THE UNIVERSITY OF WARWICK

Library Declaration and Deposit Agreement

1. STUDENT DETAILS

Please complete the following:

Full name: ……

University ID number: ……

2. THESIS DEPOSIT

2.1 I understand that under my registration at the University, I am required to deposit my thesis with the University in BOTH hard copy and in digital format. The digital version should normally be saved as a single pdf file.

2.2 The hard copy will be housed in the University Library. The digital version will be deposited in the University's Institutional Repository (WRAP). Unless otherwise indicated (see 2.3 below) this will be made openly accessible on the Internet and will be supplied to the British Library to be made available online via its Electronic Theses Online Service (EThOS) service.
[At present, theses submitted for a Master's degree by Research (MA, MSc, LLM, MS or MMedSci) are not being deposited in WRAP and not being made available via EthOS. This may change in future.]

2.3 In exceptional circumstances, the Chair of the Board of Graduate Studies may grant permission for an embargo to be placed on public access to the hard copy thesis for a limited period. It is also possible to apply separately for an embargo on the digital version. (Further information is available in the *Guide to Examinations for Higher Degrees by Research*.)

2.4 *If you are depositing a thesis for a Master's degree by Research, please complete section (a) below. For all other research degrees, please complete both sections (a) and (b) below:*

(a) Hard Copy

I hereby deposit a hard copy of my thesis in the University Library to be made publicly available to readers (please delete as appropriate) EITHER immediately OR after an embargo period of ……..... months/years as agreed by the Chair of the Board of Graduate Studies.

I agree that my thesis may be photocopied. YES / NO (*Please delete as appropriate*)

(b) Digital Copy

I hereby deposit a digital copy of my thesis to be held in WRAP and made available via EThOS.

Please choose one of the following options:

EITHER My thesis can be made publicly available online. YES / NO (*Please delete as appropriate*)

OR My thesis can be made publicly available only after…..[date] (Please give date)
YES / NO (*Please delete as appropriate*)

OR My full thesis cannot be made publicly available online but I am submitting a separately identified additional, abridged version that can be made available online.
YES / NO (*Please delete as appropriate*)

OR My thesis cannot be made publicly available online. YES / NO (*Please delete as appropriate*)

3. GRANTING OF NON-EXCLUSIVE RIGHTS

Whether I deposit my Work personally or through an assistant or other agent, I agree to the following:

Rights granted to the University of Warwick and the British Library and the user of the thesis through this agreement are non-exclusive. I retain all rights in the thesis in its present version or future versions. I agree that the institutional repository administrators and the British Library or their agents may, without changing content, digitise and migrate the thesis to any medium or format for the purpose of future preservation and accessibility.

4. DECLARATIONS

(a) I DECLARE THAT:

- I am the author and owner of the copyright in the thesis and/or I have the authority of the authors and owners of the copyright in the thesis to make this agreement. Reproduction of any part of this thesis for teaching or in academic or other forms of publication is subject to the normal limitations on the use of copyrighted materials and to the proper and full acknowledgement of its source.

- The digital version of the thesis I am supplying is the same version as the final, hard-bound copy submitted in completion of my degree, once any minor corrections have been completed.

- I have exercised reasonable care to ensure that the thesis is original, and does not to the best of my knowledge break any UK law or other Intellectual Property Right, or contain any confidential material.

- I understand that, through the medium of the Internet, files will be available to automated agents, and may be searched and copied by, for example, text mining and plagiarism detection software.

(b) IF I HAVE AGREED (in Section 2 above) TO MAKE MY THESIS PUBLICLY AVAILABLE DIGITALLY, I ALSO DECLARE THAT:

- I grant the University of Warwick and the British Library a licence to make available on the Internet the thesis in digitised format through the Institutional Repository and through the British Library via the EThOS service.

- If my thesis does include any substantial subsidiary material owned by third-party copyright holders, I have sought and obtained permission to include it in any version of my thesis available in digital format and that this permission encompasses the rights that I have granted to the University of Warwick and to the British Library.

5. LEGAL INFRINGEMENTS

I understand that neither the University of Warwick nor the British Library have any obligation to take legal action on behalf of myself, or other rights holders, in the event of infringement of intellectual property rights, breach of contract or of any other right, in the thesis.

Please sign this agreement and return it to the Graduate School Office when you submit your thesis.

Student's signature: .....…… Date: .....

The diversity and distribution of *Mycobacterium* species in varying ecological and climatic environments

Tanya Khera

A thesis submitted to the University of Warwick for the degree of Doctor of Philosophy

School of Life Sciences

University of Warwick,

Coventry, CV4 7AL

September 2012

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my supervisors Professor Elizabeth Wellington and Dr. Orin Courtenay for their guidance and support during my time at Warwick.

I am grateful to past and present members of Professor Elizabeth Wellington's research group and members of Ecology and Epidemiology for their advice, in particular Brian Oakley, Sam Mason, Selene Huntley, Vinca Russell, Nikolaos Kyratsous and Emma Travis. I wish to also thank members of the Armauer Hansen Research Institute in Ethiopia, particularly Girum Erenso, Abraham Aseffa and Rea Tschopp for their guidance and assistance during fieldwork. I would like to acknowledge members of Rob Knight's research group particularly Tony Walters for helpful discussions regarding the analysis of pyrosequencing data.

I would like to thank Simon Butler, my parents, my housemates Hannah Lee and Luci Witcomb for their encouragement and constant support. Finally I would like to thank Simon Butler, Karina Pall and Sam Mason for proof reading my thesis!

Declaration

I hereby declare that all the results presented in this thesis were obtained by myself under the supervision of Professor Elizabeth Wellington and Dr. Orin Courtenay, unless stated otherwise. This thesis has not been submitted for a degree in any other institution. All sources of information presented in this thesis have been acknowledged by a reference.

Summary

The species within the genus *Mycobacterium* are commonly detected in a variety of environments including soil, water and dust. Many species within the group are capable of causing opportunistic diseases and are hypothesised to be responsible for the reduction in BCG efficacy in tropical countries.

Consequently it is important to understand the diversity and biogeography of mycobacteria in the environment. Soil and water samples were collected from a total of 42 residential sites in 9 different climatic regions. To determine community composition, community DNA was extracted and amplicon pyrosequencing was employed to target the 16S rRNA gene of the *Mycobacterium* genus and slow-growing mycobacteria. Quantitative PCR was employed to quantify the total abundance of *Mycobacterium* species and specifically members of the *M. tuberculosis* complex.

The study revealed a greater diversity of both fast-growing and slow-growing mycobacteria than previously reported. Prevalent species in soil were closely related to the fast growers *M. neglectum*, *M. moriokaense* and the slow growers *M. malmoense* and *M. colombiense*, in contrast to water had a high abundance of sequences related to the fast growers *M. aurum sp*. ATCC 23070, *M. neoaurum* and the slow-growers *M. gordonae* and *M. colombiense*. The abundance of the *Mycobacterium* genus ranged from 3.35×10^{1} to 8.01×10^{8} gene copies per gram/ml. *M. bovis* was detected in six environmental samples using qPCR.

Biogeographical analysis demonstrated the importance of elevation and temperature for the community composition of mycobacteria in soil. A non-linear relationship was observed between elevation and the outcome variables *Mycobacterium* species richness, diversity and abundance with a peak mid-elevation. In contrast latitude was the primary factor to explain the composition and diversity of mycobacteria in water samples.

To our knowledge this is the first time that the diversity and abundance of mycobacteria has been elucidated on a large geographical scale using pyrosequencing and multivariate analyses. Results indicate ample opportunity for human exposure to mycobacteria with potentially pathogenic species in soil and water substrates. These results have implications for the risk of infection and similar biogeographical surveys on a worldwide scale may provide improved correlations with BCG vaccine efficacy

List of Abbreviations

AIC	Akaike information criterion
ATCC	American Type Culture Collection
ATP	Adenosine-5'-triphosphate
BCG	Bacillus Calmette-Guerin
bp	Base pair
BLAST	Basic Local Alignment Search Tool
BSA	Bovine serum albumen
BTB	Bovine tuberculosis
CCA	Canonical Correspondence Analysis
CCD	Charge coupled device
CC	Correlation Coefficient
CI	Confidence Interval
Ct	cycle threshold
DGGE	Denaturing Gradient Gel Electrophoresis
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
DSMZ	Deutsche Sammlung von Mikroorganismen und Zellkulturen
EM	Environmental Mycobacteria
FG	Fast-growing mycobacteria
GLM	Generalised Linear Model
H’	Shannon diversity index
MAC	*M. avium* complex
MAIS	*Mycobacterium avium- M. intracellulare- M.scrofulaceum*
ML	Maximum likelihood
NGS	Next Generation Sequencing
NIC	No inhibition control
NTM	Non-tuberculous Mycobacteria
OR	Odds-Ratio
OTU	Operational Taxonomic Unit
PAH	Polycyclic Aromatic Hydrocarbon
PCoA	Principle Coordinate Analysis
PCR	Polymerase chain reaction
PHYML	Phylogenetic estimation using Maximum Likelihood
PPD	Purified Protein Derivative
qPCR	Quantitative polymerase chain reaction
QIIME	Quantative Insights into Microbial Ecology
SG	Slow-growing mycobacteria
SOLiD	Support Oligonucleotide Ligation detection
TB	Tuberculosis
TIRF	Total Internal Reflection Fluorescence
WHO	World Health Organisation

Chapter 1: General Introduction

1.1. Taxonomy and physiology of the *Mycobacterium* genus

The *Mycobacterium* genus belongs to the family *Mycobacteriaceae*, the order *Actinomycetales*, the phylum *Actinobacteria* and kingdom *Bacteria*. According to the List of Prokaryotic names with Standing in Nomenclature (LSPN) there are currently 154 *Mycobacterium* species (Euzeby 1997). Species within the *Mycobacterium* genus are commonly termed environmental mycobacteria (EM), nontuberculous (NTM), opportunistic or atypical mycobacteria. The *Mycobacterium tuberculosis* complex (MTC) is considered a separate group belonging to the *Mycobacterium* genus.

Mycobacteria are gram-positive, aerobic, acid-alcohol fast, G-C rich, non-motile and rod-shaped (Falkinham 2009b). A cell is typically between 0.2-0.6µm wide and 1-10µm long. The cell wall is comprised of a plasma membrane, a peptidoglycan layer, a hydrophobic thick mycolate layer and outer lipids (Barry et al. 1998; Brennan & Nikaido 1995). Species are separated on the basis of growth into fast-growing mycobacteria (FG) and slow-growing mycobacteria (SG); taking 2-7 days and 15-28 days to culture on solid medium respectively (Shinnick & Good 1994). Mycobacteria can be further subdivided into four Ruyon groups; the first three contain SG (Ruyon 1959). Selected mycobacteria produce carotenoid organic pigment and this is used to characterise these species. Group 1 are the photochromogens, which only produce pigment in the presence of light. Group 2 are the scotochromogens, which produce pigment regardless of light. Group 3 are non-chromogens, which do not produce pigment or exhibit a very pale yellow pigment. Finally Group 4 contains the FG, which do not produce a pigment. The Adansonian classification of *Mycobacterium*

species enables the differentiation of species based on biochemical and physiological characteristics such as enzyme activity, mycolic acid patterns and growth temperature (Bojalil et al. 1962; Levy-Frebault & Portaels 1992; Rastogi et al. 2001).

There are distinct phylogenetic separations in the *Mycobacterium* group which are informative for classification (Figure 1.1). Differences in the 16S rRNA gene can be used to differentiate between the FG and SG. For example a large group of SG contain an extended long helix at *Escherichia coli* positions 451 to 482 whereas all the FG and some SG contain a short helix (Rogall et al. 1990; Stahl & Urbance 1990). The long helix 18 slow growers included many of the opportunistic pathogens compared to the other groups (Leclerc et al. 2003). Additionally, the majority of FG has two 16S rRNA gene copies whereas the SG has one copy. FG and SG can also be phylogenetically separated based on the 16S-23S rRNA internal transcriber spacer region (Roth et al. 1998). The *Mycobacterium tuberculosis* complex within the slow-growing group, is comprised of several pathogenic species including the primary causative agent of tuberculosis (TB); *M. tuberculosis*. The genome of the *M. tuberculosis* complex is >96% similar and species within the complex are identical in the 16S rRNA gene (Chaves et al. 2010). The species within it are known to have sequentially evolved through genomic deletions (Figure 1.2) (Brosch et al. 2002; Mostowy et al. 2002; Smith et al. 2009). The *Mycobacterium avium* complex (MAC) within the slow-growing group comprises of eight species, where 16S rRNA gene sequence similarity ranges from 96.3-99.7% (Cayrou et al. 2010; Salah et al. 2009a). The 16S rRNA gene of *Mycobacterium* species is highly similar, in some cases it is not possible to resolve to the species level (Figure 1.1). However, it is still possible to use the 16S rRNA gene to draw overall conclusions about the diversity of mycobacteria.

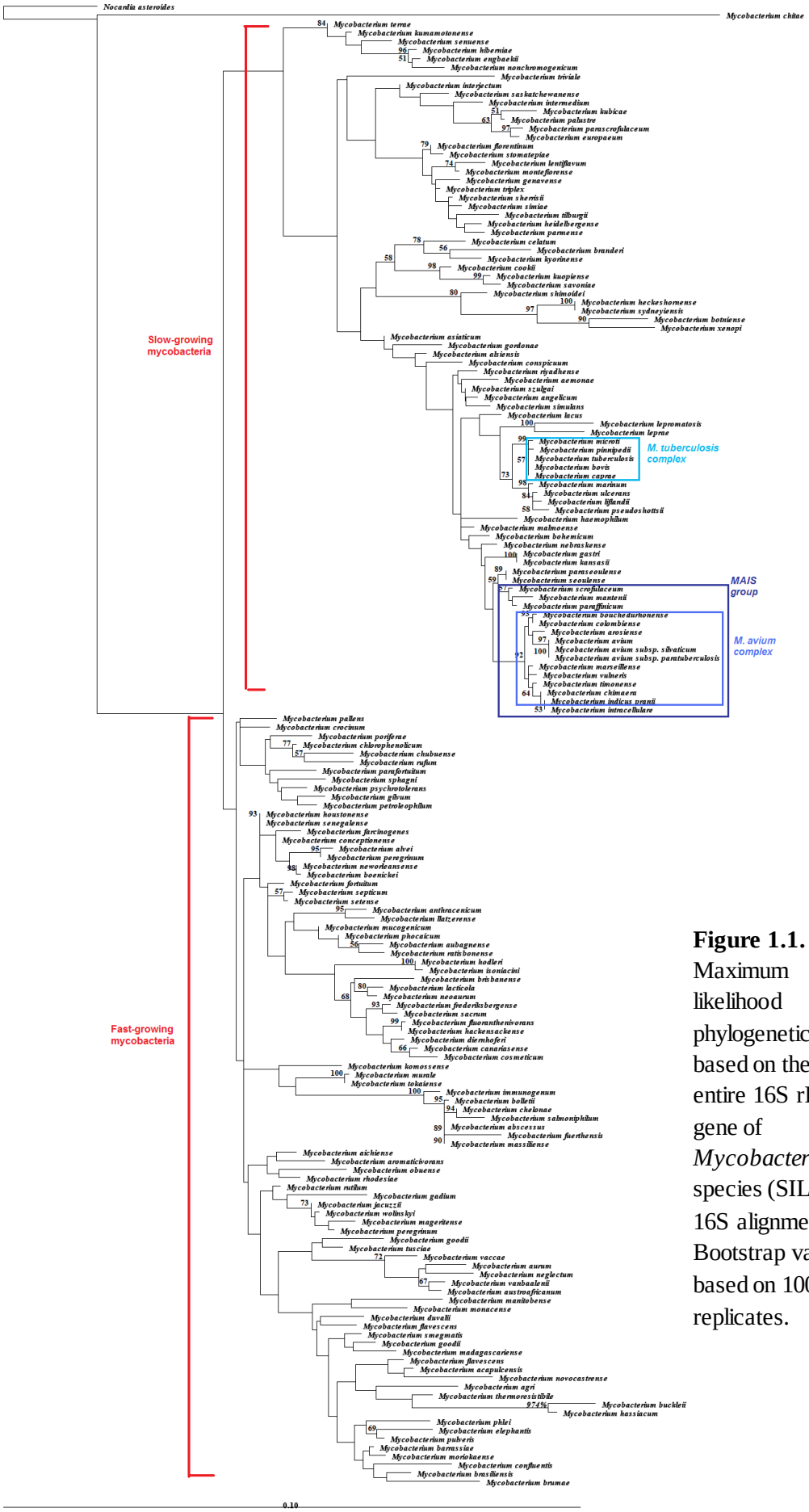


Figure 1.1. Maximum likelihood phylogenetic tree based on the entire 16S rRNA gene of *Mycobacterium* species (SILVA 16S alignment). Bootstrap values based on 1000 replicates.

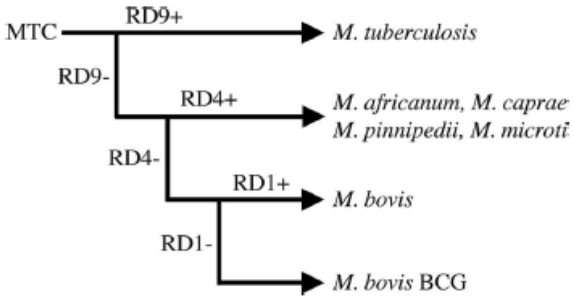


Figure 1.2. Phylogenetic evolution of the *M. tuberculosis* complex. Taken from (Pinsky & Banaei 2008).

1.2. Interactions between mycobacteria, the host and the environment

A spectrum of *Mycobacterium* species are capable of causing opportunistic disease in humans, livestock and wildlife. The concentrated interactions of hosts and the environment is thought to increase the risk of infection. Mycobacteria have several mechanisms which enable this group of species to survive in the environment some of which parallel virulence mechanisms required for intracellular survival within the human host.

1.2.1. Human clinical significance of *Mycobacterium* species

Although many *Mycobacterium* species are primarily saprophytic, a considerable number have been isolated as the causative agent of infection and can cause a wide range of clinical manifestations (Table 1.1). Disease is more prevalent in those with immunocompromising conditions, however mycobacteria can also cause infection in immunocompetent individuals (Piersimoni 2009).

Table 1.1. Types of diseases and causative agents, adapted from mycobacteria reviews; (Crump et al. 2009; Falkinham 1996, 2002; Marras 2002; Primm et al. 2004; Wagner & Young 2004; Wallace et al. 1997; Wallace et al. 1998)**.**

Clinical Disease	Common Causative species	Unusual Causative Species
Pulmonary Disease	*M.tuberculosis M.abscessus M.avium complex M.kansasii M.malmoense M.xenopi*	*M.asiaticum M.branderi M.celatum M.fortuitum M.gordonae M.haemophilum M.intermedium M.lentiflavum M.magdeburgensis M.shimodii M.simiae M.smegmatis M.szulgai*
Lymphadenitis	*M.avium complex M.malmoense M.scrofulaceum*	*M.abscessus M.bohemicum M.chelonae M.fortuitum M.haemophilum M.heidelbergense M.interjectum M.kansasii M.lentiflavum M.tusciae*
Cutaneous Disease	*M.abscessus M.chelonae M.fortuitum M. leprae M.marinum M.ulcerans*	*M.haemophilum M.kansasii M.malmoense M.smegmatis*
Disseminated Disease	*M.avium complex M.chelonae M.haemophilum M.kansasii M.scrofulaceum*	*M.abscessus M.conspicuum M.fortuitum M.genavense M.malmoense M.marinum M.sherrisii M.simiae M.triplex M.xenopi*
Nosocomial Disease	*M.abscessus M.fortuitum M.chelonae*	*M.aurum M.avium M.gordonae M.mucogenicum M.neoaurum M.simiae M.smegmatis M.xenopi*

Pulmonary disease is the most prevalent localised form of the *Mycobacterium* diseases and is characterised by chronic cough, sputum production, weight loss and fatigue. Apart from *M. tuberculosis*, many *Mycobacterium* species have been isolated from patient sputum (Morita et al. 2005). Pre-existing lung diseases such as bronchiectasis, pneumoconiosis, cystic fibrosis and emphysema have been found to predispose individuals to pulmonary infection (Cook 2010; Torrens et al. 1998; Wickremasinghe et al. 2005); in these cases the lung may become a medium for mycobacteria to proliferate. A cohort of elderly women with no predisposing factors developed MAC pulmonary disease (Lady Windermere Syndrome), it has been hypothesised that this is due to cough suppression causing inflammation and infection (Reich and Johnson 1992).

Lymphadenitis is predominantly a disease suffered by children, characterised by swelling and inflammation of the lymph nodes (Wolinsky 1992). Cases have no systemic illness and have normal chest X-rays, treatment involves excision of the infected lymph node and may include antibiotic therapy (Benson-Mitchell & Buchanan 1996).

Cutaneous diseases are generally characterised by lesions affecting the integumentary system and can be caused by a variety of SG and FG. Leprosy is a chronic granulatomous disease infecting the skin and peripheral nerves (Britton et al. 2004). It is the second most common mycobacterial disease after pulmonary tuberculosis. The causative agent *M. leprae* is slow-growing and is transmitted through respiratory droplets. Buruli Ulcer is a disease of the subcutaneous fat tissue; cases typically exhibit necrotic lesions (Stienstra et al. 2001) and it has become increasingly prevalent in West Africa and Australia in the recent decade. Unlike other *Mycobacterium* species, the causative agent *M. ulcerans* produces the toxin mycolactone which is a factor in the organisms virulence (George et al. 1999). Environmental reservoirs of *M. ulcerans* include riverine swamps, soil and vegetation (Ross et al. 1997). The species was also detected in water insects, mosquitoes and possums (Fyfe et al. 2010; Johnson & Lavender 2009; Portaels et al. 1999). It is thought that infection is caused by transmission from these environmental sources and animal reservoirs (Stienstra et al. 2001). Skin and soft tissue infections most commonly found in developed countries include those caused by the FG the *M. fortuitum* group and the *M. chelonae-M. abscessus* group which have derived from earlier posttraumatic wounds (Brown-Elliott & Wallace 2002). The slow-grower *M. marinum* is the causative agent of a cutaneous infection which is known to be transmitted from swimming pools, jacuzzis and fish tanks (Collins 1985).

Disseminated disease also referred to as extrapulmonary occurs when the responsible mycobacteria spreads via the lymphatic or blood circulation to other organs (Horsburgh et al. 1993). Disseminated diseases are becoming increasingly more common in HIV-AIDS patients (Nunn et al. 2005; Nunn et al. 2007). The HIV epidemic has changed the pattern of TB incidence worldwide. While the epidemic disease was previously pulmonary or localised to the skin and was a disease of older males (Falkinham 1996), there has been an increase in disseminated cases, primarily caused by MAC. It is estimated that there are five million individuals worldwide that are TB and HIV coinfected (Wittes 2000). Those who are immunocompromised due to HIV characterised by CD4+ T cell depletion are more at risk of TB and specifically reactivation of latent TB. Therefore HIV infection is one of the most important risk factors for TB.

Nosocomial *Mycobacterium* species infections have been predominantly recognised in developed countries, particularly among cardiac surgery patients and those that develop post-injection abscesses (Wallace et al. 1998). Hospital acquired infections are mostly caused by the FG (Cook 2010). It is thought that patients are more predisposed due to immunocompromising conditions and water systems and contaminated surgical instruments may be the source of transmission (Phillips & von Reyn 2001).

The prevalence of *Mycobacterium* species infection and disease worldwide is uncertain due to the lack of effective and rapid diagnostics. It is concerning that in some cases *M. tuberculosis* is misdiagnosed as the causative agent of smear-positive pulmonary TB cases. This has implications for treatment, as effective treatment regimes are different for *Mycobacterium* species diseases compared to *M. tuberculosis* diseases (Wallace et al. 1997). It is estimated that 10% of TB patients

are infected with mycobacteria other than *Mycobacterium tuberculosis* (Collins 1989). More recent estimates suggest that incidence rates of NTM in industrialised countries vary from 1.0-1.8 cases per 100,000 persons (Griffith et al. 2007; Horsburgh 1996).

1.2.2. Domestic and Wildlife mycobacterial disease

Bovine tuberculosis (bTB) is a chronic respiratory disease of cattle characterised by lesions present on the surface of the lung and chest cavity. The etiologic agent of bTB is *M. bovis*, a member of the *M. tuberculosis* complex. Although many countries have now eradicated bTB, such as Sweden and Norway, there are still a number of countries where it is still prevalent and an established economic problem, including the UK, Republic of Ireland and New Zealand. In the UK the national herd incidence was estimated to be 6.87% and between 2005-2006 £90 million was spent on bTB control measures (Proud 2006). Transmission of *M. bovis* is thought to occur via inhalation of aerosol droplets. It is reasoned that the persistence of bovine tuberculosis is due in part to wildlife reservoirs. Examples of wildlife reservoirs include the Eurasian badger in the UK and Ireland, the white-tail deer in Michigan USA, the brush-tail possum in New Zealand and boar and deer in France and Spain (Aranaz et al. 2004; Corner 2011; Donnelly et al. 2006; O'Brien et al. 2002) .

Johnes disease also referred to as paratuberculosis is caused by *Mycobacterium avium paratuberculosis* (MAP). It is primarily a disease of ruminants, characterised by the clinical symptoms of diarrhoea and wasting, however it is often subclinical (Manning & Collins 2001). The disease has been detected worldwide; a review of Johne's disease prevalence in Europe found estimates were as high as 24% in Germany, whereas in the UK it ranged from 2.6-3.5% (Nielsen & Toft 2009). The annual economic loss due to paratuberculosis in the UK is estimated at £12.9 million

(SAC 2002).Wildlife reservoirs include rabbit, weasels, stoat and foxes (Beard et al. 2001). Transmission is thought to occur via faecal contamination of feed or through infected milk. The species *M. avium paratuberculosis* has also been implicated in Crohn's disease; however this remains a contentious issue (Sanderson et al. 1992).

Avian mycobacteriosis causes wasting and produces nodules on the liver, spleen, intestine and bone marrow (Tell et al. 2001). The disease affects both poultry, wild and exotic birds. Causative agents include *M. genavense, M. avium* subsp. *avium* and *M. intracellulare.* In addition, *M. avium* subsp. *hominissuis* is the causative agent of lymphadenitis in pigs and in other mammals. Reports of *M. avium* disease have also been found in horses, cats, dogs, deer, goats and sheep (Thorel et al. 2001).

1.2.3. Human-Animal-Environment transmission interface

Unlike *M. tuberculosis*, the transmission of other *Mycobacterium* species is not thought to occur through person to person contact, but through contact with contaminated environments and wildlife reservoirs. It is of particular concern that there is a high risk of transmission in areas where humans, livestock and wildlife are in close contact. For example a study found that increased knowledge of mycobacterial diseases was associated with sharing water sources with wild animals (Kankya et al. 2010). Concern has been raised in countries where *M. bovis* is relatively prevalent and milk is not pasteurised; one study found an increased risk of active TB in farmers that consumed raw milk (Regassa et al. 2008).

1.2.4. Intracellular survival within protozoa and macrophages

Several *Mycobacterium* species have been shown to proliferate within amoebae in the environment (Adekambi et al. 2006b). The mechanisms of entry and replication in amoebae parallel phagocytosis by macrophages, as a result it is suggested that

virulence factors of mycobacteria have evolved from the interaction of amoebae in the environment. *Mycobacterium* species were shown to replicate within the ameobal vacuole and selected species could survive within the exocyst (Medie et al. 2011; Salah & Drancourt 2010; Thomas & McDonnell 2007). Survival within the ameobal cyst has enabled mycobacteria to resist chlorine disinfection (Salah et al. 2009b; Whan et al. 2006) and display reduced susceptibility to antimicrobials (Miltner & Bermudez 2000). In addition *M. avium* growth within *Acanthamoeba castellani* was shown to enhance its virulence compared to *M. avium* grown in broth (Cirillo et al. 1997). Specific virulence determinants have been identified that enable mycobacteria to enter and replicate within amoebae and macrophages. For example the pathogenicity island present in the *M. avium* genome was shown to be necessary for intracellular uptake of *M. avium* (Danelishvili et al. 2007). The ESX-1 secretion system encoded by the RD1 region was shown to be important for intracellular replication, where mutant *M. marinum* strains were less able to replicate in amoeba compared to wild type strains (Kennedy et al. 2012). In agreement the secretion system was shown to be involved in virulence by modulating the host immune response (DiGiuseppe Champion & Cox 2007). A total of 20 genes were shown to be upregulated in *M. avium* during ameobal infection, 8 of which are known to be upregulated during phagocytosis (Tenant & Bermudez 2006). The life cycle of mycobacteria within macrophages has been associated with many additional virulence determinants including genes involved in inhibition of apoptosis (Velmurugan et al. 2007), components such as phenolic glycolipids known to reduce levels of proinflammatory cytokines (Ehrt & Schnappinger 2007), secretory immunogenic proteins such as ESAT-6 (Gordon et al. 2009; Wards et al. 2000) and resuscitation promoting factors involved in the reactivation after dormancy (Kana et

al. 2008). The formation of spores was identified in mycobacteria (Ghosh et al. 2009; Lamont et al. 2012), which provides a mechanism for survival in extreme environments, however spore formation in mycobacteria is still an issue of contention (Traag et al. 2010). Overall, these virulent mechanisms discussed enable mycobacteria to survive successfully within the intracellular environment.

1.2.5. Human sensitisation to *Mycobacterium* species

A survey of naval recruits found higher sensitisation to mycobacteria in southern parts of the Unites States compared to northern parts (Edwards et al. 1969). This is in concordance with previous environmental surveys which found increased numbers of *Mycobacterium avium-M. intracellulare-M.scrofulaceum* (MAIS) isolates in soils and waters from south-eastern United States compared to north-eastern parts (Brooks et al. 1984; Falkinham 1980). A cross-country comparison of the sensitisation of individuals to the *M. avium* complex (MAC) found higher skin test sensitisation in Kenya (30%) compared to Trinidad (13%), the US (12%) and Finland (3%) (von Reyn et al. 1993a), suggesting higher background exposure to environmental mycobacteria in Kenya.

In a countrywide sensitivity study conducted in Malawi, widespread sensitivity to *Mycobacterium* antigens was observed in a cohort of 36,000 individuals (Fine et al. 2001). Individuals had greatest sensitivity to *M. intracellulare* and *M. scrofulaceum* antigens and the lowest sensitivity to *M. xenopi*. Sensitivity to all antigens was shown to increase with age, indicating cumulative exposure to mycobacteria over time. Individuals with prior sensitivity to antigens of FG but not slow- growers were associated with lower risks of both leprosy and TB. It may be that FG are better adjuvants for protective immune responses than SG.

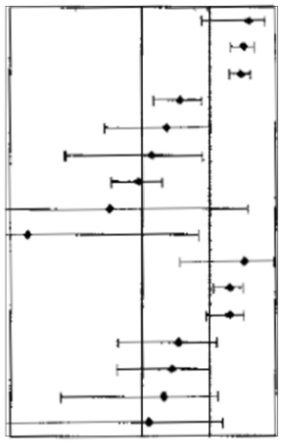
Few studies have aimed to directly correlate human exposure to mycobacteria with species diversity in the immediate locality. An attempt was made in Malawi, however the study only definitively identified *M. fortuitum* in the environment. This is thought to be reflective of the methods and not of the true *Mycobacterium* species diversity (Chilima et al. 2006). High recovery rates of *M. leprae* were detected in soil samples collected from 'leprosy patient areas' (55%) compared with 'no-patient areas' (15%), providing correlative evidence for a potential transmission route (Lavania et al. 2008). Similar *Mycobacterium* species were detected in potting soils and from patients with pulmonary infections (De Groote et al. 2006). It was suggested that potting soil aerosols were a potential source of *Mycobacterium* species exposure. Correlations were observed between mycobacteria in peat moss air and plant worker exposure (Cayer et al. 2007). This suggests the environment is a reservoir of exposure and subsequent infection to mycobacteria.

1.3. The BCG vaccine

The Bacillus Calmette-Guerin (BCG) vaccine is one of the most widely used vaccines worldwide; it has been implemented for over 75 years and over 100 million doses are administered annually. It was developed at the Institute Pasteur de Lille by Albert Calmette and Camille Guerin between 1906 and 1919 (Bonah 2005). Through 230 passages of *Mycobacterium bovis* over glycerinated bile potato medium, the live attenuated strain was no longer pathogenic but had retained its antigenicity (Liu et al. 2009; Oettinger et al. 1999). The efficacy results from vaccine trials have shown that BCG confers variable protection, and as such its effectiveness has become an issue of debate. Among many hypotheses, it is suggested that exposure to environmental mycobacteria reduces BCG vaccine efficacy.

1.3.1. BCG vaccine efficacy and global administrative policy

Meta-analysis studies for TB have shown that the vaccine has a protective effect of 50% (Brewer 2000; Colditz et al. 1994). Studies on the duration of protection suggest that the efficacy of BCG wanes over a period of 10-14 years (Sterne et al. 1998; Weir et al. 2008b). Meta-analysis studies for childhood miliary tuberculosis and meningitis have all shown that BCG confers good protection (efficacy 75-86%) (Rodrigues et al. 1993). It has been estimated that the vaccine prevented 41,215 childhood tuberculosis cases in 2002 (Trunz et al. 2006). BCG is also shown to provide good protection against leprosy (Ponnighaus et al. 1992). However, the vaccine has caused great controversy due to its variable protection against adult pulmonary TB (0-80%) in different geographical areas (Wilson et al. 1995). This is illustrated in Figure 1.3, which shows that BCG is found to confer little to no protection in countries that are closer to the equator against adult pulmonary TB compared to those which are further from the equator. It was calculated that latitude explained 41% of the variance observed between efficacy estimates (Colditz et al. 1994).



Study location	Study Type	Vaccine strain	Total cases
Haiti	CT	Frappier	18
England	CT	Copenhagen	203
Native Americans, USA	CT	Phippa	249
Puerto Rico	CT	Park	295
Mandanapalle	CT	Madras	80
Georgia, USA	CT	Tice	60
Chingleput, India	CT	Copen & Pasteur	533
Georgia, USA	CT	Tice	7
Illinois, USA	CT	Tice?	19
England*	CC	Glaxo	33
Cameroon	CC	Pasteur	213
Argentina*	CC	Mixed	131
Indonesia	CC	Japan & Pasteur	88
Kenya	CC	Glaxo	126
Colombia*	CC	Copenhagen	178
Karonga, Malawi	COH	Glaxo	23

Figure 1.3. Estimates of BCG efficacy for pulmonary disease from clinical trials (CT), case control (CC) and cohort (COH) studies. Taken from (Fine et al. 1999). The BCG vaccine efficacy with 95% confidence intervals are shown from -100% to 100%. Asterisks denote studies of pulmonary disease in children.

Consequently, the immunisation policies for the BCG vaccine vary globally. A review of 180 countries found that 157 recommend routine use of the BCG vaccine and 23 have either stopped the universal BCG vaccination entirely or have never initiated widespread vaccination (Zwerling et al. 2011). Additional differences between countries include the age of administration, the use of a booster dose and the strain of the BCG vaccine. Primarily developing countries are included within the Expanded Program on Immunisation (EPI), which routinely uses BCG vaccination at birth. Countries such as the Czech Republic, Ukraine and Croatia recommend vaccination at birth and a subsequent booster. Low incidence countries such as the UK and France no longer universally administer the BCG vaccine, but only vaccinate high risk groups. Other locations have never implemented a universal BCG program such as the USA and Canada; these countries rely on prevention and treatment.

1.3.2. Hypotheses for reduced BCG vaccine efficacy

Reasons for the variable BCG efficacy remain uncertain, however several hypotheses have been proposed. These include the variation of BCG strains, differences in administration of vaccine, the genetics and nutrition of different populations and exposure to mycobacteria (Fine 2001).

Attempts to propagate the strain under different laboratory conditions worldwide have produced phenotypically and genomically different daughter strains. Currently the most commonly used sub-strains include Pasteur, Copenhagen, Glaxo, Tokyo, Russian and Moreau strains (Fine et al. 1999). Genomic analyses of strains have found variations between ‘early’ strains such as Russia and ‘late strains’ such as Glaxo, for example variation in tandem duplications DU1 and DU2. As such it is hypothesised that the attenuation has reduced the efficacy of BCG (Brosch et al. 2007). However, the comparative human vaccine trial in Malawi and the UK used

the same vaccine strain in both countries but found differences in the protection conferred (Black et al. 2002). In addition, two different strains of BCG were used in the Chingleput trial and neither conferred protection (IMCR 2006).

Vaccine factors such as dose, boosters and route of administration may contribute to the variation in efficacy that is observed. For example, the storage of BCG in tropical countries may affect its efficacy, where sunlight is thought to decrease the effectiveness (Edwards & Tolderlund 1952). Variation in the routes of BCG administration ranging from the oral route to the intradermal needle, jet injection and multiple puncture methods may affect its efficacy (Roche et al 1995). Doses of BCG have been found to change the type of immunity induced; lower BCG doses were found to induce a Th1 type response whereas higher doses were found to induce a mixed Th1/Th2 response (Power et al. 1998).

