dronda, F., Alonso-Sanz, M., Noriega, A.R. (1999) Evidence of exogenous reinfection and mixed infection with more than one strain of *Mycobacterium tuberculosis* among Spanish HIV-infected inmates. *AIDS*. 13:615-620

Chevrel-Dellagi, D., Abderrahman, A., Haltiti, R., Koubaji, H., Gicquel, B., Dellagi, K. (1993) Large-Scale DNA Fingerprinting of *Mycobacterium tuberculosis* Strains as a Tool for Epidemiological Studies of Tuberculosis. *J Clin Microbiol.* 31:2446-2450

Colditz, G.A., Brewer, T.F. Berkey, C.S., Wilson, M.E., Burdick, E., Fineberg, H.V., Mosteller, F. (1994) Efficacy of BCG vaccine in the prevention of tuberculosis: meta-analysis of the published literature. *JAMA*. 271:698-702

Cole, S.T. (1999) Learning from the genome sequence of *Mycobacterium tuberculosis* H37Rv. *FEBS Lett*. 452:7-10

Cole, S.T., Barrell, R.G. (1998) Analysis of the genome of *Mycobacterium tuberculosis* H37Rv. *In* Genetics and Tuberculosis (Novartis Foundation Symposium 217) *Eds.* Chadwick, D.J. and Cardew, G. John Wiley and Sons. pp 160-172

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry III, C.E., 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, M.A., Rajandream, M-A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J.E., Taylor, K., Whitehead, S., Barrell, B.G. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*. 393:537-544

Cole, S.T., Eiglmeier, K., Parkhill, J., James, K.D., Thomson, N.R., Wheeler, P.R., Honoré, N., Garnier, T., Churcher, C., Harris, D., Mungall, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R.M., Devlin, K., Duthoy, S., Feltwell, T., Fraser, A., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Lacroix, C., Maclean, J., Moule, S., Murphy, L., Oliver, K., Quail, M.A., Rajandream, M.-A., Rutherford, K.M., Rutter, S., Seeger, K., Simon, S., Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Taylor, K., Whitehead, S., Woodward, J.R., Barrell, B.G. (2001) Massive gene decay in the leprosy bacillus. *Nature*. 409:1007-1011

Collins, D.M., Stephens, D.M. (1991) Identification of an insertion sequence, IS*1081*, in *Mycobacterium bovis*. *FEMS Microbiol Lett.* 67:11-15

Collins, C.M., Gutman D.M. (1992) Insertional Inactivation of an *Escherichia coli* Urease Gene by IS*3411*. *J Bacteriol.* 174:883-888

Comstock, G.W., Daniel, T.M., Snider, D.E., Edwards, P.Q., Hopewell, P.C., Vandiviere, H.M. (1981) The tuberculin skin test. *Am Rev Respir Dis.* 124:356-363

Condos, R., Rom, W.N., Liu, Y.M., Schluger, N.W. (1998) Local Immune Responses Correlate with Presentation and Outcome in Tuberculosis. *Am J Respir Crit Care Med*. 157:729-735

Cooper, A.M., Dalton, D.K., Stewart, T.A., Griffen, J.P., Russell, D.G., Orme, I.M. (1993) Disseminated tuberculosis in IFN-γ gene-disrupted mice. *J Exp Med*. 178:2243-2248

Coronado, V.G., Beck-Sagué, Hutton, M.D., Davis, B.J., Nicholas, P., Villareal, C., Woodley, C.L., Kilburn, J.O., Crawford, J.T., Frieden, T.R., Sinkowitz, R.L., Jarvis, W.R. (1993) Transmission of Multidrug-Resistant *Mycobacterium tuberculosis* among Persons with Human Immunodeficiency Virus Infection in an Urban Hospital: Epidemiologic and Restriction Fragment Length Polymorphism Analysis. *J Infect Dis.* 168:1052-1055

Craig, N.L. (1996) Transposition. In: *Escherichia coli* and *Salmonella typhimurium*: Cellular and Molecular Biology. (Editors: F.C. Neidhart *et al.*), 2nd edition, Volume 2, pp 2339-2362. American Society for Microbiology, Washington, DC

Cree, I.A., Nurbhai, G., Milne, G., Beck, J.S. (1987) Cell death in granulomata: the role of apoptosis. *J Clin Pathol.* 40:1314-1319

Cutler, C.W., Jotwani, R., Pulendran, B. (2001) Dendritic Cells: Immune Saviors or Achilles' Heel. *Infect Immun.* 69:4703-4708

D

Dale, J.W. (1995) Mobile genetic elements in mycobacteria. *Eur Respir J.* 8(Suppl. 20) 633s-648s

Dale, J.W., Tang, T.H., Wall, S., Zainuddin, Z.F., Plikaytis, B. (1997) Conservation of IS*6110* sequence in strains of *Mycobacterium tuberculosis* with single and multiple copies. *Tuber Lung Dis.* 78:225-227

Dale, J.W., Brittain, D., Cataldi, A.A., Cousins, D., Crawford, J.T., Driscoll, J., Heersma, H., Lillebaek, T., Quitugua, T., Rastogi, N., Skuce, R.A., Sola, C., Van Soolingen, D., vincent, V. (2001) Spacer oligonucleotide typing of bacteria of the *Mycobacterium tuberculosis* complex: recommendations for standaradised nomenclature. *Int J Tuberc Lung Dis.* 5:216-219

Dannenberg, A.M. (1982) Pathogenesis of pulmonary tuberculosis. *Am Rev Respir Dis.* 125:25-29

Dannenberg, A.M., Rook, G.A.W. (1994) Pathogenesis of Pulmonary Tuberculosis: an Interplay of Tissue-Damaging and Macrophage –Activation responses – Dual Mechanisms That control Bacillary Multiplication. In *Tuberculosis: Pathogenesis, Protection and Control.* pp. 459-483. Edited by B. Bloom, ASM Press, Washington, DC

Daniel, T.M., Bates, J.H., Downes, K.A. (1994) History of Tuberculosis. In *Tuberculosis: Pathogenesis, Protection and Control.* pp. 13-23. Edited by B. Bloom, ASM Press, Washington, DC.

David, H.L., Newman, C.M. (1971) Some observations on the genetics of isoniazid resistance in the tubercle bacilli. *Am Rev Respir Dis.* 104:508-15

De Boer, A.S, Borgdorff, M.W., De Haas, P.E.W., Nagelkerke, N.J.D., Van Embden, J.D.A., Van Soolingen, D. (1999) Analysis of Rate of Change of IS*6110* RFLP Patterns of *Mycobacterium tuberculosis* Based on Serial Patient Isolates. *J Infect Dis.* 180:1238-1244

Delogu, G., Brennan, M.J. (2001) Comparative Immune Response to PE and PE_PGRS Antigens of *Mycobacterium tuberculosis*. *Infect Immun*. 69:5606-5611

De Voss, J.J., Rutter, K., Schroeder, B.G., Barry, C.E. (1999) Iron Acquisition and Metabolism by Mycobacteria. *J Bacteriol.* 181:4443-4451

Dickinson, J.M., Aber, V.R., Mitchison, D.A. (1977) Bactericidal activity of streptomycin, isoniazid, rifampicin, ethambutol and pyrazinamide alone and in combination against *Mycobacterium tuberculosis*. *Am Rev Respir Dis*. 116:627-635

Dillon, D.C., Alderson, M.R., Hay, C.H., Lewinsohn, D.M., Coler, R., Bement, T., Campos-Neto, A., Skeiky, Y.A.W., Orme, I.M., Roberts, A., Steen, S., Dalemans, W., Badaro, R., Reed, S.G. (1999) Molecular Characterization and Human T-Cell Responses to a Member of a Novel *Mycobacterium tuberculosis mtb39* Gene Family. *Infect Immun.* 67:2941-2950

Doenhoff, M.J. (1997) A role for granulomatous inflammation in the transmission of infectious disease:schistosomiasis and tuberculosis. *Parasitology*. 115:S113-S125

Doolittle, W.F., Sapienza, C. (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature*. 284:601-603

Doran, T.J., Hodgson, A.L.M., Davies, J.K., Radford, A.J. (1992) Characterisation of a novel repetitive DNA sequence from *Mycobacterium bovis*. *FEMS Microbiol Lett.* 75:179-185

Du Plessis, D.G., Warren, R., Richardson, M., Joubert, J.J., Van Helden, P.D. (2001) Demonstration of reinfection and reactivation in HIV-Negative Autopsied cases of secondary tuberculosis: multilesional genotyping of *Mycobacterium tuberculosis* utilizing IS*6110* and other repetitive element-based DNA fingerprinting. *Tuber Lung Dis.* 81:211-220

Dybvig, K. (1993) DNA rearrangements and phenotypic switching in prokaryotes. *Mol Microbiol.* 10:465-471

Dye, C., Scheele, S., Dolin, P., Pathania, V., Raviglione, M.C. (1999) Global Burden of Tuberculosis: Estimated Incidence, Prevalence, and Mortality by Country. *JAMA.* 282:677-686

E

Eisenach, K.D. (1999) Molecular diagnostics. In *Mycobacteria*: *Molecular Biology and Virulence,* pp. 161-196. Edited by C. Ratledge and J. Dale. Blackwell Science Ltd.