Genetic susceptibilities are hypothesised to contribute to the differences observed between different populations with regard to the incidence of TB disease and BCG efficacy. Several genes that control the cellular immune mechanisms, for example the IFN-γ receptor polymorphisms and SLC11A1 (NRAMP1) polymorphisms, were shown to affect the susceptibility to TB (Blackwell et al. 1994; Dorman & Holland 1998). Nutritional differences are considered to affect the protectiveness normally afforded by BCG, for example vitamin D is suggested to provide a immunoregulatory role (Lalor et al. 2011).

There is considerably more evidence to support the hypothesis that exposure to mycobacteria is responsible for the variable efficacy of the BCG vaccine (Fine 2001). Two ways in which *Mycobacterium* species are hypothesised to interfere with BCG efficacy are (1) blocking where previous immunity induced by *Mycobacterium*

species restricts growth of the BCG and (2) masking where BCG is unable to provide additional immunity to that already induced by mycobacteria (Fine 1995). Exposure to mycobacteria may cause inappropriate priming of the immune system leading to subsequent poor vaccine efficacy. Both animal and human studies have endeavoured to prove this.

1.3.3. Animal models show reduced BCG vaccine efficacy

Early guinea pig experiments suggested that exposure to environmental mycobacteria provided some level of protection and masked the protection afforded by BCG; *M. fortuitum*, *M. avium* and *M. kansasii* imparted 15%, 50% and 85% as much protection as BCG respectively (Fine 1995; Palmer & Long 1966). Brandt and colleagues sensitised mice to several *Mycobacterium* species and the mice were then subjected to BCG vaccination (Brandt et al. 2002). It was shown that multiplication of BCG in the liver and spleen was reduced by ten to thirty times. This demonstrated that *Mycobacterium* species are capable of blocking the protective immune response normally induced by BCG. Further research has aimed to model the current EPI guidelines by the vaccination of mice before subsequent gavages of *M. avium*, in order to emulate vaccination at birth followed by environmental exposure to mycobacteria (Flaherty et al. 2006). Results demonstrated an increase in the number of *M. tuberculosis* CFU present on the lungs in *M. avium* exposed BCG vaccinated mice compared to control BCG vaccinated mice, suggesting BCG is no longer protective.

Attempts to understand the immunological mechanisms have measured several postulated protective markers. Murine studies have shown that *M. avium* exposure before BCG vaccination reduces IFN-γ production compared to control BCG vaccinated mice, where IFN-γ is associated with a Th1 protective cell mediated

response (Demangel et al. 2005; Young et al. 2007). Additional research has shown that mice sensitised to *M. chelonae* before BCG vaccination induced a $CD4^+$ T cell cytotoxic effect, so that the number of *M. bovis* BCG cells in the lung were reduced by six-fold compared to non-sensitised mice (Ho et al. 2009). It is suggested that mice are unable to establish a protective immune response to *M. tuberculosis* infection due to the depletion of BCG cells. Further research by the same group found that *M. chelonae* sensitised mice which were subsequently BCG vaccinated had higher levels of the anti-inflammatory cytokine IL-10 and $CD4^+$ $CD25^+$ regulatory T cells which also suppressed the inflammatory response to BCG challenge (Ho et al. 2010). This provides an immunological mechanism for which mycobacteria may change the host response to BCG.

Variable BCG efficacy estimates have been observed in the cattle population; New Zealand calves were found to have high IFN-γ responses to *M.* avium protein purified derivative (PPD) suggesting prior exposure to mycobacteria (Buddle *et al.* 2002). After BCG vaccination, cattle were found to have minimal protection against virulent *M. bovis* compared to the non-vaccinated group.

1.3.4. Human vaccine trials show variable BCG efficacy

An early trial in Chingleput, Madras, India began in 1968 and was instigated by the Council of Medical Research and World Health Organisation (WHO). The randomised control trial included over 260,000 individuals receiving either one of two BCG strains (French or Danish) or a placebo. After 7.5 years BCG conferred no protection, with a similar number of pulmonary TB cases observed in each group (Anonymous 1979). It was found that 95% of cases were reactors to *Mycobacterium* species before vaccination, leading to the suggestion that the low efficacy of BCG was due to the effect of exposure to mycobacteria (Smith et al. 2000). After a 15 year

follow up, BCG was still not protective in the population, however it was found to confer a small amount of protection (32%) in individuals who were not reactors to *Mycobacterium* species (IMCR 2006).

As part of a long-term study in Karonga, Malawi ("Karonga Prevention Study") several aspects of mycobacterial disease epidemiology were elucidated, including the low efficacy of the BCG vaccine and human exposure to *Mycobacterium* species (Crampin et al. 2009). A randomised control study in the UK and Malawi, where BCG vaccine efficacy is 50%-80% and 0% respectively, studied the comparative immunology before and one year after vaccination (Black et al. 2002). Prior to vaccination Malawians (61%) had a higher IFN-γ response to tuberculin PPD compared to the UK (23%). Post vaccination the BCG attributable increase in IFN-γ responses was higher in the UK (60%) compared to Malawi (17%). This demonstrated the correlation between prior high sensitisation to mycobacteria and BCG efficacy. Other immunological correlates, such as the anti-inflammatory cytokine IL-10, were higher in response to tuberculin PPD in Malawi compared to the UK after BCG vaccination (Weir et al. 2004). It was suggested that IL-10 prevents the release of IFN-γ, which in turn is necessary for protection against TB, suggesting a reason why Malawi has poorer efficacy estimates. Higher background exposure to mycobacteria in Malawi compared to the UK was further demonstrated by IFN-γ responses to *Mycobacterium* species PPDs (Weir et al. 2006). A randomised control trial based solely in Malawi compared the effects of BCG vaccination in those with a naturally high exposure to MAIS to those with a lower exposure (Black et al. 2001). The study found those with higher MAIS exposure had lower protective IFN-γ responses to tuberculin PPD. In agreement with the between country comparisons, the within country comparisons also demonstrated the high

levels of mycobacteria exposure negatively effects the protectiveness of the BCG vaccine.

1.4. Importance of *Mycobacterium* species in bioremediation

Many FG are well known polycyclic aromatic hydrocarbons (PAH) microbial degraders (Miller et al. 2004). These PAH compounds are produced from incomplete combustion of organic matter. Sources of PAH production include forest fires, indoor heating, burning of waste and fossil fuels. The organic compounds contain two or more benzene aromatic rings, and due to their low aqueous solubility are commonly found in the soil environment. These are environmental pollutants which persist in the environment and are carcinogenic, teratogenic and mutagenic (Haritash & Kaushik 2009; Kanaly & Harayama 2000). Thus microbes play an essential role in the degradation of these pollutants. Mycobacteria are commonly found in PAH contaminated sites, and those capable of biodegradation include *M. flavescens, M. frederiksbergense, M. holderi* and *M. vanbaalenii* (Dean-Ross & Cerniglia 1996; Kleespies et al. 1996; Stingley et al. 2004; Willumsen et al. 2001). Therefore there is great potential for the use of mycobacteria for bioremediation purposes.

1.5. *Mycobacterium* species in the environment

The environment is of interest as a source of exposure, infection and for bioremediation. Several reviews have documented the wide range of reservoirs of mycobacteria and demonstrate that mycobacteria are commonly detected in the soils, water, aerosols and dust (Dailloux et al. 1999; Falkinham 2009a; van Ingen et al. 2009a; Vaerewijck et al. 2005). The commonly detected species as determined by a review of 25 published studies in both soil and water include *M. fortuitum, M. flavescens, M. chelonae, M. terrae* and *M. kansasii* (Table 1.2). However there are

some differences between soil and water communities, for example *M. gordonae* was commonly detected in water studies (74% - 14/19) compared to soil studies (23% - 3/13) (Table 1.2).

A high prevalence and diversity of mycobacteria was detected in boreal forest and alpine soils (Iivanainen et al. 1997; Kopecky et al. 2011; Niva et al. 2006). Estimates for the total *Mycobacterium* species load in boreal soils range from 10^7 to 10^8 cells per gram of soil (Nieminen et al. 2006). PAH contaminated soils also displayed a high diversity of mycobacteria and this has highlighted the importance of FG in biodegradation (Leys et al. 2005; Miller et al. 2004; Uyttebroek et al. 2006). A limited number of diversity studies were done in lower latitude countries; both the Malawian and Indian study identified *M. fortuitum* as the predominant isolate recovered from soil and dust (Chilima et al. 2006; Kamala et al. 1994).

Mycobacteria have been isolated from a wide variety of water sources, for example Parisian and South African studies identified species in water distribution systems (Le Dantec et al. 2002b; September et al. 2004; Torvinen et al. 2004). A unique study which compared water distribution systems in four countries, the US, Finland, Zaire and Kenya, found that MAC isolation rates were higher in Finland and the US compared to Zaire and Kenya (Von Reyn et al. 1993b). Hospital water systems have been shown to contain a high diversity of mycobacteria, particularly of FG known to cause opportunistic disease (Chang et al. 2002; Galassi et al. 2003; Hussein et al. 2009). Mycobacteria have also been isolated from swimming pools and hot tubs, particularly the species *M. marinum* and *M. jacuzzi* (Leoni et al. 1999). Water damaged buildings and cooling towers were shown to have a high *Mycobacterium* genus abundance (10^3 to 10^6 cells per litre) this suggests this is a niche for

mycobacteria (Adrados et al. 2011; Torvinen et al. 2006). The diversity of mycobacteria has also been ascertained in natural water sources such as freshwater rivers, ground water, lakes (Bland et al. 2005; Covert et al. 1999; Lee et al. 2008; Niva et al. 2006). Mycobacteria abundance was quantified from coastal estuary waters and ranged from 0 to 10^3 per ml (Jacobs et al. 2009).

Table 1.2. A review of 25 published studies; 19 studies include water and 13 studies include soil. Numbers refer to the number of published studies that have identified the *Mycobacterium* species. The table shows the 25 most identified *Mycobacterium* species identified in published studies. (Bland et al. 2005; Chang et al. 2002; Chilima et al. 2006; Covert et al. 1999; De Groote et al. 2006; Eddyani et al. 2008; Galassi et al. 2003; Ghaemi et al. 2006; Hussein et al. 2009; Kamala et al. 1994; Kankya et al. 2011; Le Dantec et al. 2002b; Lee et al. 2008; Leoni et al. 1999; Leys et al. 2005; Mendum et al. 2000; Michel et al. 2007; Narang et al. 2009; Niva et al. 2006; Norby et al. 2007; Parashar et al. 2009; Rahbar et al. 2010; September et al. 2004; Thorel et al. 2004; Torvinen et al. 2004)

***Mycobacterium* spp.**	**Soil Studies**	**Water Studies**	**Sum of Studies**
M. fortuitum	10	14	24
M. gordonae	3	14	17
M. flavescens	7	6	13
M. chelonae	5	6	11
M. terrae	5	6	11
M. kansasii	4	5	9
M. smegmatis	4	4	8
M. intracellulare	2	5	7
M. nonchromogenicum	2	5	7
M. avium	4	2	6
M. phlei	2	4	6
M.avium complex	2	3	5
M. abscessus	3	2	5
M. peregrinum	2	3	5
M. asiaticum	2	2	4
M. gadium	2	2	4
M. lentiflavum	1	3	4
M. mucogenicum	1	3	4
M. scrofulaceum	1	3	4
M. simiae	1	3	4
M. szulgai	2	2	4
M. thermoresistible	3	1	4
M. vaccae	1	3	4
M. gastri	1	2	3
M. septicum	2	1	3

1.6. Biogeography of *Mycobacterium* species

Perhaps the most documented environmental correlate for diversity is pH. Specific surveys of the MAIS complex found significantly higher counts with lower pH in acid brown swamps and soils from Eastern United States (Brooks et al. 1984; Kirschner 1992). In agreement, higher numbers of isolates were detected in acidic environments such as alpine forest soils, brook sediments and pasture soils (Iivanainen et al. 1999; Norby et al. 2007; Thorel et al. 2004).

An additional important environmental correlate for diversity and abundance is seasonal temperature. Community profiles in acidic forest soils were significantly different in the summer compared to the winter season (Kopecky et al. 2011). Higher numbers of *Mycobacterium* species isolates were obtained from UK water samples in June (16°C-19°C) compared to December (11°C-16°C), and it was reasoned that active multiplication occurs at warmer temperatures (Donoghue et al. 1997). However, a South Indian study found reduced numbers of isolates in the summer (35°C-40°C) (Kamala et al. 1994), and it was suggested that mycobacteria can not thrive at extreme temperatures. In agreement, a prevalence study of a freshwater river environment found a negative correlation between temperature and total *Mycobacterium* species counts (CC = -0.462, P = 0.023) (Bland et al. 2005).

The effects of moisture are not well researched, however a Malawian survey found higher numbers of mycobacteria positive samples in the dry season compared to the wet season (Chilima et al. 2006). In addition, mycobacteria were recovered from alpine soil with low moisture content (Thorel et al. 2004). However, mycobacteria were commonly detected in moisture damaged buildings (Torvinen et al. 2006).

Other abiotic factors such as the chemical properties of soil and water environments have been correlated with *Mycobacterium* abundance. High carbon levels in brook sediments were significantly associated with an increased number of isolates (Iivanainen et al. 1999). In coastal water estuaries, higher *Mycobacterium* species load was correlated with increased nitrogen content, lower salinity and reduced concentrations of dissolved oxygen (Jacobs et al. 2009).

In summary, lower pH, intermediate warmer drier seasons and high nutrient environments are associated with higher abundance and diversity of mycobacteria. It is important to acknowledge that species within the *Mycobacterium* genus have to some extent different optima for abiotic factors and this may explain some of the differences between studies.

1.7. Methods for assessing *Mycobacterium* species diversity

Currently there is no single standard protocol for the isolation and identification of mycobacteria present in environmental samples. It is problematic to identify mycobacteria due to the complexity of the environmental sample, the presence of many other species and inhibition factors such as humic acids. Methods used for diversity studies in the environment include traditional culture, biochemical tests and molecular methods. There have also been more recent advances that have enabled a more comprehensive insight into microbial diversity such as next generation sequencing (NGS) technologies.

1.7.1. Traditional Cultivation

Cultivation of *Mycobacterium* species from the environment frequently involves decontamination and growth on Lowenstein-Jenson slants. Cultivation can be time consuming given that speed of growth can vary from 3 days to several months.

Selective decontamination is carried out before cultivation and includes the addition of chemicals such as sodium hydroxide and oxalic acid to the environmental sample, Decontamination is necessary to ensure the clearance of competing rapidly growing bacteria. However a balance is required as harsh treatment removes less resistant mycobacteria. For example it was shown that using 3% sodium dodecyl sulphate and 1% sodium hydroxide for decontamination reduced the number of mycobacteria to 1% of the original number (Le Dantec et al. 2002a). A separate assessment found the addition of sodium thiosulfate reduced the number of positive mycobacteria samples by 9.1% (Thomson et al. 2008). Therefore when developing a protocol which includes culturing mycobacteria it is appropriate to vary concentrations of decontaminants to ensure maximum survival of isolates.

Biochemical and physiological characteristics are often used to differentiate between different environmental isolates; for example characterisation using tests for pigmentation, speed of growth, nitrate reduction and Tween 80 hydrolysis (Donoghue et al. 1997; Kamala et al. 1994; Rastogi et al. 2001). Biochemical tests have been under scrutiny for several reasons including the likelihood of erroneous identification and inability to differentiate between undescribed taxa (Springer et al. 1996). Discrepancies were observed between the results of biochemical and molecular tests of 20/34 clinical mycobacteria isolates. It was found that this was due to inaccuracies of biochemical methods and interpretation (Springer et al. 1996). The accuracy and the time-consuming nature of cultivation techniques have made way for new methods to be developed.

1.7.2 Molecular Methods

DNA amplification techniques such as Denaturing Gradient Gel Electrophoresis (DGGE), Restriction Fragment Length Polymorphism (PCR-RFLP), PCR-Sanger

sequencing and mycolic acid separation using High Performance Liquid Chromatography (HPLC) have been used extensively in microbial ecology. Molecular genotypic methods are now widely used for rapid diagnostics of clinical samples and for environmental samples.

DGGE is a powerful technique that has been successfully applied in environmental microbiology to assess microbial diversity, spatial-temporal changes and environmental conditions (Ferris et al. 1996; Heuer et al. 1997; Muyzer & Smalla 1998; Ward et al. 1998). DGGE enables the separation and subsequent differentiation of PCR amplified DNA products that are the same size but differ in sequence (Fischer & Lerman 1983). This is dependent on the electrophoretic potential of the lowest melting domain of the DNA sequence. Migration through the polyacrylamide gel will dramatically slow as the DNA molecule reaches the denaturant concentration required for melting and due to the change in conformation. A 40bp guanine-cytosine rich portion (GC clamp) is added to one of the primers to serve as the highest melting domain. This prevents complete separation of the double helix as the DNA molecule migrates through the gel (Myers et al. 1985). DGGE has been used to determine the mycobacteria diversity in environmental samples; the studies successfully identified a wide diversity of known species and environmental isolates (Leys et al. 2005; Niva et al. 2006).

PCR-RLFP relies on restriction enzymes that digest DNA into smaller fragments, which can subsequently be separated by gel electrophoresis on an agarose gel. The technique has been widely used for the identification of *Mycobacterium* species, particular from clinical samples. *Mycobacterium* genomic targets for restriction analysis have included the gene encoding heat shock protein 65-kDa protein (*hsp65*), the 16S-23S rRNA internal transcriber spacer (16S-23S ITS) region and the 16S

rRNA gene (Parashar et al. 2009; Telenti et al. 1993; Vaneechoutte et al. 1993). Using the 16S-23S ITS region has advantages over using *hsp65* target sequence as it shows higher intraspecies stability in RLFP pattern (Roth et al. 2000). The 16S rRNA region is the most commonly used sequence for species level identification and as such the sequence database for the 16S rRNA gene is most comprehensive. However, the 16S rRNA gene is known to have a high degree of conservation in mycobacteria and therefore in some cases are indistinguishable, for example the *M. tuberculosis* complex has an identical 16S rRNA gene (Stinear et al. 2004).

Chain termination Sanger sequencing determines the genomic sequence of the DNA which has been used as an accurate method to identify *Mycobacterium* species. It has been widely used in *Mycobacterium* diversity studies of the soil and water environments, in which DNA is often extracted from the environmental cultured isolates (Le Dantec et al. 2002b; September et al. 2004). In both studies the relatively low number of positive samples and diversity found in water samples may be attributed to decontamination methods. It is therefore attractive to bypass decontamination by extracting DNA directly from the soil. Chilima compared PCR carried out directly on soil/water samples and decontamination/culturing of the samples before PCR (Chilima et al. 2006). Detection levels for both methods were similar with the total number of positive samples being 48% for direct PCR methods and 51% for the decontamination and culture methods. Therefore in this case the molecular and isolation methods were in broad agreement. However, only strains of *M. fortuitum* were detected in the soil. The low diversity found may be due to a number of factors; it may be that the cultivation techniques favoured the isolation of *M. fortuitum* and therefore other species were not detected.

HPLC has been used to differentiate between mycobacteria by the analysis of patterns generated by mycolic acids. The mycolic acids are high molecular weight ß-hydroxyl fatty acids with long alkyl chains present on the cell wall of mycobacteria (Barry et al. 1998). This technique is more commonly employed in the clinical setting with the identification of clinical isolates in pure culture (Butler et al. 1991; Glickman et al. 1994; Thibert & Lapierre 1993; Tortoli et al. 2001). However, members of the *M. tuberculosis* complex are indistinguishable using this method (Butler et al. 1991). The application of this technique is limited as it requires a high abundance of one species or a few species, this is not often the case in environmental samples.

As with other techniques the molecular methods discussed have limitations for example the methods are qualitative/semi-quantitative compared. It is possible that less prevalent species in the sample will not be detected due to PCR amplification bias towards more abundant species. However, one study showed the DGGE method could be used to detect species that were 1% abundant (Muyzer et al. 1993).

1.7.3. Introduction to high throughput sequencing technologies

Next generation sequencing (NGS) technologies have become widely accessible to researchers over the past twelve years, before this traditional Sanger sequencing had dominated the genomics field. The introduction of massive parallel sequencing enables millions of sequences to be read, in contrast to Sanger sequencing which is limited by 96 capillary arrays (Hert et al. 2008). NGS increased the daily throughput by a factor of between 100-1000 and the reduced cost of sequencing one million nucleotides to 0.1% to 4% of Sanger sequencing (Kircher & Kelso 2010).

1.7.3.1. 454 Roche pyrosequencing

Pyrosequencing next generation technologies was developed by Mostafa Ronaghi & Pal Nyren in 1996 at the Royal Institute of Technology, Stockholm (Ronaghi et al. 1996). The pyrosequencing platform is currently licensed to 454 Life Sciences, Roche.

Pyrosequencing begins with the construction of a DNA template library. A PCR reaction creates correct size specific amplicons and fusion primers enable the adapters and ligators required for pyrosequencing to be incorporated into the sequences. The amplicons are then denatured and each single stranded DNA is bound to a single bead which is 28µm in diameter. Emulsion PCR is carried out for clonal amplification of the single stranded DNA; this occurs in a water-oil emulsion which enables each bead to be within a microreactor PCR reaction (Hori et al. 2007). The beads are then loaded onto a PicoTitre plate, so that one bead containing millions of the clonally amplified single stranded templates are present in one well (Margulies et al. 2005). There are two million wells on a PicoTitre Plate and each of these wells is loaded with beads containing sulphurylase and luciferase (Figure 1.4).

DNA sequencing begins with the addition of four deoxynucleotide triphosphates (dNTPs) to the wells sequentially (Ronaghi 2001; Ronaghi et al. 1998). DNA polymerase will catalyse the incorporation of the base provided it is complementary to the single stranded template. Once the dNTP is incorporated pyrophosphate is released (proportional to the number of bases added), which catalyses a series of enzymatic reactions (Figure 1.4). Pyrophosphate is converted into ATP and this is catalysed by ATP sulphurylase. ATP in turn catalyses the subsequent reaction which converts luciferin to oxyluciferin. This omits visible light which can be measured using a charge-coupled device camera (CCD). A thorough washing step between the

additions of a new base removes excess reagents. Apyrase degrades excess unincorporated nucleotides and excess ATPs, and once this is complete the next dNTPs and enzyme can be added. Bioinformatic analysis can begin with the flowgrams of the sequencing run shown by the signal intensity of each incorporation event.

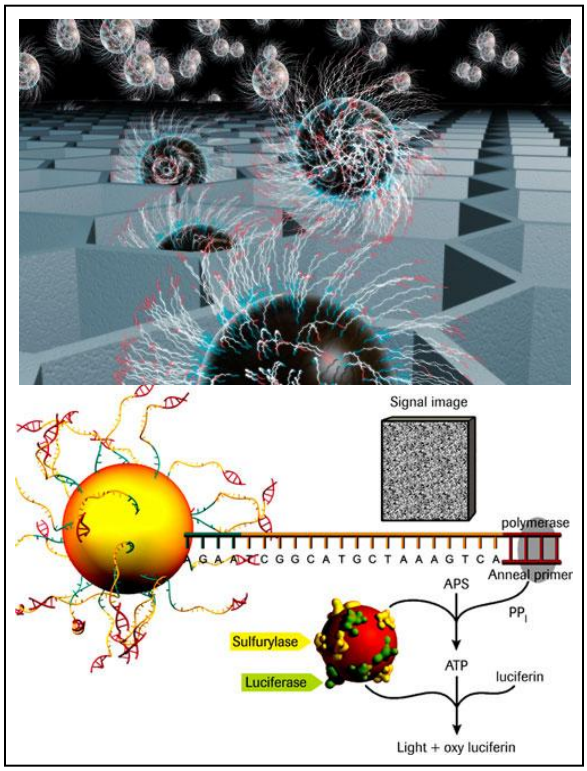


Figure 1.4. Pyrosequencing schematic of protocol taken from Roche 454 http://my454.com/products/technology.asp

Pyrosequencing has already been applied to many areas of research with four main applications, (1) the use of whole genome sequencing to generate accurate genome assemblies from sequenced fragments of DNA, (2) targeted resequencing of particular regions to look for germline or somatic mutations associated with cancer and other diseases, (3) metagenomic studies to study the diversity of an environment using a specific region, the majority of microbial studies use the 16S/18S rRNA gene and (4) Transcriptomics is the study of mRNAs using *de novo* assembly of sequences.

Limitations of the pyrosequencing technologies include the sequencing of homopolymers, especially those that are more than three bases long due to errors in signal intensity (Hert et al. 2008). In addition, it is known that there is signal noise from neighbouring wells on the PicoTitrePlate, however this has been greatly reduced by using Titanium reagents. Pyrosequencing also produces a smaller quantity of data compared with other next generation technologies such as Illumina (Table 1.3). Advantages of pyrosequencing include higher accuracy with an error every 1 in 1000-10,000 base calls and longer read lengths (400-500bp) compared with other NGS technologies (Table 1.3).

1.7.3.2. Alternative high-throughput sequencing technologies

Other next generation sequencing technologies include Solexa GAII (Illumina), SOLiD (Applied Biosystems) and Heliscope (Helicos Biosciences), each with different library preparation, sequencing and imaging steps (Ansorge 2009; Metzker 2010). The Solexa Genome Analyser (GA)II by Illumina works by producing a DNA library; DNA is fragmented and adaptors are ligated (Bentley et al. 2008). The single stranded DNA is then covalently bound to the flow cell surface which is also covered with oligonucleotides, clonal bridge amplification which occurs in clusters containing approximately 1000 sequences. After library preparation, sequencing can begin from millions of clonal clusters by cyclic reversible termination. All four different nucleotides are added to the plate each with a different labelled dye. When a base is successfully incorporated it is detected by total internal reflection fluorescence (TIRF) and imaged. After the incorporation of nucleotides, the terminator is removed and this is followed by a washing step so that the sequencing reaction can continue. Illumina technologies produce an enormous amount of data (5000Mb/day), however it is limited by a read length of 100 nucleotides (Table 1.3).

Other limitations include base call errors between A/C and G/T as there are only two lasers and due to interference from chemistry crystals and dust (Kircher & Kelso 2010). Phasing can occur due to bases not incorporating or inefficient blocking which results in inaccurate base calls from the clusters.

Unlike pyrosequencing and illumina, SOLiD (Support Oligonucleotide Ligation detection) uses ligation technologies compared to the traditional polymerase approach (Valouev et al. 2008). Library preparation like pyrosequencing relies on emulsion PCR, however the beads are bound to a glass slide, a universal primer is bound to the adaptor and subsequently octamer oligonucleotides are added to the reaction. When the fourth and the fifth nucleotides in the octamer oligonucleotides bind, the rest of the oligonucleotide is cleaved and the flourophore is detected. These steps are repeated to extend the sequence. After this, new primers are added which are sequentially one base closer and the octamer additions occur again in order to sequence the 'gaps'. Errors can occur through the random placement of beads on the glass slide, where dust can be mistaken for beads and the variable distances between beads can cause low quality base calls.

HeliScope (Helicos Biosciences) was the first example of commercialised single molecule sequencing. Libraries are made by shearing the DNA and attaching a poly (dA)-tails, and these are then bound to poly (dT) oligonucleotides which are attached to glass slides (Harris et al. 2008). Nucleotides are sequentially added to the reaction, and fluorescence is detected using TIRF technology. This is followed by cleavage of inhibiting groups before the reaction can continue. HeliScope has the advantage of not requiring a PCR amplification step; often the PCR step can introduce biases, particularly in template representation. However, it is limited by the sequence lengths

of only 32 nucleotides and the high error rate (1 in 100) which can occur as the signal strength of base calls is weak (Table 1.3).

Table 1.3: Comparison of high throughput technologies, taken from (Kircher & Kelso 2010)

Technology	Throughput (Mb/day)	Length (nt)	Quality	Costs ($/Mb)
Sanger	6	800	10^{-4}-10^{-5}	500
455/Roche	750	400	10^{-3}-10^{-4}	20
Illumina	5000	100	10^{-2}-10^{-3}	0.5
SOLiD	5000	50	10^{-2}-10^{-3}	0.5
Helicos	5000	32	10^{-2}	<0.50

1.7.4. Bioinformatic analysis of high throughput data

The generation of millions of sequences using NGS has introduced the challenge of bioinformatic analysis. Challenges include storage and management of the data; it was estimated that it is more expensive to store the raw data from Illumina Solexa GA II and SOLiD than it is to generate it (Batley & Edwards 2009). It was estimated that an Illumina Solexa GAII run produces 115,200 Tiff image files making up a terabyte of data (Richter & Sexton 2009). Data can be deposited in centralised databases such as GenBank, DNA Databank of Japan and the European Molecular Biology Laboratory. Analysis of pyrosequencing data provides additional challenges, for example there is no standardised method to analyse data which in turn makes comparisons between datasets difficult. However, most pyrosequencing bioinformatic pipelines include quality control procedures, OTU classification, taxonomic assignment, phylogenetic analysis and alpha and beta diversity analysis.

1.7.4.1. In-house bioinformatics pipeline

The in-house pipeline is an open source project (http://www2.warwick.ac.uk/fac/sci/lifesci/research/thermophyl/pipeline/) developed by Dr. Brain Oakley, University of Warwick (Calvo-Bado et al. 2011; Oakley et al. 2012). It consists of a series of Perl scripts to perform quality control procedures,

implements CD-HIT-EST (Li & Godzik 2006) to cluster the sequences into OTUs and contains R scripts (R Development Core Team 2011) to generate summary statistics and graphics. Perl and R are programming languages which can run on a Linux platform. Advantages of the pipeline include the capabilities to modify the scripts for specific dataset requirements and additional shell scripts which enable an automated approach.

1.7.4.2. QIIME

Quantative Insights into Microbial Ecology **(**QIIME) is an open source project (http://qiime.sourceforge.net/) developed by the Knight Laboratory, University of Colorado, Boulder (Caporaso et al. 2010b). The QIIME bioinformatic pipeline enables the user to run the pipeline through a QIIME Virtual Box installation. The Python scripts allow Sanger, Roche/454 pyrosequencing and Illumina data to be processed. Essential steps of the QIIME pipeline include quality control of the sequences, OTU assignment, taxonomic assignment and alignment of sequences, construction of phylogenetic trees and the calculation of diversity metrics. Advantages of QIIME include that it is continually updated to incorporate new methods of metadata analysis; it offers a wide range of programs to cluster, align, build phylogenetic trees and assign taxonomy.

1.7.4.3. Comparisons to alternative bioinformatic pipelines

Many other bioinformatic pipelines exist including the Ribosomal Database Project (RDP) pyrosequencing pipeline (Cole et al. 2009). The pipeline includes; pre-processing stages, the alignment of sequences, RDP classification of species and community analysis. This web-based pipeline has automated procedures whereby data is uploaded, this has the advantage of being relatively easy to use but also provides less flexibility for analysis. PANGEA (pipeline for analysis of next

generation of amplicons) is similar to the in-house pipeline in that it is a series of perl scripts, uses CD-HIT-EST for clustering and contains shell scripts (Giongo et al. 2010). Advantages include more flexibility provided by ability to modify the source code. mothur was developed by Dr. Pat Schloss, University of Michigan (Schloss et al. 2009); this open source program was written in C++ and includes preprocessing steps, community analysis and comprehensive online web pages for support.

1.8. Case-study: Ethiopia

Ethiopia has a population of approximately 84.5 million. The country is ranked 8th worldwide for TB; the incidence of TB is 300 per 100 000 population/year (WHO 2011b). 36% of all incident TB cases are extra pulmonary (Berg et al. 2009) one of the highest rates in the world.

The HIV epidemic in Africa has dramatically changed the epidemiological pattern of TB, where clinical TB disease progression and reactivation of latent disease has rapidly increased (Converse 2000). It was estimated in 2004 that in the African region 34% of all new adult cases of TB worldwide were infected with HIV (Lonnroth et al. 2006). Approximately 19% of TB cases are HIV positive, however estimates have been as high as 30% in Ethiopian urban areas (WHO 2011b; Yassin et al. 2004). Diagnostics of TB infection in HIV positive individuals pose a serious problem as the tuberculin skin test and smear culture technique are less sensitive in these persons (Tegbaru et al. 2006; Yassin et al. 2004). Predisposing helminth infection is prevalent in Ethiopia (Elias et al. 2006), in addition to malnutrition and poverty which are serious public health issues and risk factors for TB.

There is little understanding of the extent of bovine tuberculosis in Ethiopia, but it is estimated that it is widespread, albeit a low prevalence (Berg et al. 2009). It is

estimated to have one of the largest livestock populations in Africa with approximately 33 million cattle (Shitaye et al. 2007). Transmission of *M. bovis* between humans and animals has been found to occur readily (Regassa et al. 2008). *M. bovis* infection in humans is normally extra pulmonary and is characteristically cervical lymphadenitis (Shitaye et al. 2007). Transmission to humans occurs via unpasteurised milk, raw meat, animal blood consumption and inhalation of cough spray (Cosivi et al. 1999). It is estimated that 85% of cattle and 82% of humans in Africa live in areas where bovine TB is only partly controlled or not controlled at all (Ameni et al. 2006). Although *M. tuberculosis* is known to be the principal cause of tuberculosis, control of bovine tuberculosis is essential, especially in the light of increased immunocompromised individuals with HIV/AIDS.

Ethiopia is an exemplar region where mycobacterial infection and disease is of concern due to the number of immunocompromised persons. BCG coverage was estimated to be 72% in 2007, and that figure has been above 50% since 1994 (WHO-UNICEF 2011). The efficacy of BCG in Ethiopia is unknown, however countries with similar climates and demographic status have shown BCG confers little protection. Although there are no estimates for BCG efficacy within Ethiopia, both human and animal studies have shown widespread sensitivity to mycobacteria (Berg et al. 2009; Lema & Stanford 1984; Tschopp et al. 2010a).

1.9. Aims and Hypotheses

Aims:

1. To optimise culture independent molecular methods to target *Mycobacterium* species (Chapter 3).

2. To determine the diversity and prevalence of mycobacteria within ecologically diverse locations and different sample types (Chapter 4).

3. To use a multivariate approach to determine the effect of environmental and spatial parameters on *Mycobacterium* species diversity and prevalence across ecologically diverse regions (Chapter 5).

4. To quantify the *Mycobacterium* genus abundance and pathogenic *Mycobacterium* species in environmental samples (Chapter 6).

Hypotheses:

1. Molecular methods will provide a more comprehensive insight into mycobacteria diversity and abundance in the environment compared to traditional cultivation methods.

2. Environmental variables such as pH, moisture, temperature and elevation will change the diversity and abundance of mycobacteria.

3. The environment is a source of exposure and transmission of opportunistic mycobacteria.

Chapter 2: Materials & Methods

2.1. Environmental sample collection of Pilot Soils

Samples were collected from sites in and around the market town of Mehal Meda, Ethiopia (Figure 2.1) and from Cryfield field, University of Warwick, UK. Soil sample 1108, described as very compacted, dry soil was collected from a well-used by people and cattle in Mehal Meda town (10°31'N, 39°66'E). Soil sample 1109 was collected from the compacted soil floor at the entrance to a house near Mehal Meda (10°33'N, 39°74'E). Soil sample 1110 was collected from a ploughed field in Mezozo (9°97'N 39°74'E). Soil sample 1111 was collected from a house yard on a roadside in Sembo (9°42'N 39.35E). The samples were collected by Dr. Orin Courtenay in October 2008 during the dry season. The UK soil Cryfield (52°36'N 1°56'E) was collected in August 2008.

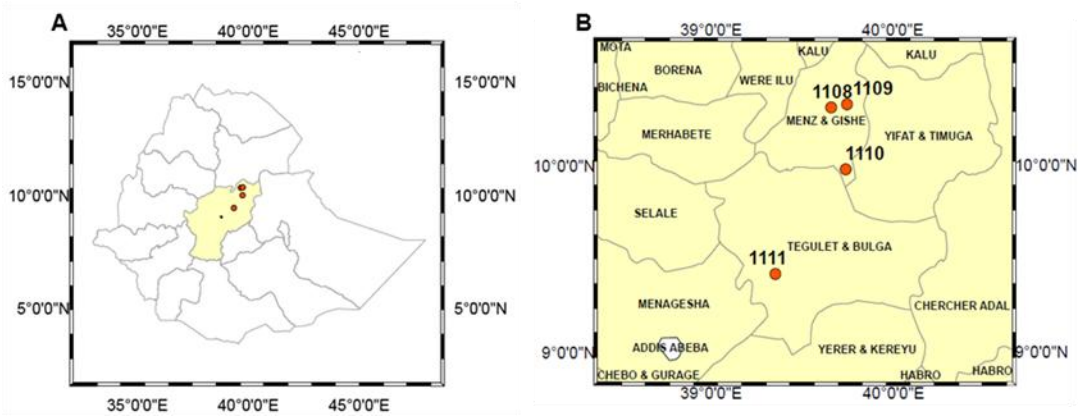


Figure 2.1. Sampling locations of the pilot soils 1108, 1109, 1110 and 1111. **(A)** Map of Ethiopia, the orange dots are the sampling sites and black dot is Addis Ababa **(B)** Zoomed in sampling locations in relation to Addis Ababa.

2.2. Ethiopia Fieldwork

Environmental samples were collected from a total of 42 villages located between the latitudes 4°42'-12°46'N and longitudes 34°15 – 37°52'E (Figure 2.2). Villages were sampled from a total of nine zones; Gonder (abbrev. Go), Woldiya (Wo), Bale (Ba), Jinka (Ji), Omorate (Om), Turmi (Tu), Butajira (Bu), Hossana (Ho) and Gambella (Ga) (Figure 2.2 & 2.3). In the text villages belonging to the same zones are represented by the first two letters of the zones, followed by the number of the village (Table 2.1).

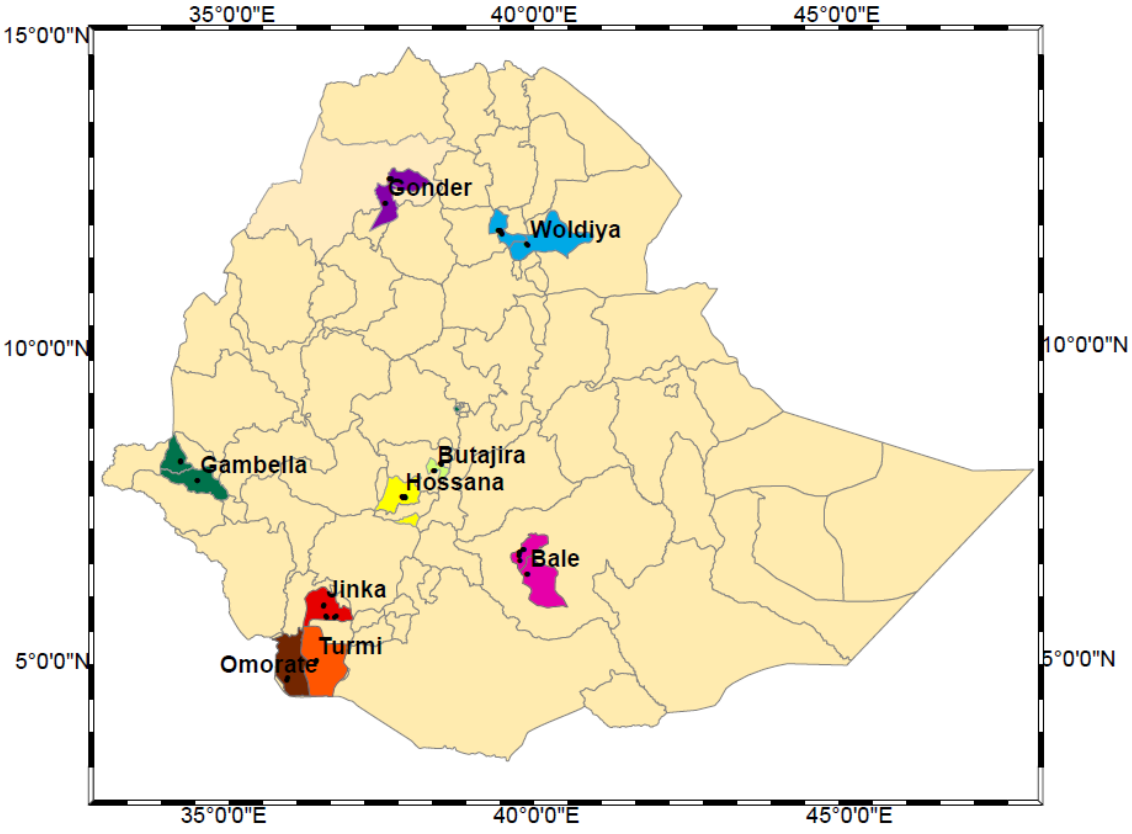


Figure 2.2. Map of Ethiopia to show the sampling locations, colour shaded areas represent the nine regions sampled

Table 2.1. Climate description of the regions sampled and number of villages sampled within each region

Region	No. of villages sampled	Site Code	Description Summary
Omorate	2	Om1-Om2	Hot semi-arid lowlands
Turmi	2	Tu1-Tu2	Hot semi-arid lowlands
Jinka	6	Ji1-Ji6	Hot/warm dense woodland
Gambella	4	Ga1-Ga4	Tropical hot sub-humid lowlands
Hossana	4	Ho1-Ho4	Warm sub-humid cultivation
Butajira	4	Bu1-Bu4	Warm sub-humid cultivation
Gonder	4	Go1-Go4	Hot sub-humid cultivation midlands
Bale	8	Ba1-Ba8	Warm moist lowlands to sub-humid mixed forest highlands
Woldiya	8	Wo1-Wo8	Warm moist lowlands to tepid sub-humid cultivated highlands

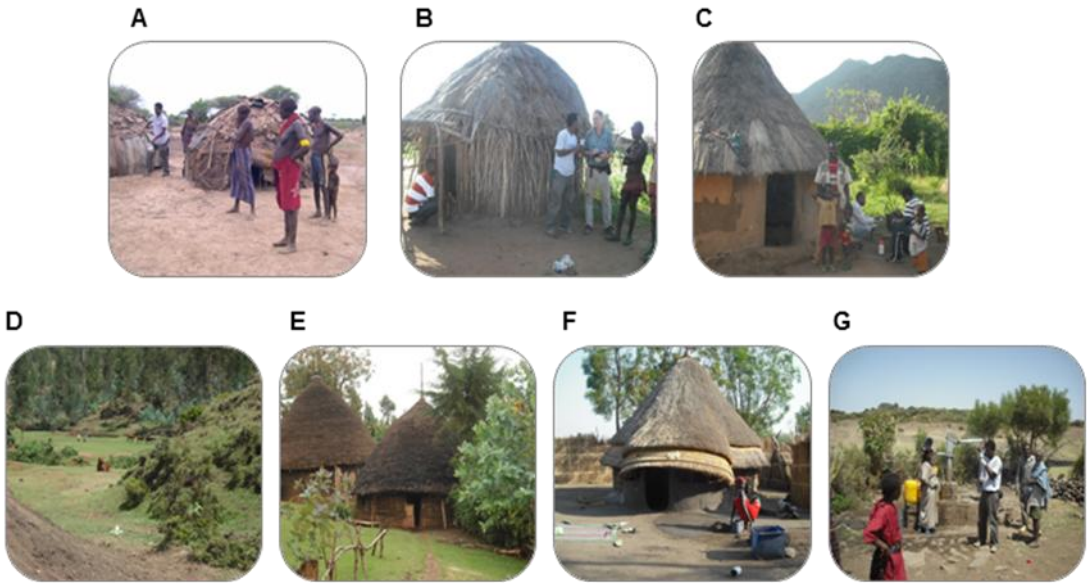


Figure 2.3: Pictures of sampling regions **(A)** Omorate: semi-desert and shrubland, population are pastoralists and ethnicity is Dasenech. **(B)** Turmi: dry shrubland, population are Agropastoralists and ethnicity is Hamer. **(C)** Jinka: acacia-commiphora woodlands, population are agriculturalists and ethnicity is Ari. **(D)** Butajira: dense green vegetation and grassland, population are agriculturalists and ethnicity is Miskan agriculturalists **(E)** Hossana: moderately cultivated, population are agriculturalists and ethnicity is Hadiya. **(F)** Gambella: wooded grassland, population are agriculturalists and ethnicity is Anuak and Neuer **(G)** Gonder: dry shrubland and cultivation, population are Amhara agriculturalists and ethnicity is Amhara.

2.2.1. Description of sampling regions

The administrative divisions of Ethiopia from largest to smallest include region, zone, woreda, kebele and village (Figure 2.4). Samples were collected from January 2010 to March 2010, during the dry season. Mean annual temperatures range from <7.5°C to >27.5°C, with tepid to cool areas in central parts and hotter areas towards the borders of Ethiopia (Figure 2.5). The mean annual rainfall in Ethiopia ranges from <200mm to 2200mm. Western areas have the highest mean annual rainfall of 1800-2000mm and southern and eastern areas have estimates of less than 200mm (Figure 2.6). Ethiopia has a wide-ranging terrain; elevation ranges from -165-4550m and is highest near the Great Rift Valley (Figure 2.7). Land use ranges from cultivation in northern to central parts, swampland in the west and grassland in the south-eastern parts of Ethiopia (Figure 2.8). Major Agro Ecological Zones (MAEZ) range from humid lowlands in the East, moist lowlands to highlands in the North, arid lowlands in the East and South (Figure 2.9 & Table 2.2). Zones were chosen to represent a wide range of environmental areas.

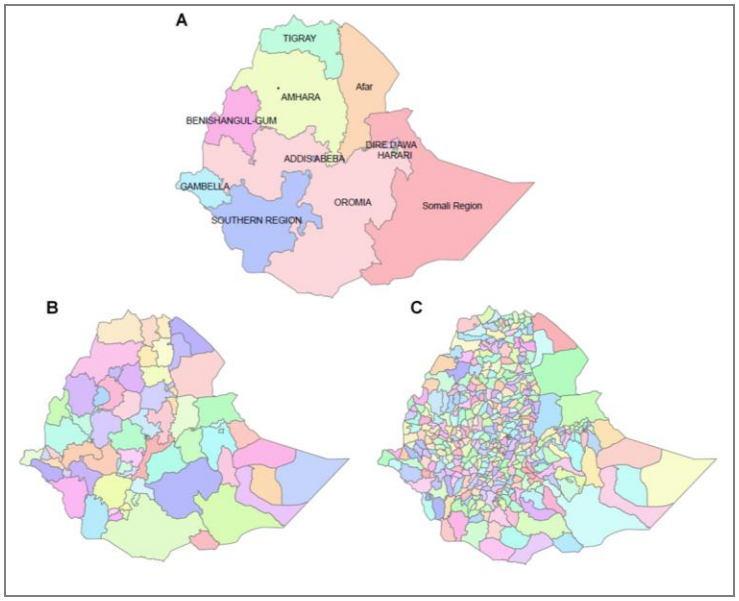


Figure 2.4. Administrative areas of Ethiopia obtained from the Ministry of Agricultural and Rural Development (MARD). **(A)** 11 administrative regions **(B)** 76 administrative zones **(C)** 466 administrative woredas

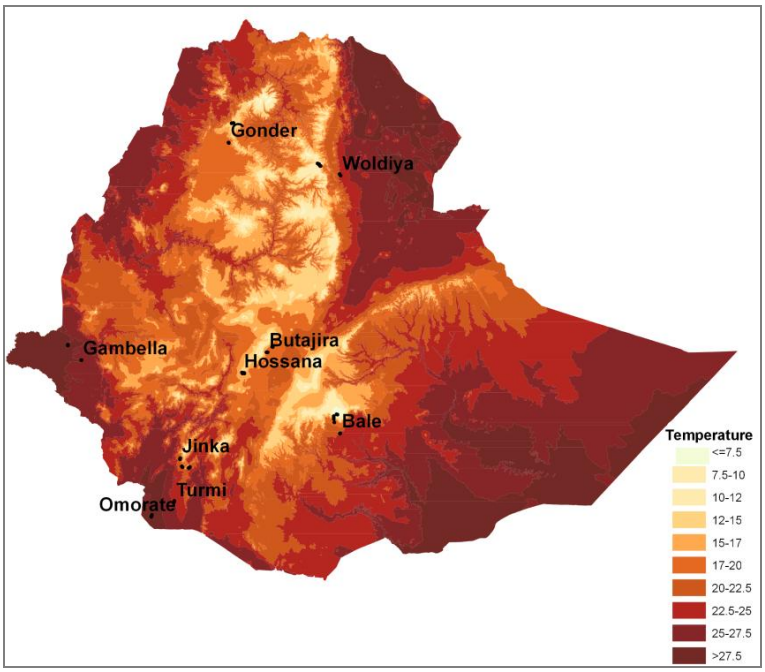


Figure 2.5. Annual temperature records of Ethiopia obtained from MARD

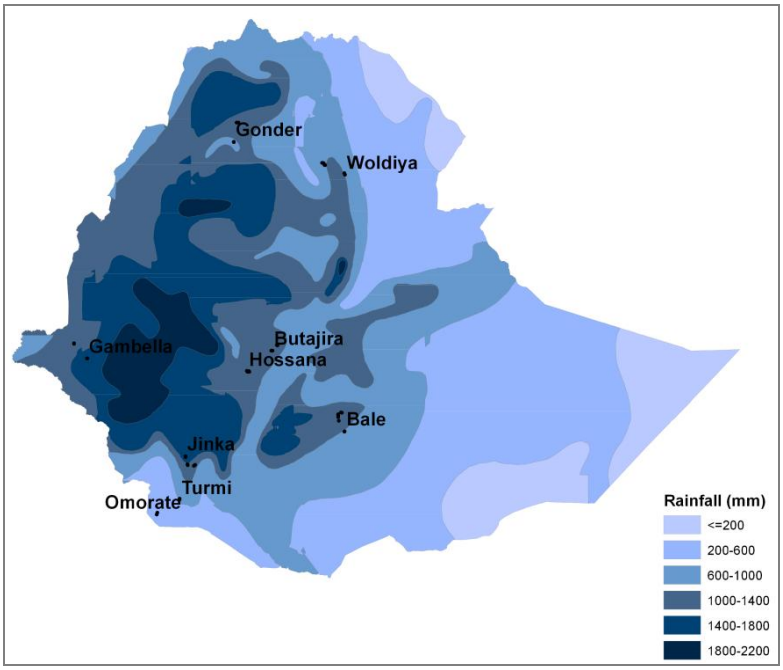


Figure 2.6. Annual rainfall records of Ethiopia obtained from MARD

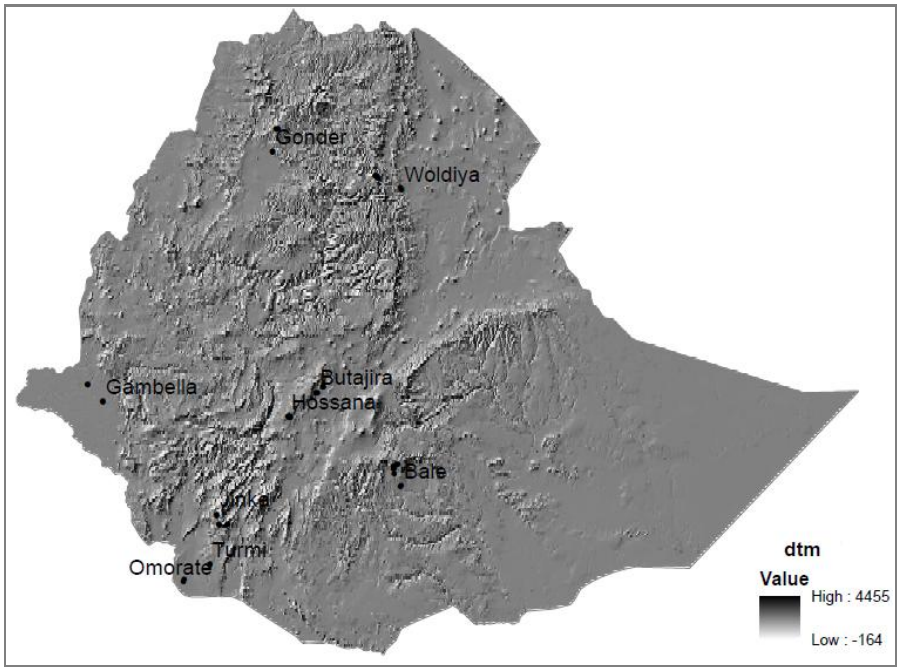


Figure 2.7. Digital elevation model (DTM) for Ethiopia in metres obtained from MARD

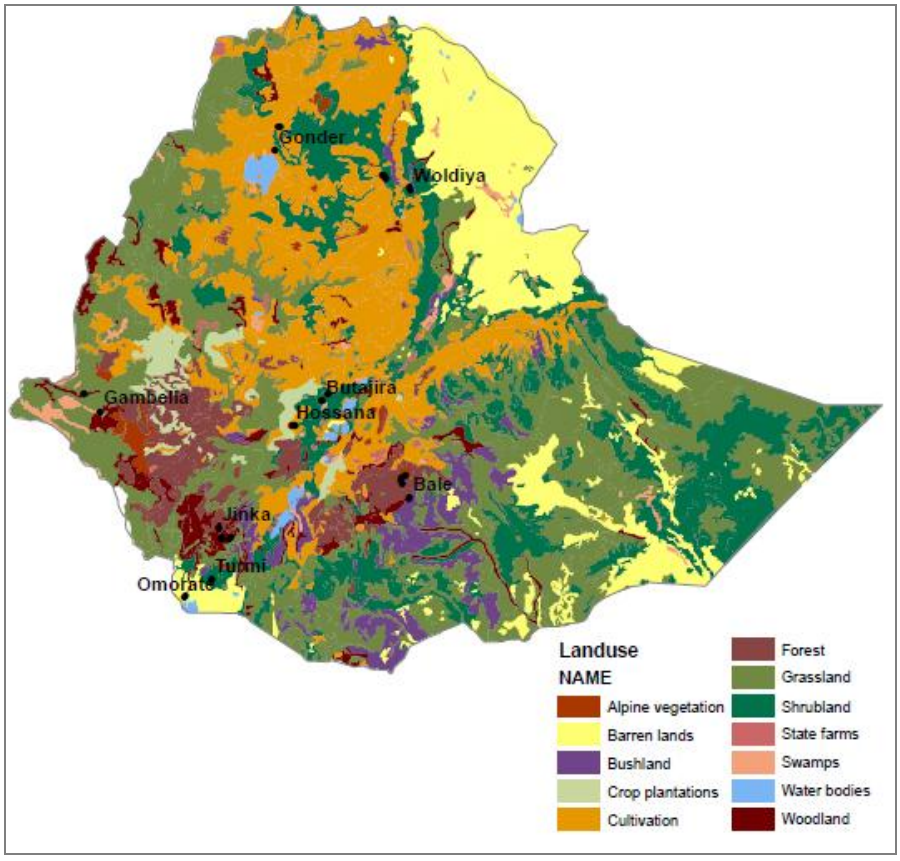


Figure 2.8. Land use data for Ethiopia. Data obtained from the International Livestock Research Institute (ILRI) GIS services (http://192.156.137.110/gis/default.asp) accessed June 2010.

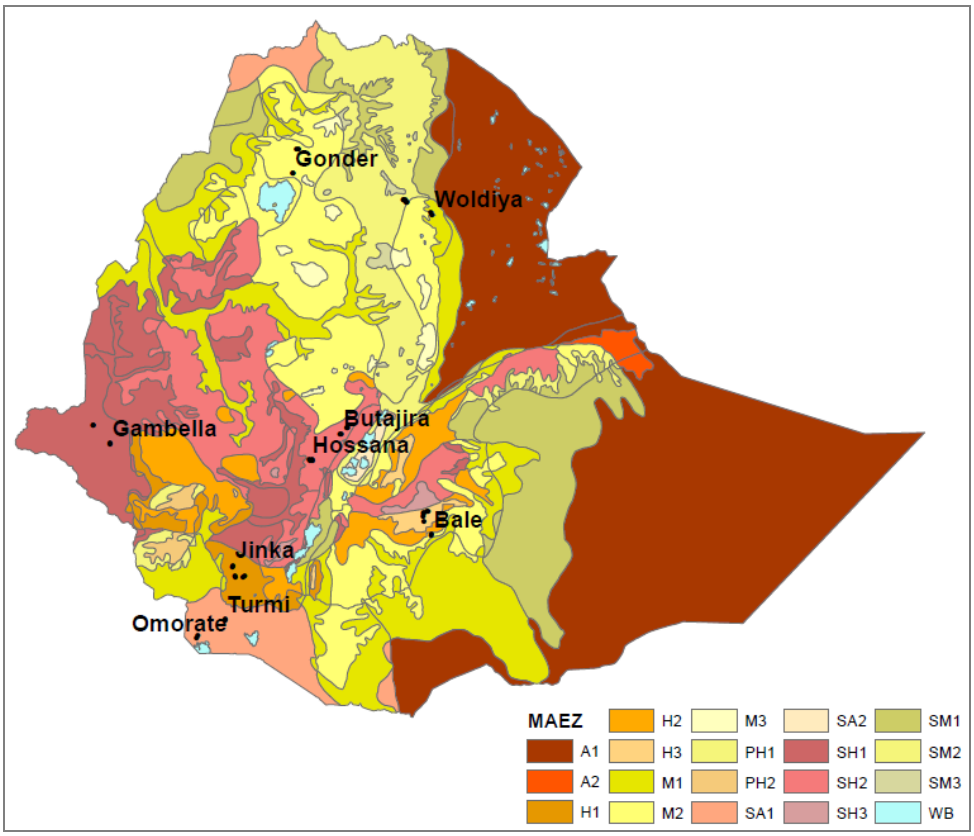


Figure 2.9. Major Agricultural Ecological Zones of Ethiopia obtained from MARD.

Table 2.2. Key for Figure 2.7

Symbol	Major Agro-ecological Zones
A1	Hot arid lowland plains
A2	Warm arid lowland plains
SA1	Hot semi-arid lowlands
SA2	Warm semi-arid lowlands
SM1	Hot sub-moist lowlands
SM2	Warm sub-moist lowlands
SM3	Tepid sub-moist mid highlands
M1	Hot moist lowlands
M2	Warm moist lowlands
M3	Tepid moist mid highlands
SH1	Hot sub-humid lowlands
SH2	Warm sub-humid lowlands
SH3	Tepid sub-humid mid highlands
H1	Hot humid lowlands
H2	Warm humid lowlands
H3	Tepid humid mid highlands
PH1	Hot per-humid lowlands
PH2	Warm per-humid lowlands
WB	Water Body

2.2.2. Elevation transects

In two zones (Bale and Woldiya) samples were taken from villages across a wide range of elevations within a relatively short distance (Figure 2.10 & 2.11). In Bale eight villages were sampled within a distance of 54km with a range of elevation from 1224-3997m and in Woldiya eight villages within a distance of 61km were sampled with a range of elevation from 1414m-3396m (Figure 2.12).

Figure 2.10: Bale elevation transect. Population are Oromyia agriculturalists and pastoralists **(A)** Dola Mena: <1000m dry lowlands **(B)** Harenna: <2000m woodland forest **(C)** Rira: <3000m cultivation **(D)** Sannetti Plateau: <4000m afro-alpine vegetation

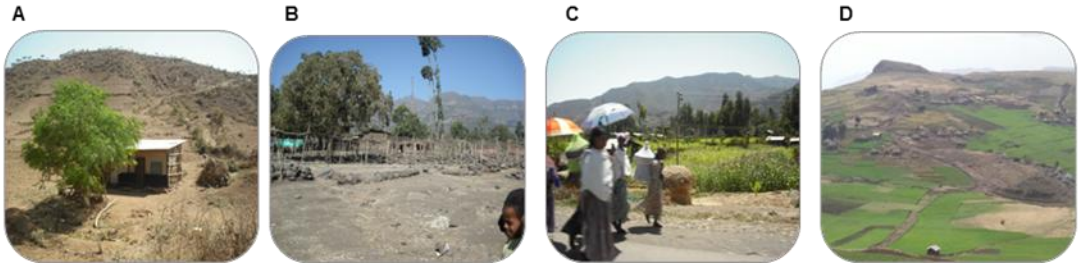


Figure 2.11: Woldiya elevation transect. Population are Amhara agriculturalists A) No. 24 kebele: <1800m dry shrubland B) Sankar: <2400 open shrubland C) Debot: <3000 cultivation D) No. 28 kebele: <3600 highland cultivation

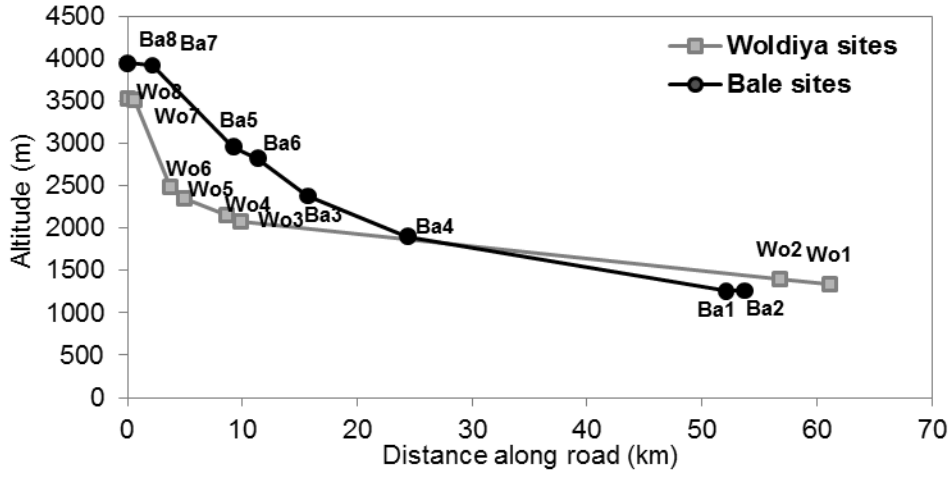


Figure 2.12: Distance and elevation of the eight villages from regions Bale and Woldiya

2.2.3. Soil and water collection

In each of the 42 villages, soil was sampled from five sites including two households, two household yards and one communal site such as a school or market (Figure 2.13), where the top $3cm^3$ of soil and vegetation was removed prior to sample collection to reduce the effect of UV irradiation. At each of the five sites (<2 kilometres apart), soil was collected using a sterile trowel into a plastic bag from three areas (metres apart) and immediately pooled to make a composite sample. The plastic bag was shaken and three replicates of approximately 5g taken immediately and stored in bijou tubes. The composite samples were stored at -20°C.

For water sample collection, two sites for each of the 42 villages were chosen comprising of one drinking source and one bathing source (often these were interchangeable) (Figure 2.13). Samples were collected from a variety of sources including groundwater, drinking pumps, springs, rivers and lakes. At each sampling point, 100ml of water was collected from the surface and was filtered using a 50ml sterile plastic syringe and the MicrofilV filtration device with 0.22µm mixed cellulose esters white gridded filters (Millipore, MA, USA). After collection the filters were removed from the plastic holder using sterile forceps and air-dried. The filters were then coiled and stored in 5ml bijou tubes in a cool box with ice packs (4°C) for transporting to the laboratory where the filters were stored at -20°C.

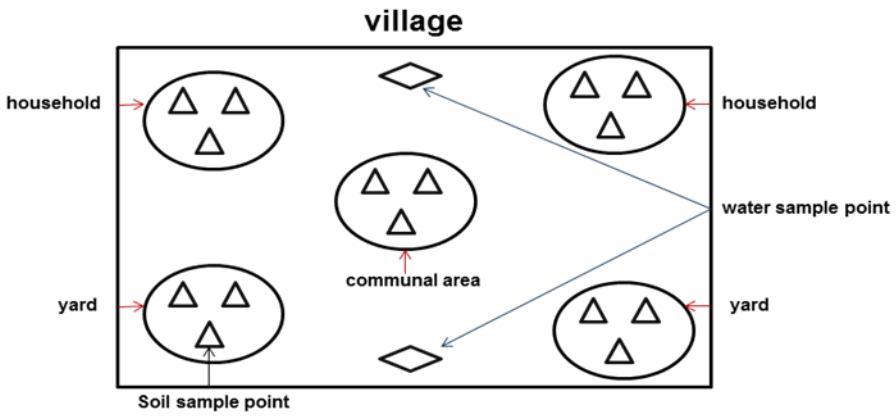


Figure 2.13. Sampling schematic of soil and water collection points per village

2.2.4. Environmental readings

Latitude, longitude and elevation readings were taken at each site using an eTrex Legend H Outdoor GPS (Garmin Ltd., UK). Readings for moisture was recorded at each soil site using the SM200 Soil Moisture Sensor (Delta-T Devices Ltd., UK) and the HH2 Moisture Meter (Delta-T Devices Ltd., UK) with the output as volumetric water content expressed as a percentage. Temperature and pH readings for soil and water sites were taken using a Field Scout SoilStik (Spectrum technologies, Inc., IL, USA). The average reading was calculated for soil sites belonging to a village and water sites belonging to a village, and these averages were used for downstream analysis. The variables for multivariate analysis were temperature, elevation, moisture, pH, latitude and longitude.

2.2.5. Tuberculosis Health Record & Ethiopia Climate data collection

Data was collected from health centres within the administrative regions of locations where environmental sampling took place. Data was obtained from a total of eight health centres and it included the number of new TB cases, HIV-TB coinfected cases, smear-positive cases, smear-negative cases and extrapulmonary TB cases for the year 2008-2009. Additional data was obtained from the Federal Ministry of Health of Ethiopia (MoH). Every year since 1987 in the Ethiopian Calendar (1994/1995 General Calendar) the Federal Ministry of Health has published Health and Health related indicators. These indicators are mostly recorded at the regional level, and provide information regarding population demographics and incidences of infectious diseases MoH data was accessed on Oct 2011, http://www.ethiomedic.com/index.php/national-guidlines/category/1-health-and-health-related-indicators-for-ethiopia.html?start = 5. MoH data included new TB

cases, smear-positives TB cases, smear negative TB cases and extrapulmonary cases for the year 2008-2009.

2.3.1 Bacterial strains and growth conditions

The strains used in this study are listed in Table 2.3. All species were stored at -80°C as glycerol stocks. The stocks were resuscitated on Middlebrook 7H10 agar (BD, Oxford, UK) and grown at 37°C, with exception of *M. fortuitum, M. intracellulare, M. vaccae* and *M. xenopi* which were cultivated on Lowenstein Jenson slants (BD, Oxford, UK). A single colony was then inoculated onto Middlebrook 7H9 broth (BD, Oxford, UK) and incubated in a water bath at 37°C with shaking at 150rpm. For *Mycobacterium avium* subspecies *paratuberculosis* a supplement of ferric mycobactin J (Allied Monitor, Fayette, Missouri, USA) was added to both the media and broth.

Table 2.3. *Mycobacterium* species strains used in this study

Taxon	Strain
Mycobacterium abscessus	DSM 44196T*
Mycobacterium avium subsp. *avium*	DSM 44156T*
Mycobacterium avium subsp. *paratuberculosis*	DSM 44133T*
Mycobacterium fortuitum	DSM 46621T*
Mycobacterium intracellulare	DSM 43223T*
Mycobacterium kansasii	DSM 44162T*
Mycobacterium marinum	DSM 44344T*
Mycobacterium scrofulaceum	DSM 43992T*
Mycobacterium vaccae	DSM 43292T*
Mycobacterium xenopi	DSM 43995T*
Mycobacterium agri	****
Mycobacterium aichense	****
Mycobacterium aurum	ATCC 23366***
Mycobacterium bovis BCG	Pasteur
Mycobacterium chubense	****
Mycobacterium duvalii	****
Mycobacterium fortuitum	ATCC 14468***
Mycobacterium gilvum	****
Mycobacterium gordonae	ATCC 14470***
Mycobacterium nonchromogenicum	ATCC 19530***
Mycobacterium obuense	****
Mycobacterium peregrinum	****
Mycobacterium phlei	ATCC 354***
Mycobacterium smegmatis	ATCC 13578***

***** Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH Mascheroder Weg 1b, 38124 Braunschweig, Germany

*** American Type Culture Collection (ATCC), P.O. Box 1549, Manassas, VA 20108, USA

****Supplied by John Magee, Regional Centre for Mycobacteriology, Newcastle, UK

2.3.2 DNA Extraction

DNA was extracted from *Mycobacterium* species strains grown in liquid broth using the DNeasy Blood & Tissue Kit (QIAGEN, Ltd., Crawley, UK) as per manufacturer's instructions.

Total community DNA was extracted from 0.5g-0.6g of soil from each of the 210 soil sites (5 sites x 42 villages). Based on a published kit comparison test for soil DNA extractions (Pontiroli et al. 2011), FastDNA® Spin Kit for Soil (MP Biomedicals, OH, USA) was chosen as a suitable extraction method for all soil DNA extractions in this study and DNA was extracted according to manufacturer's instructions. To summarise, 0.5-0.6 grams of soil was added to lysing matrix tubes containing silica particles and sodium phosphate buffer for mechanical lysis using the Ribolyser Instrument Precelleys. DNA is then separated from the cell debris and using ethanol it is precipitated and purified using centrifugation steps.

Total community DNA was extracted from the 0.22µm filters from each of the 84 water sites (2 sites x 42 villages) using the PowerWater® DNA Isolation Kit (MoBio Laboratories, Inc., CA, USA) according to manufacturer's instructions. In brief, both chemical and mechanical cell lysis occurs and centrifugation separates the DNA supernatant from beads and proteins. The subsequent addition of a high salt solution is necessary for the binding of DNA to the silica membrane, the salt is then washed off using an alcohol solution and DNA is eluted from the membrane. Three DNA extraction methods from water were compared and the PowerWater® DNA Isolation Kit (MoBio Laboratories, Inc., CA, USA) was shown to be the most sensitive as *M. bovis* was detected at lower concentrations of spiking (section 3.2).

2.4. Primer Design and PCR: targeting *Mycobacterium* 16S rRNA gene

The 16S rRNA gene sequences belonging to species of the *Mycobacterium* genus were retrieved from SILVA 16S rRNA gene database (Pruesse et al. 2007). To target the genus *Mycobacterium* several published primer sets were compared to find a primer set that would amplify a region suitable for PCR-DGGE analysis. The primers used for PCR-DGGE and pyrosequencing are listed in Table 2.4. A previously published primer set was chosen which specifically targeted the *Mycobacterium* genus JSY16S (Young et al. 2005). The PCR programme for JSY16S was 94^{o}C for 5 min, followed by 35 cycles of 94 ^{0}C, 55 ^{0}C for 1 min, 65 ^{0}C for 1 min, and finally a single extension step of 65 ^{0}C for 5 min. The second set of primers was specific to SG; the reverse primer APTK16SR targets the long helix 18 present only in certain slow growing mycobacteria at the *Escherichia coli* positions 451 to 482 (Leclerc et al. 2003; Rogall et al. 1990; Stahl & Urbance 1990; Young et al. 2005). The forward primer APTK16SF was designed using the programme PrimerBLAST, it targets a conserved region of the *Mycobacterium* genus 16S rRNA encoding gene. For APTK16S the thermal protocol is as follows 94^{o}C for 5 min, followed by 35 cycles of 94 ^{0}C, 59.9 ^{0}C for 1 min, 67 ^{0}C for 1 min, and finally a single extension step of 67 ^{0}C for 5 min. A GC clamp (5'CGCCC GCCGCGCGCGGCGGGCGGGGCGGGGGCACGGGGGG3') was added to the reverse primers JSY16SR and APTK16SR for PCR-DGGE methods. A nested approach was applied using the universal primer set pA pH to amplify the entire 16S rRNA gene (Edwards et al. 1989). The PCR cycle for pA pH was 94.0°C for 1 min, followed by 40 cycles of 94.0°C for 1 min, 62.0°C for 1min, 72.0°C for 1.30 min and a single extension step of 72°C for 7mins. All PCR reactions were performed in a 50µl reaction containing 1µl of DNA, 25µl of PCR master mix (Promega, (UK) Ltd.,

Southampton), 1µl (10umol) of each primer, 2.5µl DMSO, 2µl BSA, and 17.5µl of sterile distilled water using the Eppendorf Mastercycler® PCR machine (Eppendorf, AG, Hamburg, Germany).

Table 2.4. Primers used to target mycobacteria for diversity analysis

Primer Name	Sequence	Target group	Target gene	Reference
pA	AGA GTT TGA TCC TGG CTC AG	Eubacteria	Entire 16S encoding gene	(Edwards et al. 1989)
pH	AAG GAG GTG ATC CAG CCG CA	Eubacteria	Entire 16S encoding gene	(Edwards et al. 1989)
JSY16SF	TGG GAA ACT GGG AAA CTG GGT CTA ATA	*Mycobacterium* genus	16S encoding gene	(Young et al. 2005)
JSY16SR	CCC GCA CGC CCA AGT TAA GCT GTG AG	*Mycobacterium* genus	16S encoding gene	(Young et al. 2005)
APTK16SF	GCT TAA CAC ATG CAA GTC GAA CGG AAA GG	Slow growing *Mycobacteria*	16S encoding gene	This thesis
APTK16SR	GTC AAT CCG AGA GAA CCC GGA CCT TCG TCG	Slow growing *Mycobacteria*	16S encoding gene	This thesis
APTK16SF-Pyro	GTC AAT CCG AGA GAA CCC GGA C	Slow growing *Mycobacteria*	16S encoding gene	This thesis
APTK16SR-Pyro	GCT TAA CAC ATG CAA GTC GAA CG	Slow growing *Mycobacteria*	16S encoding gene	This thesis

2.5. Real-time quantative PCR

All real-time quantative PCR (qPCR) reactions were loaded on to MicroAmp 96-Well Reaction Plates (Applied Biosystems Inc., CA, USA) which were sealed with MicroAmp Optical Adhesive Film (Applies Biosystems) and centrifuged for 1 min at 1250 RPM. Reactions were run on an ABI 7500 Fast Real-Time PCR System (Applied Biosystems Inc., CA, USA). The standard curve (absolute quantification) method was used. Each qPCR plate run included 7 standards separated by 10^{-1} dilutions. The amount of DNA was estimated using the dsDNA copy number calculator (http://www.uri.edu/research/gsc/resources/cndna.html), which used the

concentration of DNA (ng) and the size of the genome (bp) to determine the number of gene copies per µl. Results for all qPCR assays were analysed using the 7500 Fast System SDS Software (Applied Biosystems). In the Analysis Settings, the Auto Ct option was used to calculate the Ct values for the *Mycobacterium* genus qPCR. For the *M. bovis* qPCR and the *M. tuberculosis* complex multiplex, on the Manual Ct the AutoBaseline was set to 0.01 as there were many samples just below the Auto setting baseline and all negative template controls (NTCs) were negative. Samples were only considered positive if all three of the replicates were positive and subsequently the mean number of gene copies was calculated.