Emile, J.-F., Patey, N., Altare, F., Lamhamedi, S., Jouanguy, E., Boman, F., Quillard, J., Lecomte-Houcke, M., Verola, O., Mousnier, J.-F., Dijoud, F., Blanche, S., Fischer, A., Brousse, N., Casanova, J.-L. (1997) Correlation of granuloma structure with clinical outcome defines two types of idiopathic disseminated BCG infection. *J Pathol.* 181:25-30

Enarson, D.A. (1991) Principles of IUATLD Collaborative National Tuberculosis Programmes. *Bull Int Union Tuberc.* 66:195-200

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

F

Fang, Z., Forbes, K.J. (1997) A *Mycobacterium tuberculosis* IS*6110* Preferential Locus (ipl) for Insertion into the Genome. *J Clin Microbiol.* 35:479-481

Fang, Z., Morrison, N., Watt, B., Doig, C., Forbes, K.J. (1998) IS*6110* Transposition and Evolutionary Scenario of the Direct Repeat Locus, in a Group of Closely Related *Mycobacterium tuberculosis* Strains. *J Bacteriol.* 180:2102-2109

Fang, Z., Doig, C., Kenna, D.T., Smittipat, N., Palittapongarnpim, P., Watt, B., Forbes, K.J. (1999a) IS*6110*-Mediated Deletions of Wild-Type Chromosomes of *Mycobacterium tuberculosis. J Bacteriol.* 181:1014-1020

Fang, Z., Doig, C., Morrison, N., Watt, B., Forbes, K.J. (1999b) Characterization of IS*1547*, a New Member of the IS*900* Family in the *Mycobacterium tuberculosis* Complex, and Its Association with IS*6110*. *J Bacteriol.* 181:1021-1024

Fayet, O., Ramond, P., Polard, P., Prère, M.F., Chandler, M. (1990) Functional similarities between retroviruses and the IS*3* family of bacterial insertion sequences? *Mol Microbiol.* 4:1771-1777

Fedoroff, N.V. (1999) Transposable elements as a molecular evolutionary force. *Ann N Y Acad Sci.* 870:251-264

Felsenstein, J. (1985) Confidence limits on phylogenies: an approachnusing the bootsrap. *Evolution.* 39:783-793

Fenhalls, G., Geyp, M., Dent, D.M., Parker, M.I. (1999) Breast tumour cell-induced down regulation of type I collagen mRNA in fibroblasts. *Br J Cancer.* 81:1142-1149

Fenhalls, G., Wong, A., Dezuidenhout, J., Van Helden, P. Bardin, P., Lukey, P.T. (2000) In Situ Production of Gamma Interferon, Interleukin-4 and Tumor Necrosis Factor Alpha mRNA in Human Lung Tuberculous Granulomas. *Infect Immun.* 68:2827-2836

Filliol, I., Sola, C., Rastogi, N. (2000a) Detection of a previously unamplified spacer within the DR locus of *Mycobacterium tuberculosis*: epidemiological implications. *J Clin Microbiol.* 38:1231-4

Filliol, I., Ferdinand, S., Negroni, L., Sola, C., Rastogi, N. (2000b) Molecular Typing of *Mycobacterium tuberculosis* Based on Variable Number of Tandem DNA Repeats Used Alone and in Association with Spoligotyping. *J Clin Microbiol.* 38:2520-2524

Fine, P.E.M. (1995) Variation in protection by BCG: Implications of and for heterologous immunity. *Lancet.* 346:1339-1345

Flesch, I.E., Kaufman, S.H. (1991) Mechanisms involved in mycobacterial growth inhibition by gamma-interferon-activated bone marrow macrophages: role of reactive nitrogen intermediates. *Infect Immun.* 59:3213-3218

Flynn, J.L., Goldstein, M.M., Triebold, K.J., Koller, B., Bloom, B.R. (1992) Major histocompatability complex class I-restricted T cells are required for resistance to *Mycobacterium tuberculosis* infection. *Proc Natl Acad Sci USA.* 89:12013-12017

Flynn, J.L., Chan, J., Triebold, K.J., Dalton, D.K., Stewart, T.A., Bloom, B.R. (1993) An essential role for interferon-γ in resistance to *Mycobacterium tuberculosis* infection. *J Exp Med.* 178:2249-2254

Flynn, J.L., Goldstein, M.M., Chan, K.J., Triebold, K.J., Pfeffer, C.J., Lowenstein, C.J., Schreiber, R., Mak, T.W., Bloom, B.R. (1995) Tumor necrosis factor-α is required in the protective immune response against *M. tuberculosis* in mice. *Immunity.* 2:561-572

Flynn, J.L., Bloom, B.R. (1996) Role of T1 and T2 Cytokines in the Response to *Mycobacterium tuberculosis. Annals of the New York Academy of Sciences.* 795:137-146

Fomukong, N.G., Dale, J.W. (1993) Transpositional activity of IS*986* in *Mycobacterium smegmatis. Gene.* 130:99-105

Fomukong, N.G., Tang, T.H., Al-Maamary, S., Ibrahim, W.A., Ramayah, S., Yates, M., Zainudin, Z.F., Dale, J.W. (1994) Insertion sequence typing of *Mycobacterium tuberculosis*: characterization of a widespread subtype with a single copy of IS*6110*. *Tuber Lung Dis.* 75:435-440

Fomukong, N., Beggs, M., El-Hajj, H., Templeton, G., Eisenach, K., Cave, M.D. (1998) Differences in the prevalence of IS*6110* insertion sites in *Mycobacterium tuberculosis* strain: low and high copy number of IS*6110*. *Tuber Lung Dis.* 78:109-116

Frieden, T.R., Sterling, T., Pablos-Mendez, A., Kilburn, J.O., Cauthen, G.M., Dooley, S.W. (1993) The Emergence of Drug-Resistant Tuberculosis in New York City. *N Engl J Med.* 328:521-526

Friedland, J.S. (1994) Chemotactic cytokines and tuberculosis. *Biochem Soc Trans.* 22:310-312

Friedman, C.R., Stoeckle, M.Y., Kreswirth, B.N., Johnson, W.D., Manoach, S.M., Berger, J., Sathianathan, K., Hafner, A., Riley, L.W. (1995) Transmission of Multidrug-Resistant Tuberculosis in a Large Urban Setting. *Am J Respir Crit Care Med.* 152:355-359

Frothingham, R., Hills, H.G., Wilson, K.H. (1994) Extensive DNA Sequence Conservation throughout the *Mycobacterium tuberculosis* Complex. *J Clin Microbiol.* 32:1639-1643

G

Galas, D.J., Chandler, M. (1989) Bacterial insertion sequences, p. 109-162. *In* D.E. Berg and M. Howe (ed.), Mobile DNA. American Society for Microbiology, Washington, D.C

Gey van Pittius, N.C., Gamieldien, J., Hide, W., Brown, G.D., Siezen, R.J., Beyers, A.D. (2001). Structure and phylogeny of the *Mycobacterium tuberculosis* ESAT-6 gene cluster; a gene cluster present in the high G+C gram-positive bacteria and multiple duplicated in the mycobacteria. *Genome Biol.* 2:0044.1-0044.18

Ghanekar, K., McBride, A., Dellagostin, O., Thorne, S., Mooney, R., McFadden, J. (1999) Stimulation of transposition of the *Mycobacterium tuberculosis* insertion sequence IS*6110* by exposure to a microaerobic environment. *Mol Microbiol.* 33:982-993

Gillespie, S.H., Dickens, A., McHugh, T.D. (2000) False Molecular Clusters due to Nonrandom Association of IS*6110* with *Mycobacterium tuberculosis*. *J Clin Microbiol.* 38:2081-6

Glickman, M.S., Cox, J.S., Jacobs, W.R. (2000) A novel mycolic acid cyclopropane synthetase is required for cording, presistence, and virulence of *M. tuberculosis*. *Mol Cell.* 5:717-727

Glynn, J.R., Bauer, J., de Boer, A.S., Borgdorff, M.W., Fine, P.E.M., Godfrey-Faussett, P., Vynnycky, E. (1999) Interpreting DNA fingerprint clusters of *Mycobacterium tuberculosis*. *Int J Tuberc Lung Dis.* 3:1055-1060

Godfrey-Faussett, Stoker, N.G., Scott, J.A.G., Pasvol, G., Kelly, P., Clancy, L. (1993) DNA Fingerprints of *Mycobacterium tuberculosis* do not change during the development of rifampicin resistance. *Tuber Lung Dis.* 74:240-243

Gonzalez-Juarrero, M., Turner, O.C., Turner, J., Marietta, P., Brooks, J.V., Orme, I.M. (2001) Temporal and Spatial Arrangment of Lymphocytes within Lung Granulomas Induced by Aerosol Infection with *Mycobacterium tuberculosis*. *Infect Immun.* 69:1722-1728

Gordon, A.H., D'Arcy Hart, P., Young, M.R. (1980) Ammonia inhibits phagosome-lysosome fusion in macrophages. *Nature.* 286:79-80

Gordon, S.V., Brosch, R., Billault, A., Garnier, T., Eiglmeier, Cole, S.T. (1999a) Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Mol Microbiol.* 32:643-655

Gordon, S.V., Heym, B., Parkhill, J., Barrell, B., Cole, S.T. (1999b) New insertion sequences and a novel repeated sequence in the genome of *Mycobacterium tuberculosis* H37Rv. *Microbiology.* 145:881-892

Gordon, S.V., Eiglmeier, K., Garnier, T., Brosch, R., Parkhill, J., Barrell, B., Cole, S.T., Hewinson, R.G. (2001) Genomics of *Mycobacterium bovis*. *Tuber Lung Dis.* 81:157-163

Goyal, M., Ormerod, L.P., Shaw, R.J. (1994a) Epidemiology of an outbreak of drug-resistant tuberculosis in the U.K. using restriction fragment length polymorphism. *Clin Sci.* 86:749-751

Goyal, M., Young, D., Zhang, Y., Jenkins, P.A., Shaw, R.J. (1994b) PCR Amplification of Variable Sequence Upstream of *katG* Gene to Subdivide Strains of *Mycobacterium tuberculosis* Complex. *J Clin Microbiol.* 32:3070-3071

Goyal, M., Saunders, N.A., Van Embden, J.D.A., Young, D.B., Shaw, R.J. (1997) Differentiation of *Mycobacterium tuberculosis* Isolates by Spoligotyping and IS*6110* Restriction Fragment Length Polymorphism. *J Clin Microbiol.* 35:647-651

Gravekamp, C., Rosner, B., Madoff, L.C. (1998) Deletion of repeats in the alpha C protein enhances the pathogenicity of group B streptococci in immune mice. *Infect Immun.* 66:4347-4354

Groenen, P.M.A., Bunschoten, A.E., Van Soolingen, D., Van Embden, J.D.A. (1993) Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. *Mol Microbiol.* 10:1057-1065

Gruft, H., Johnson, R., Claflin, R., Loder, A. (1984) Phage-typing and drug-resistance patterns as tools in mycobacterial epidemiology. *Am Rev Respir Dis.* 130:96-97

Guilhot, C., Otal, I., Van Rompaey, I., Martin, C., Gicquel, B. (1994) Efficient transposition in mycobacteria: construction of *Mycobacterium smegmatis* insertional mutant libraries. *J Bacteriol.* 176:535-9

H

Hall, B.G. (1998) Activation of the *bgl* Operon by Adaptive Mutation. *Mol Biol Evol* 15:1-5

Hallet, B., Sherratt, D.J. (1997) Transposition and site-specific recombination: adapting DNA cut-and-paste mechanisms to a variety of genetic rearrangements. *FEMS Microbiol Rev*. 21:157-178

Hammerschmidt, S., Hilse, R., van Putten, J.P.M., Gerardy-Schahn, R., Unkmeir, A., Frosch, M. (1996) Modulation of cell surface sialic acid expression in *Neisseria meningitidis* via a transposable genetic element. *EMBO J* 15:192-198

Hansen, J. E., Engelbrecht, J., Bohr, H., Nielsen, J. O., Hansen, J-E. S. & Brunak, S. (1995) Prediction of O-glycosylation of mammalian proteins: Specificity patterns of UDP-GalNAc:-polypeptide N-acetylgalactosaminyltransferase. *Biochem J* 308:801-813

Hass, D.W., Des Prez, R.M. (1994) Tuberculosis and AIDS: a historical perspective on recent developments. *Am J Med*. 96:439-450

Havlir, D.V. (1994) *Mycobacterium avium* complex: advances in therapy. *Eur J Clin Microbiol Infect Dis*. 13:915-924

Hellyer, T.J., DesJardin, L.E., Hehman, G.I., Cave, M.D., Eisenach, K.D. (1999) Quantitative analysis of mRNA as a marker for viability of *Mycobacterium tuberculosis*. *J Clin Microbiol*. 37:290-295

Henderson, R.A., Watkins, S.C., Flynn, J.L. (1997) Activation of human dendritic cells following infection with *Mycobacterium tuberculosis*. *J Immunol*. 159:635-643

Hermans, P.W.M., van Soolingen, D., Bik, E.M., De Haas, P.E.W., Dale, J.W., Van Embden, J.D.A. (1991) Insertion Element IS*987* from *Mycobacterium bovis* BCG Is Located in a Hot-spot Integration Region for Insertion Elements in *Mycobacterium tuberculosis* Complex Strains. *Infect Immun*. 59:2695-2705

Hermans, P.W.M., Van Soolingen, D., Dale, J.W., Schuitema, A.R.J., McAdam, R.A., Catty, D., Van Embden, J.D.A. (1990) Insertion Element IS*986* from *Mycobacterium tuberculosis*: a Useful Tool for Diagnosis and Epidemiology of Tuberculosis. *J Clin Microbiol*. 28:2051-2058

Hermans, P.W.M., Van Soolingen, D., Van Embden, J.D.A. (1992) Characterization of a Major Polymorphic Tandem Repeat in *Mycobacterium tuberculosis* and Its Potential Use in the Epidemiology of *Mycobacterium kansasii* and *Mycobacterium gordonae*. *J Bacteriol*. 174:4157-4165

Hermans, P.W.M., Messadi, F., Guebrexabher, H., Van Soolingen, D., De Haas, P.E.W., Heersma, H., De Neeling, H., Ayoub, A., Portaels, F., Frommel, D., Zribi, M., Van Embden, J.D.A. (1995) Analysis of the Population Structure of *Mycobacterium tuberculosis* in Ethiopia, Tunisia and The Netherlands. Usefulness of DNA Typing for Global Epidemiology. *J Infect Dis*. 171:1504-1513

Hernandez-Pando, R., Bornstein, Q.L., Leon, D.A., Orozco, E.H., Madrigal, V.K. (2000a) Inflammatory cytokine production by immunological and foreign body multinucleated giant cells. *Immunology*. 100:352-358

Hernandez-Pando, R., Jeyanathan, M., Mengistu, G., Aguilar, D., Orozco, H., Harboe, M., Rook, G.A.W., Bjune, G. (2000b) Persistence of DNA from *Mycobacterium tuberculosis* in superficially normal lung tissue during latent infection. *Lancet*. 356:2133-2138

Herrmann, J.L., O'Gaora, P.O., Gallagher, A., Thole, J.E.R., Young, D.B. (1996) Bacterial glycoproteins: a link between glycosylation and proteolytic cleavage of a 19kDa antigen from *Mycobacterium tuberculosis*. *EMBO J*. 15:3547-3554

Ho, T.B.L., Robertson, B.D., Taylor, G.M., Shaw, R.J., Young, D.B. (2000) Comparison of *Mycobacterium tuberculosis* genomes reveals a 20 kb variable region in clinical isolates. *Yeast*. 17:272-282

Hollingshead, S.K., Fischetti, V.A., Scott, J.R. (1987) Size variation in group A streptococcal M protein is generated by homologous recombination between intragenic repeats. *Mol Gen Genet*.207:196-203

Hopewell, P.C. (1994) Overview of Clinical Tuberculosis. In *Tuberculosis: Pathogenesis, Protection and Control*. pp. 25-46. Edited by B. Bloom, ASM Press, Washington, DC

Hopwood, D. A., Bibb, M. J., Chater, K. F., Kieser, T., Bruton, C. J., Kieser, H. M., Lydiate, D. J., Smith, C.P., Ward, J. M. & Schrempf, H. (1985). Genetic manipulation of Streptomyces, a laboratory manual, John Innes Foundation

Howard, W.L., Kopfenstein, M.D., Steininger, W.J., Woodruff, C.E. (1970) The loss of tuberculin reactivity in certain patients with active pulmonary tuberculosis. *Chest*. 57:530-534

Hu, Y., Coates, A.R.M. (1999) Transcription of the Stationary-Phase-Associated *hspX* Gene of *Mycobacterium tuberculosis* Is Inversely Related to Synthesis of the 16-Kilodalton Protein. *J Bacteriol*. 181:1380-1387

Hubner, A., Hendrickson, W. (1997) A fusion promoter created by a new insertion sequence, IS1490, activates transcription of 2,4,5-trichlorophenoxyacetic acid catabolic genes in Burkholderia cepacia AC1100. *J Bacteriol*. 179:2717-2723

Huh, Y.J., Ahn, D.I., Kim, S.J. (1995) Limited variation of DNA fingerprints (IS*6110* and IS*1081*) in Korean strains of *Mycobacterium tuberculosis*. *Tuber Lung Dis*. 76:324-329

I

Ishiguro, N., Sato, G. (1984) Spontaneous Deletion of Citrate-Utilizing Ability Promoted by Insertion Sequences. *J Bacteriol.* 160: 642-650

Ivanyi, J., Thole, J. (1994) Specifity and Function of T- and B-Cell Recognition in Tuberculosis. In *Tuberculosis: Pathogenesis, Protection and Control.* pp. 459-483. Edited by B. Bloom, ASM Press, Washington, DC

J

Jones, W.D., Kubica, G.P. (1968) Fluorescent antibody techniques with mycobacteria. III. Investigation of the five serological homogeneous groups of mycobacteria. *Zentralbl Bakteriol Orig. A.* 207:58-62

Jones, W.D., Greenberg, J. (1978) Modification of methods used in bacteriophage typing of *Mycobacterium tuberculosis* isolates. *Am Rev Respir Dis.* 7:467-469

Jones, K.F., Hollingshead, S.K., Scott, J.R., Fischetti, V.A. (1988) Spontaneous M6 protein size mutants of group A streptococci display variation in antigenic and opsonogenic epitopes. *Proc Natl Acad Sci USA.* 85:8271-8275

Jou, N-T., Yoshimori, R.B., Mason, G.R., Louie, J.S., Liebling, M.R. (1997) Single-Tube, Nested, Reverse Transcriptase PCR for Detection of Viable *Mycobacterium tuberculosis. J Clin Microbiol.* 35:1161-1165

Jouanguy, E., Altare, F., Lamhamedi, S., Revy, P., Emile, J.F., Newport, M., Levin, M., Blanche, S., Seboun, E., Fischer, A., Casanova, J.L. (1996) Interferon-gamma-receptor deficiency in an infant with fatal bacille Calmett-Guérin infection. *N Engl J Med.* 335:1956-1961

K

Kamerbeek, J., Schouls, L., Kolk, A., Van Agterveld, M., Van Soolingen, D., Kuijper, S., Bunschoten, A., Molhuizen, H., Shaw, R., Goyal, M., Van Embden, J. (1997) Simultaneous Detection and Strain Differentiation of *Mycobaterium tuberculosis* for Diagnosis and Epidemiology. *J Clin Microbiol.* 35(4)907-914

Kapur, V., Whittam, T.S., Musser, J.M. (1994) Is *Mycobacterium tuberculosis* 15,000 Years Old? *J Infect Dis.* 170:1348-1349

Kato-Maeda, M., Rhee, J.T., Gingeras, T.R., Salamon, H., Drenkow, J., Smittipat, N., Small, P.M. (2001) Comparing Genomes within the Species *Mycobacterium tuberculosis. Genome Res.* 11:547-554

Kleckner, N., Reichardt, K., Botstein, D. (1979) Inversions and deletions of the Salmonella chromosome generated by the translocatable tetracycline resistance element Tn*10*. *J Mol Biol.* 127:89-115

Kobzik, L., Schoen, F.J. (1994) The Lung. In Pathologic Basis of Disease, pp. 673-735. Edited by Cotran, R.S., Kumar, V.K., Robbins, S.L., Schoen, F.J. W.B. Saunders Company, Philadelphia, Pennsylvania.