2.5.1. *Mycobacterium* genus qPCR

A real-time qPCR assay was employed to determine the *Mycobacterium* genus load in all environmental samples. The previously developed *Mycobacterium* genus primers which target the internal transcriber spacer and partial 23S gene was used for quantification (Coppenraet et al. 2004; Jacobs et al. 2009). For each reaction, the total volume was 25µl, comprising of 12.5µl of TaqMan Environmental master mix 2.0 (Applied Biosystems Inc., CA, USA), 1µl (10µmole) of the forward primer 5'-GGGGTGTGGTGTTTGAG-3', 1µl (10µmole) of the reverse primer, 5'-CTCCCACGTCCTTCATC-3', 1µl (10µmole) of the probe, 5'-(6-carboxyfluorescein)-TGGATAGTGGTTGCGAGCATC-(black hole quencher 1)-3', 2.5µl of a 10mg/ml bovine serum albumin (BSA) solution, 6µl of sterile water and 1µl of the total community DNA. Reactions were run on an ABI 7500 Fast Real-Time PCR System (Applied Biosystems Inc., CA, USA) using the conditions of 2 min at 50°C, followed by 10 mins at 95°C, then by 40 cycles of 15 secs at 95°C and 62°C for 1 min. *Mycobacterium tuberculosis* DNA dilutions were used to generate a standard curve which ranged from 5.8×10^5 gene copies/µl to 0.58 gene copies/µl.

2.5.2. *M. bovis* qPCR

qPCR was employed to ascertain the amount of *M. bovis* present in all Ethiopian environmental samples. The qPCR assay was previously designed and employed to target the RD4 region (Sweeney et al. 2007). For each reaction, the total volume was 20µl, comprising of 12.5µl of TaqMan Environmental master mix 2.0 (Applied Biosystems Inc., CA, USA), 1µl (20pmole) of the forward primer 5'-TGTGAATTCATACAAGCCGTAGTCG-3', 1µl (20pmole) of the reverse primer, 5'-CCCGTAGCGTTACTGAGAAATTGC-3', 1µl (20pmole) of the probe, 5'-(6-carboxyfluorescein)–AGCGCAACACTCTTGGAGTGGCCTAC-(tetramethyl-6-carboxyrhodamine)-3', 2.5µl of a 10mg/ml bovine serum albumin (BSA) solution, 2µl of sterile water and 1µl of the total community DNA. Reactions were run on an ABI 7500 Fast Real-Time PCR System (Applied Biosystems Inc., CA, USA) using the conditions of 2 min at 50°C, followed by 10 min at 95°C, then by 40 cycles of 15 secs at 95°C and 58°C for 1 min. *Mycobacterium bovis* DNA dilutions were used to generate a standard curve which ranged from 8.5 x 10^5 gene copies/µl to 0.85 gene copies/µl.

2.5.3. *M. tuberculosis* and MTC multiplex qPCR

A multiplex qPCR assay was employed to determine the *M. tuberculosis* and the *M, tuberculosis* complex (MTC) load. The previously designed assay targets the *lepA* gene to detect members of the *Mycobacterium tuberculosis* complex and the *wbbl1* gene to quantify *M. tuberculosis* and *M. cannetti* from the rest of the MTC (Reddington et al. 2011). For each reaction, the total volume was 30µl, comprising of 15µl of TaqMan Environmental master mix 2.0 (Applied Biosystems Inc., CA, USA), 0.6µl of each primer and probe (Table 2.5), 3µl of a 10mg/ml bovine serum albumin (BSA) solution, 6.8µl of sterile water and 1µl of the total community DNA.

Reactions were run on an ABI 7500 Fast Real-Time PCR System (Applied Biosystems Inc., CA, USA) using the conditions of 2 min at 50°C, followed by 10 min at 95°C, then by 40 cycles of 15 secs at 95°C and 58C for 1 min. *Mycobacterium tuberculosis* DNA dilutions were used to generate a standard curve which ranged from 5.8 x 10^5 gene copies/µl to 0.58 gene copies/µl.

Table 2.5. Primers and probes used for the MTC and *M. tuberculosis/M. cannetti* multiplex quantative PCR. The *wwbl1* gene specific for *M. tuberculosis* and *M. cannetti*.

Name	Concentration (µmole)	Sequence (5' to 3')
MTC Fw	25	AGACCGTGCGGATCTTG
MTC Rv	25	CATGGAGATCACCCGTGA
MTC probe	10	HEX-ACGGATTGGTCACCCGGATT-BHQ1
*wbbl1*_Fw	25	TACCAGCTTCAGTTTCCGT
*wbbl1*_Rv	25	GCACCTATATCTTCTTAGCCG
*wbbl1*_probe	10	FAM-ATGGTGCGCAGTTCACTGC-BHQ1

2.6. Denaturing gradient gel electrophoresis (DGGE) and sequencing

PCR-DGGE was performed using the DCodeTM Universal Detection System according to manufacturer's instructions (Bio-Rad Laboratories, CA, USA). The PCR products were loaded onto an 8% polyacrylamide gel (acrylamide-bis-acrylamide 37.5: 1) in 0.5 X TAE buffer. The denaturing gradient for the *Mycobacterium* genus specific JSY16S PCR products was 45%-55%. The denaturant gradient for the SG specific APTK16S PCR products was 40%-60%. 100% denaturant corresponds to 7M Urea and 40% [vol/vol] deionized formamide (Muyzer et al. 1993). PCR products (22µl) with 8µl of Fermentas 6x DNA loading dye were loaded into each lane of the gel. The gels were then run at 60V for 16.5 h and maintained at a constant temperature. Electrophoresis was carried out initially for 10 minutes at 150V and subsequently for 960 mins at 60V at 60^{o}C in 7 litres of 0.5X TAE buffer (40mM Tris-acetate and 1mM EDTA, pH8.0). The gels were stained with Ethidium bromide (10mg/ml) for 15 mins and destained with distilled water for

20 minutes and photographed (GeneFlash Gel Documentation System, (Syngene, Cambridge, UK). Major bands were excised and then reamplified with either the JSY16S primer set or the APTK16S primer set. The PCR products were purified using the QIAquick Gel Extraction Kit (QIAGEN, Ltd., Crawley, UK) purification kit according to the manufacturer's instructions, and then sequenced using an ABI PRISM 3130xl Genetic Analyser. All sequences were aligned using ClustalW (www.ebi.ac.uk/Tools/msa/clustalw2) and compared with the gene database using local BLAST (http://www.ncbi.nih.gov/) as described in section 2.8.4.

2.6.1. PCR-DGGE specificity

The DNA extracted from *Mycobacterium* species listed in Table 2.3 was used to assess the gradient required for sufficient differentiation of species as determined by band positions using the *Mycobacterium* genus (JSY16S) and SG (APTK16S) primer sets.

2.6.2. PCR-DGGE sensitivity

Cryfield soil was spiked with water (control) or with *M. bovis* BCG cells (10^1-10^6 cellcopies/ml) in duplicate. The *M. bovis* BCG culture was quantified using an RD4 specific qPCR (section 2.5.2) and the culture was then used to make the dilution series. Spiked soils were left overnight at room temperature and then 0.5g of soil was extracted using the FASTDNA Spin Kit as per manufacturer's instructions. A nested and non-nested approach was compared for both primer sets JSY16S and APTK16S (section 2.5.1), and DGGE was performed.

2.6.3. Water DNA extraction kit comparison using PCR-DGGE

Tap water was spiked with *M. bovis* BCG cells (10^2-10^5 cell copies/ml) and filtered using a 50ml sterile plastic syringe and the MicrofilV filtration device with 0.22µm mixed cellulose esters white gridded filters (Millipore, MA, USA). Six sets of dilution series were made, intended for three different extraction methods, and the comparison of immediate DNA extraction and DNA extraction from filters after 14 days at -20°C. The three different DNA extraction methods included the commercial kit PowerWater® DNA Isolation Kit (MoBio Laboratories, Inc., CA, USA), the manual adapted Griffiths method (Griffiths et al. 2000) and the manual adapted Pickup method (Pickup et al. 2004). The extracted metagenomic DNA was used as templates for nested PCR reactions using primers JSYS16S and APTK16S (section 2.4) and comparative PCR-DGGEs were carried out (section 2.6.1)

2.7. Pyrosequencing

From the 210 (42 villages x 5 sites) extracted soil DNA samples, DNA was pooled into the respective villages. Similarly the 84 water DNA samples (42 villages x 2 sites) were pooled at the village level. Samples were normalised using readings from the NanoDrop 1000 spectrophotometer (NanoDrop products, Wilmington, DE, USA) to ensure approximately equal amounts of nucleic acid (ng) from each site contributed to the final composite metagenomic DNA in each village sample. DNA was pooled before PCR amplification for soil and water from each village so that mycobacteria DNA communities could be analysed on the village level.These composite samples were sent to the Research and Testing Laboratory (RTL) (Lubbock, TX) for PCR amplification and pyrosequencing.

2.7.1. Pyrosequencing PCR

Two sets of primers were used for the Bacterial tag-encoded FLX Titanium amplicon pyrosequencing PCR method; JSY16S and APTK16S-Pyro (Table 2.4). The HotStarTaq Plus Master Mix Kit (Qiagen) was used for PCR under the conditions 94^{o}C for 5 min, followed by 35 cycles of 94 ^{0}C, 55 ^{0}C for 1 min, 65 $^{\circ}$C for 1 min, and finally a single extension step of 65 ^{0}C for 5 min for JSY16S. For APTK16S the PCR was carried out under conditions 94^{o}C for 5 min, followed by 35 cycles of 94 ^{0}C, 59.9 ^{0}C for 1 min, 67 $^{\circ}$C for 1 min, and finally a single extension step of 67 ^{0}C for 5 min. A secondary PCR was performed which incorporates the tags and linkers into the primary amplicon for use in FLX amplicon sequencing. The tags were a sample specific sequence used to identify the sequence from a bulk reaction. The amplicon products were then mixed in equal volumes and purified using Agencourt Ampure Beads (Agencourt Bioscience, Beverley, MA).

2.7.2. Pyrosequencing method

An FLX sequencing run was performed on a 70x75 GS PicoTiterPlate (PTP) using the Genome Sequencer FLX system (Roche) and a Titanium platform. Titanium reagents and a mixture of HotStart and HotStar high fidelity Taq polymerases in addition to a one-step PCR with 30 cycles were used to reduce chimera formation. Pyrosequencing was performed at the Research and Testing Laboratory (RTL) (Lubbock, TX) based upon RTL protocols (www.researchandtesting.com).

2.8. Pyrosequencing bioinformatic in-house pipeline

An in-house bioinformatics pipeline designed by Dr. B. Oakley facilitated the quality control of sequences and the constructed alpha and beta diversity statistics and graphics (Oakley et al. 2012). Alpha diversity is defined as the diversity present within a village such as species richness. Beta diversity is defined as the differences

in diversity between different villages such as those illustrated in CCA and PCA plots. The in-house pipeline was used to quality control the pilot dataset presented in Chapter 3. The pipeline was also employed to create additional figures for Ethiopian datasets presented in Chapters 4 and 5.

2.8.1. Quality control of sequences

The pipeline consists of Perl and BioPerl scripts on a Linux platform which were used to filter, rename and trim sequences based on three main criteria: (1) The sequence matched the forward proximal primer, (2) the length of the amplicon was between the expected sizes of 466-472bp and 420-424bp for JSY16S and APTK16S respectively and (3) the sequence contained no ambiguous bases. This criteria was used to improve the quality of the data (Huse et al. 2007). APTK16S dataset sequences were reverse complemented so that the sequences were the correct 5' to 3' orientation. Perl scripts 1-4 (Table 2.6) were used for these quality control procedures. The final output file was a single fasta formatted file including all sequences which have passed quality control. This fasta file was then used for downstream analysis.

2.8.2. Similarity-based OTU classification

Quality controlled sequences were clustered into Operational Taxonomic Units (OTUs) using the CD-HIT-EST package (Li and Godzik 2006). The user-defined cut-off was 97% sequence similarity. CD-HIT-EST orders the sequences according to the length; from longest to shortest. The longest sequence is the representative sequence for the first cluster; the subsequent sequences ordered by length are either clustered with a previous group or constitute a representative sequence for an additional cluster. Default options were chosen apart from the –g parameter was set

to 1, so that sequences were clustered with the most similar OTU (slow mode) and do not cluster with the first OTU that meets the similarity threshold (fast mode).

Table 2.6: Names of Perl & BioPerl scripts and explanations of the function

Number	Name of script	Purpose of script
1	my_fasta_file_concatenator_renamer.pl	Removes original sequence name, renames using file name and joins files
2	Pyro2_one_primer_screen.pl	Removes any sequences that do not contain the primer sequence and removes bases preceding it
3	Pyro3_length_summary_and_trimming.pl	Removes sequences that are not the specified length in bases and removes sequences with ambiguous N bases
4	reverse_complements.pl	Reverse complements sequences in a fasta file
5	my_cluster_cleaner.pl	Converts output file from CD-HIT into a single data table, input for R
6	my_cd-hit_to_mothur_converter.pl	Converts output file from CD-HIT into a MOTHUR formatted file
7	my_group_file_maker.pl	Lists the sequence IDs and sample names
8	make_rarefaction_curves_color.r	Produces rarefaction curves from the cluster cleaner output file
9	make_richness_diversity_stats.r	Produces richness estimates from the cluster cleaner output file
10	mantel_cca_stats_tk.r	Produces a text file of the mantel and CCA results and significance
11	mantel_cca_plotting_tk.r	Produces the CCA plots/figures
12	randomise_sequences.R	Randomly selects a specified number of sequences from fasta file

2.8.3. Alpha & Beta diversity analysis

The output from CD-HIT-EST was then reformatted using the Perl scripts 5-7 (table 2.6) to provide the input for R (R Development Core Team 2011) and mothur (Schloss et al. 2009) downstream rarefaction curves and to calculate alpha diversity estimates such as nonparametric Shannon and Simpson diversity indices (Table 2.6

scripts 8-9). Dendrograms were calculated from pairwise comparisons using the Jaccard Index. The Jaccard Index calculates the number of shared OTUs in two communities divided by the sum of total number of OTUs in community A and community B, minus the number of shared OTUs.

Mantel and Partial Mantel tests were calculated using R scripts (Table 2.6, scripts 10-11). The Mantel test assesses the relationships between species composition and environmental variables by a pairwise regression analysis and as a result determines the variation in species composition that can be explained by environmental and spatial factors. Bray-Curtis dissimilarity matrices are constructed for the species data and the environmental variables and using regression analysis a correlation coefficient matrix is generated. The significance of the correlation coefficient matrix is determined by 999 permutation tests. Partial Mantel tests determine whether there is residual variation after controlling for an environmental factor.

Canonical Correspondence Analysis (CCA) plots based on an OTU frequency table were constructed using R scripts (Table 2.6, scripts 10-11). CCA is a direct ordination technique in which the axes are linear combinations of the environmental variables (Ter Braak 1986). It is calculated using reciprocal averaging: at each stage of the averaging process a multiple regression is performed of the sample scores on the environmental variables. New scores are generated and this process continues until the scores stabilise. The results represent the most variability in the species data which is restricted by linear combinations of the environmental variables and the species data. The significant environmental variables were shown as arrows pointing in the direction in which the variables cause the most change. Longer arrows represent the variables which were more important in explaining the distribution of samples. Arrows pointing in the same direction suggest that these variables were

positively correlated whereas arrows pointing in the opposite direction were negatively correlated.

2.8.4. Local BLAST for taxonomic matches

Taxonomic assignment of the all sequences was carried out using local BLAST. *Mycobacterium* species were retrieved from the SILVA 16S database SSU ref release 108 (Accessed 16th August 2011) (Pruesse et al. 2007); this included 1019 unique *Mycobacterium* species. Using local BLASTn (Altschul et al. 1990), sequences with a BLAST match with an e-value of ≤ 0.001 were retained for analysis. JColorGrid was used to visualise the prevalence of the BLAST matches (Joachimiak et al. 2006).

2.8.5. Identification of potential Chimeras using ChimeraSlayer

Sequences were first formatted using the NAST-iEr utility which uses the Nearest Alignment Space Termination (NAST) algorithm to create multiple sequence alignments (DeSantis et al. 2006). In more detail, the sequences were aligned with a Greengenes database which has approximately 10,000 non-chimeric 16S rRNA gene bacteria and archaea sequences. The programme ChimeraSlayer was subsequently used to identify chimeras by first determining potential parents and then checking whether the sequence had greater homology to the computed chimeras compared to the reference database (Haas et al. 2011).

2.8.6 Random Resampling

To assess the effect of uneven numbers of sequences per sample, a random resample of sequences was taken from each sample to control for sampling effort. This was used to assess whether random resampling produced any significant differences in alpha and beta diversity compared to the analysis of all sequences. An R script (Table 2.6, script 12) was used to randomly select a specified number of sequences

from a fasta file and the in-house pipeline alpha-diversity and beta diversity analysis was repeated for the randomised sequences.

2.9. The QIIME pyrosequencing bioinformatics pipeline

Quantative Insights into Microbial Ecology (QIIME) is an alternative bioinformatics pipeline to analyse pyrosequencing data. Comparisons between the programs and analysis of the QIIME pipeline relative to the in house pipeline are detailed in Figure 2.14. The QIIME pipeline was used to quality control the Ethiopian datasets presented in Chapters 4 and 5.

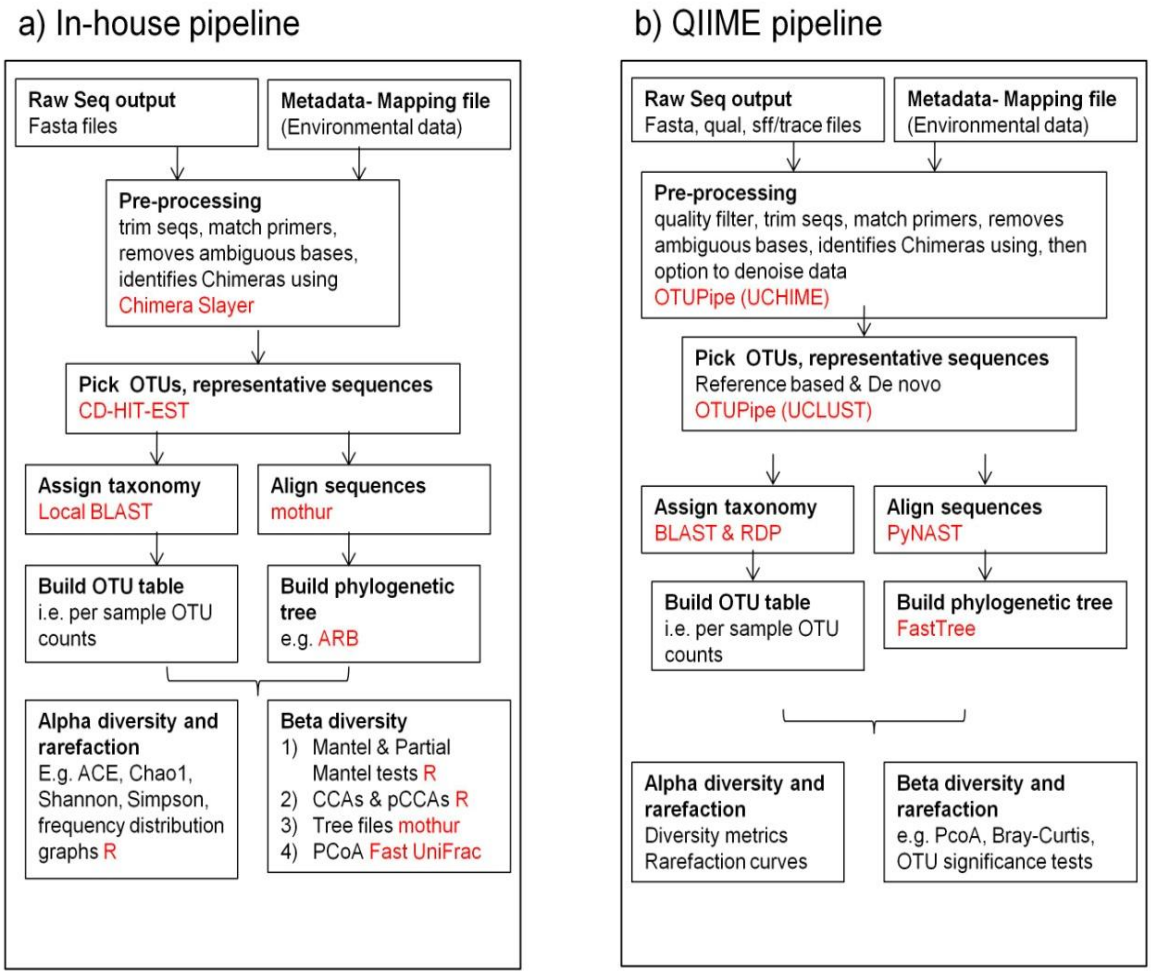


Figure 2.14. Comparison of **(A)** the in-house Warwick pyrosequencing bioinformatic pipeline compared to **(B)** the QIIME pyrosequencing pipeline. The steps are highlighted bold and the programs employed are highlighted in red.

2.9.1. QIIME Preparation of data

To run the QIIME pipeline using pyrosequencing 454 data, sff files or fasta and quality files are required in addition to an environmental metadata. The sff file provided from the sequencing centre contained a fasta (fna file) of all the sequences, a quality file and a flowgram. The QIIME pipeline also required a mapping file containing information of all the samples, barcodes, primers and environmental data.

2.9.2. Quality control procedures

The python script *split libraries.py* was used to filter the sequences based on user specified criteria. Criteria included trimming the length of the sequences, matching the proximal primer and quality score assessment. For *Mycobacterium* genus (JSY16S) pyrosequencing dataset, sequences <400bp were removed. Sequences had to match the forward proximal primer with a maximum of 1 mismatch and had to have an average quality score of ≥25. The length of >400bp was chosen based on the histograms of raw data and positions of variability across the amplicon. For the SG (APTK16S) pyrosequencing dataset, sequences that had a sequence length of <420bp were removed. Sequences had to match the forward proximal primer with a maximum of 1 mismatch and had to have an average quality score of ≥25. The minimum length of 420bp was chosen based on the frequency histogram of raw sequences and had to be long enough to include the long helix 18 insert necessary for the discrimination of SG. Sequences were also reverse complemented so that the sequences were in the 5' to 3' orientation.

2.9.3. OTUPipe: Clustering, chimera check and denoising

Datasets were checked for chimeras and denoised using OTUPipe (Edgar 2011a). Sequences were first sorted according to length, then dereplicated and subsequently clustered into Operational Taxonomic Units (OTUs) using UCLUST (Edgar 2010)

which created "seed" representative sequences for each cluster. Each sequence was then compared for >97% similarity to the seed sequences; those with <97% became the seed sequence for the next cluster and so forth. Chimera checks were carried out in UCHIME within OTUPipe (Edgar et al. 2011b) using both a *de novo* abundance based method and a reference based method against 10,000 reference chimera-free archaeal and bacterial sequences (the Gold Genomes Online Database 3.0). Following this, the data was denoised by removing OTU clusters which had 1 sequence per cluster with subsequent iterative reclustering. The most abundant sequence from each OTU became the representative sequence. After quality control procedures, samples with less than 385 sequences per sample were removed for the purpose of analysis. The program UCLUST removes singleton data in accordance with research which has shown that this improves the quality of the data (Zhou et al. 2011; Behnke et al. 2011).However, it is possible that some legitimate data is excluded and therefore a conservative dataset remains.

2.9.4. QIIME taxonomic assignments

A BLAST application within the QIIME virtualbox was used to assign taxonomic species to representative sequences. *Mycobacterium* species were retrieved from the SILVA 16S database SSU ref release 108 (Accessed 16th August 2011) (Pruesse et al. 2007) and subsequently the representative sequences were matched to sequences using BLAST. The maximum e-value for the assignment to be recorded was 0.001.

2.9.5. Alignments and phylogenetic trees

To align the representative sequences a python implemented NAST (Nearest Alignment Space Termination) algorithm (PyNAST) was used (Caporaso et al. 2010a; DeSantis et al. 2006). Sequences were aligned to a "Greengenes core set" which contains 4938 non-chimeric candidate sequences which were already PyNAST

aligned. The representative sequence alignment was filtered using a dynamic lanemask, which filters gap positions if they occur in >80% of the sequences with entropy threshold set at 10% with the top 10% most variable positions filtered. Alternatively, sequences were aligned using the align.seqs function in mothur (Schloss et al. 2009), using the customised template database constructed from SILVA containing *Mycobacterium* species 16S rRNA genes. Aligned databases were imported into ARB for the construction of phylogenetic trees using the maximum-likelihood algorithm in ARB (Ludwig et al. 2004).

2.9.6. Alpha & Beta diversity analysis

The OTU summary table containing the frequencies of sequences in each OTU was the input for rarefaction plots and alpha diversity metric the phylogeny based metric Phylogenetic Diversity (PD). The PD metric takes into account the total phylogenetic branch length belonging to each sample from the input phylogenetic tree of representative sequences, the higher the number the more diverse the sample is (Faith 1992). The OTU table and the phylogenetic tree containing all representative sequences were used to generate principal coordinate analysis plots (PCoA) to display beta diversity. Unweighted and weighted analyses were both carried out; unweighted analysis looks at the presence/absence of OTUs whereas the weighted analysis looks at the prevalence of OTUs. Pearson correlation tests were calculated to test the significance between the continuous environmental and spatial data and OTU prevalence. OTU prevalence was only accepted as significant after Bonferroni correction ($P<0.05$). Random resampling was also carried out within QIIME to assess differences in alpha and beta diversity.

2.10. Additional statistical analysis

Generalised linear models (GLM) were calculated using STATA 11.0 (Stata Corporation, College Station, TX, USA). The outcome dependent variables were the Shannon diversity estimates and the abundance of the *Mycobacterium* genus as determined by qPCR. Outcome variables were tested for normal distributions using the Shapiro-Wilk test. The independent explanatory variables were the six environmental and spatial variables. Explanatory variables were individually tested; variables that were significant (P <0.05) were entered into multivariate analysis which was then reduced to the minimum adequate model by backward stepwise elimination of non-significant variables. One-way ANOVA was calculated using STATA 11.0 and was implemented to assess the relationship between regions and *Mycobacterium* genus qPCR abundance and the number of different BLAST matches. Linear and quadratic regression models were constructed to determine the relationships between the *Mycobacterium* genus abundance, diversity and richness estimates. Akaike's Information Criteria (AIC) was used to identify the most appropriate model; the model with the lowest AIC value was presented.

2.11. ArcGIS mapping

ESRI map shapefiles were obtained from the Ministry of Agriculture and Rural Development (MARD) containing information regarding administrative structure and climatic data. Selected shapefiles were visualised using ArcGIS 10.0 (ESRI, Redlands, CA, USA). The TB disease incidence data collected from the health centres and obtained from the Ministry of Health (MoH) (section 2.2.6) was added to the attribute table and plotted onto the administrate shapefile map of Ethiopia.

Chapter 3: Optimisation of approaches to analyse *Mycobacterium* species diversity in soil and water

3.1. Abstract

In order to study the diversity of mycobacteria in the environment, it was necessary to compare methods for DNA extraction and detection. Cultivation approaches are known to be insensitive for the detection of mycobacteria and therefore was not considered appropriate for this diversity study. The molecular methods PCR-DGGE and pyrosequencing were optimised to target the 16S rRNA gene of mycobacteria and separately a group of SG. A combination of DNA extraction methods and PCR-DGGE were shown to have a detection limit of 10^5 cells per gram for soil and 10^2 cells per ml for water samples. The pilot study of five soil samples demonstrated an unprecedented diversity of FG and SG. The PCR-DGGE and pyrosequencing methods provided a consensus of the diversity of mycobacteria, for example a high relative abundance of pyrosequencing reads and DGGE bands were matched to *M. moriokaense*, *M. colombiense* and *M. riyadhense*. Overall, the pyrosequencing approach provided a comprehensive assessment of the diversity of mycobacteria and has the capacity to be employed for larger scale surveys.

3.2. Introduction

The challenge for microbial ecologists is to accurately determine the diversity and abundance of environmental microbial communities. It is accepted that there are approximately 1.5 $x10^{10}$ cells of bacteria per gram of soil and that this is underestimated by cultivation techniques, which only successfully isolate 1% of cells (Torsvik et al. 1990). Cultivation techniques are known to skew the profile of isolates obtained and provide inaccurate estimations of species richness and evenness. As such molecular methods have been extensively employed, requiring a robust method

for metagenomic DNA extraction from environmental samples. However this procedure also presents its own challenges, for example soil samples can contain inhibitory compounds such as humic and fulvic acids and it is difficult to extract DNA as it tends to adsorb to soil particles.

There is currently no standard method for direct DNA extraction from soil, but there are many published methods that use similar procedures in order to obtain sufficient DNA yield and purity (Frostegard et al. 1999; Griffiths et al. 2000; Krsek & Wellington 1999; Petric et al. 2011; Tsai & Olson 1991; Towe et al. 2011). Two main approaches exist: (1) separation of bacteria from soil before DNA extraction; and (2) direct DNA extraction from soil (Robe 2003). The first approach is achieved using centrifugation to separate the bacteria from soil, DNA is then extracted by cell lysis and purified (Faegri et al. 1977). The first approach is more advantageous as the extract contains less humic acids and the less DNA from non-bacterial sources. However, the second approach provides a higher DNA yield and is less time consuming. This procedure generally includes physical lysis, for example using a bead-beating technique or freeze thaw cycles and the addition of a detergent such as sodium dodecyl sulfate (SDS) for chemical lysis. DNA is subsequently precipitated and purified by the addition of alcohol, silica membrane separation or by using agarose gel electrophoresis (Schneegurt et al. 2003).

A comparison of five DNA extraction methods from soil used qPCR to assess *M. bovis* recovery from spiked soils (Pontiroli et al. 2011). The comparison found that the commercial FastDNA spin kit and the manual adapted Griffiths method produced the highest analytical sensitivity of $8.5x10^4$ cells per gram. This finding was supported by an evaluation of three soil DNA extraction methods, which found that the FastDNA spin sample kit (Bio101, La Jolla, Calif) yielded the highest amounts of

16S rDNA product in several different soil types (Martin-Laurent et al. 2001). Consequently the FastDNA spin kit was chosen as a suitable DNA extraction method for soil samples, as it was shown to be most sensitive and included a harsh mechanical step required to breakdown the tough cell wall of gram-positive bacteria such as mycobacteria.

Community DNA extraction from water has been less well studied, however steps generally include filtration, centrifugation, physical and chemical lysis, followed by DNA precipitation and purification. Published methods include the Somerville method; designed for aquatic environments, it filters the water sample (<1 litre) before cell lysis is achieved using SDS and proteinase K (Somerville et al. 1989). The more recent Weinbauer method is similar to the Griffiths method in that it relies on both mechanical bead beating and centrifugation procedures for the simultaneous extraction of both DNA and RNA (Weinbauer et al. 2002). An alternative freeze-thaw extraction method was used to detect *Mycobacterium* species from water supply systems (Kawai et al. 2004). Methods for metagenomic DNA extraction from water follow similar protocols to that used for soil samples, however little has been done to compare different DNA extraction methods from water.

This study has evaluated three methods of DNA extraction; one commercial method PowerWater DNA Isolation Kit (MO BO Laboratories, Inc.) and two manual methods adapted Griffiths and Pickup methods (Griffiths et al. 2000; Pickup et al. 2004). All methods began with a filtration step through a 0.22µm-polycarbonate filter. The PowerWater kit and the adapted Griffiths method use mechanical bead-beating to lyse the cells. Subsequently both the adapted Griffiths and adapted Pickup methods have centrifugation steps and in contrast the PowerWater kit uses silica membrane separation to isolate the DNA.

After successful isolation of metagenomic DNA, the identification of suitable genomic regions was required for the assessment of diversity. In this study the target had to include conserved regions specifically for the *Mycobacterium* genus and contain enough variability within the target for the differentiation of species. Various genomic regions have been targeted to study *Mycobacterium* species diversity such as the RNA polymerase B-subunit (*rpoB)* (Adekambi et al. 2006a; Lim et al. 2008; Salah et al. 2008), heat-shock protein 65 (*hsp65)* (Kim et al. 2010; Pourahmad et al. 2009; Ringuet et al. 1999), DNA gyrase B-subunit (*gyrB*) (Chimara et al. 2004; Kasai et al. 2000) and the 16S-23S internal transcriber spacer region (16S-23S ITS) (Ngan et al. 2011; Roth et al. 1998; Xiong et al. 2006). However the 16S rRNA gene is the most commonly used sequence for species level identification (Boddinghaus et al. 1990; Dobner et al. 1996; Kirschner et al. 1993; Kox et al. 1995; Nieminen et al. 2006). As a result the Genbank database entries for *Mycobacterium* 16S rRNA gene are by far the most comprehensive (5930 entries) compared to *rpoB* (2377 entries), 16S-23S ITS (710 entries) and *gyrB* (406 entries). However it should be acknowledged that the 16S rRNA gene has a high degree of conservation in mycobacteria and therefore in some cases species are indistinguishable. Despite this limitation, the 16S rRNA gene for *Mycobacterium* is the most widely studied, and was therefore chosen for this project. It has been shown previously that the *Mycobacterium* genus 16S rRNA gene primer set (JSY16S) employed in this study predominantly amplifies fast-growing species, perhaps due to their higher abundance in the environment or that many FG have two 16S rRNA gene copies compared to one copy (Young et al. 2005). Therefore a second set of primers was designed to target SG containing the long helix 18. The molecular methods DGGE and

pyrosequencing have been used extensively for the evaluation of microbial diversity and were employed here to target *Mycobacterium* species diversity.

3.3. Aims

i) To optimise PCR-DGGE to assess *Mycobacterium* species diversity using primers to target the whole *Mycobacterium* genus and a primer set specifically for a group of SG.

ii) To validate an in-house bioinformatics pipeline with pyrosequencing runs for both the *Mycobacterium* genus and SG from five pilot soils.

iii) To assess the sensitivity of three different DNA extraction methods from water samples.

iv) To compare PCR-DGGE and pyrosequencing approaches for estimating *Mycobacterium* species diversity using five pilot soil samples.

3.4.1. PCR-DGGE Specificity: Differentiation of *Mycobacterium* species

DNA from cultures of *Mycobacterium* species were used to optimise the DGGE conditions needed for adequate separation of species. The *Mycobacterium* genus DGGE was optimised to a gradient of 45%-55% to obtain the greatest resolution and separation of bands as possible. From a total of 11 species, there were 8 bands which electrophoresed to the same position on the gel this included *M. aichense* and *M. gilvum, M. fortuitum* and *M. phlei* and *M. agri* and *M. peregrinum* (Figure 3.1A). The DGGE specifically for the SG was optimised to a gradient of 40%-60%, five bands electrophoresed to the same position on the gel, this included *M. marinum* and *M. xenopi* (Figure 3.1B).

3.4.2 PCR-DGGE Sensitivity: BCG spiked soils

Sensitivity was assessed by the detection of *M. bovis* spiked soil at different concentrations (10^1-10^6 cells per gram) after DNA extraction DGGEs were performed for the *Mycobacterium* genus and the SG (section 2.6.3). The sensitivity of the *Mycobacterium* genus DGGE was 10^5 cells per gram (Figure 3.2A), and this did not improve using a nested approach (Figure 3.2B). The DGGE profiles were the same for the non-nested and the nested approach, suggesting the nesting procedure does not change the community structure. The sensitivity of the SG PCR-DGGE was much higher; the non-nested approach detection threshold was 10^3 cells per gram (Figure 3.3A) and this was improved on with the nested approach to 10^2 cells per gram (Figure 3.3B). The nested approach appeared to be more sensitive as an extra band was present, however this band was not present at spiking concentrations above 10^4 cells per gram perhaps due to the bias in the PCR reaction toward higher concentrations of *M. bovis* template.

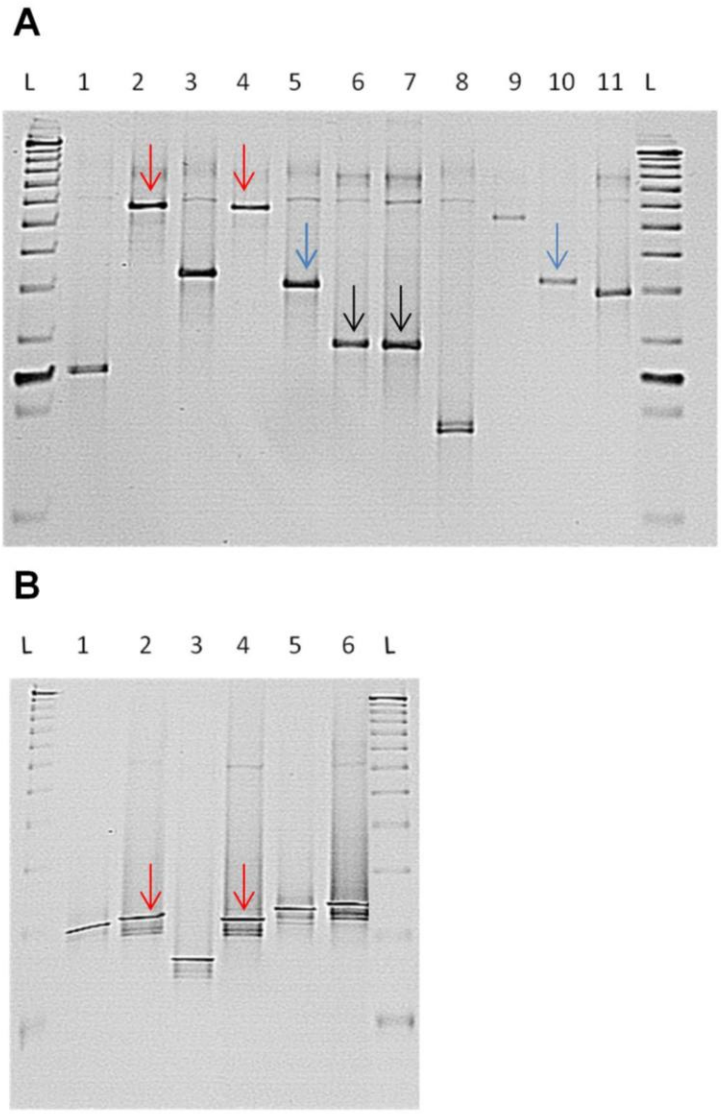


Figure 3.1. DGGE profiles to show separation of mycobacteria 16S rDNA PCR amplicons. **(A)** using the *Mycobacterium* genus specific primer set (JSY16S). L is a reference ladder, lanes 1-11 are respectively; *M. smegmatis*, *M. aichiense*, *M. aurum*, *M. gilvum*, *M. phlei*, *M. agri*, *M. peregrinum*, *M. duvalii*, *M. abscesses*, *M. fortuitum* and *M. vaccae* **(B)** SG specific primer set (APTK16S). L is a reference ladder, lanes 1-6 respectively; *M. intracellulare*, *M. marinum*, *M. kansasii*, *M. xenopi*, *M. avium paratuberculosis*, *M. bovis BCG*. Coloured arrows show identical positions.

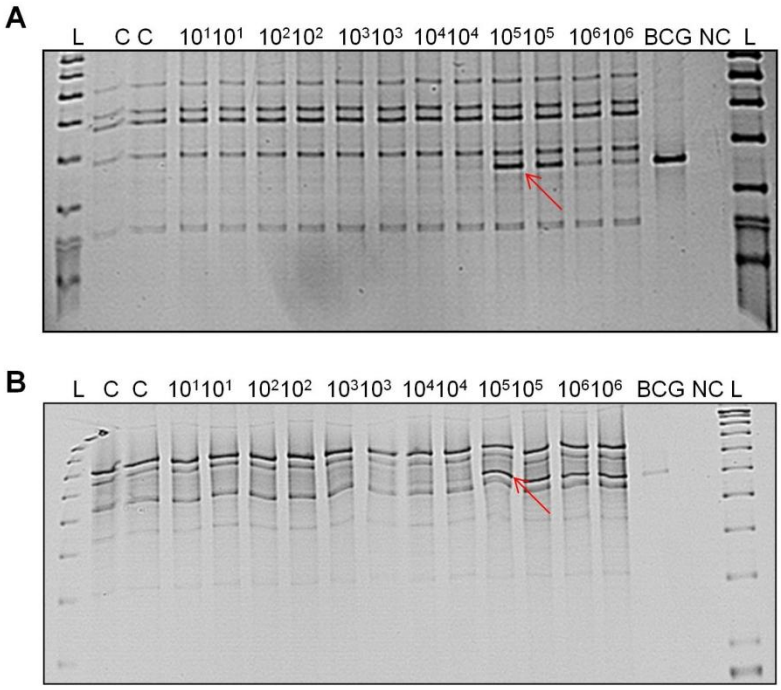


Figure 3.2. *Mycobacterium* genus JSY16S sensitivity DGGEs. Control is Cryfield soil spiked with water. Dilutions of *M. bovis* BCG range from 10^1 to 10^6 cell equivalents of genomic DNA in duplicate. **(A)** Non-nested PCR approach **(B)** Nested approach. Arrows show limit of detection.

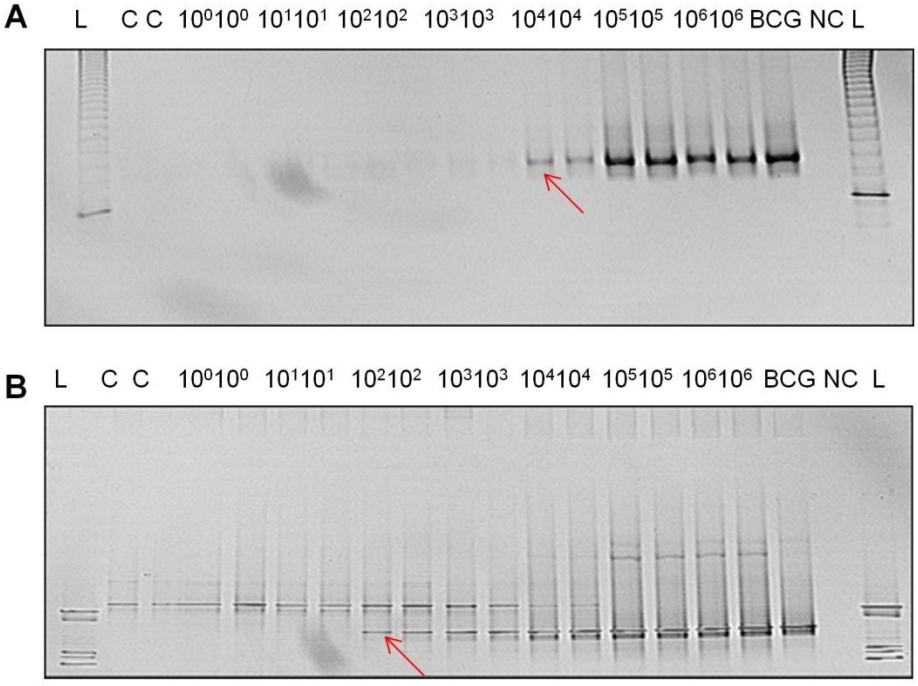


Figure 3.3. SG (APTK16S) sensitivity DGGEs. Control is Cryfield soil spiked with water. Dilutions of *M.bovis* BCG range from 10^1 to 10^6 cell equivalents of genomic DNA in duplicate. **(A)** non-nested PCR approach **(B)** nested PCR approach. Arrows show limit of detection.

3.5. Water DNA extraction kit comparison

Methods for DNA extraction from water samples were assessed for sensitivity using spiked water samples with known concentrations of *M. bovis* BCG. DGGE was used to assess the recovery of *M. bovis* BCG. One commercial kit PowerWater DNA Isolation Kit (MoBio Laboratories, Inc., CA, USA) was compared to two manual methods; the adapted Griffiths (Griffiths et al. 2000) and Pickup (Pickup et al. 2004) methods. The extraction of DNA 14 days after water filtration was compared to immediate DNA extraction to mimic the delay from collection to extraction in a field study. The *Mycobacterium* genus DGGE showed PowerWater was the most sensitive method detecting 10^2 *M. bovis* cells per ml, compared to the Griffiths (10^4 cells per ml) and Pickup methods (10^5 cells per ml) (Figure 3.4A). The 14 day delay of DNA extraction did not change the sensitivity of the PowerWater or the Griffiths methods, but did change the sensitivity of the Pickup method so that it no longer detected 10^5 cells per gram (Figure 3.4B). The higher sensitivity of the PowerWater method was further demonstrated by the SG DGGE whereby sensitivity remained at 10^2 cells per gram even after the delay of 14 days (Figure 3.5). The Griffiths method was the second most sensitive method at 10^4 cells per gram and sensitivity appeared to increase to 10^2 cells per gram after 14 days of storage (Figure 3.5B). The Pickup method was the least sensitive, detecting 10^5 cells per gram after immediate DNA extraction but failing to detect this after 14 days of storage (Figure 3.5). Overall, the commercial PowerWater kit appeared to be the most sensitive method and the least affected by 14 days of storage before DNA extraction. It might be suggested that the mechanical cell lysis bead-beating step was more efficient, as the Pick-up method did not include this step.

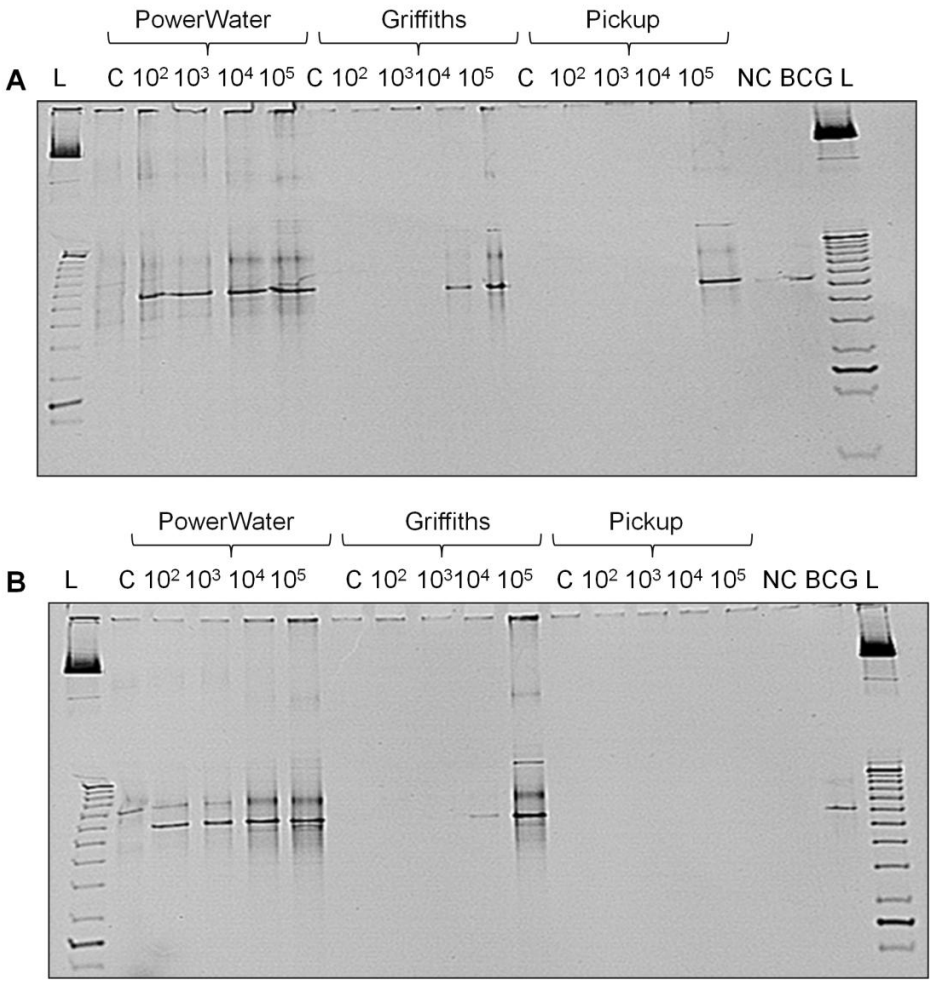


Figure 3.4. Detection of *M. bovis* using three water DNA extraction kits; PowerWater, Griffiths and Pickup using *Mycobacterium* genus primers. Dilutions of *M. bovis* BCG range from 10^2 to 10^5 cell equivalents of genomic DNA per ml. C is control Cryfield soil spiked with water, L is the marker ladder, NC is the negative control PCR reaction without *M. bovis* BCG, BCG is the positive control PCR reaction with *M. bovis* BCG **(A)** Immediate DNA extraction **(B)** DNA extraction after 14 days.

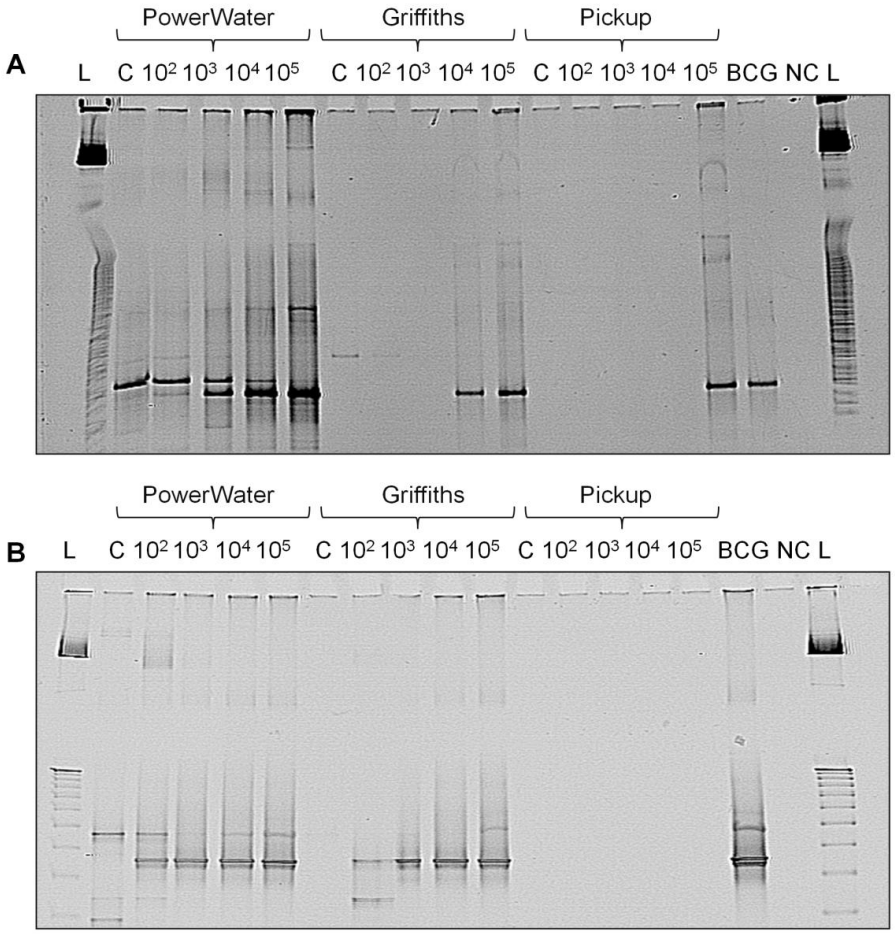


Figure 3.5. Detection of *M. bovis* using three water DNA extraction kits; PowerWater, Griffiths and Pickup using SG primers. Dilutions of *M. bovis* BCG range from 10^2 to 10^5 cell equivalents of genomic DNA per ml. C is control Cryfield soil spiked with water, L is the marker ladder, NC is the negative control PCR reaction without *M. bovis* BCG, BCG is the positive control PCR reaction **(A)** Immediate DNA extraction **(B)** DNA extraction after 14 days

3.6. Pilot study: Community analysis using PCR-DGGE

The description of pilot soil sites is provided in section 2.1. DGGE gels were analysed used the GelCompar software which calculated band intensity and band positions which were used to generate Shannon diversity indices and cluster dendrograms. DGGE profiles of four Ethiopian soils and one UK soil for *Mycobacterium* species have shown very different community profiles for all of the soils suggesting high diversity (Figure 3.6). The *Mycobacterium* genus DGGE showed unique profiles for all the soil samples with differences in band positions and intensities suggestive of different species and abundances (Figure 3.6A). The SG DGGE displayed fewer bands for each soil sample compared to the *Mycobacterium* genus DGGE, which was expected as a smaller group of mycobacteria were targeted (Figure 3.6B). Cluster analysis of the *Mycobacterium* genus DGGE based on the position of bands and band intensity suggested that Cryfield and 1110 are most closely related, followed by 1109 and 1111 (Figure 3.7A). However, cluster analysis of the SG suggested 1110 and 1111 were most closely related followed by 1108 and Cryfield (Figure 3.7B). The Shannon diversity indices (H') suggested that Cryfield was the most diverse in the *Mycobacterium* genus DGGE, but was the least diverse for SG (Table 3.1) perhaps due to a higher diversity of FG that outcompete the SG. However, soil 1108 had the highest SG H' estimate and the second highest *Mycobacterium* genus H' index estimate (Table 3.1).

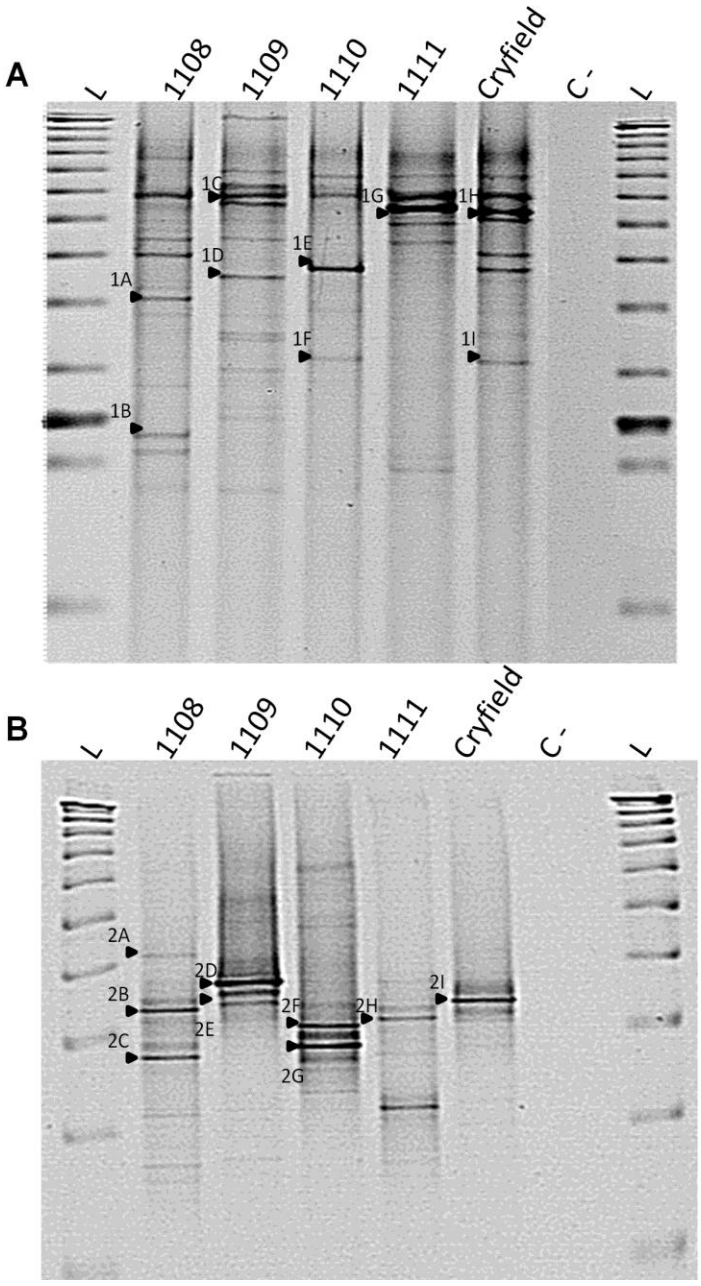


Figure 3.6. DGGE profiles of environmental mycobacteria communities in soil. L is a reference ladder, C- is the negative PCR control, 1108-1111 are the four Ethiopian soils and Cryfield is the UK soil. Lane 7 is the negative PCR control. The arrows (A-I) represent bands that were excised and sequenced (Table 3.2 &3.3). **(A)** using *Mycobacterium* genus primers (JSY16S) **(B)** using SG primers (APTK16S)**.**

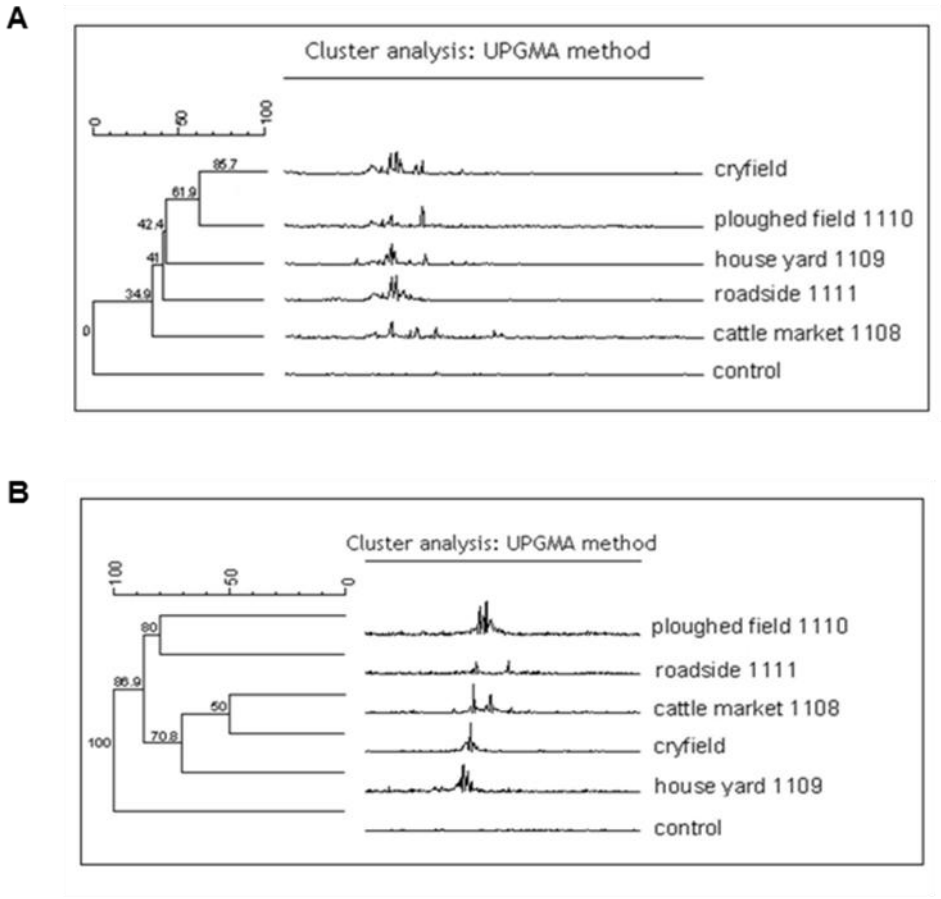


Figure 3.7. Dendrogram cluster analysis of the DGGE profiles of environmental mycobacteria communities in soil **(A)** using *Mycobacterium* genus 16S rDNA specific primers (JSY16S) corresponding to Fig 3.6A. **(B)** using the SG specific 16S rDNA specific primers (APTK16S) corresponding to Figure 3.6B.

Table 3.1. The Shannon Diversity Index (H') calculated from the number and intensities of the DGGE bands. GelCompar software (Applied Maths, Inc. TX, USA) for the PCR-DGGE gels for the five soils (Figure 6)

Soil	*Mycobacterium* genus (H')	SG (H')
Cattle market (1108)	1.851	1.565
House yard (1109)	1.848	1.334
Ploughed field (1110)	1.231	1.339
Roadside (1111)	1.49	0.691
Cryfield	1.864	0.578

The *Mycobacterium* species composition of the soil samples was determined by extraction of bands and Sanger sequencing. From the *Mycobacterium* genus DGGE sequenced bands none of the soils shared the same species, even though several bands were excised from the same position (Table 3.2). All species identified were

FG, which suggested that the group are more abundant than SG or that the primer set is biased toward the detection of FG perhaps owing to the fact that the majority of the group have two 16S rRNA gene copies whereas SG have one copy. *Mycobacterium* sp. L47, *Mycobacterium* sp. DCY42 and *Mycobacterium* sp. JS623 have been commonly detected in the environment and the species *M. brasiliensis* and *M. moriokaense* are known rare opportunistic pathogens. As two sequences matching *M. moriokaense* were sequenced from two different positions, it might be suggested that the species has multiple migrating positions. As one match was 98.9% similar and the other 97.36% similar to *M. moriokaense,* it would suggest that they are both closely related environmental strains. Sequences related to the *M. avium* complex member *M. colombiense* were detected in three out of five soils, noticeably not in the Cryfield soil (Table 3.3). *M. riyadhense* was only detected in sample 1110 and *M. haemophilum* was only detected in Cryfield (Table 3.3).

Table 3.2. Sequence comparisons of bands excised from the *Mycobacterium* genus PCR-DGGE (Figure 3.6A).

DGGE band	Site	Identity (%)	E value	Match accession number	Match name
A	1108	98.08	0	DQ249999	*Mycobacterium sp. L47*
B	1108	99.73	0	EU165538	*Mycobacterium brasiliensis*
C	1108	99.44	0	X93029	*Mycobacterium sp.*
D	1109	100	<0.001	EF019937	*uncultured Mycobacteriaceae bacterium*
E	1109	97.36	0	AY859686	*Mycobacterium moriokaense*
F	1110	98.91	0	AY859686	*Mycobacterium moriokaense*
G	1110	98.52	0	FJ605266	*Mycobacterium sp. DCY42*
H	1111	98.31	0	AY162028	*Mycobacterium sp. JS623*
I	Cryfield	97.53	<0.001	FJ538898	*Mycobacterium sp. GR-2001-270*

Table 3.3. Sequence comparisons of bands excised from the SG PCR-DGGE (Figure 3.6B).

DGGE band	Site	Identity (%)	E value	Match accession number	Match name
A	1108	99.42	0	GQ153275	*Mycobacterium colombiense*
B	1108	97.28	<0.001	GQ153275	*Mycobacterium colombiense*
C	1109	99.71	0	FJ794352	*Mycobacterium sp. NLA000202017*
D	1109	99.71	0	GQ153275	*Mycobacterium colombiense*
E	1110	99.71	0	GQ153275	*Mycobacterium colombiense*
F	1110	99.13	0	EU274642	*Mycobacterium riyadhense*
G	1111	99.09	0	EU274642	*Mycobacterium riyadhense*
H	Cryfield	99.71	0	GQ153275	*Mycobacterium colombiense*
I	Cryfield	98.85	0	U06638	*Mycobacterium haemophilum*

3.7. Pyrosequencing of pilot soils

Pyrosequencing was performed on metagenomic DNA of five pilot soils using the *Mycobacterium* genus and SG primer sets. The pilot pyrosequencing data was analysed using the in-house pipeline. The *Mycobacterium* genus dataset produced a total of 60,278 sequences. Sequences were retained if the amplicon length was between 468-472bp; this was based on the expected amplicon length and plotted frequencies of sequence counts (Figure 3.8). Overall 36.5% of sequences were retained after quality control procedures (Table 3.4). The SG dataset produced a total of 18,713 sequences. Sequences were retained if the amplicon length was between 420-424bp; this was based on the expected amplicon length and plotted frequencies of sequence counts (Figure 3.9). In total 53.3% of sequences remained after quality control procedures (Table 3.5), suggesting that this pyrosequencing run was more efficient than pyrosequencing using the *Mycobacterium* genus primer set. The rarefaction curves for the number of Operational Taxonomic Units (OTU) at a 97% cut-off for both primer sets indicate that sampling effort was sufficient as shown by the sample curves which plateau at <1000 sequences (Figure 3.10).

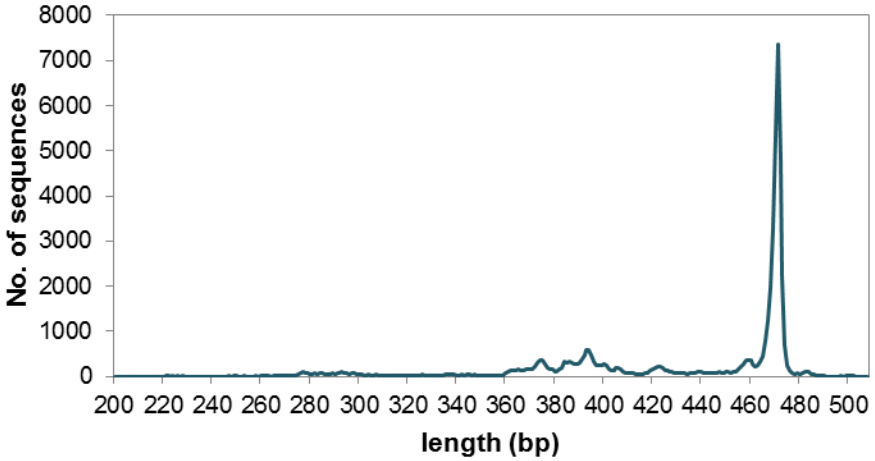


Figure 3.8. Histogram to show the length of the raw sequences for the *Mycobacterium* genus pilot soil dataset.

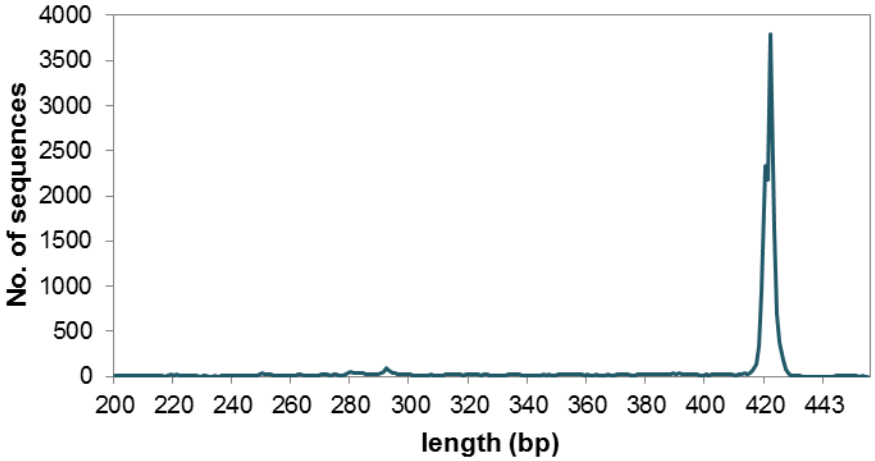


Figure 3.9. Histogram to show the length of the raw sequences for the SG pilot soil dataset.

Table 3.4. Sequence counts at steps of quality control pipeline for *Mycobacterium* genus dataset

Site	1108	1109	1110	1111	Cryfield	Total
Raw sequences	17647	14839	6634	5161	15997	60278
F primer found	14718	12975	5909	4623	13853	52078
After trimming/remove N bases	6873	5984	1067	864	7799	22587
No. of Chimeras	126	386	18	17	46	593
Final seq count	6747	5598	1049	847	7753	21994
Percentage of seqs retained	38	38	16	16	48	36

Table 3.5. Sequence counts at steps of quality control pipeline for SG dataset

Site	1108	1109	1110	1111	Cryfield	Total
Raw sequences	2066	5399	3259	5037	2952	18713
F primer found	1814	4948	2904	4574	2746	16986
After trimming/remove N bases	1441	4000	1206	1201	2191	10039
No. of Chimeras	18	8	17	11	2	56
Final seq count	1423	3992	1189	1190	2189	9983
Percentage of seqs retained	69	74	38	24	74	53

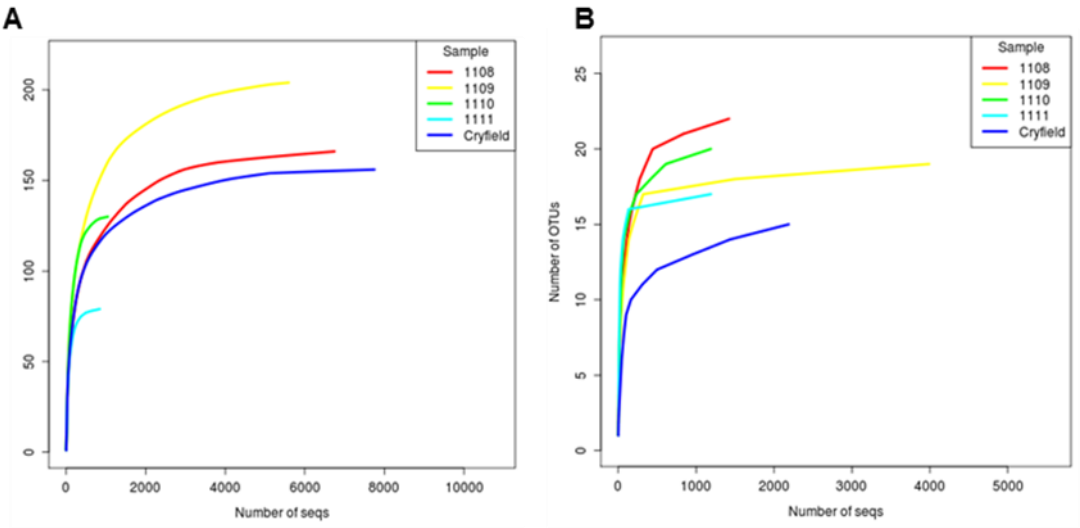


Figure 3.10. Rarefaction curves for the quality controlled sequences of the five soils clustered into Operational Taxonomic Units (OTUs) at a 97% cut-off **(A)** using the *Mycobacterium* genus primers **(B)** using the SG primers.

3.7.1. Alpha diversity of mycobacteria in pilot soils

As expected and demonstrated by all three measures the *Mycobacterium* genus dataset was more diverse and species-rich than the SG dataset. Within the *Mycobacterium* genus dataset, soil 1109 was the most diverse and 1111 was the least diverse as shown by OTU richness and the relative abundance of OTUs (H' and Simpson indices) (Table 3.6). A random resample of the data (847 sequences per sample) showed similar trends where sample 1111 had the lowest number of OTUs and Shannon diversity estimate, however 1110 appeared to be the most diverse sample closely followed by soil 1109 (Table 3.6). Differences between the Shannon

and Simpson diversity estimates were not statistically significant between all sequences and the random resample (T-Test $P>0.05$). Soil sample 1111 was also the least diverse for SG and soil 1108 the most diverse as demonstrated by the Shannon index (Table 3.7). The random resample (1189 sequences per sample) also showed 1108 had the highest Shannon diversity estimate, however in contrast 1109 was found to have the lowest diversity estimate (Table 3.7). Differences in the SG diversity estimates between all sequences and the random resample were also not significant (T-Test $P>0.05$). Differences in diversity from all sequences and the random resample were observed, this is because the relative abundance of sequences is an estimate from the PCR and pyrosequencing reaction.

Table 3.6. Diversity richness estimates for the *Mycobacterium* genus datasets

	All sequences			Random resample		
Sample	Shannon	Simpson	No. of OTUs	Shannon	Simpson	No. of OTUs
1108	3.84	0.961	166	3.30	0.936	68
1109	4.19	0.973	204	3.51	0.948	84
1110	3.73	0.942	130	3.62	0.957	89
1111	3.02	0.893	79	2.86	0.896	59
Cryfield	3.52	0.930	156	2.95	0.891	63

Table 3.7. Diversity richness estimates for the SG datasets

	All sequences			Random resample		
Sample	Shannon	Simpson	No. of OTUs	Shannon	Simpson	No. of OTUs
1108	1.83	0.742	22	1.59	0.707	11
1109	1.04	0.522	19	0.31	0.104	11
1110	1.66	0.706	20	1.29	0.646	12
1111	0.58	0.209	17	0.59	0.222	11
Cryfield	1.76	0.777	15	1.27	0.614	10

3.7.2. Beta-Diversity of mycobacteria in pilot soils

A total of 273 OTUs were present in soils analysed with the *Mycobacterium* genus primer set; 13.2% (36 OTUs) were in all five samples whereas 31.9% (87 OTUs) were found in only one sample. Similarities in OTU community composition were observed on a Canonical Correspondence Analysis (CCA) plot for samples 1111 and 1109, and 1108 appeared to be an outlier (Figure 3.11A). In agreement with the CCA, the community composition as depicted by the OTU relative abundance graphs illustrated the similarity of 1109 and 1111 and the similarity of 1110 and Cryfield (Figure 3.12A). However, the Jaccard class dendrograms which only take into account the presence/absence of shared OTUs suggested that Cryfield was closely related to 1108 and 1109 (110 shared OTUs) (Figure 3.13A).

Soils analysed with the SG primers resulted in a total of 33 different OTUs, of which 18.2% of OTUs (6 OTUs) were shared and 30.3% (10 OTUs) were unique to one sample. The CCA for the SG dataset showed 1108 and 1109 have similar community compositions and separately 1110 and Cryfield clustered together (Figure 3.11B). The OTU relative abundance graph also depicted the similarity between 1108 and 1109 and similarity between 1110 and Cryfield (Figure 3.12B). The JClass dendrograms are in agreement in that 1108 and 1109 share similar OTU community structure (14 shared OTUs) (Figure 3.13B).