Koch, R. (1882) Die Aetiologie der Tuberculos. *Ber. Klin. Wochenschr.* 19:21 [English translation in: Koch, R. (1982) The aetiology of tuberculosis. *Rev Infect Dis* 4:1270-1274]

Koch, R. (1891) Weitere Mitteilung über das Tuberkulin. *Dtsch Med Wochenschr.* 43:1189-1192

Kremer, K., Van Soolingen, D., Frothingham, R., Haas, W.H., Hermans, P.W.M., Martín, C., Palittapongarnpim, P., Plikaytis, B.B., Riley, L.W., Yakrus, M.A., Musser, J.M., Van Embden, J.D.A. (1999) Comparison of Methods Based on Different Molecular Epidemiologic Markers for Typing of *Mycobacterium tuberculosis* Complex Strains: Interlaboratory Study of Discriminatory Power and Reproducibility. *J Clin Microbiol.* 37:2607-2618

Kurepina, N.E., Sreevatsan, S., Plikaytis, B.B., Bifani, P.J., Connell, N.D., Donnelly, R.J., Van Sooligen, D., Musser, J.M., Kreiswirth, B.N. (1998) Characterization of the phylogenetic distribution and chromosomal insertion sites of five IS*6110* elements in *Mycobacterium tuberculosis*: non-random integration in the *dnaA-dnaN* region. *Tuber Lung Dis.* 79:31-42.

L

Labandeira-Rey, M., Baker, E.A., Skare, J.T. (2001) VraA (BBI16) Protein of *Borrelia burgdorferi* Is a Surface-Exposed Antigen with a Repetitive Motif That Confers Partial Protection against Experimental Lyme Borreliosis. *Infect Immun.* 69:1409-1419

Lachenauer, C.S,, Creti, R., Michel, J.L., Madoff, L.C. (2000) Mosaicism in the alpha-like protein genes of group B streptococci. *Proc Natl Acad Sci USA*. 97:9630-9635

Lalvani, A., Pathan, A.A., Durkan, H., Wilkinson, K.A., Whelan, A., Deeks, J.J., Reece, W.H., Latif, M., Pasvol, G., Hill, A.V. (2001) Enhanced contact tracing and spatial tracking of *Mycobacterium tuberculosis* infection by enumeration of antigen-specific T cells. *Lancet*. 357:2017-21

Lawrence, J.G., Dykhuizen, D.E., DuBose, R.F., Hartl, D.L. (1989) Phylogenetic analysis using insertion sequence fingerprinting in *Escherichia coli*. *Mol Biol Evol*. 6:1-14

Leão, S.C., Rocha, C.L., Murillo, L.A., Parra, C.A., Patarroyo, M.E. (1995) A Species-Specific Nucleotide Sequence of *Mycobacterium tuberculosis* Encodes a Protein that Exhibits Hemolytic Activity when Expressed in *Escherichia coli*. *Infect Immun*. 63:4301-4306

Levinson, G., Gutman, G.A. (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol*. 4:203-221

Lin, R., Bernard, E.M., Armstrong, D., Chen, C., Riley, L.W. (1996) Transmission patterns of tuberculosis in Taiwan: analysis by restriciton fragment length polymorphism. *Int J Infect Dis*. 1:18-21

Lloyd, R.G., Low, K.B. (1996) Homologous recombination. In: *Escherichia coli* and *Salmonella typhimurium*: Cellular and Molecular Biology. (Editors: F.C. Neidhart *et al*.), 2nd edition, Volume 2, pp 2236-2255. American Society for Microbiology, Washington, DC

Lounis, N., Truffot-Pernot, C., Grosset, J., Gordeuk, V.R., Boelaert, J.R. (2001) Iron and *Mycobacterium tuberculosis* infection. *J Clin Virol*. 20:123-126

Lysnyansky, I., Sachse, K., Rosenbusch, R., Levisohn, S., Yogev, D. (1999) The vsp locus of *Mycoplasma bovis:* gene organization and structural features. *J Bacteriol*. 181:5734-41.

M

Mahairas, G.G., Sabo, P.J., Hickey, M.J., Singh, D.C., Stover, K. (1996) Molecular Analysis of Genetic Differences between *Mycobacterium bovis* BCG and Virulent *M. bovis*. *J Bacteriol*. 178:1274-1282

Mahillon, J., Chandler, M. (1998) Insertion Sequences. *Microbiol Mol Biol Rev*. 62:725-774

Mazurek, G.H., Cave, M.D, Eisenach, K.D., Wallace, R.J., Bates, J.H., Crawford, J.T. (1991) Chromosomal DNA Fingerprint Patterns Produced with IS*6110* as Strain-Specific Markers for Epidemiologic Study of Tuberculosis. *J Clin Microbiol*. 29:2030-2033

McAdam, R.A., Hermans, P.W.M., Van Soolingen, D., Zainuddin, Z.F., Catty, D., Van Embden, J.D.A., Dale, J.W. (1990) Characterization of a *Mycobacterium tuberculosis* insertion sequence belonging to the IS*3* family. *Mol Microbiol*. 4: 1607-1613

McAdam, R.A., Guilhot, C., Gicquel, B. (1994) Transposition in Mycobacteria. In *Tuberculosis: Pathogenesis, Protection and Control*. pp. 199-216. Edited by B. Bloom, ASM Press, Washington, DC.

McAdam, R.A., Weisbrod, T.R., Martin, J., Scuderi, J.D., Brown, A.M., Cirillo, J.D., Bloom, B.R., Jacobs, W.R. Jr. (1995) *In vivo* growth characteristics of leucine and methionine auxotrophic mutants of *Mycobacterium bovis* BCG generated by transposon mutagenesis. *Infect Immun*. 63:1004-12.

McClintock, B. (1948) Mutable loci in maize. *Carnegie Int Wash Year Book*. 47:155-169

McHugh, T.D., Gillespie, S.H. (1998) Nonrandom Association of IS*6110* and *Mycobacterium tuberculosis*: Implications for Molecular Epidemiological Studies. *J Clin Microbiol*. 36:1410-1413

McKinney, J.D., Höner zu Bentrup, K., Muñoz-Elias, E.J., Miczak, A., Chen, B., Chan, W-T., Swenson, D., Sacchettini, J.C., Jacobs, W.R., Russell, D.G. (2000) Persistence of *Mycobacterium tuberculosis* in macrophages and mice requires the glyoxylate shunt enzyme isocitrate lyase. *Nature*. 406:735-738

McNeil, M.R., Brennan, P.J. (1991) Structure, function and biogenesis of the cell envelope of mycobacteria in relation to bacterial physiology, pathogenesis and drug resistance; some thoughts and possibilities arising from recent structural information. *Res Microbiol*. 142:451-463

Mendiola, M.V., Martín, C., Otal, I., Gicquel, B. (1992) Analysis of the regions responsible for IS*6110* RFLP in a single *Mycobacterium tuberculosis* strain. *Res Microbiol*. 143:767-772

Metzgar, D., Wills, C. (2000) Evolutionary changes in mutation rates and spectra and their influence on the adaptation of pathogens. *Microb Infect*. 2:1513-1522

Mohan, V.P., Scanga, C.A., Yu, K., Scott, H., Tanaka, K.E., Tsang, E., Tsai, M.C., Flynn, J.L., Chan, J. (2001) Effects of Tumor Necrosis Factor Alpha on Host Immune Response in Chronic Persistent Tuberculosis: Possible Role for Limiting Pathology. *Infect Immun*. 69:1847-1855

Molhuizen, H.O.F., Bunschoten, A.E., Schouls, L.M., Van Embden, J.D.A. (1999) Rapid detection and Simultaneous Strain Differentiation of *Mycobacterium tuberculosis* Complex Bacteria by Spoligotyping. In *Methods in Mol Microbiol, Vol 101: Mycobacteria Protocols*, pp. 129-143. Edited by T. Parish and N. Stoker. Humana Press Inc.

Moss, A.R., Alland, D., Telzak, E., Hewlett, D., Sharp, V., Chiliade, P., LaBombardi, V., Kabus, D., Hanna, B., Palumbo, L., Brudney, K., Weltman, A., Stoeckle, K., Chirgwin, K., Simberkoff, M., Moghazeh,

S., Eisner, W., Lutfey, M., Kreiswirth, B. (1997) A city-wide outbreak of a multiple-drug-resistant strain of *Mycobacterium tuberculosis* in New York. *Int J Tuberc Lung Dis.* 1:115-121

Mossman, T.R., Coffman, R.L. (1989) Heterogeneity of cytokine secretion patterns and functions of helper T cells. *Adv Immunol.* 46:111-147

Moxon, E.R. (1992) *J Infect Dis.* 165:S77-S81

Moxon, E.R., Rainey, P.B., Nowak, M.A., Lenski, R.E. (1994) Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr Biol.* 4:24-33

Murray, P.J. (1999) Defining the requirements for immunological control of mycobacterial infections. *TIM.* 7:366-371

Musser, J.M., Amin, A., Ramaswamy, S. (2000) Negligible Genetic Diversity of *Mycobacterium tuberculosis* Host Immune System Protein Targets: Evidence of Limited Selective Pressure. *Genetics.* 155:7-16

N

Naas, T., Blot, M., Fitch, W.M., Arber, W. (1994) Insertion Sequence-Related Genetic Variation in Resting *Escherichia coli* K-12. *Genetics.* 136:721-730

Nash, H.A. (1996) Site-specific recombination: integration, excision, resolution and inversion of defined DNA segments. In: *Escherichia coli* and *Salmonella typhimurium*: Cellular and Molecular Biology. (Editors: F.C. Neidhart *et al.*), 2nd edition, Volume 2, pp 2363-2376. American Society for Microbiology, Washington, DC

Newport, M.J., Huxley, C.M., Huston, S., Hawrylowicz, C.M., Oostra, B.A., Williamson, R., Levin, M. (1996) A mutation in the interferon-γ-receptor gene and susceptibility to mycobacterial infection. *N Engl J Med.* 335:1941-1949

Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 10:1-6

Niemann, S., Richter, E., Rüsch-Gerdes, S. (1999a) Stability of *Mycobacterium tuberculosis* IS*6110* Restriction Fragment Length Polymorphism Patterns and Spoligotypes Determined by Analyzing Serial Isolates from Patients with Drug-Resistant Tuberculosis. *J Clin Microbiol.* 37:409-412

Niemann, S., Richter, E., Rüsch-Gerdes, S. (1999b) Stability of IS*6110* Restriction Fragment Length Polymorphism Patterns of Multidrug-Resistant *Mycobacterium tuberculosis* Strains. *J Clin Microbiol.* {Letter} 37:3078-3079

Niemann, S., Rüsch-Gerdes, S., Richter, E., Thielen, H., Heykes-Uden, H., Diel, R. (2000) Stability of IS*6110* Restriction Fragment Length Polymorphism Patterns of *Mycobacterium tuberculosis* Strains in Actual Chains of Transmission. *J Clin Microbiol.* 38:2563-2567

Noordhoek, G.T., Kolk, A.H.J., Bjune, G., Catty, D., Dale, J.W., Fine, P.E.M., Godfrey-Faussett, P., Cho, S-N., Shinnick, T., Svenson, S.B., Wilson, S., Van Embden, J.D.A. (1994) Sensitivity and Specificity of PCR for Detection of *Mycobacterium tuberculosis*: a Blind Comparison Study among Seven Laboratories. *J Clin Microbiol.* 32:277-284

North, R.J., Izzo, A.A. (1993) Granuloma Formation in Severe Combined Immunodeficient (SCID) Mice in Response to Progressive BCG Infection. *Am J Pathol.* 142:1959-1965

O

O'Brien, L., Roberts, B., Andrew, P.W., (1996) *In vitro* interaction of *Mycobacterium tuberculosis* and macrophages: activation of anti-mycobacterial activity of macrophages and mechanisms of anti-mycobacterial activity. *Curr Top Microbiol Immunol.* 215:97-130

O'Brien, J.K., Sandman, L.A., Kreiswirth, B.N., Rom, W.N., Schluger, N.W. (1997) DNA fingerprints from *Mycobacterium tuberculosis* isolates of patients confined for therapy noncompliance show frequent clustering. *Chest.* 112:387-392

O'Brien, R., Flynn, O., Costello, E., O'Grady, D., Rogers, M. (2000) Identification of a Novel DNA Probe for Strain Typing *Mycobacterium bovis* by Restriction Fragment Length Polymorphism Analysis. *J Clin Microbiol.* 38:1723-1730

Olakanmi, O., Britigan, B.E., Schlesinger, L.S. (2000) Gallium disrupts iron metabolism of mycobacteria residing within human macrophages. *Infect Immun.* 68:5619-5627

Orgel, L.E., Crick, F.H.C. (1980) Selfish DNA: the ultimate parasite. *Nature.* 284:604-607

Otal, I., Martín, C., Vincent-Lévy-Frebault, V., Thierry, D., Gicquel, B. (1991) Restriction Fragment Length Polymorphism Analysis Using IS*6110* as an Epidemiological Marker in Tuberculosis. *J Clin Microbiol.* 29:1252-1254

P

Palittapongarnpim, P., Luangsook, P., Tansuphaswadikul, S., Chuchottaworn, C., Prachaktam, R., Sathapatayavongs, B. (1997) Restriction fragment length polymorphism study of *Mycobacterium tuberculosis* in Thailand using IS*6110* as probe. *Int J Tuberc Lung Dis.* 1:370-376

Papadimitriou, J.M., Van Bruggen, Y. (1986) Evidence that multinucleate giant cells are exampls of mononuclear phagocytic differentiation. *J Pathol.* 148:149-157

Papadopoulos, D., Schneider, D., Meier-Eiss, J., Arber, W., Lenski, R.E., Blot, M. (1999) Genomic evolution during a 10,000-generation experiment with bacteria. *Proc Natl Acad Sci USA.* 96:3807-3812

Parish, T., Mahenthiralingam, E., Draper, P., Davis, E.O., Colston, M.J. (1997) Regulation of the inducible acetamidase gene of *Mycoacterium smegmatis. Microbiology.* 143:2267-2276

Parish, T., Stoker, N.G. (1999a) Mycobacteria: Bugs and Bugbears (Two steps forward and one step back). *Mol Biotechnol.* 13:191-200

Parish, T. & Stoker, N.G. (1999b) Electroporation of Mycobacteria. In *Methods in Mol Microbiol, Vol 101: Mycobacteria Protocols*, pp. 129-143. Edited by T. Parish and N. Stoker. Humana Press Inc

Parish, T. & Wheeler, P. R. (1999). Preparation of Cell-Free extracts from Mycobacteria. In *Methods in Mol Microbiol, Vol 101: Mycobacteria Protocols*, pp. 77-89. Edited by T. Parish and N. Stoker. Humana Press Inc

Park, S.F., Purdy, D., Leach, S. (2000) Localized reversible frameshift mutation in the flhA gene confers phase variability to flagellin gene expression in *Campylobacter coli. J Bacteriol.* 182:207-10.

Pedrosa, J., Saunders, B.M., Appelberg, R., Orme, I.M., Silva, M.T., Cooper, A.M. (2000) Neutrophils Play a Protective Nonphagocytic Role in Systemic *Mycobacterium tuberculosis* Infection of Mice. *Infect Immun.* 68:577-583

Pelicic, V., Jackson, M., Reyrat, J.M., Jacobs, W.R. Jr, Gicquel, B., Guilhot, C. (1997) Efficient allelic exchange and transposon mutagenesis in *Mycobacterium tuberculosis. Proc Natl Acad Sci USA.* 94:10955-60.

Petes, T.D., Hill, C.W. (1988) Recombination between repeated genes in microorganisms. *Ann Rev Genet.* 22:147-168

Pfuetze, K.H., Pyle, M.M., Hinshaw, H.C., Feldman, W.H. (1955) The first clinical trial of streptomycin in human tuberculosis. *Am Rev Tuberc.* 71:752-754

Philipp, W.J., Poulet, S., Eiglmeier, K., Pascopella, L., Balasubramanian, V., Heym, B., Bergh, S., Bloom, B., Jacobs, W.J., Cole, S.T. (1996) An integrated map of the genome of the tubercle bacillus *Mycobacterium tuberculosis* H37Rv, and comparison with *Mycobacterium leprae. Proc Natl Acad Sci USA.* 93:3132-3137

Placido, R., Mancino, A., Amendola, A., Mariani, S., Vendetti, S., Piacentenni, M., Sanduzzi, A., Bocchino, M.L., Zembala. M., Colizzi, V. (1997) Apoptosis of human monocytes/macrophages in *Mycobacterium tuberculosis* infection. *J Pathol.* 181:31-38

Podglajen, I., Breuli, J., Coliatz, E. (1994) Insertion of a novel DNA sequence, IS*1186*, upstream of the silent carbapenemase gene *cfiA*, promotes expression of carbapenam resistance in clinical isolates of *Bacterioides fragilis. Mol Microbiol.* 12:105-114

Polard P, Prere MF, Chandler M, Fayet O. (1991) Programmed translational frameshifting and initiation at an AUU codon in gene expression of bacterial insertion sequence IS911. *J Mol Biol.* 222:465-77

Poulet, S., Cole, S.T. (1995) Characterization of the highly abundant polymorphic GC-rich-repetitive sequence (PGRS) present in *Mycobacterium tuberculosis. Arch Microbiol.* 163:87-95

Pym, A.S., Brosch, R. (2000) Tools for the Population Genomics of the Tubercle Bacilli. *Genome Res.* 10:1837-1839

R

Rado, T.A., Bates, J.H., Engel, H.W.B., Mankiewicz, E., Murohashi, T., Mizuguchi, Y., Sula, L. (1975) World Health Organization Studies on Bacteriophage Typing of Mycobacteria. *Am Rev Respir Dis.* 111:459-468

Radolf, J.D., Chamberlain, N.R., Clausell, A., Norgard, M.V. (1988) Identification and Localization of Integral Membrane Proteins of Virulent *Treponema pallidum* subsp. *pallidum* by Phase Partitioning with the Nonionic Detergent Triton X-114. *Infect Immun.* 56:490-498

Ramakrishnan, L., Federspiel, N.A., Falkow, S. (2000) Granuloma-Specific Expression of Mycobacterium Virulence Proteins from the Glycine-Rich PE-PGRS Family. *Science.* 288:1436-1439

Reyrat, J.M., Berthet, F.X., Gicquel, B. (1995) The urease locus of *Mycobacterium tuberculosis* and its utilization for the demonstration of allelic exchange in *Mycobacterium bovis* bacillus Calmette-Guerin. *Proc Natl Acad Sci USA.* 92:8768-72.

Rhee, J.T., Piatek, A.S., Small, P.M., Harris, L.M., Chaparro, S.V., Kramer, F.R., Alland, D. (1999) Molecular Epidemiologic Evaluation of Transmissibility and Virulence of *Mycobacterium tuberculosis*. *J Clin Microbiol.* 37:1764-1770

Rigouts, L., Portaels, F. (1994) DNA fingerprints of *Mycobacterium tuberculosis* do not change during the development of resistance to various anti-tuberculosis drugs. *Tuber Lung Dis.* 75:160

Riley, R.L., Mills, C.L., Nyka, W., Weinstock, N., Storey, P.B., Sultan, L.K., Riley, M.C., Wells, W.F. (1959) Aerial dissemination of pulmonary tuberculosis: a two year study of contagion in a tuberculosis ward. *Am J Hyg.* 70:185

Rindi, L. Lari, N., Garzelli, C. (1999) Search for Genes Potentially Involved in *Mycobacterium tuberculosis* Virulence by mRNA Differential Display. *Biochem Biophys Res Commun.* 258:94-101

Rivera-Marrero, C.A., Burroughs, M.A., Masse, R.A., Vannberg, F.O., Leimbach, D.L., Roman, J., Murtagh, J.J. (1998) Identification of genes differentially expressed in *Mycobacterium tuberculosis* by differential display PCR. *Microb Pathog.* 25:307-316

Robertson, B.D., Meyer, T.F. (1992) Genetic variation in pathogenic bacteria. *TIG.* 8:422-427

Robertson, H.E. (1933) Persistence of tuberculous infection. *Am J Pathol.* 9:711

Rodrigues, L., Smith, P. (1990) Tuberculosis in developing countries and methods for its control. *Trans R Soc Trop Med Hyg.* 1990:739-744.