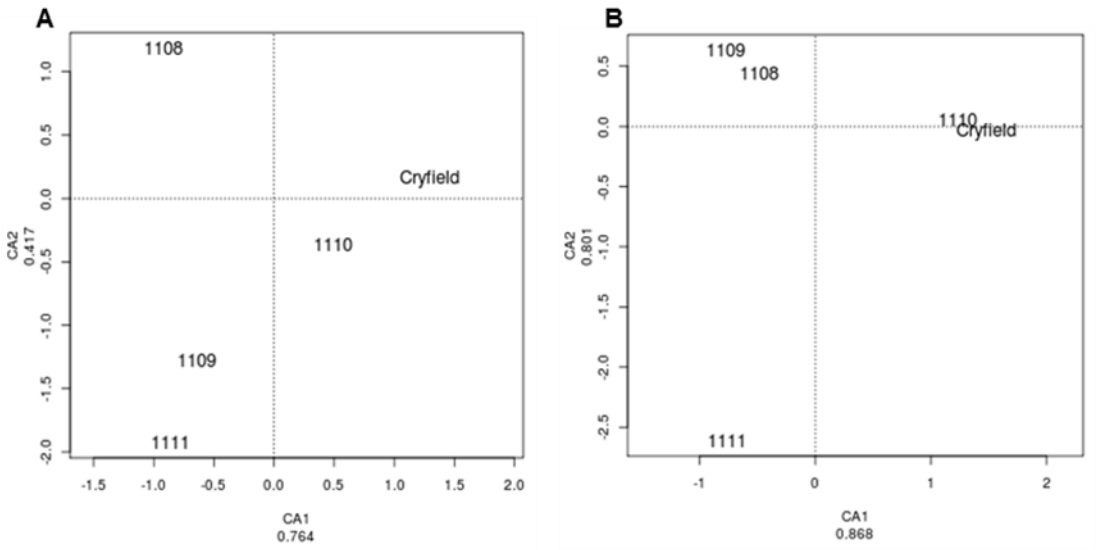


Figure 3.11. Canonical Correspondence Analysis (CCA) for the five soil samples based on shared and unique OTUs **(A)** using the *Mycobacterium* genus primers **(B)** using the SG primers

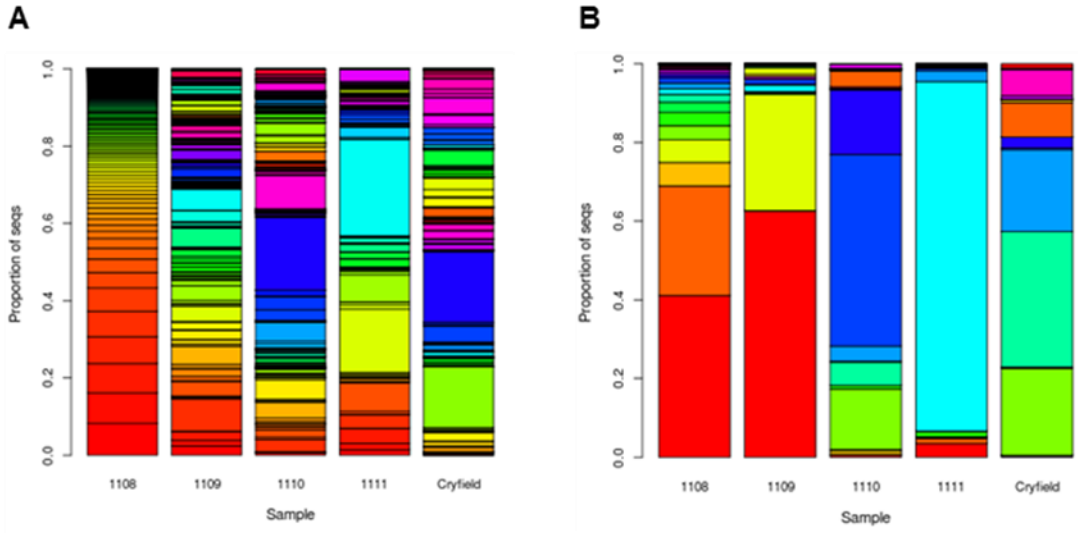


Figure 3.12. OTU distribution graphs show the prevalence of OTUs in each sample. **(A)** *Mycobacterium* genus primer set **(B)** SG primer set

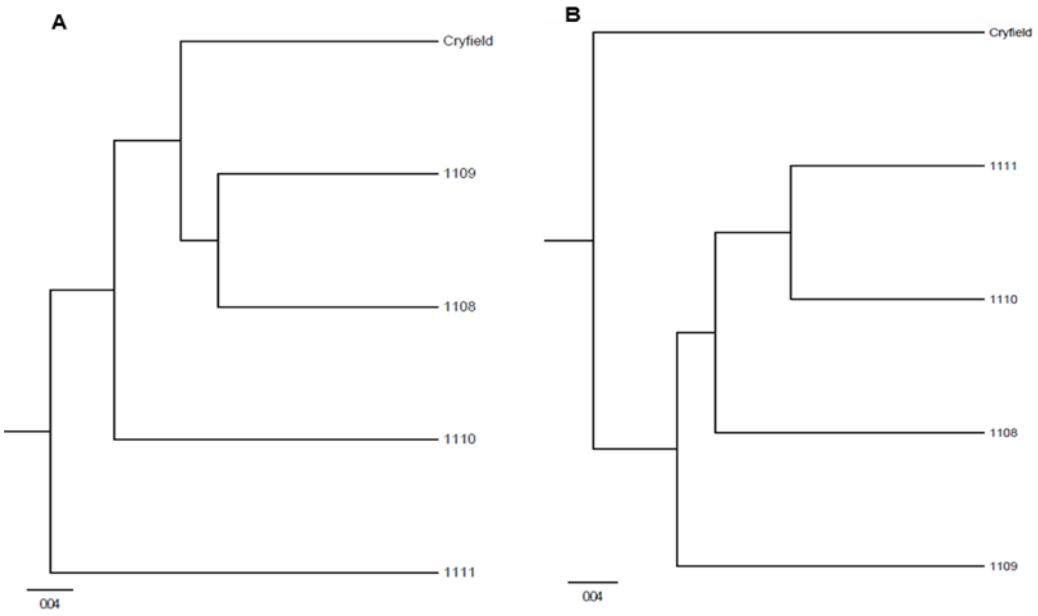


Figure 3.13. Jclass dendrogram similarity trees show similarities between different sites. **(A)** *Mycobacterium* genus primer set **(B)** SG primer set

3.7.3. Diversity of *Mycobacterium* species in pilot soil samples

Local BLAST was performed on all sequences from the *Mycobacterium* genus and SG pilot soil datasets against a reference downloaded SILVA database (section 2.8.4). A total of 151 different *Mycobacterium* species were matched at an e-value of <0.001 from the *Mycobacterium* genus dataset; these included matches to cultured species, environmental and clinical strains. A total of 52.5% sequences were matched to environmental or clinical strains, suggesting a wide diversity of uncharacterised *Mycobacterium* species. A total of 27.8% sequences were matched to 35 different polycyclic aromatic hydrocarbon (PAH) degrading *Mycobacterium* species and only 2.0% of all sequences were matched to 14 different characterised cultured SG. The SG represent a small proportion of sequences, perhaps due to the group being less abundant in the environment or due to primer bias. Overall the most prevalent BLAST matches to known cultured *Mycobacterium* species included the FG *M. neglectum* (12.2%), *M. moriokaense* (7.4%), *M. insubricum* (2.9%), *M. chlorophenicum* (2.4%), *M. holsaticum* (2.3%) and *M. pyrenivorans* (1.8%). Marked

differences were observed between soil samples, for example *M. neglectum* and *M.* sp. *Ellin 148* appeared to be more prevalent in all four Ethiopian soils 1108-1111 compared to Cryfield (Figure 3.12). In contrast the BLAST matches to *M.* sp. *JS624, M.* sp. *JS621, M. holsaticum, M.* sp. *KAR6* and *M.* sp. *CNJ859 PL04* were relatively more abundant in Cryfield compared to the Ethiopian soils (Figure 3.14). Both 1110 and Cryfield displayed a higher prevalence of *M. moriokaense* and *M. insubricum*. However, many species were equally prevalent in all samples; *M.* sp. A14N, *M.* sp. GR-2009-164, *M. gadium*, *M.* sp. T103 and *M. tuscaie* (Figure 3.14).

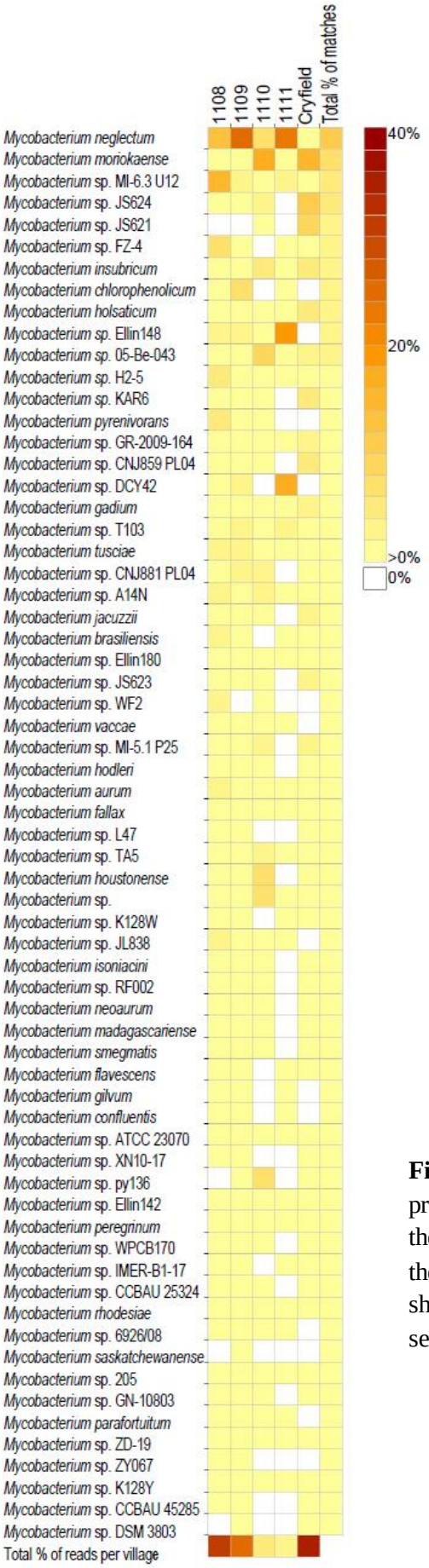


Figure 3.14. Heatmap of the prevalence (%) of BLAST matches in the *Mycobacterium* genus dataset of the five soils. The BLAST matches shown represent the 95% of all sequences.

The local BLAST search for sequences from the pilot soils with the SG primer set generated 27 different BLAST matches, of which 70.4% (19/27) were known opportunistic pathogens and 29.6% (8/27) of species were also identified using the *Mycobacterium* genus primer set. However it was unclear as to why 0.05% of sequences matched the FG *M. salmoniphilum*, perhaps due to non-specific binding. Differences were observed between soil samples, for example *M. riyadhense* was more prevalent in 1110, *M. aemonae* more prevalent in 1111 and *M. gordonae* was more prevalent in soil 1108 and Cryfield (Figure 3.15). BLAST matches to an *M. avium* complex member *M. colombiense* were most prevalent in Ethiopian soils 1109 and 1108, representing 94.9% and 32.8% of all sequences respectively. Several species were present in all samples such as *M. asiaticum*, *M. colombiense*, *M. gordonae* and *M. riyadhense* (Figure 3.15).

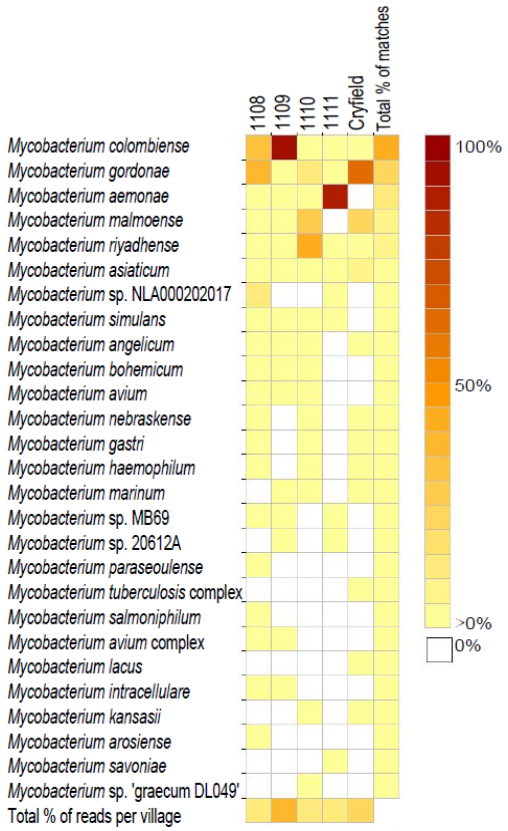


Figure 3.15. Heatmaps of the prevalence (%) of BLAST matches in the SG dataset of the five soils.

3.7.4. Phylogenetic Diversity

Phylogenetic diversity was assessed by the alignment of OTU representative sequences and DGGE bands A-I with known cultured *Mycobacterium* species downloaded from SILVA followed by the construction of a maximum likelihood phylogenetic trees (section 2.9.5). Relatively abundant OTUs (>1%) from the *Mycobacterium* genus dataset and *Mycobacterium* genus DGGE bands A-I clustered with numerous FG (Figure 3.16). Only one clade consisting of 2 OTUs and 2 DGGE bands did not cluster with any known *Mycobacterium* species, perhaps representing a novel environmental species. Overall each OTU represented less than 7.7% of all sequences suggesting a large diversity of *Mycobacterium* species. In agreement with the BLAST results, abundant OTUs were matched to prevalent BLAST matches. For example three OTUs clustered close to *M. neglectum*, representing 8.5% of all sequences, and two OTUs clustered with *M. pyrenivorans*, representing 3.6% of sequences. Consistency between BLAST matches of the *Mycobacterium* genus DGGE bands (Table 3.2) and positions on the phylogenetic tree were observed, for example B_1108 clustered next to *M. brasiliensis* and an OTU representing 1.4% of the sequences and was also BLAST matched to *M. brasiliensis* (Figure 3.16). There were some discrepancies, for example one OTU that clustered with *M. aichense* was not detected in the pyrosequencing BLAST results and 5 OTUs (8.7%) clustered with *M. fortuitum* and *M. septicum* but only represented 0.08% of the BLAST results.

The wide diversity of SG was observed in the maximum likelihood phylogenetic tree of all OTUs from the SG dataset and known *Mycobacterium* species (Figure 3.17). In concordance with the pyrosequencing BLAST results, 9 OTUs clustered with the *M. avium* complex, which represented 45.5% of all sequences and 12 OTUs clustered with *M. gordonae* representing 24.8% of all sequences. In support, SG DGGE bands

that were BLAST matched to *M. colombiense* also clustered with the *Mycobacterium avium* complex. DGGE bands F_1110 and G_1110 were BLAST matched and clustered with *M. riyadhense*. However there was some discrepancy observed where an OTU grouped with *M. shinjukuense* but was not present in the BLAST results. One OTU grouped with FG, but only represented 0.06% of all sequences. Overall, strong agreement was observed between BLAST and maximum likelihood methods and between DGGE and pyrosequencing,

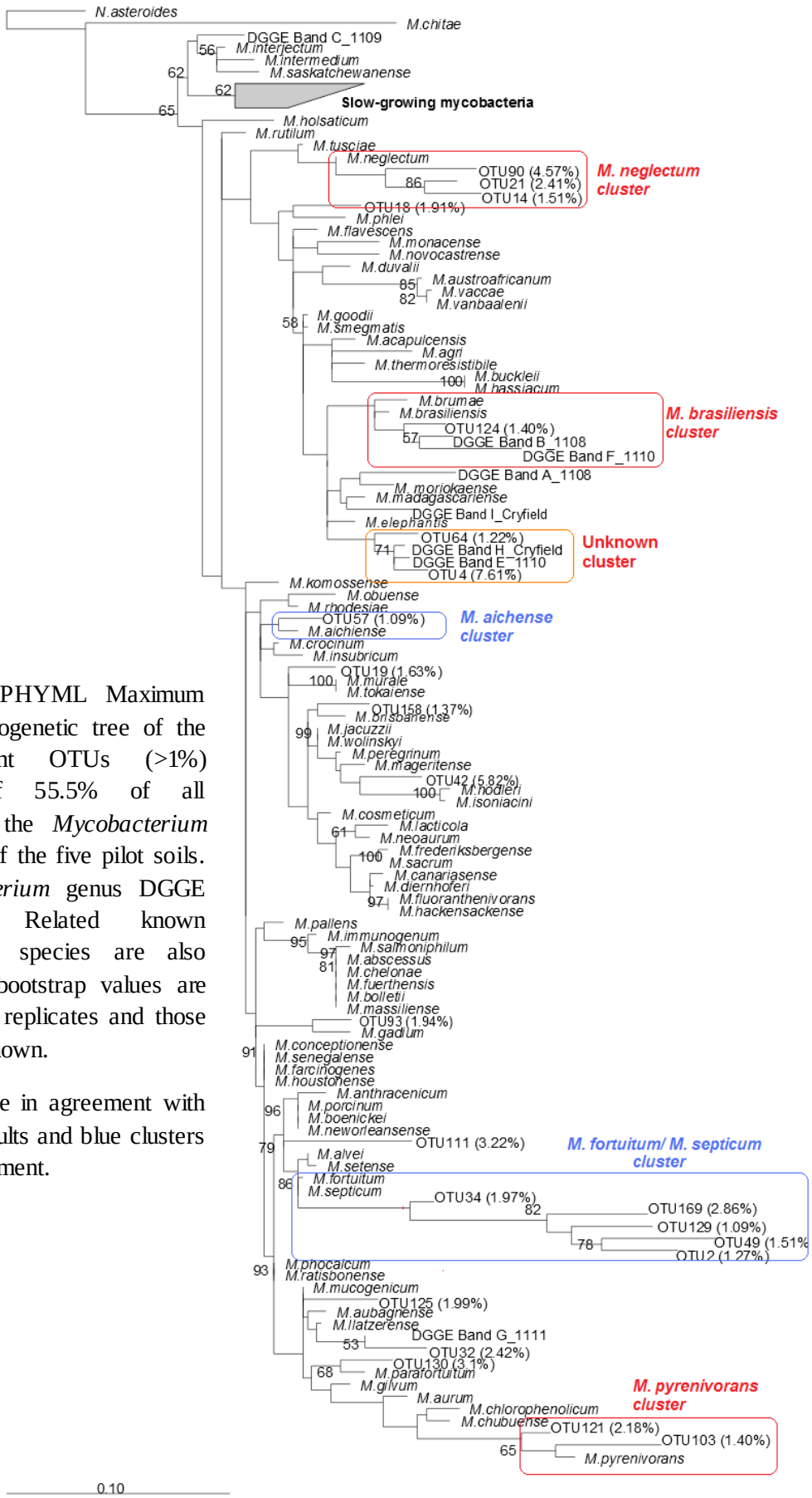


Figure 3.16. PHYML Maximum likelihood phylogenetic tree of the most prevalent OTUs (>1%) comprising of 55.5% of all sequences of the *Mycobacterium* genus dataset of the five pilot soils. and *Mycobacterium* genus DGGE bands A-I. Related known *Mycobacterium* species are also shown. The bootstrap values are based on 1000 replicates and those over 50% are shown.

Red clusters are in agreement with the BLAST results and blue clusters are not in agreement.

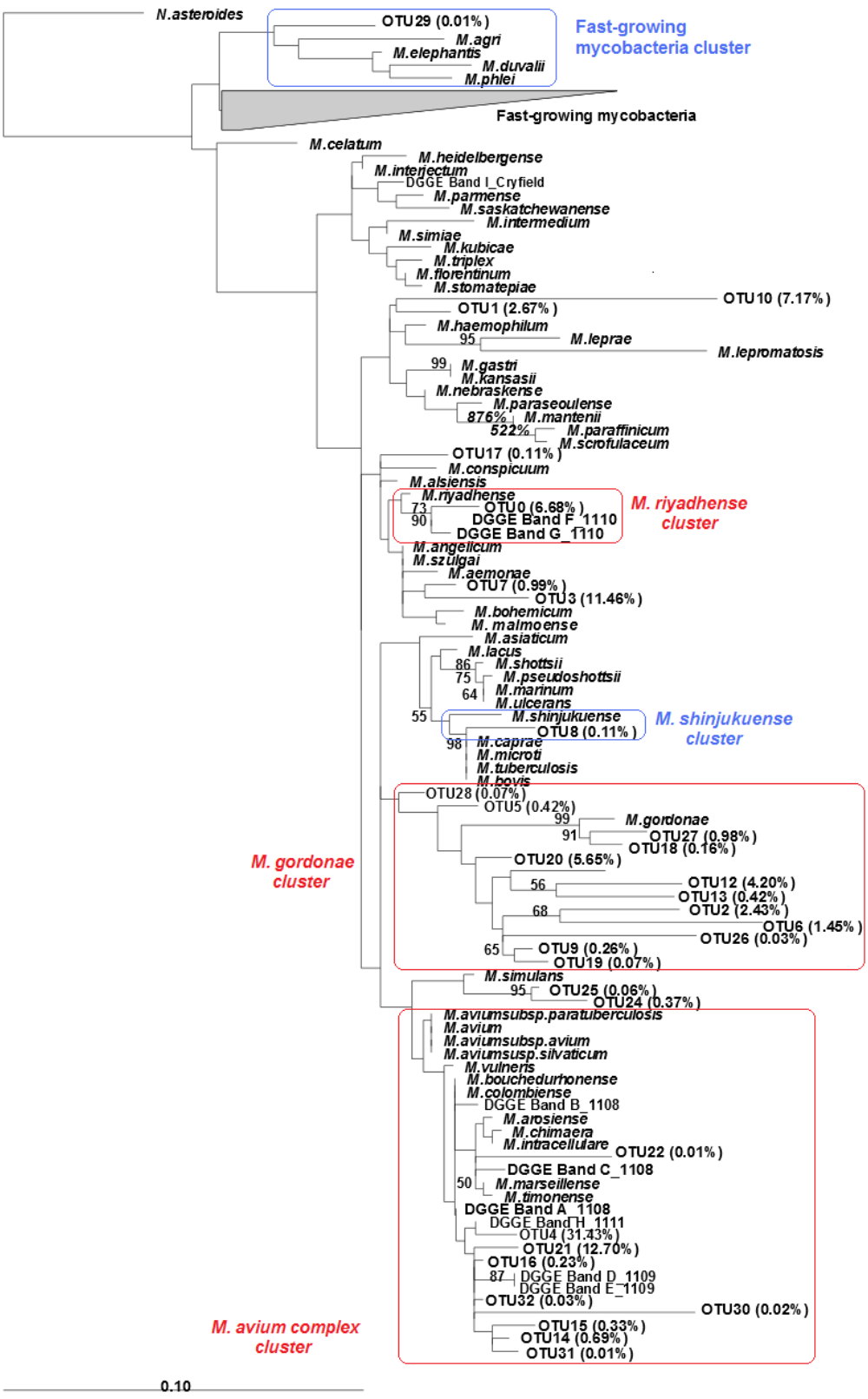


Figure 3.17. PHYML Maximum likelihood phylogenetic tree of the all 33 OTU representing 100% of the SG dataset and the SG DGGE bands A-I. Related known *Mycobacterium* species are also shown. The bootstrap values are based on 1000 replicates and those over 50% are shown. Red clusters are in agreement with the BLAST results and blue clusters are not in agreement.

3.8. Discussion

In order to assess *Mycobacterium* species diversity in environmental samples it was necessary to evaluate techniques for the extraction of community DNA and to assess the efficiency of specific target primers in combination with a molecular detection method. DGGE was optimised for the detection of the *Mycobacterium* genus and the long helix 18 containing SG. The DGGE method has been previously employed to determine the diversity of mycobacteria (Leys et al. 2005; Niva et al. 2006; Uyttebroek et al. 2006), however this is the first time it has been used specifically to examine SG diversity.

In this study the specificity of both primer sets was evaluated using extracted DNA from cultured *Mycobacterium* species. However it was not possible to differentiate between a number of *Mycobacterium* species as some shared identical band positions. Consequently further steps, such as sequencing of the bands extracted from the DGGE, were required as the method could not solely rely on a library of reference bands corresponding to *Mycobacterium* species. DGGE was shown to have a detection limit of 10^5 cells per gram and 10^2 cells per gram for the *Mycobacterium* genus and SG primer sets using a nested approach. This was comparable to a *Mycobacterium* 16S rRNA gene specific DGGE which had a detection limit of 10^6 CFU per gram of soil and was improved to 10^2 CFU per gram using a nested approach (Leys et al. 2005). The SG primer set was likely to have had a better sensitivity compared to the *Mycobacterium* genus primer set as it targets a smaller group with less diversity, therefore *M. bovis* was not likely to be outcompeted by other species in the PCR reaction.

A comparison of three water DNA extraction methods clearly demonstrated that the commercial PowerWater kit was the most sensitive with a limit of detection of at

least 10^2 cells per ml and was not affected by the storage of filters for 14 days. The adapted Griffiths method had a detection limit of 10^4 cells per ml and the adapted Pickup method was the least sensitive with a limit of 10^5 cells per ml. It is possible that the Pickup method did not perform as well due to the inefficiency of the centrifugation procedure for cell lysis and the lack of a mechanical bead-beating step. The detection limit for water extraction using the PowerWater kit is comparable to, if not better than, the detection limit for soil extraction using the FastDNA spin kit. This was expected as water is less complex than soil, containing less inhibitory compounds and less microbial diversity (Torsvik 2002). As a result of the findings from this study the PowerWater kit was employed for DNA extractions from subsequent water samples.

The pilot study has shown a high diversity of *Mycobacterium* species in soil samples. Both methods rely on PCR of metagenomic DNA which has inherent limitations such as the bias of primer sets towards the amplification of certain species, especially where there are mismatches in the primer sequence and where species are more abundant than others (Hong et al. 2009; Kanagawa 2003; von Wintzingerode et al. 1997). Therefore a second primer set targeting SG was employed to target *Mycobacterium* diversity, as although the *Mycobacterium* genus primer set was shown to amplify DNA from cultures of SG this group was not commonly detected in environmental community DNA.

The DGGE and pyrosequencing methods were both used to determine the alpha diversity of the mycobacteria present in the pilot soil samples. There was some agreement between Shannon diversity estimates for both methods, where sample 1111 consistently had lower diversity estimates for both datasets and sample 1108 had the highest SG diversity estimates. However there were some discrepancies, for

example Cryfield had the highest Shannon diversity estimate for the *Mycobacterium* genus PCR-DGGEs but had one of the lowest for pyrosequencing. One would expect differences in the alpha diversity estimates between DGGE and pyrosequencing due to the differences in sampling scale, where DGGE is likely to underestimate the alpha diversity. Similarities were observed between beta-diversity results for DGGE and pyrosequencing. For the *Mycobacterium* genus the pyrosequencing CCA and the UPMGA DGGE cluster analysis were in complete agreement, with similarities found between 1110 and Cryfield and 1109 and 1111. However, discrepancies were observed for the SG CCA; only 1110 and 1111 were consistently similar for both methods, as estimated by UPMGA DGGE cluster analysis and Jaccard pyrosequencing cluster analysis.

Close agreement was ascertained from the BLAST results of the DGGE bands and the pyrosequencing reads. All sequences from the DGGE bands were also detected in the pyrosequencing, apart from the DGGE sequence matched to *uncultured Mycobacteriaceae bacterium*. Agreement between the DGGE approach and pyrosequencing has been observed previously, the study found consistency between several molecular methods however overall pyrosequencing provided a several orders of magnitude more data than the other methods (Oakley et al. 2010). Overall it is generally accepted that DGGE mostly detects the relatively abundant species (Chan et al. 2002) and this was the case in the pilot study. For example *M. moriokaense* was matched to a DGGE band from sample 1110 and this species was the most prevalent pyrosequencing BLAST match for sample 1110. The environmental species *Mycobacterium spp. DCY42* was matched to a DGGE band from sample 1111 and represented 17.71% of sample 1111 pyrosequencing BLAST matches. The SG *M. colombiense* was matched to DGGE sequences from samples

1108, 1109 and 1111 and represented 32.8%, 94.9% and 3.7% of pyrosequencing BLAST matches respectively. Therefore *M. colombiense* was found to be abundant using both methods, particularly in samples 1108 and 1109. *M. riyadhense* was matched to a DGGE sequence from sample 1110 and was also the most abundant pyrosequencing BLAST match for this sample (46.9%). However, although *M. haemophilum* was detected in Cryfield using both pyrosequencing and DGGE, it only represented 0.1% of pyrosequencing BLAST matches. As not all DGGE bands were sequenced it is possible that other more abundant pyrosequencing BLAST matches to *Mycobacterium* species were present. Overall there was a good level of agreement between the two molecular methods; the relatively most abundant pyrosequencing BLAST matches also were detected using DGGE.

In summary the combination of molecular methods and the agreement between them particularly for the BLAST matches has provided a clearer and more confident picture of the *Mycobacterium* species diversity present within the pilot soils. The PCR-DGGE provided a snapshot of the relatively abundant *Mycobacterium* species. However, the limitation of DGGE is that less abundant but potentially important species are unlikely to be detected. In contrast the pyrosequencing has provided a far more comprehensive insight into the diversity of *Mycobacterium* species and as such the pilot study has illustrated this method would be most appropriate for a large-scale environmental survey.

Chapter 4: *Mycobacterium* species diversity in the Ethiopian environment

4.1 Abstract

The species within the genus *Mycobacterium* are commonly detected in a variety of environments, many are important in soil degradative processes however some are opportunistic pathogens. Soil and water source samples were collected from a total of 42 villages and amplicon pyrosequencing was employed to target the 16S rRNA gene of mycobacteria and separately SG. After quality control procedures, a comprehensive characterization of the relative abundance and distribution of mycobacteria was achieved. The study revealed a greater diversity of both FG (>200 species) and SG (>50 species) than previously reported. A marked heterogeneity in community structure was observed between different environmental sources, where soil samples had a more even community structure compared to water samples. A high relative abundance of sequences related to polycyclic aromatic hydrocarbon degraders were observed in both soil and water (>20%). Sequences related to SG opportunistic pathogens were commonly detected in the environment, a particularly high relative abundance of the *M. avium* complex was observed in soil samples (28%). Water samples were dominated by sequences related to *M. gordonae* (45%). The number of sequences related to the *M. tuberculosis* complex was shown to be relatively abundant in water samples (3.1%). Results indicate ample opportunity for human exposure to opportunistic pathogenic mycobacteria in the environment.

4.2. Introduction

The assessment of microbial communities using high throughput technologies such as pyrosequencing requires quality control of raw sequences before subsequent assessment of diversity. Sequencing errors can occur from incomplete extension,

single base insertions and single base deletions (Margulies et al. 2005). These errors can arise due to the position on a PicoTitre plate, the sequence length and the presence of homopolymers (Gilles et al. 2011). The accuracy of pyrosequencing the 16S rRNA gene variable region V6 was estimated to be 99.5% and the performance of quality control procedures improved this to 99.75% (Huse et al. 2007). Steps which were shown to improve the quality of the data included the removal of sequences with (1) ambiguous base calls (2) read lengths outside the expected amplicon length (3) reads that did not match the proximal primer and (4) low quality scores (average <25).

The primers employed for pyrosequencing and the length of the amplicon were found to influence species richness and evenness estimates; however the greatest variation in these estimates was associated with different variable regions within the gene chosen for pyrosequencing (Engelbrektson et al. 2010). This was consistent with the findings of an additional experiment which found that the variable regions V6, V1 and V2 were more likely to inflate diversity compared to the variable regions V3, V7 and V8 (Youssef et al. 2009). A comparison of several pyrosequencing datasets found that different bacterial community compositions were dependent on which variable regions were targeted (V1-V3 and V7-V9) (Kumar et al. 2011). These reports suggest primer sets targeting different variable regions should be employed for the accurate determination of diversity and composition.

Specific issues, such as the formation of chimeric sequences during PCR amplification as a result of incomplete extension and mispriming, can lead to inflated diversity estimates and the identification of false novel species. The program ChimeraSlayer identifies chimeras by comparing sequences to a reference set of chimeric-free sequences in order to identify potential parents (Haas et al. 2011). The

more accurate and sensitive UCHIME program identifies chimeras using both a reference database and a *de novo* abundance based method (Edgar et al. 2011b). The employment of a chimeric detection program to identify and remove chimeras was shown to improve the quality of the data (Schloss et al. 2011). Early pyrosequencing microbial diversity studies suggested a "rare biosphere" existed due to the presence of a high number of singleton OTUs (Sogin et al. 2006). However, it is now generally accepted that singleton OTUs can occur due to PCR single base substitutions and sequencing error. The accuracy of pyrosequencing results from technical replicates was improved by the removal of singleton sequences (Behnke et al. 2011; Zhou et al. 2011), and clustering 16S rRNA amplicons at ≥97% cut-off absorbs this type of error (Kunin et al. 2010). Denoising programs such as AmpliconNoise and USEARCH (OTUPipe) are used to identify pyrosequencing errors by flowgram clustering and/or alignment of sequences (Edgar 2011a; Quince et al. 2011). The quality control procedures discussed are necessary to ensure that more accurate estimates of diversity are obtained.

In order to assess microbial diversity it is essential to define the unit of analysis. Often the number of species is employed as a measure of the diversity of microbial communities. However an issue of contention is the definition of a bacterial species; although there is no consensus as to what defines a bacterial species, a polyphasic approach is often employed using phenotypic characteristics, genetic information and ecological niches for classification. Among the genomic definitions of a bacterial species are that strains should share at least 70% of their genomes as assessed by DNA reassociation (Wayne et al. 1987) and have at least 97% similarity in the 16S rRNA gene (Stackebrandt & Goebel 1994). Whereas the phylo-phenetic species concept defines a species as a monophyletic group of organisms that share common

characteristics and can be discriminated based on phenotypic properties (Rossello-Mora & Amann 2001). The more recent genomic-phylogenetic species concept based on the differences in DNA, RNA and protein has been put forward owing to the advent of newer genomic techniques (Staley 2006). Finally a shared ecological niche can also be used to define a species, such as a group that has evolved together with similar ecological resources (Cohan 2002). *Mycobacterium* species classification and taxonomy is discussed in section 1.1. The limitations of defining species for the analysis of bacterial diversity can often be avoided by using Operational Taxonomic Units (OTU); this method has the advantage of a specified cut-off for sequence similarity which enables the separation of sequences into species-like groups.

Microbial diversity can be defined as the richness and evenness of a given community with richness defined as the number of species/OTUs and evenness as the relative abundance of the species/OTUs. The methods for assessing the diversity of microbial communities include parametric, nonparametric and phylogenetic approaches (Bohannan & Hughes 2003). Parametric methods include those that assume the distribution of the data and fit the data to a particular model. For example the biological niche based models such as dominance decay, random assortment or dominance pre-emption and the statistical models such as log series or log normal (Curtis et al. 2002; Dumbrell et al. 2010; Magurran 2004). However this approach relies on a theoretical species distribution. Non-parametric methods do not assume the distribution of the data, and include the Shannon and Simpson estimates. The calculations often take into account the relative abundance of species/OTUs. This approach is limited as it can often underestimate diversity. The phylogenetic method directly assesses the evolutionary relationships between species/OTUs and metrics such as the Phylogenetic Diversity (PD) and the UniFrac metric. This is achieved by

assigning a discrete number as an indicator of the shared phylogenetic branch length between different samples (Faith 1992; Lozupone et al. 2011). The employment of BLAST to assign the closest taxonomic relative and the construction of phylogenetic trees of OTUs with known *Mycobacterium* species allows for inferences to be made according to what is already known about these species. A combination of all of these approaches should facilitate a valid estimation and insight into the diversity of *Mycobacterium* species in the environment.

Mycobacterium species diversity has previously been assessed in a number of different environments including alpine, boreal forests, arable plots, tropical and polycyclic aromatic hydrocarbon (PAH)-contaminated environments (Chilima et al. 2006; Kamala et al. 1994; Leys et al. 2005; Mendum et al. 2000; Niva et al. 2006; Thorel et al. 2004). However, the majority of these studies have employed cultivation techniques known to skew the profile of isolates and provide a poor estimation of the true diversity. The sheer number of sequences produced using the pyrosequencing approach should provide a more accurate insight into the *Mycobacterium* species diversity. There is particular interest in the SG due to their role as opportunistic pathogens, whereas the relevance of FG lies in their hypothesised role in BCG interference (Fine et al. 2001) and because the group are biodegraders of polycyclic aromatic hydrocarbons (PAHs) (Miller et al. 2004).

4.3. Aims

i) To quality control the pyrosequencing data of 42 soil samples and 42 water samples using two primer sets to target the (1) *Mycobacterium* genus and (2) SG.

ii) To determine the alpha diversity of *Mycobacterium* species in soil and water samples using OTU richness and evenness estimates.

iii) To assess the diversity of opportunistic pathogenic *Mycobacterium* species and Polycyclic Aromatic Hydrocarbon (PAH) degraders using phylogenetic and BLAST approaches.