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

Romagnani, S. (1997) The Th1/Th2 paradigm. *Immunol Today.* 18:263-266

Romain, F., Horn, C., Pescher, P., Namane, A., Rivierre, M., Puzo, G., Barzu, O., Marchal, G. (1999) Deglycosylation of the 45/47-kilodalton antigen complex of *Mycobacterium tuberculosis* decreases its capacity to elicit *in vivo* or *in vitro* cellular immune responses. *Infect Immun.* 67:5567-5572

Román, M.C., Sicilia, M.J.L. (1984) Preliminary investigation of *Mycobacterium tuberculosis* biovars. *J Clin Microbiol.* 20:1015-1016

Ross, B.C., Raios, K., Jackson, K., Dwyer, B. (1992) Molecular Cloning of a Highly Repeated DNA Element from *Mycobacterium tuberculosis* and Its Use as an Epidemiological Tool. *J Clin Microbiol.* 30:942-946

Roth, J.R. (1996) Rearrangements of the bacterial chromosome formation and applications. In: *Escherichia coli* and *Salmonella typhimurium*: Cellular and Molecular Biology. (Editors: F.C. Neidhart *et al.*), 2nd edition, Volume 2, pp 2256-2276. American Society for Microbiology, Washington, DC

S

Salamon, H., Kato-Maeda, M., Small, P.M., Drenkow, J., Gingeras, T.R. (2000) Detection of Deleted Genomic DNA Using a Semiautomated Computational Analysis of GeneChip Data. *Genome Res.* 10:2040-2050

Sambrook, J., Fritsch, E.F., Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual, Second Edition. (Cold Spring Harbor Laboratory)

Sampson, S.L., Warren, R.M., Richardson, M., Van der Spuy, G.D., Van Helden, P.D. (1999) Disruption of coding regions by IS*6110* insertion in *Mycobacterium tuberculosis*. *Tuber Lung Dis.* 79:349-359

Sampson, S.L., Warren, R.M., Richardson, M., Van der Spuy, G.D., Van Helden, P.D. (2001) IS*6110* insertions in *Mycobacterium tuberculosis*: predominantly into coding regions. *J Clin Microbiol.* 39:3423-3424

Saunders, N.A. (1999) Strain Typing of *Mycobacterium tuberculosis*. *J Infect.* 38:80-86

Saunders, B., Cooper, A.M. (2000) Restraining mycobacteria: Role of granulomas in mycobacterial infections. *Immunol Cell Biol.* 78:334-341

Saunders, B.M., Frank, A.A., Orme, I.M. (1999) Granuloma formation is required to contain bacillus growth and delay mortality in mice chronically infected with *Mycobacterium tuberculosis*. *Immunology.* 98:324-328

Schluger, N.W., Rom, W.N. (1998) The Host Immune Response to Tuberculosis. *Am J Respir Crit Care Med.* 157:679-691

Schulzer, M., Fitzgerald, J.M., Enarson, D.A., Grzybowski, S. (1992) An estimate of the future size of the tuberculosis problem in sub-Saharan Africa resulting from HIV infection. *Tuber Lung Dis.* 73:52-58

Seah, G.T., Scott, G.M., Rook, G.A.W. (2000) Type 2 Cytokine Gene Activation and Its Relationship to Extent of Disease in Patients with Tuberculosis. *J Infect Dis.* 181:385-9

Seibert, F.B., Glenn, J.T. (1941) Tuberculin purified protein derivative. Preparation and analyses of a large quantity for standard. *Am Rev Tuberc.* 44:9-25

Seifert, H.S., Wright, C.J., Jerse, A.E., Cohen, M.S., Cannon, J.G. (1994) Multiple gonococcal pilin antigenic variants are produced during experimental human infections. *J Clin Invest.* 93:2744-2749

Serkin, C.D., Seifert, H.S. (1998) Frequency of pilin antigenic variation in *Neisseria gonorrhoeae*. *J Bacteriol.* 180:1955-1958.

Shinnick, T.M. (1987) The 65-kilodalton antigen of *Mycobacterium tuberculosis*. *J Bacteriol.* 169:1080-1088

Shinnick, T.M., Good, R.C. (1994) Mycobacterial Taxonomy. *Eur J Clin Microbiol Infect Dis.* 13:884-901

Singh, K.K., Zhang, X., Patibandla, A.S., Chien, P., Laal, S. (2001) Antigens of *Mycobacterium tuberculosis* Expressed during Preclinical Tuberculosis: Serological Immunodominance of Proteins with Repetitive Amino Acid Sequences. *Infect Immun.* 69:4185-4191

Skeiky, Y.A.W., Ovendale, P.J., Jen, S., Alderson, M.R., Dillon, D.C., Smith, S., Wilson, C.B., Orme, I.M., Reed, S.G., Campos-Neto, A. (2000) T Cell Expression Cloning of a *Mycobacterium tuberculosis* Gene Encoding a Protective Antigen Associated with the Early Control of Infection. *J Immunol.* 165:7140-7149

Small, P.M., Moss, A. (1993) Molecular Epidemiology and the New Tuberculosis. *Infectious Agents and Disease*. 2:132-138

Small, P.M., Hopewell, P.C., Singh, S.P., Paz, A., Parsonnet, J., Ruston, D.C., Schecter, G.F., Daley, C.L., Schoolnik, G.K. (1994) The epidemiology of tuberculosis in San Francisco. A Population-Based Study Using Conventional and Molecular Methods. *N Engl J Med.* 330:1703-1709

Smith, P.G., Moss, A.R., (1994) Epidemiology of Tuberculosis. In *Tuberculosis: Pathogenesis, Protection and Control.* pp. 47-59. Edited by B. Bloom, ASM Press, Washington, DC.

Smith, D., Hänsch, H., Bancroft, G., Ehlers, S. (1997) T-cell-independent granuloma formation in response to *Mycobacterium avium*: role of tumour necrosis factor-α and interferon-γ. *Immunology*. 92:413-421

Smith, D., Wiegeshaus, E., Balasubramanian, V. (2000) An Analysis of Some Hypotheses Related to the Chingelput Bacille Calmette-Guérin Trial. *Clin Infect Dis.* 31:S77-80

Smithwick, R.W. (1976) *Laboratory Manual for Acid-Fast Microscopy*. Center for Disease Control, Atlanta

Snapper, S. B., Melton, R. E., Mustafa, S., Kieser, T. & Jacobs, W. R. (1990) Isolation and characterisation of efficient plasmid transformation mutants of *Mycobacterium smegmatis*. *Mol Microbiol* 4:1911-1919

Sola, C., Filliol, I., Gutierrez, M.C., Mokrousov, I., Vincent, V., Rastogi, N. (2001) Spoligotype Database of *Mycobacterium tuberculosis*: Biogeographic Distribution of Shared Types and Epidemiologic and Phylogenetic Perspectives. *Emerg Infect Dis*. 7:390-396.

Sonnhammer, E.L.L., von Heijne, G., Krogh, A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. In *Proc. Sixth Int. Conf. on Intelligent Systems for Molecular Biology*. pp. 175-182, Edited by: J. Glasgow *et al*. AAAI Press.

Sreevatsan, S., Pan, X., Stockbauer, K.E., Connell, N.D., Kreiswirth, B.N., Whittam, T.S., Musser. J.M. (1997) Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci USA*. 94:9869-9874

Stead, W.W., Eisenach, K.D., Cave, M.D., Beggs, M.L., Templeton, G.L., Thoen, C.O., Bates, J.H. (1995) When Did *Mycobacterium tuberculosis* Infection First Occur in The New World? *Am J Respir Crit Care Med.* 151:1267-1268

Stead, W.W. (1997) The origin and erratic global spread of tuberculosis. *Clin Chest Med.* 18:65-77

Strässle, A., Putnik, J., Weber, R., Fehr-Merhof, A., Wüst, J., Pfyffer, G.E. (1997) Molecular Epidemiology of *Mycobacterium tuberculosis* Strains Isolated from Patients in a Human Immunodeficiency Virus Cohort in Switzerland. *J Clin Microbiol.* 35:374-378

Sturgill-Koszycki, S., Schlesinger, P.H., Chakraborty, P., Haddix, P.L, Collins, H.L., Fok, A.K., Allen, R.D., Gluck, S.L., Heuser, J., Russell, D.G. (1994) Lack of acidification in *Mycobacterium tuberculosis* phagosomes produced by exclusion of the vesicular proton-ATPase. *Science*. 263:678-681

Sugawara, I., Yamada, H., Kazumi, Y., Do, N., Otomo, K., Aoki, T., Mizuno, Udagawa, T., Tagawa, Y., H., Iwakura, Y. (1998) Induction of granulomas in interferon-γ gene-disrupted mice by avirulent but not by virulent strains of *Mycobacterium tuberculosis*. *J Med Microbiol.* 47:871-877

Supply, P., Magdalena, J., Himpens, S., Locht, C. (1997) Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. *Mol Microbiol.* 26:991-1003

T

Tekaia, F., Gordon, S.V., Garnier, T., Brosch, R., Barrell, B.G., Cole, S.T. (1999) Analysis of the proteome of *Mycobacterium tuberculosis* in silico. *Tuber Lung Dis.* 79:329-342

Thierry, D., Cave, M.D., Eisenach, K.D., Crawford, J.T., Bates, J.H., Gicquel, B., Guesdon, J.L. (1990a) IS*6110*, an IS-like element of *Mycobacterium tuberculosis* complex. *Nucleic Acids Res*. 18:188

Thierry, D., Brisson-Noël, A., Vincent-Lévy-Frébault, V., Nguyen, S., Guesdon, J.-L., Gicquel, B. (1990b) Characterization of a *Mycobacterium tuberculosis* Insertion Sequence, IS*6110*, and Its Application in Diagnosis. *J Clin Microbiol.* 28:2668-2673

Thierry, D., Chavarot, P., Marchal, G., Le Thi, K.T., Ho, M.L., Nguyen, N.L., Le, N.V., Ledru, S., Fumoux, F., Guesdon, J-L. (1995) *Mycobacterium tuberculosis* strains unidentified using the IS*6110* probe can be detected by oligonucleotides derived from the Mt308 sequence. *Res Microbiol.* 146:325-328

Triccas, J.A., Parish, T., Britton, W., J., Gicquel, B. (1998) An inducible expression system permitting the efficient purification of a recombinant antigen from *Mycobacterium smegmatis. FEMS Microbiol Lett.* 167:151-156

Tripathy, S.P. (1979) Trial of BCG vaccines in South India for tuberculosis prevention: First Report. *Bulletin of the World Health Organization*. 57:819-827

U

Upton, A., Mushtaq, A., Victor, T., Sampson, S., Smith, D.-M., van Helden, P., Sim, E. (2001) Arylamine N-acetyltransferase of *Mycobacterium tuberculosis* is a polymorphic enzyme and a site of Isoniazid Metabolism. *Mol Microbiol.* 42:309-317

V

Van Belkum, A., Scherer, S., Van Alphen, L., Verbrugh, H. (1998) Short-sequence DNA Repeats in Prokaryotic Genomes. *Microbiol Mol Biol Rev.* 62:275-293

Van Crevel, R., Karyadi, E., Preyers, F., Leenders, M., Kullberg, B.-J., Nelwan, R.H.H., van der Meer, J.W.M. (2000) Increased Production of Interleukin 4 by CD4$^+$ and CD8$^+$ T Cells from Patients with Tuberculosis Is Related to the Presence of Pulmonary Cavities. *J Infect Dis.* 181:1194-1197

Van der Woude, M., Braaten, B., Low, D. (1996) Epigenetic phase variation of the *pap* operon in *Escherichia coli. TIM*. 4:5-9

Van Embden, J.D.A., Cave, M.D., Crawford, J.T., Dale, J.W., Eisenach, K.D., Gicquel, B., Hermans, P., Martin, C., McAdam, R., Shinnick, T.M., Small, P.M. (1993) Strain Identification of *Mycobacterium tuberculosis* by DNA Fingerprinting: Recommendations for a Standardized Methodology. *J Clin Microbiol.* 31:406-409

Van Embden, J.D.A., Van Gorkom, T., Kremer, K., Jansen, R., Van der Zeijst, B.A.M., Schouls, L.M. (2000) Genetic Variation and Evolutionary Origin of the Direct Repeat Locus of *Mycobacterium tuberculosis* Complex Bacteria. *J Bacteriol.* 182:2393-2401

Van Rie, A., Warren, R.M., Beyers, N., Gie, R.P., Classen, C.N., Richardson, M., Sampson, S.L., Victor, T.C., Van Helden, P.D. (1999a) Transmission of a Multidrug-Resistant *Mycobacterium tuberculosis* Strain Resembling "Strain W" among Noninstitutionalized, Human Immunodeficiency Virus-Seronegative Patients. *J Infect Dis.* 180:1608-1615

Van Rie, A., Warren, R.M., Richardson, M., Victor, T.C., Gie, R.P., Enarson, D.A., Beyers, N., Van Helden, P.D. (1999b) Exogenous reinfection as a cause of recurrent tuberculosis after curative treatment. *N Engl J Med.* 341:1174-1179

Van Soolingen, D., Hermans, P.W.M., De Haas, P.E.W., Soll. D.R., Van Embden, J.D.A. (1991) Occurence and stability of Insertion Sequences in *Mycobacterium tuberculosis* Complex Strains: Evaluation of an Insertion Sequence-Dependent DNA Polymorphism as a tool in the Epidemiology of Tuberculosis. *J Clin Microbiol.* 29:2578-2586

Van Soolingen, D., De Haas, P.E.W., Hermans, P.W.M., Groenen, P.M.A., Van Embden, J.D.A. (1993) Comparison of Various Repetitive DNA Elements as Genetic Markers for Strain Differentiation and Epidemiology of *Mycobacterium tuberculosis*. *J Clin Microbiol.* 31:1987-1995

Van Soolingen, D., De Haas, P.E.W., Hermans, P.W.M., Van Embden, J.D.A. (1994) DNA fingerprinting of *Mycobacterium tuberculosis*. *Methods Enzymol*. 235:196-205

Van Soolingen, D., Hermans, P.W.M. (1995) Epidemiology of tuberculosis by DNA fingerprinting. *Eur Respir J.* 8:649s-656s

Van Soolingen, D., Qian, L., De Haas, P.E.W., Douglas, J.T., Traore, H., Portaels, F., Qing, H.Z., Enkhsaikan, D., Nymadawa, P., Van Embden, J.D.A. (1995) Predominance of a Single Genotype of *Mycobacterium tuberculosis* in countries of East Asia. *J Clin Microbiol.* 33:3234-3238

Van Soolingen, D., De Haas, P.E.W., Blumenthal, R.M., Kremer, K., Sluitjer, M., Pijnenburg, J.E., Schouls, L.M., Thole, J.E.R., Dessens-Kroon, M.W.G., P.M.A., Van Embden, J.D.A., Hermans, P.W.M. (1996) Host-Mediated Modification of *Pvu*II Restriction in *Mycobacterium tuberculosis*. *J Bacteriol.* 178:78-84

Vega-López, F., Brooks, L.A., Dockrell, H.M., De Smet, K.A.L., Thompson, J.K., Hussain, R., Stoker, N.G. (1993) Sequence and Immunological Characterization of a Serine-Rich Antigen from *Mycobacterium leprae*. *Infect Immun.* 61:2145-2153

Vera-Cabrera, L., Howard, S.T., Laszlo, A.., Johnson, W.M. (1997) Analysis of genetic polymorphism in the phospholipase region of *Mycobacterium tuberculosis. J Clin Microbiol.* 35:1190-1195

Victor, T.C., Jordaan, A.M., Van Rie, A., Van der Spuy, G.D., Richardson, M., Van Helden, P.D. (1999) Detection of mutations in drug resistance genes of *Mycobacterium tuberculosis* by a dot-blot hybridization strategy. *Tuber Lung Dis.* 79:343-348

W

Wall, S., Ghanekar, K., McFadden, J., Dale, J.W. (1999) Context-sensitive transposition of IS*6110* in mycobacteria. *Microbiology.* 145:3169-3176

Wangoo, A, Sparer, T., Brown, I.N., Snewin, V.A., Janssen, R., Thole, J., Cook, H.T., Shaw, R.J., Young, D.B. (2001) Contribution of Th1 and Th2 Cells to Protection and Pathology in Experimental Models of Granulomatous Lung Disease. *J Immunol.* 166:3432-3439

Warren, R., Hauman, J., Beyers, N., Richardson, M., Schaaf, H.S., Donald, P., Van Helden, P. (1996a) Unexpectedly high strain diversity of *Mycobacterium tuberculosis* in a high-incidence community. *SAMJ.* 86:45-49

Warren, R., Richardson, M., Sampson, S., Hauman, J.H., Beyers, N., Donald, P.R., Van Helden, P.D. (1996b) Genotyping of *Mycobacterium tuberculosis* with Additional Markers Enhances Accuracy in Epidemiological Studies. *J Clin Microbiol.* 34:2219-2224

Warren, R., Richardson, M., Van Der Spuy, G., Victor, T., Sampson, S., Beyers, N., Van Helden, P. (1999) DNA fingerprinting and molecular epidemiology of tuberculosis: Use and interpretation in an epidemic setting. *Electrophoresis.* 20:1807-1812

Warren, R.M., Sampson, S.L., Richardson, M., Van Der Spuy, G.D., Lombard, C.J., Victor, T.C., Van Helden, P.D. (2000) Mapping of IS*6110* flanking regions in clinical isolates of *Mycobacterium tuberculosis* demonstrates genome plasticity. *Mol Microbiol.* 37:1405-1416.

Warren, R.M., Richardson, M., Sampson, S., van der Spuy, G.D., Bourn, W., Hauman, J.H., Heersma, H., Hide, W., Beyers, N., van Helden, P.D. (*In press*) Molecular evolution of *Mycobacterium tuberculosis*: Phylogenetic reconstruction of clonal expansion. *Tuberculosis*

Wayne, L.G. (1976) Dynamics of submerged growth of *Mycobacterium tuberculosis* under aerobic and microaerophilic conditions. *Am Rev Respir Dis.* 114:807-811

Wayne, L.G. (1994) Dormancy of *Mycobacterium tuberculosis* and Latency of Disease. *Eur J Clin Microbiol Infect Dis.* 13:908-914

Webb, V., Davies, J. (1999) Antibiotics and antibiotic resistance in mycobacteria. In *Mycobacteria*: *Molecular Biology and Virulence,* pp. 287-306. Edited by C. Ratledge and J. Dale. Blackwell Science Ltd.

Weill-Hallé, B., Turpin, R. (1925) Premiers essais de vaccination antituberculeuse de l'enfant par le Bacille Calmette-Guérin (BCG). *Bull Mem Soc Med l'Hosp de Paris 1925.* 49:1589-1601

Weiser, J.N., Love, J.M., Moxon, E.R.(1989) The molecular mechanism of phase variation of *H. influenzae* lipopolysaccharide. *Cell. 59:657-665*

Wells, W.F. (1955) *Airborne Contagion and Air Hygiene.* Harvard University Press, Cambridge, Mass

White-Ziegler, C.A., Villapakkam, A., Ronaszeki, K., Young, S. (2000) H-NS controls *pap* and *daa* fimbrial transcription in *Escherichia coli* in response to multiple environmental cues. *J Bacteriol.* 182:6391-6400.