4.4.1. Pyrosequencing Quality Control

The pyrosequencing raw data was separated into four datasets based on the primer set used and the environmental substrate, and each dataset initially contained 42 samples. The four datasets were (1) *Mycobacterium* genus soils (2) *Mycobacterium* genus water (3) SG soils (4) SG water. The *Mycobacterium* genus soils dataset had a total of 288,619 raw sequences, with quality control procedures removing sequences that had <400bp. This cut-off was chosen based on the frequencies of sequence length (Figure 4.1), the expected amplicon size of 465bp and inclusion of variable regions 2-4. There was a high abundance of sequences that were less <300bp, which could be due to carry forward incomplete extension (CAFIE) during the sequencing reaction or due to a homopolymer stretch which caused the signal to fall off. After the initial quality control procedures samples with <385 sequences were not included in the analysis (Table 4.1), leaving 23 samples with a total of 72,170 sequences (range 1,037-4,416 per sample). The *Mycobacterium* genus water dataset had a total

of 291,490 raw sequences. Using the same criteria as above, sequences were removed if they had <400bp (Figure 4.1). After all quality control procedures, including chimera checks and removal of singleton clusters, there were 142,547 sequences (48.9%) and all 42 samples were retained (range 385-15,533 per sample).

A total of 313,062 raw sequences were obtained from the soil dataset using the SG primers. Based on frequency histograms of sequence length, the expected amplicon length and targeting the long-helix insertion, sequences were removed if they had <420bp (Figure 4.2). After quality control procedures samples with <500 sequences were removed from the analysis, leaving a total of 32 soil samples with 45.6% (142,727) of the sequences remaining (range 514-13,129 per sample) (Table 4.2). The SG water dataset had a total of 366,296 raw sequences and using the same criteria as above for quality control, 42 samples remained with a total of 220,648 sequences (60.2%) (range 950-18,808 per sample). A summary of the quality control procedures and the total number of sequences is shown in Table 4.3.

Overall all sequences were analysed after quality control procedures, however the low number of sequences in selected samples highlighted in red (Table 4.1 & 4.1) made the analysis meaningless so these samples were excluded.

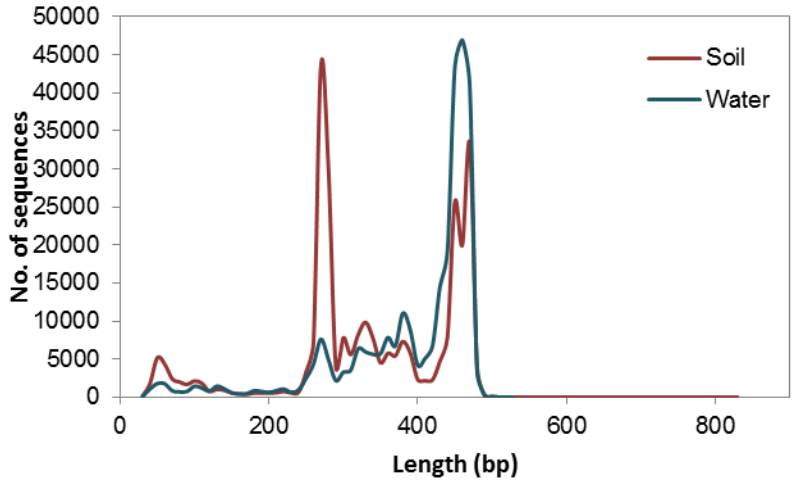


Figure 4.1. Histogram to show the length of the raw sequences for the *Mycobacterium* genus soil dataset and the water dataset. The minimum length of 400bp was chosen based on this histogram and the targeting sufficient length for differentiation and regions of variability.

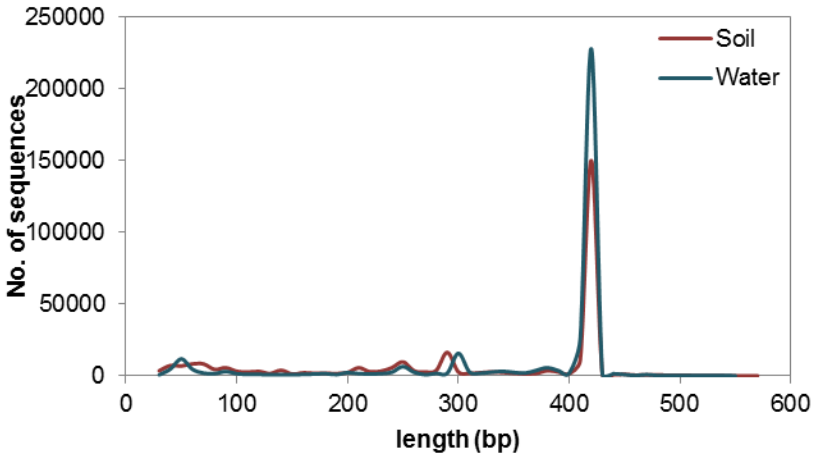


Figure 4.2. Histogram to show the length of the raw sequences for the SG soils dataset and water dataset. The minimum length of 420bp was chosen based on this histogram and the position of the long helix insertion needed for discrimination of SG

Table 4.1. *Mycobacterium* genus pyrosequencing data before and after quality control (QC). The samples highlighted in red contained less than 385 sequences after quality control and have been omitted from the analysis.

	Soil		Water	
Village	raw sequence count	sequence count after QC	raw sequence count	sequence count after QC
Ji1	6234	129	5919	3705
Ji2	9285	63	4630	2873
Ji3	3719	52	9090	5299
Ji4	5912	82	25366	15533
Ji5	4445	198	6550	3911
Ji6	2226	55	12610	7288
Om1	23921	5	5365	3035
Om2	3817	42	6314	3583
Tu1	3329	16	4497	2551
Tu2	28686	2935	12305	1589
Ho1	3559	102	3553	2184
Ho2	4728	111	7796	4587
Ho3	4555	14	4261	385
Ho4	11841	90	7988	4496
Bu1	12409	1715	9618	1628
Bu2	6960	59	13830	8093
Bu3	7404	86	5565	3368
Bu4	6418	66	5409	3364
Ga1	4881	31	5892	3293
Ga2	4877	101	7117	4549
Ga3	6570	59	4707	2345
Ga4	3020	1724	5931	3193
Ba1	6750	4000	6888	2923
Ba2	4653	2782	3900	2041
Ba3	7233	4393	7736	3411
Ba4	7836	4416	8729	4036
Ba5	5663	3528	6624	3006
Ba6	7058	4083	9830	4513
Ba7	3701	2350	6581	3119
Ba8	4865	2979	4369	2086
Wo1	4754	2681	4482	2668
Wo2	5266	2819	2521	1340
Wo3	4408	2612	1769	1017
Wo4	4094	2626	3560	1899
Wo5	4918	2899	3698	2064
Wo6	6005	3480	1887	1012
Wo7	5781	3624	12334	1850
Wo8	9338	1027	7056	3581
Go1	7068	3949	2801	1661
Go2	7144	4108	12454	3272
Go3	5572	3097	5904	3641
Go4	7716	4343	4054	2555
Total	**288619**	**72170**	**291490**	**142527**

Table 4.2. SG pyrosequencing data before and after quality control (QC). The samples highlighted in red contained less than 500 sequences after quality control and have been omitted from the analysis.

	Soil		Water	
Village	raw sequence count	sequence count after QC	raw sequence count	Sequence count after QC
Ji1	7168	3215	20996	4768
Ji2	1988	643	5348	3594
Ji3	21437	7929	7028	4868
Ji4	6830	3595	7704	5884
Ji5	21169	2392	22779	16231
Ji6	10578	594	9753	7023
Om1	12238	2806	4642	3402
Om2	1221	11	8213	5785
Tu1	192	45	6858	4095
Tu2	2492	1314	3313	1320
Ho1	24951	12466	2586	1192
Ho2	4209	2224	6716	4903
Ho3	2263	852	31510	18808
Ho4	2125	538	6267	4292
Bu1	1343	178	2260	1577
Bu2	993	237	11122	7514
Bu3	1113	203	1389	950
Bu4	1132	371	13204	9746
Ga1	136	21	5757	2921
Ga2	478	51	14488	9722
Ga3	3319	1813	4287	2707
Ga4	3625	2246	3547	2067
Ba1	6658	514	10963	7022
Ba2	2195	247	8866	5494
Ba3	13330	9959	5220	3133
Ba4	22998	5952	10394	5811
Ba5	8099	6509	13157	7479
Ba6	16712	13129	13401	8213
Ba7	14917	6500	5551	3279
Ba8	8485	4778	6318	3947
Wo1	3483	238	4649	2445
Wo2	5962	769	2958	1128
Wo3	14686	8426	9037	5404
Wo4	8690	4379	2487	1461
Wo5	7645	3512	16943	9843
Wo6	7561	5644	9323	6178
Wo7	7080	5637	5502	2551
Wo8	5763	4576	6516	3960
Go1	6572	4150	12201	5829
Go2	4231	2872	7971	4732
Go3	8091	6270	6697	3864
Go4	8904	6524	8285	5506
Total	**313062**	**142727**	**366206**	**220648**

Table 4.3. The number of reads and OTUs after quality control procedures using QIIME to QC the data.

	Mycobacterium genus		SG	
	Soil	Water	Soil	Water
No. of samples retained	23	42	32	42
Raw sequences in total (all samples)	288619	291490	313062	366206
Raw sequences	163938	291490	300776	366206
No. of raw clusters	2596	1940	359	383
No. of abundance chimera clusters	906	475	69	77
No. of reference based chimera clusters	166	140	68	57
No. of singleton clusters	1002	889	157	177
Sequences after QC	72170	142547	142727	220648
no. of OTUS	522	436	65	72

4.4.2. Inhibition RD4-GFP qPCR assays

The internal control inhibition real-time qPCR assay was employed to determine whether the samples contained inhibitory compounds, which would provide an explanation as to why some samples had less successful sequencing. Community DNA from each sample was seeded with a plasmid containing GFP sequence flanked by *M. bovis* RD4 scar primer sites (Pontiroli et al. 2011). The cycle threshold (Ct) was recorded and compared with the no inhibition control (NIC). Delta Ct (ΔCt) was calculated from the difference between the average sample Ct value and the average NIC Ct value. A ΔCt value of >1 was considered extensively inhibited, negligible to moderate inhibition was indicated by a ΔCt of 0-1 and no inhibition was indicated by a ΔCt of less than zero. From all 42 soil samples and 42 water samples, only four soil samples were considerably inhibited (Bu1, Bu2, Bu3 and Ho4) and one water sample (Ga2) (Figure 4.3). These soil samples did have less effective sequencing and samples Bu1-Bu3 were removed from both the *Mycobacterium* genus and SG soil datasets due to low sequence numbers. Inhibitory compounds within the soil samples might explain the less effective sequencing in these samples. However, water sample Ga2 had a good sequencing count for both pyrosequencing primer runs (≥4549

sequences) and numerous samples with similar or fewer sequences per sample did not show inhibition. There were 36 water samples and 17 soil samples which had a ΔCt between 0-1, but this did not appear to correlate with the number of sequences obtained. It follows that there must be additional reasons for less effective sequencing.

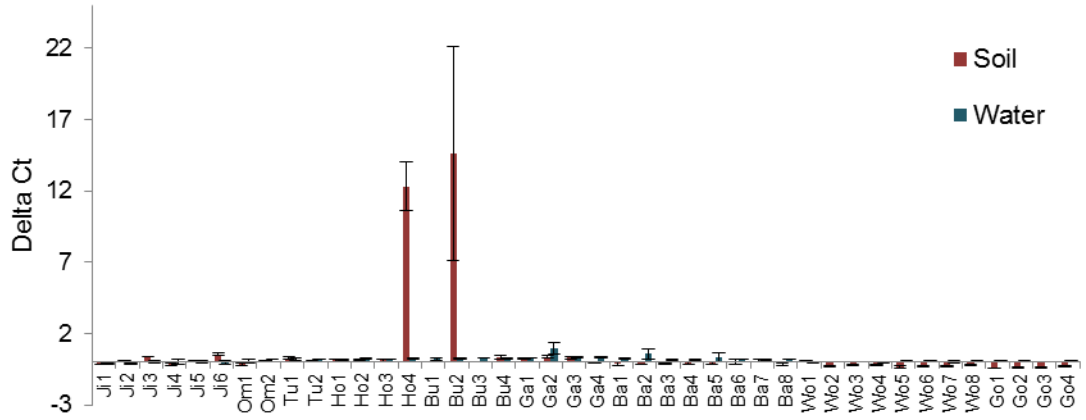


Figure 4.3. Real-time qPCR inhibition assay with plasmid RD4-GFPpCR®1.2 of 42 soil samples and 42 water samples. Error bars represent the standard deviation around the mean counts from three replicates per sample. Samples Bu1 and Bu3 failed to amplify.

4.4.3. Sufficient sampling effort

The rarefaction curves were constructed to assess sampling effort and richness. The *Mycobacterium* genus soil dataset had 23 samples that reached a plateau at less than 1000 sequences; these had between 143-241 different OTUs per sample (Figure 4.4A). The remaining 19 samples were removed from the analysis as the number of high quality sequences was comparably poor and the samples did not reach a plateau (range 5-198 sequences per sample) (Figure 4.4B). Samples Go1 and Go4 had the highest OTU counts and Ba7, Ba8 and Wo1 had the lowest OTU counts. The rarefaction curve for the *Mycobacterium* genus water dataset also illustrated that samples reached a plateau at less than 1000 sequences suggesting sufficient sampling effort was achieved (Figure 4.5). The OTU rich sites included Ji6 and Ba4 and the OTU poor sites included Ho3 and Wo3.

The variation in the number of *Mycobacterium* genus OTUs was significantly greater between source types compared to within source types ($F(1,63) = 18.0$, $P < 0.0001$). The mean number of OTUs was highest in soil samples (200) compared to water samples (148) suggesting soil samples have a higher richness of *Mycobacterium* species. The variation in the number of OTUs was not significantly different between soil samples from different regions ($F(5,17) = 1.38$, $P = 0.28$). However, this was significantly different between regions for water sample ($F = (8,33) = 5.3$, $P<0.005$). The mean number of OTUs was highest in Bale (206) and lowest in Turmi (63).

Overall the SG datasets had lower OTU counts compared to the *Mycobacterium* genus datasets; this was expected as this primer set only targets a subset of the *Mycobacterium* species. The 32 samples that were retained for analysis had a range of 13-36 OTUs per sample. Ho1 and Ba7 were the most OTU rich and Wo2 and Ji6 had the least number of OTUs (Figure 4.6A). The remaining 10 soil samples (<500 sequences per sample) did not plateau and were not included in the analysis (Figure 4.6B). The SG water dataset had a range of 13-47 OTUs per samples. OTU rich areas included Ho3 and Ji5 and low numbers of OTUs were observed in Wo1 and Ho1 (Figure 4.7). The variation in the number of OTUs between regions compared to within regions was not significant for both soil ($F(7,24) = 1.2$, $P = 0.33$) and water ($F(8,33) = 1.34$, $P = 0.26$). This suggests no regional trends in OTU richness. There was also no significant differences between environmental sources ($F(1,72) = 3.0$, $P>0.08$).

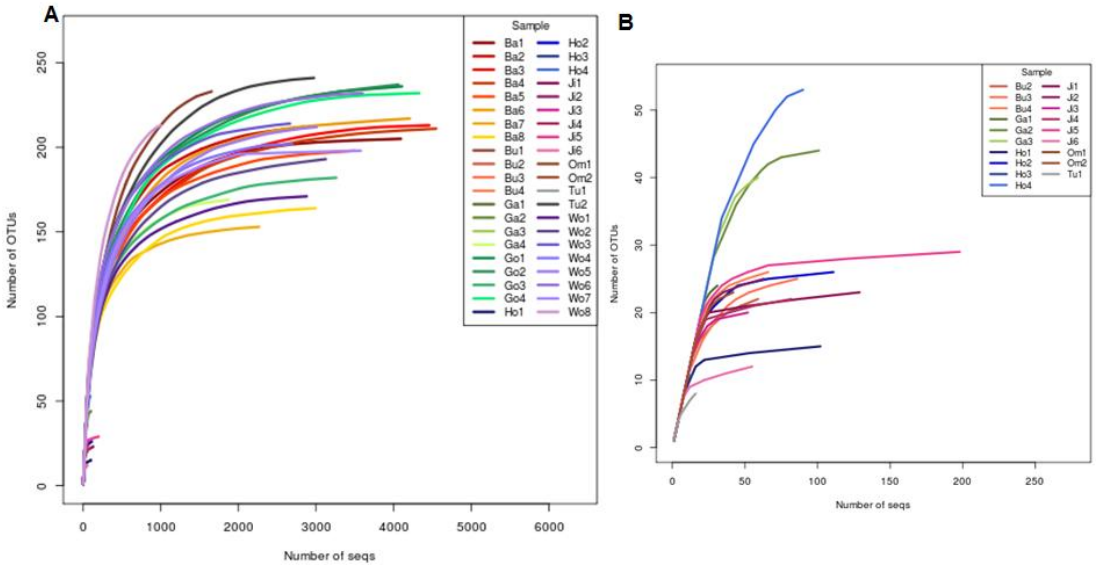


Figure 4.4. Rarefaction curves for the *Mycobacterium* genus soil dataset **(A)** all 42 soil samples **(B)** 19 soil samples with less than 1000 sequences which were removed from the analysis.

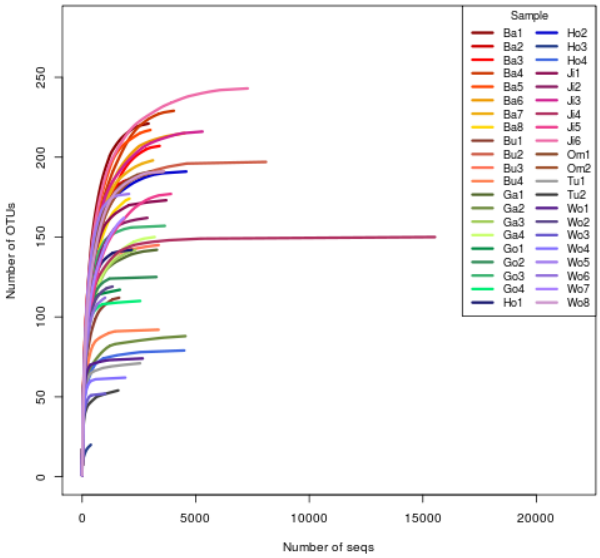


Figure 4.5. Rarefaction curve for the *Mycobacterium* genus water dataset. All 42 samples were kept for analysis after quality control procedures; the range of sequences per sample was 385-15533.

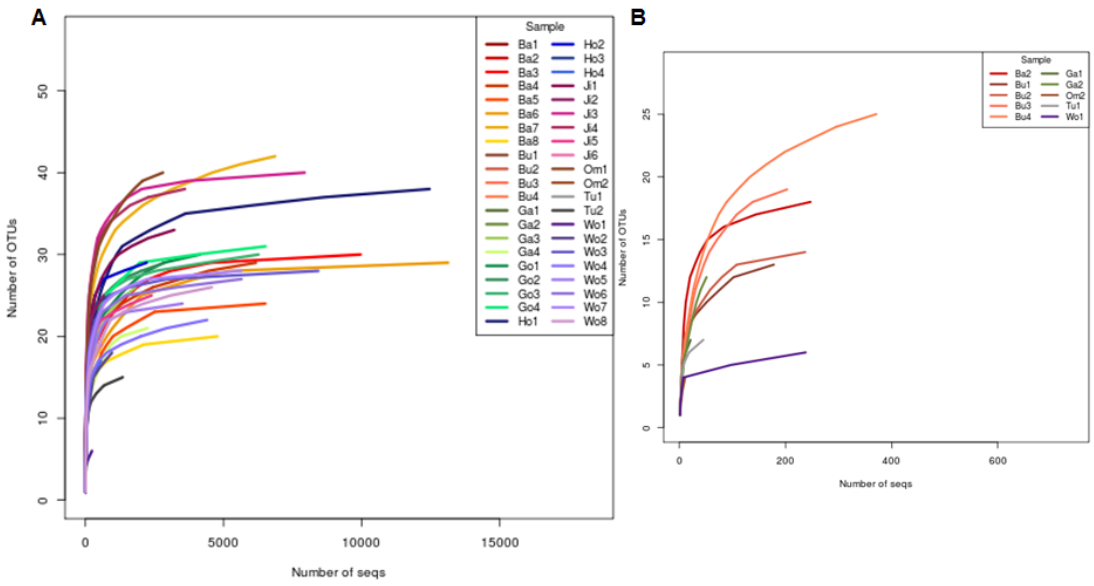


Figure 4.6. Rarefaction curves for the SG soil dataset. **(A)** All 42 soil samples after initial quality control **(B)** 10 soil samples with a low sequence count.

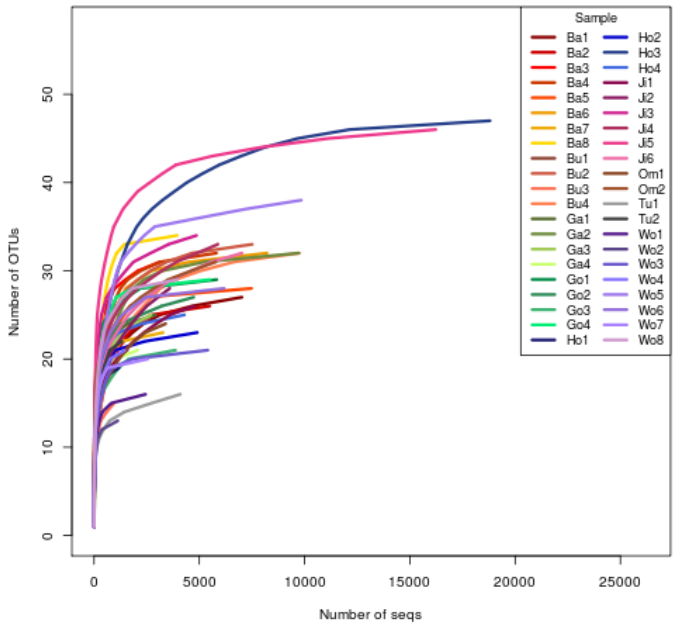


Figure 4.7. Rarefaction curve for the SG water dataset. After quality control procedures the range of sequences per sample was 950-18808.

4.5. The Alpha Diversity of *Mycobacterium* species

The Shannon index was used to assess the diversity of OTUs in the samples. This process takes into account the relative abundance of all OTUs but is known to be particularly sensitive to sample sizes. For the *Mycobacterium* genus soil dataset Go2 had the highest richness of OTUs and Wo1 was the least rich, and the Shannon

diversity estimate was highest for Wo8 and lowest for Wo7 (Table 4.4). For the *Mycobacterium* genus water dataset the Shannon diversity estimate was highest for Ba4. The *Mycobacterium* genus Shannon index for soil samples was higher on average compared to the water samples (average 4.2 and 3.6 respectively), which suggests that soils have a higher diversity of *Mycobacterium* species compared to water (section 5.6). The SG soil dataset had higher Shannon index estimate for Ba7 and lowest for the lowland area Tu2. For the SG water dataset, Ji2 had higher Shannon index estimate and Wo2 had the lowest estimate. Overall, the SG Shannon index for soil samples and water samples were similar (average 1.8 and 1.9 respectively) (section 5.7).

Table 4.4. OTUs and species richness estimates from both the *Mycobacterium* genus and SG datasets. Shannon index (H') is determined by the relative frequencies of sequences in OTUs and is an estimate of diversity.

		***Mycobacterium* genus**				**Slow-growing mycobacteria**			
		soil		**Water**		**Soil**		**water**	
Description	**Site**	**OTU**	**H'**	**OTU**	**H'**	**OUT**	**H'**	**OTU**	**H'**
Hot semi-arid low lands	Om1	-	-	191	4.12	33	2.49	24	2.18
Hot semi-arid low lands	Om2	-	-	191	3.99	-	-	31	2.36
Hot semi-arid low lands	Tu1	-	-	71	2.32	-	-	16	1.63
Hot semi-arid low lands	Tu2	220	4.28	54	2.67	15	1.38	22	2.05
Tropical hot sub-humid low lands	Ga1	-	-	142	3.49	-	-	25	1.95
Tropical hot sub-humid low lands	Ga2	-	-	88	2.78	-	-	32	1.89
Tropical hot sub-humid low lands	Ga3	-	-	142	3.95	27	2.1	25	2.13
Tropical hot sub-humid low lands	Ga4	168	4.16	150	3.88	20	2.07	21	1.82
Warm moist low lands	Ba1	200	3.84	221	4.32	16	1.41	27	2.09
Warm moist low lands	Ba2	201	4.3	186	4.26	-	-	26	1.92
Warm moist low lands	Wo1	152	3.61	74	1.74	-	-	16	1.16
Warm moist low lands	Wo2	184	4.01	119	3.64	13	1.62	13	1.15
Hot/w arm dense w oodland	Ji1	-	-	173	3.49	28	2.15	26	2.34
Hot/w arm dense w oodland	Ji2	-	-	162	3.78	21	2	28	2.58
Hot/w arm dense w oodland	Ji3	-	-	216	4.02	34	2.06	34	2.23
Hot/w arm dense w oodland	Ji4	-	-	150	1.68	33	2.08	33	2.49
Hot/w arm dense w oodland	Ji5	-	-	177	4.13	22	1.65	46	2.18
Hot/w arm dense w oodland	Ji6	-	-	243	4.03	15	1.61	32	2.53
Hot sub-humid cultivation midlands	Go1	233	4.26	117	3.01	27	2	29	1.42
Hot sub-humid cultivation midlands	Go2	227	4.31	125	2.06	21	1.96	27	1.99
Hot sub-humid cultivation midlands	Go3	185	3.92	157	2.91	24	2.08	21	1.54
Hot sub-humid cultivation midlands	Go4	241	4.38	110	2.36	26	2.07	29	1.34
Warm sub humid cultivation	Ho1	-	-	142	3.4	36	2.19	19	1.26
Warm sub humid cultivation	Ho2	-	-	191	3.67	28	1.8	23	1.51
Warm sub humid cultivation	Ho3	-	-	20	2.41	22	1.77	47	2.46
Warm sub humid cultivation	Ho4	-	-	79	2.24	21	1.82	25	1.65
Warm sub humid cultivation	Bu1	230	4.64	112	3.63	-	-	21	2.1
Warm sub humid cultivation	Bu2	-	-	197	2.95	-	-	33	2.15
Warm sub humid cultivation	Bu3	-	-	145	3.48	-	-	15	1.65
Warm sub humid cultivation	Bu4	-	-	92	2.09	-	-	32	2.36
Warm moist midlands	Wo3	209	4.2	52	1.84	24	1.74	21	1.22
Warm moist midlands	Wo4	184	4.17	62	1.49	21	1.65	21	1.8
Tepid sub-humid midlands	Wo5	194	4.36	177	3.7	19	1.58	38	2.19
Tepid sub-humid midlands	Wo6	226	4.36	112	3.79	23	1.7	28	1.58
Sub-humid mixed forest midlands	Ba3	213	4.08	207	4.42	25	1.61	31	2.1
Sub-humid mixed forest midlands	Ba4	210	4.17	229	4.45	28	1.97	32	1.97
Sub-humid mixed forest highlands	Ba5	216	4.15	217	4.38	25	1.45	28	1.42
Sub-humid mixed forest highlands	Ba6	220	4.14	215	4.18	25	1.48	32	1.87
Tepid sub-humid highlands	Wo7	204	3.57	162	4.41	24	1.67	20	1.07
Tepid sub-humid highlands	Wo8	205	4.78	191	3.87	23	2.08	29	1.77
Cold tepid highland areas	Ba7	143	3.61	198	4.25	36	2.51	23	1.6
Cold tepid highland areas	Ba8	154	3.79	174	4.38	22	1.49	34	1.64

4.6. OTU abundance distributions

Another component of diversity is the evenness of species community assemblages. Evenness is defined as the similarity of the relative abundance of OTUs, whereby high evenness describes samples with equally abundant OTUs and high dominance depicts samples that have few species which are more abundant than the remaining OTUs (Magurran 2004). This gives an indication of the niche space properties, for example dominant communities are more likely to exist in hostile environments and

even communities in more optimum conditions. Whittaker plots were employed to compare the commonness and rarity of *Mycobacterium* species between regions The *Mycobacterium* genus soil dataset demonstrated that samples from Gambella had a more dominant community assemblage with few abundant OTUs, whereas the region Butajira displayed high evenness as many OTUs had similar relative abundances (Figure 4.8). The Whittaker plots for the *Mycobacterium* genus water dataset displayed high evenness from Bale and more dominance within the community in Turmi (Figure 4.9). The rank abundance plots show there are no regional trends that are the same for both soil and water sources. The SG communities were more dominant in the soil samples from Turmi and more even in samples from Omorate (Figure 4.10). The assemblages in the water samples displayed higher dominance in Turmi and an even assemblage in Hossana (Figure 4.11). Overall the region Turmi had more dominance in the water samples community for both the *Mycobacterium* genus and slow-growing mycobacteria communities. This is perhaps because Turmi is a relatively hostile environment which may promote the growth of more resilient mycobacteria. In addition, for the *Mycobacterium* genus dataset the water samples displayed more dominance within the community compared to soil samples. This is perhaps because water represents a more hostile environment for mycobacteria compared to soil.

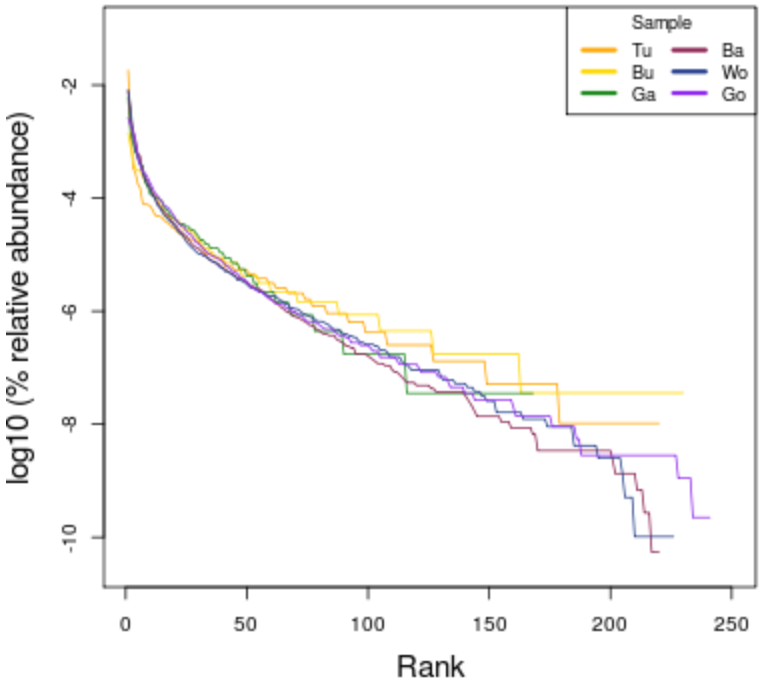


Figure 4.8. Whittaker plot of the *Mycobacterium* genus soil dataset.

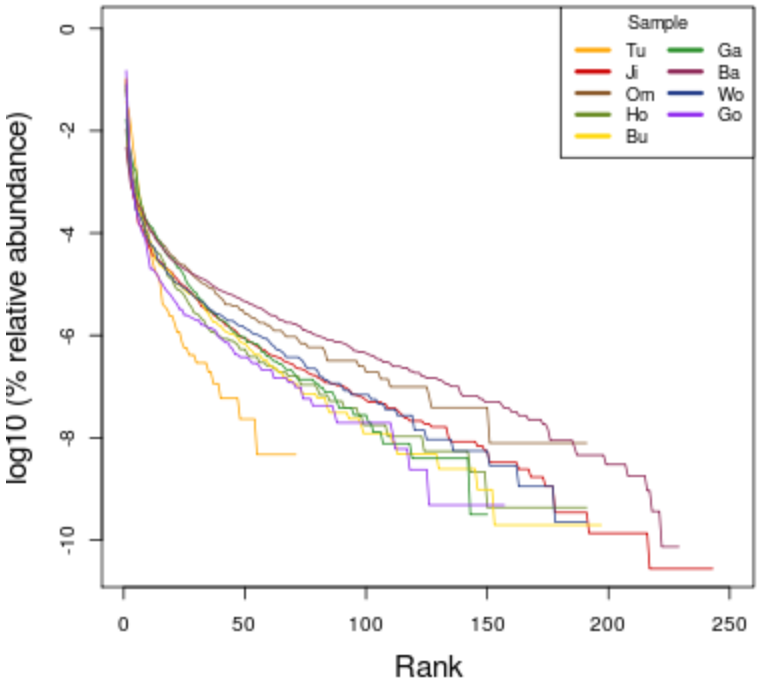


Figure 4.9. Whittaker plot of the *Mycobacterium* genus water dataset.

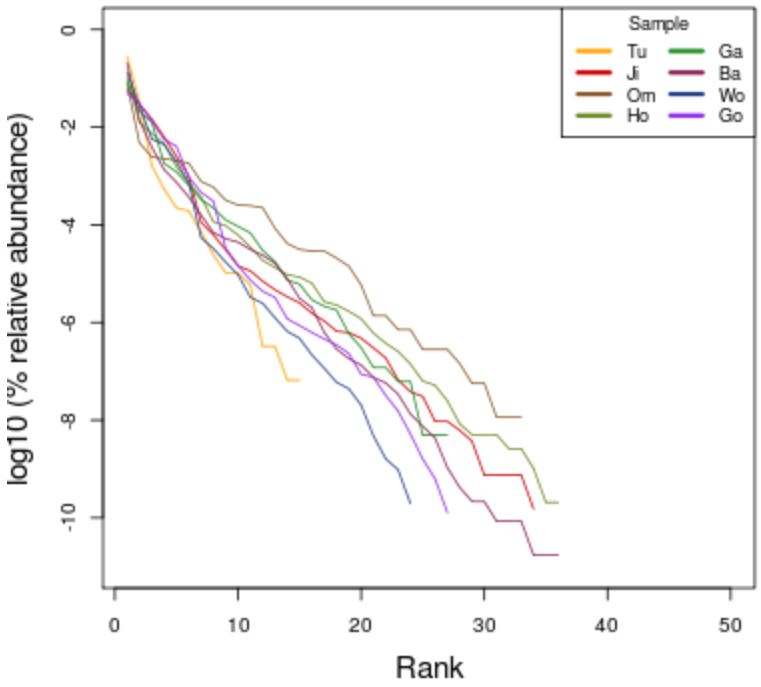


Figure 4.10. Whittaker plot of the SG soil dataset.

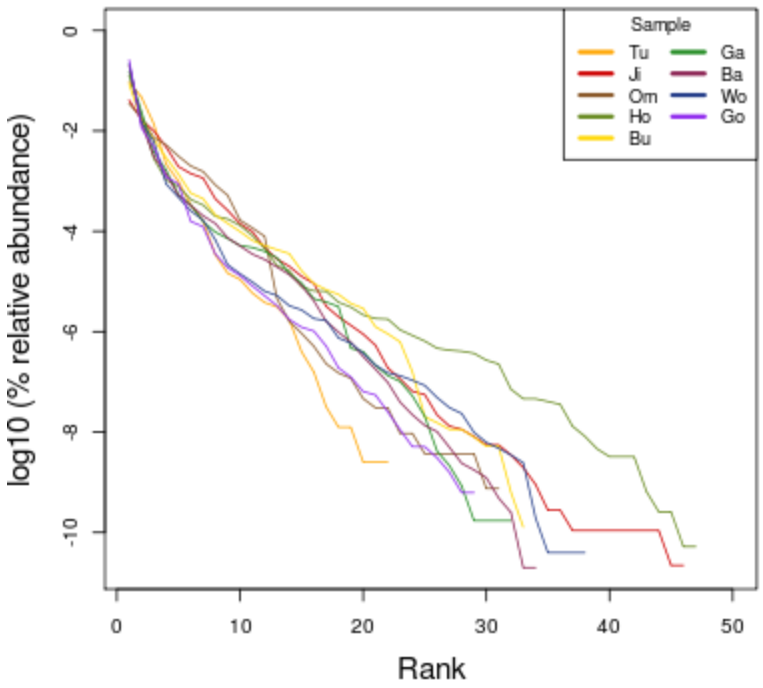


Figure 4.11. Whittaker plot of the SG water dataset.

4.7. Relative abundance of *Mycobacterium* species

A total of 222 different BLAST matches to *Mycobacterium* species were identified with an e-value of <0.001 from all 72,170 sequences of the *Mycobacterium* genus

soil dataset. Of these 60 were known FG, 46 were known SG and the remaining 116 were environmental uncharacterised species. The *Mycobacterium* genus water dataset had a total of 221 BLAST matches to *Mycobacterium* species from 142,547 sequences, and of these 61 were known FG, 44 were known SG and 116 were environmental uncharacterised *Mycobacterium* species. For both soil and water samples the top 60% most abundant species were all FG apart from the SG *M. gordonae*. This suggests that FG are more abundant in the environment. The most abundant species identified in soil compared to water were considerably different, with only *M. neglectum*, *M. neoaurum*, *M. vaccae*, and *M.* sp. T103 in the top 60% most abundant for both (Figure 4.12). However all *Mycobacterium* species in the top 60% for either soil or water were also present in the other, albeit at a lower abundance. This suggests that the *Mycobacterium* species communities are readily exchanged between soil and water environments, however the properties of the different substrates promote the proliferation of different species. The most abundant BLAST match for all water samples was *Mycobacterium aurum* ATTC 23070 with a relative abundance of 20.2% of all the *Mycobacterium* species BLAST matches (Figure 4.12B). In contrast, the most abundant species match for soil was *Mycobacterium* sp., which had a relative abundance of 6.1%, this was most closely related to *M. moriokaense* (Figure 4.12A). This suggests communities in soil have a greater degree of uncharacterised diversity and that water communities have a more dominant structure compared to soil.