Williams, D.L., Waguespack, C., Eisenach, K., Crawford, J.T., Portaels, F., Salfinger, M., Nolan, C.M., Abe, C., Sticht-Groh, V., Gillis, T.P. (1994) Characterization of Rifampin Resistance in Pathogenic Mycobacteria. *Antimicrob Agents Chemother.* 38:2380-2386

World Health Organization / International Union Against Tuberculosis and Lung Disease. (1997) Anti-tuberculosis drug resistance in the world. WHO/TB/97.229., Geneva

World Health Organization (1998a) *WHO Weekly Epidemiological Record.* 73:40

World Health Organization. (1998b) Services in Tuberculosis control: Part II: Microscopy. WHO/TB/98.258., Geneva

Y

Yang, Z.H., de Haas, P.E.W., van Soolingen, D., van Embden, J.D.A., Andersen, Å.B. (1994) Restriction Fragment Length Polymorphism of *Mycobacterium tuberculosis* Strains Isolated from Greenland during 1992: Evidence of Tuberculosis Transmission between Greenland and Denmark. *J Clin Microbiol.* 32:3018-3025

Yang, Z.H., Mtoni, I., Chonde, M., Mwasekaga, M., Fuursted, K., Askgård, D.S., Bennedsen, J., De Haas, P.E.W., Van Soolingen, D., Van Embden, J.D.A., Andersen, Å.B. (1995) DNA Fingerprinting and

Phenotyping of *Mycobacterium tuberculosis* Isolates from Human Immunodeficiency Virus (HIV)-Seropositive and HIV-Seronegative Patients in Tanzania. *J Clin Microbiol.* 33:1064-4069

Yang, Z., Chaves, F., Barnes, P.F., Burman, W.J., Koehler, J., Eisenach, K.D., Bates, J.H., Cave, M.D. (1996) Evaluation of Method for Secondary DNA Typing of *Mycobacterium tuberculosis* with pTBN12 in Epidemiologic Study of Tuberculosis. *J Clin Microbiol.* 34:3044-3048

Yang, Z., Barnes, P.F., Chaves, F., Eisenach, K.D., Weis, S.E., Bates, J.H., Cave, M.D. (1998) Diversity of DNA Fingerprints of *Mycobacterium tuberculosis* Isolates in the United States. *J Clin Microbiol.* 36:1003-1007

Yang, Z.H., Ijaz, K., Bates, J.H., Eisenach, K.D., Cave, M.D. (2000) Spoligotyping and Polymorphic GC-Rich Repetitive Sequence Fingerprinting of *Mycobacterium tuberculosis* Strains Having Few Copies of IS*6110*. *J Clin Microbiol.* 38:3572-3576

Yeh, R.W., Ponce de Leon, A., Agasino, C.B., Hahn, J.A., Daley, C.L., Hopewell, P.C., Small, P.M. (1998) Stability of *Mycobacterium tuberculosis* DNA Genotypes. *J Infect Dis.* 177:1107-1111

Yuen, L.K.W., Ross, B.C., Jackson, K.M., Dwyer, B. (1993) Characterization of *Mycobacterium tuberculosis* Strains from Vietnamese Patients by Southern Blot Hybridization. *J Clin Microbiol.* 31:1615-1618

Z

Zainuddin, Z.F., Dale, J.W. (1989) Polymorphic Repetitive DNA sequences in *Mycobacterium tuberculosis* Detected with a Gene Probe from a *Mycobacterium fortuitim* Plasmid. *J Gen Microbiol.* . 135:2347-2355

Zhang, Y., Young, D. (1994) Strain variation in the *katG* region of *Mycobacterium tuberculosis*. *Mol Microbiol.* 14:301-308

Zhang, M., Lin, Y., Iyer, D.V., Gong, J., Abrams, J.S., Barnes, P.F. (1995) T-cell cytokine responses in human infection with *Mycobacterium tuberculosis*. *Infect Immun.* 63:3231-3234.

Zhang, M., Gong, J., Yang, Z., Samten, B., Cave, M.D., Barnes, P.F. (1999) Enhanced Capacity of a Widespread Strain of *Mycobacterium tuberculosis* to Grow in Human Macrophages. *J Infect Dis.* 179:1213-7

Zhang, Q., Wise, K.S. (2001) Couple Phase-Variable Expression and Epitope Masking of Selective Surface Lipoproteins Increase Surface Phenotypic Diversity in *Mycoplasma hominis*. *Infect Immun.* 69:5177-5181

Zheng, X, Teng, L.J., Watson, H.L., Glass, J.I., Blanchard, A., Cassell, G.H. (1995) Small repeating units within the *Ureaplasma urealyticum* MB antigen gene encode serovar specificity and are associated with antigen size variation. *Infect Immun.* 63:891-8

Zumárraga, M., Bigi, F., Alito, A., Romano, M.I., Cataldi, A. (1999) A 12.7 kb fragment of the *Mycobacterium tuberculosis* genome is not present in *Mycobacterium bovis*. *Microbiology*. 145:893-897

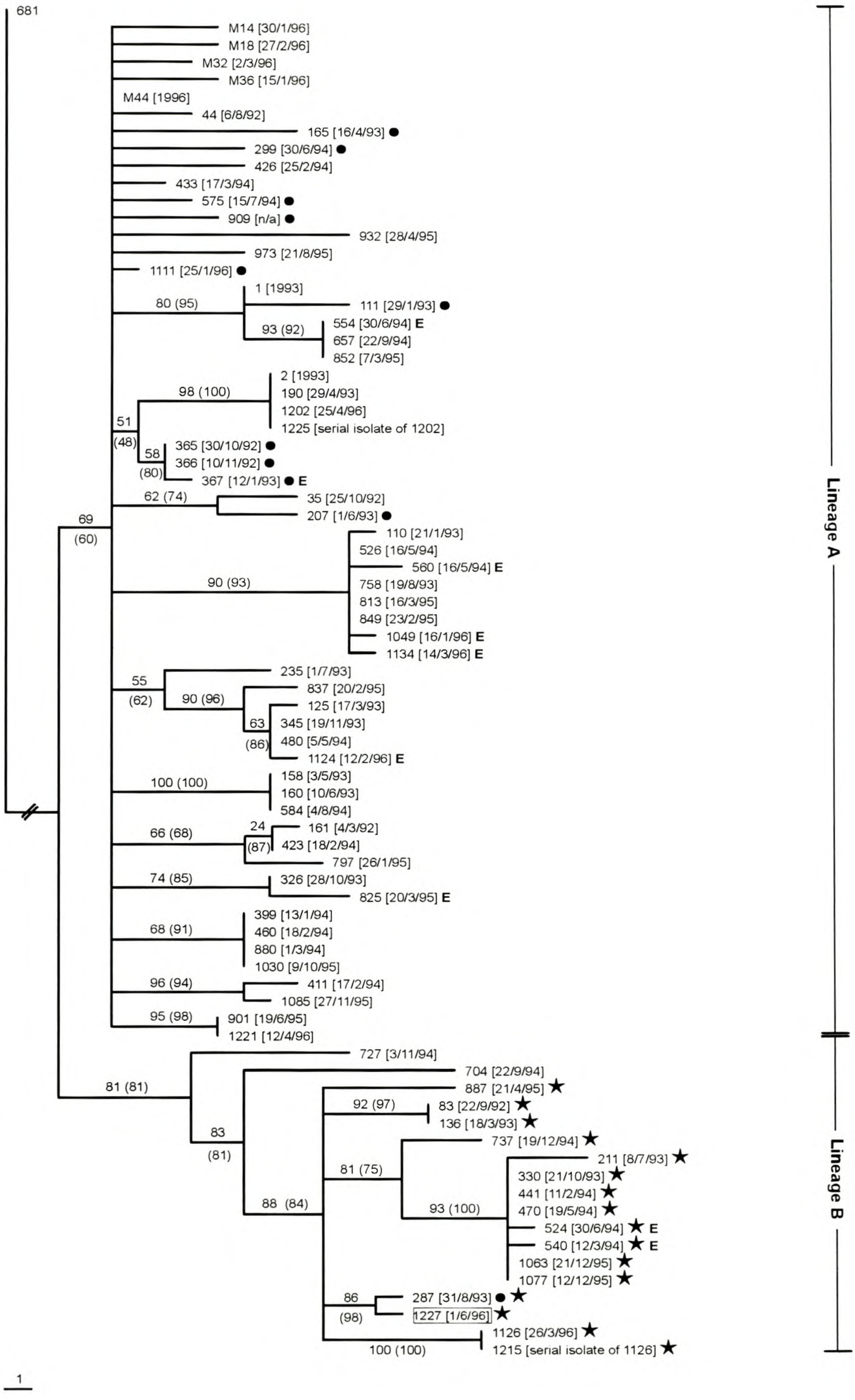
Appendix: Phylogenetic tree calculated from strain family 28 RFLP mutational data

(Reproduced from Warren *et al.*, 2001)

Majority rule consensus tree, rooted to outgroup isolate 681, was calculated using the heuristic algorithm or neighbour-joining distance algorithm (PAUP* 4.0). The tree length was 251 and the consistency index was 0.713 for both algorithms used. Bootstrap values are shown as internal labels, the values in brackets are the bootsrap values calculated using the neighbour-joining distance distance algorithm, while the non-bracketed values are the bootsrap values calculated using the heuristic parsimony algorithm. Nodes were only scored as significant if >50% of the trees contained the node. Branch lengths are proportional to the number of evolutionary steps (scale represents a single evolutionary step). Included at each peripheral node is the isolate number and the isolation date. Isolates originating from the distant rural community are labeled with the prefix M, while isolates from patients who reside outside of the urban study community but attended the primary health care clinic in the community are labeled with a closed circle. Isolates shown to be methylation negative phenotype are indicated with a star. Isolates showing concurrent genomic evolution with isolation date are labeled E.

The predecessor strain 1227 is boxed.

Appendix: Phylogenetic tree calculated from strain family 28 RFLP mutational data.



With magic, you can turn a frog into a prince.
With science, you can turn a frog into a Ph.D.,
and you still have the frog you started with

- The Science of Discworld

Terry Pratchett, Ian Stewart, Jack Cohen