The BLAST results of 142,727 sequences from the SG soil dataset generated a total of 59 different *Mycobacterium* species and the SG water dataset consisted of 220,648 sequences which generated 58 different matches. The top 95% most abundant SG present in the soil and water were very similar (Figure 4.13) this suggests that

mycobacteria from the soil are washing into the water. In the top 95%, 14/15 species detected in water were also in the top 95% for soil samples. This suggests that the community structure for soil and water was dominated by the same SG. From the top 95% most abundant matches, 12/17 and 10/15 SG species were opportunistic pathogenic species in soil and water respectively. The SG *M. gordonae* was the most abundant species (45.1%) in water samples (Figure 4.13B) whereas *M. malmoense* was the most abundant (24.59%) in soil samples (Figure 4.13A). This set of findings were consistent with the *Mycobacterium* genus BLAST results as it illustrated by more dominance within the water sample communities compared to soil samples. The *Mycobacterium avium* complex (MAC) was relatively abundant in soil (28.2%) and in water samples (16.2%). The most abundant member of the MAC was *M. colombiense*, which represented 22.6% and 12.8% of all BLAST matches from soil and water respectively (Figure 4.13).

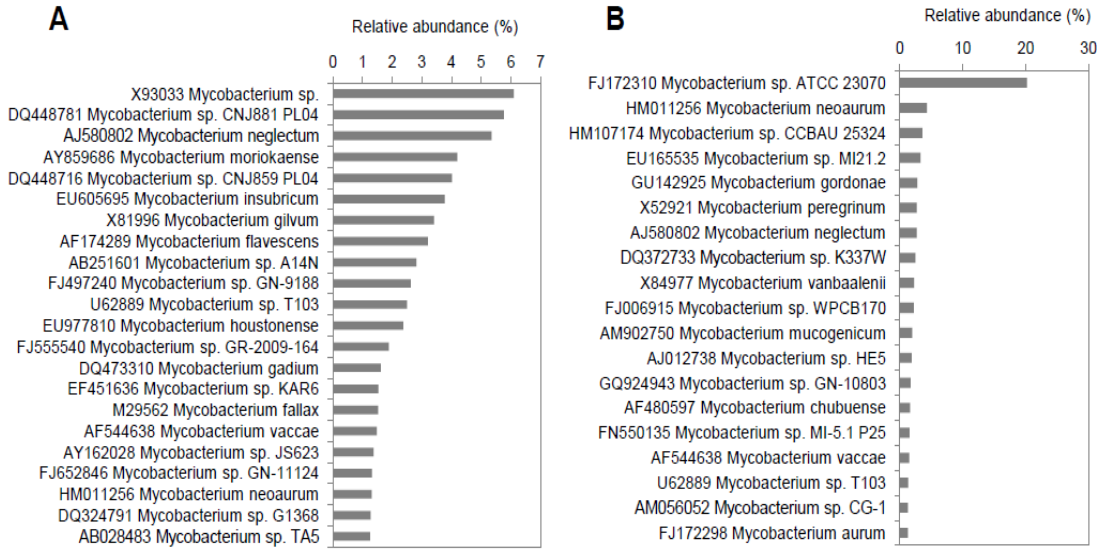


Figure 4.12. The relative abundance of the top 60% most relatively abundant species using the *Mycobacterium* genus dataset. **(A)** soils **(B)** water

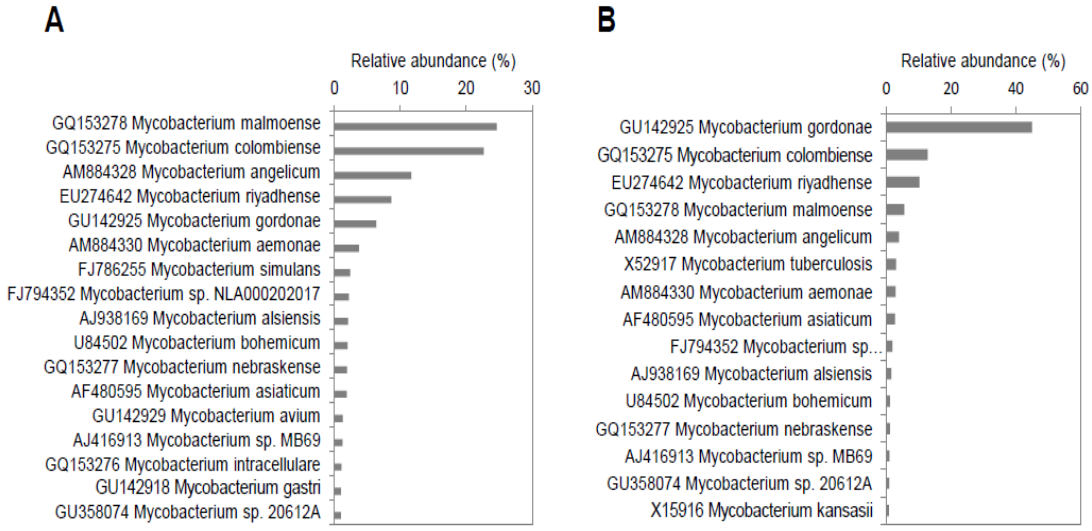


Figure 4.13. The relative abundance of the top 95% most relatively abundant species using the SG dataset. **(A)** soils **(B)** water

4.8. Phylogenetic diversity of the *Mycobacterium* genus across all sites

Maximum likelihood (ML) phylogenetic trees were constructed to (1) compare the representative sequences with known *Mycobacterium* species downloaded from SILVA and (2) compare the OTUs with each other. The ML phylogenetic trees constructed for the *Mycobacterium* genus soil dataset included OTUs that represented >0.5% of all sequences. Each OTU represented <4.2% of all sequences, which suggests a high diversity of *Mycobacterium* species and even community assemblages. The ML phylogenetic tree consisting solely of OTUs demonstrated that the representative sequences that clustered together were BLAST matched to the same species (Figure 4.14), suggesting genetic variation in OTUs closely related to the same species. The alternative approach was to align OTU representative sequences with known *Mycobacterium* species (Figure 4.15) and these results were consistent with OTU BLAST matches. Both approaches demonstrated that the relative abundant OTU1 (3.58%) was closely related to *M. neoaurum* (Figure 3.14 & 3.15). In addition, OTUs 33 and 7 representing 3.04% of all sequences were both BLAST matched (Figure 4.14) and present in the same clade (Figure 4.15) as *M.*

neglectum. The OTUs 6 and 17 were both BLAST matched and clustered close to the SG *M. bohemicum* and *M. parrafinicum* (Figure 4.14 & 4.15). However there were some discrepancies between both approaches; for example OTUs 21, 51 and 11 were BLAST matched to *M. jacuzzii* (Figure 4.14) but did not cluster with *M. jacuzzii* (Figure 4.15).

The same approach was taken to assess the phylogenetic diversity of mycobacteria in the water samples. The most abundant OTU represented 19.5% of all sequences, indicating that water displayed more dominance within the community. The most abundant OTU 0 (19.52%) was BLAST matched to the environmental species *Mycobacterium* sp. WPCB170 (Figure 4.16) and clustered close to *M. augbanense*, *M. celatum* and *M. mucogenicum* (Figure 4.17). The OTU 3 was in the same clade (Figure 4.16) and BLAST matched (Figure 4.17) to *M. gordonae*. Both approaches illustrated that OTUs 11 and 416 were closely related to *M. neoaurum* representing 1.97% of the sequences (Figure 4.16 & 4.17). However, two OTUs that were BLAST matched to *M. sentense* and OTU matched to *M. mucogenicum* (Figure 4.16) did not cluster with these species (Figure 4.17). Overall the alignment method and BLAST matches to OTUs were in close agreement, providing further support for the identification of species.

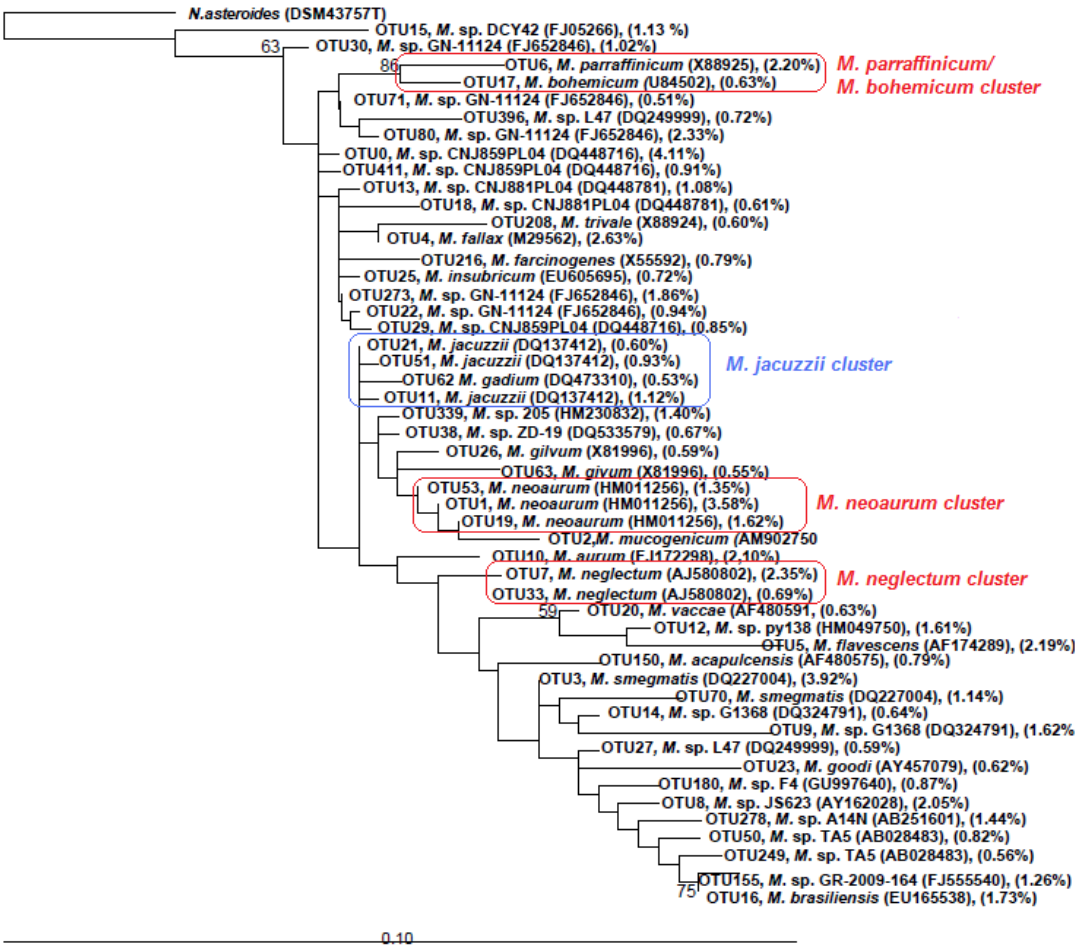


Figure 4.14. PHYML tree of the most relatively abundant OTUs (>0.5%) for all soil samples analysed using the *Mycobacterium* genus primer set. Fifty most prevalent OTUs, representing 70.1% of the dataset. Each entry reads name of the OTU, BLAST match of the OTU and the relative abundance of the OTU. The phylogenetic tree was bootstrapped with 1000 replicates.

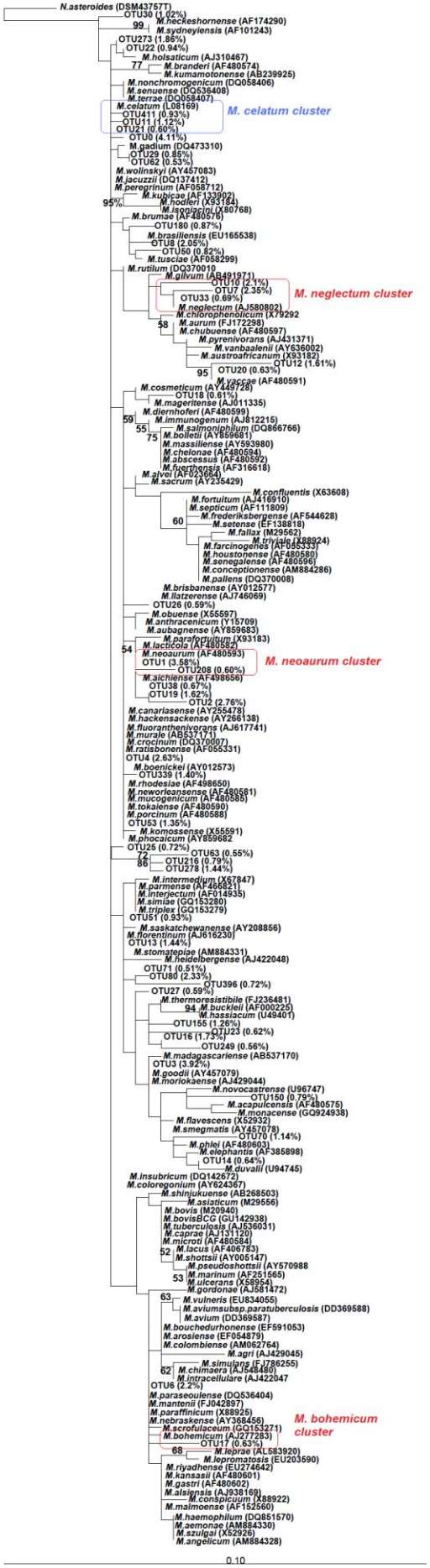


Figure 4.15. PHYML phylogenetic tree of prevalent clusters (>0.5%) for all soil samples and known *Mycobacterium* species. Bootstrapped with 1000 replicates, the fifty most prevalent OTUs aligned with known *Mycobacterium* species, representing 70.1% of the dataset.

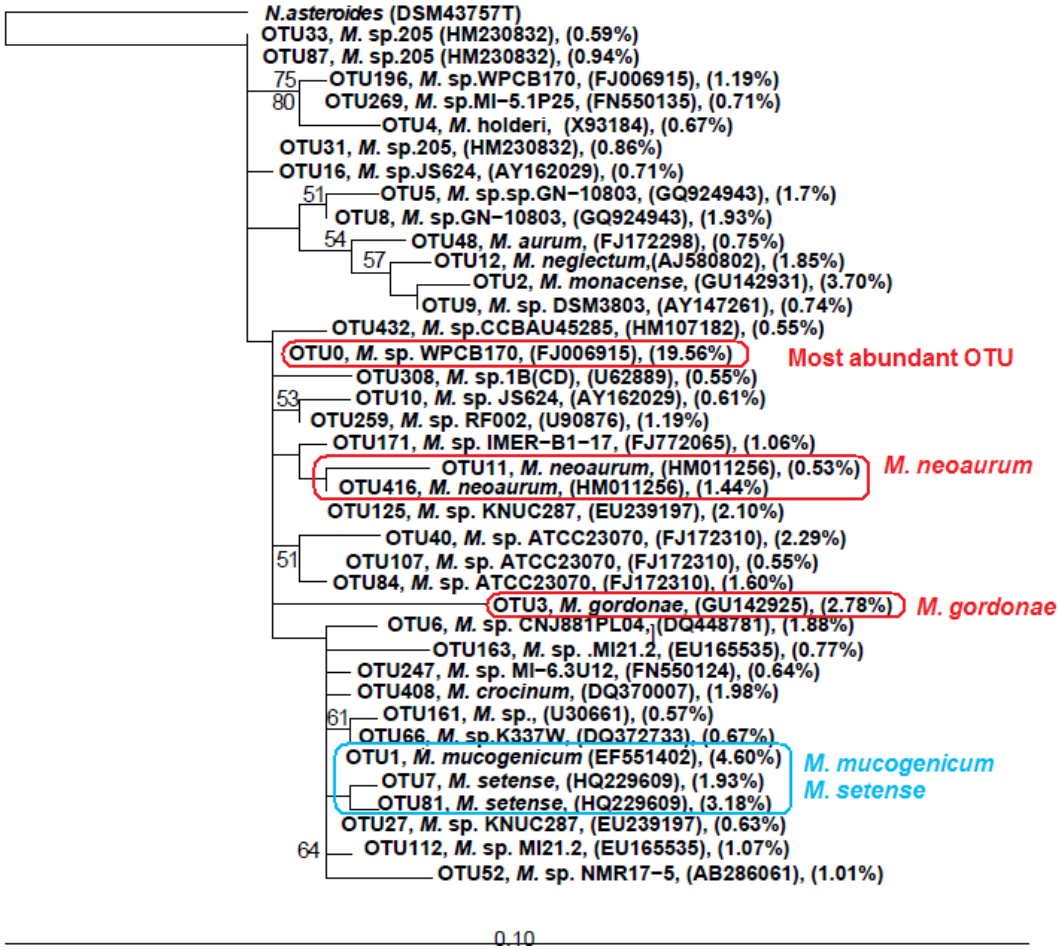


Figure 4.16. PHYML tree of the most relatively abundant OTUs (>0.5%) for all water samples analysed using the *Mycobacterium* genus primer set. The phylogenetic tree was bootstrapped with 1000 replicates. Thirty-eight most prevalent OTUs, representing 70.1% of the dataset. Each entry reads name of the OTU, BLAST match of the OTU and the relative abundance of the OTU.

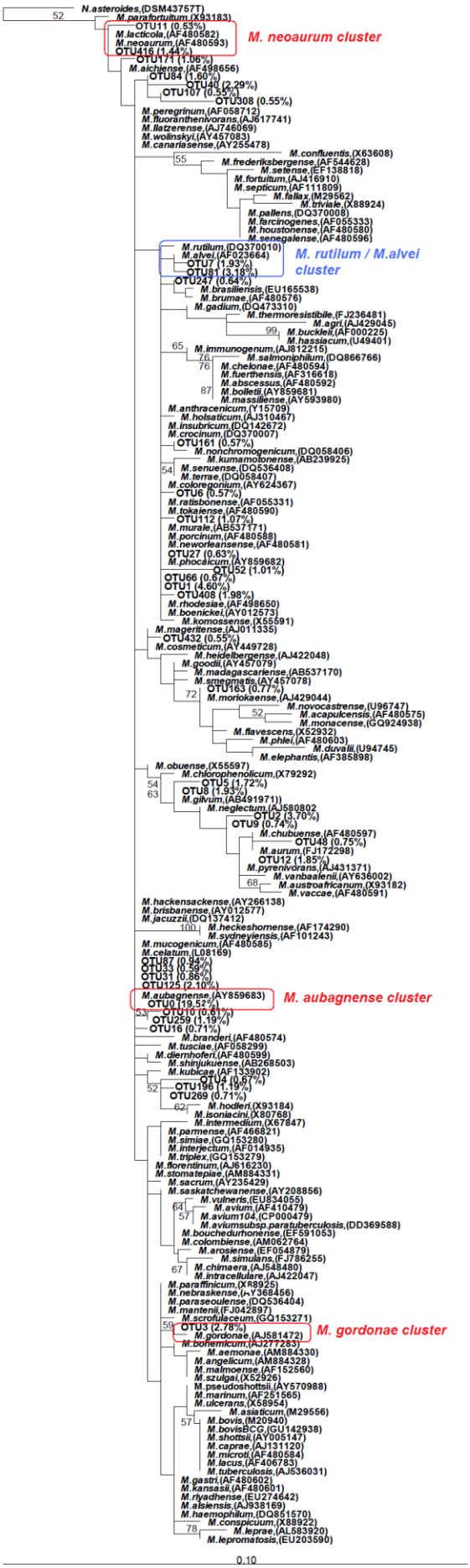


Figure 4.17. PHYML phylogenetic tree of prevalent clusters (>0.5%) for all water samples and known *Mycobacterium* species. The phylohenetic tree was bootstrapped with 1000 replicates. Thirty-eight most prevalent OTUs aligned with known *Mycobacterium* species, representing 70.1% of the dataset.

4.9. Phylogenetic diversity of SG across all sites

The representative sequences of most abundant OTUs (>0.1%) present within soil samples were used to construct ML phylogenetic trees with and without known *Mycobacterium* species. Five OTUs were shown to cluster with the *M. avium* complex, representing 22.2% of sequences (Figure 4.19). In concordance with these findings the same OTUs also clustered together and four were BLAST matched to *M. avium* complex members (Figure 4.18). An additional seven OTUs were BLAST matched to *M. riyadhense* (Figure 4.18) and three of these OTUS were clustered close to *M. riyadhense* (Figure 4.19). The OTU 3 (6.7%) was BLAST matched to *M. gordonae* (Figure 4.18) and present in the same clade as *M. gordonae* (Figure 4.19). However there were several BLAST matches to OTUs, such as those matched to *M. malmoense* and *M. marinum* (Figure 4.18), that did not cluster with these species (Figure 4.19).

The SG communities in water samples were dominated by *M. gordonae*, illustrated by eleven OTUs that were BLAST matched to this species (Figure 4.20) and seven OTUs that clustered close to this species; these represented 35.8% of all sequences (Figure 4.21). Interestingly, abundant OTUs that clustered within the *M. tuberculosis* complex were only predominant in water samples comprising of 2.97% of the sequences (Figure 4.21) and these OTUs were also BLAST matched to *M. tuberculosis*. Both approaches demonstrated that OTUs 1 and 29 were closely related to *M. conspicuum* representing 6.3% of the sequences (Figure 4.20 & 4.21). Three OTUs were BLAST matched to the *M. avium* complex member *M. colombiense* (Figure 4.20) and these OTUs clustered within the MAC (Figure 4.21). However, the OTUs 17 and 3 that clustered with *M. malmoense* represented 4.66% of the sequences (Figure 4.21), however these were BLAST matched to *M. aemonae*

(Figure 4.20). For both soils and waters, 6 OTUs clustered with the FG *M. chitae* which suggests some non-specificity (Figure 4.19 & 4.20), of interest *M. chitae* is the one of the most distantly related *Mycobacterium* species (Stahl & Urbance 1990). Overall the phylogenetic profiles of the soil and water samples displayed a number of different OTU clusters; however, it is notable that the main differences are the relative abundances of the predominant clusters.

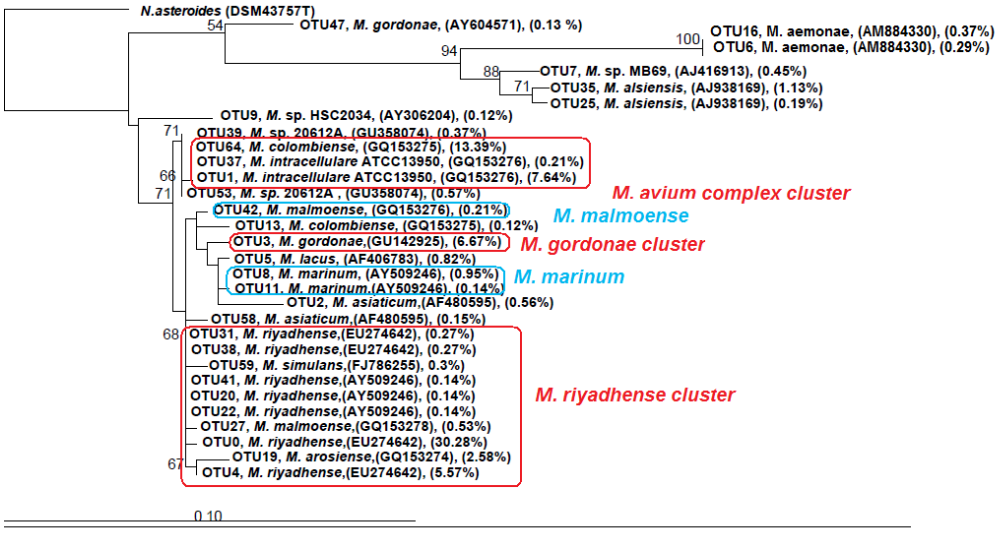


Figure 4.18. PHYML tree of the most relatively abundant OTUs (>0.1%) for all soil samples analysed using the SG primer set. The phylogenetic tree was bootstrapped with 1000 replicates Thirty most prevalent OTUs, representing 99% of the dataset. Each entry reads name of the OTU, BLAST match of the OTU and the relative abundance of the OTU.

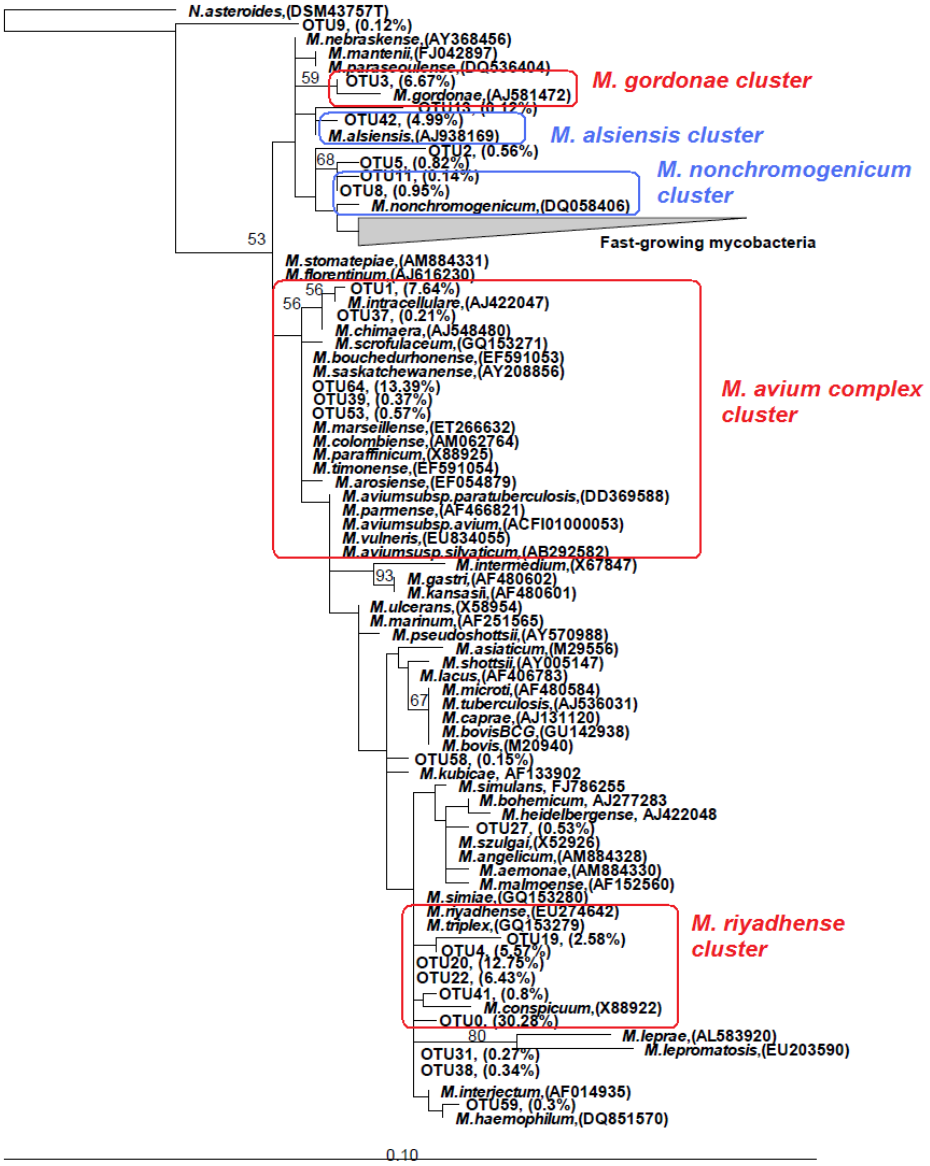


Figure 4.19. PHYML tree of prevalent clusters (>0.1%) for all soil samples and known SG. The phylogenetic tree was bootstrapped with 1000 replicates. Thirty most prevalent OTUs aligned with known *Mycobacterium* species, representing 99% of the dataset.

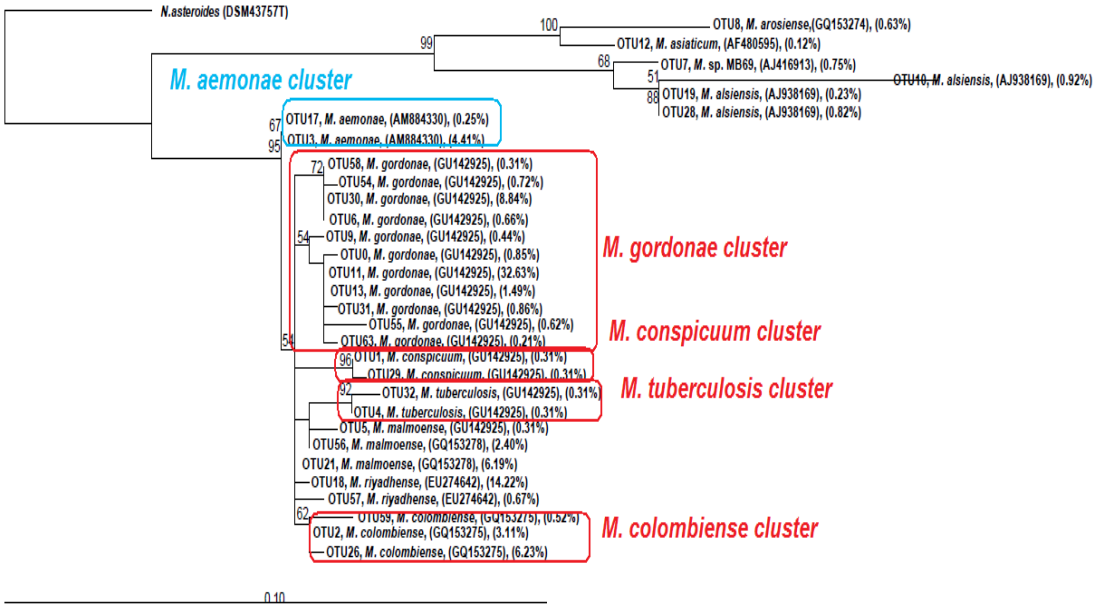


Figure 4.20. PHYML tree of the most relatively abundant OTUs (>0.1%) for all water samples analysed using the SG primer set. The phylogenetic tree was bootstrapped with 1000 replicates Thirty-one most prevalent OTUs, representing 99% of the dataset. Each entry reads name of the OTU, BLAST match of the OTU and the relative abundance of the OTU.

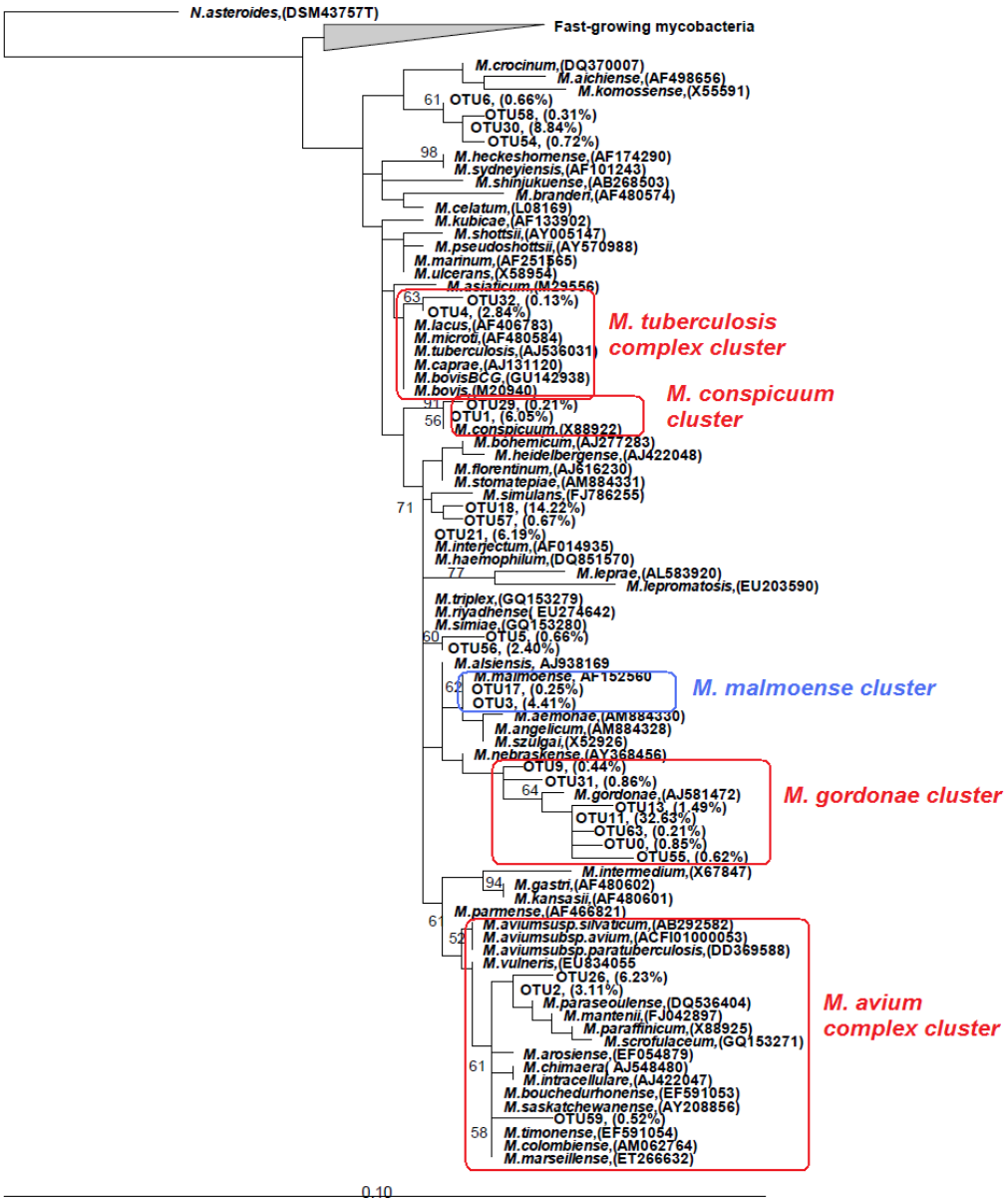


Figure 4.21. PHYML phylogenetic tree of prevalent clusters (>0.1%) for all water samples and known SG. The phylogenetic tree was bootstrapped with 1000 replicates. Thirty-one most prevalent OTUs aligned with known *Mycobacterium* species representing 99% of the data.

4.10. Discussion

This is the first time that *Mycobacterium* species diversity has been characterised on a countrywide scale using pyrosequencing technologies. Through the use of a molecular pyrosequencing approach to bypass the limitations of culture-dependent methods it was demonstrated that Ethiopian soil and water environments have an

exceptionally high diversity of FG and SG. Alpha diversity estimates of soil and water environments revealed a greater diversity of FG compared to SG; this suggests FG are more successful in the environment. Soil samples were shown to have a significantly higher diversity of mycobacteria (no. of OTUs) compared to water, perhaps due to the higher organic content of soil. However this environmental trend was not observed for the SG dataset. A significantly higher variation was observed between regions for the diversity (no. of OTUs) of the *Mycobacterium* genus water dataset, suggesting that regions may be characterised according to the particular species and lineages belonging to a particular area. However, this regional clustering was not significant for the soil samples or for the soils and waters of the SG dataset. The Whittaker plots and the relative abundance of BLAST matches indicated that the water samples community structure shows more dominance. In contrast to soil samples which contained a greater number of even *Mycobacterium* species assemblages. Phylogenetic analysis of prevalent OTUs and BLAST results demonstrated that soil and water have many of the same *Mycobacterium* species, perhaps due to the ease of transmission between the soil and water environments. However soil and water displayed different relative abundances of *Mycobacterium* species, possibly due to different properties of soil and water which promote different mycobacteria to profilerate.

The phylogenetic analysis, BLAST results of all sequences and OTUs presented similar findings in relation to the identity of the *Mycobacterium* species. By utilising multiple approaches in this way it is possible to be more confident about the community structure present in the soil and water samples. The most abundant species in the *Mycobacterium* genus pyrosequencing assay for soil were all FG apart from one SG, *M. gordonae*. Identified relatively abundant species included *M.*

moriokaense, *M. flavescens, M. houstonense* and *M. neoaurum* which rarely cause disease , in addition to a further five species which were first described as clinical isolates (*M. gilvum*, *M. insubricum*, *M.* sp. GN-11124, *M.* sp. GN-9188 and *M.* sp. GR-2009-164) and six species known for the degradation of polycyclic aromatic hydrocarbon (PAH) *M. gilvum* (strain PYR-GCK), *M. flavescens, M* sp. A14N, *M.* sp. JS623, *M.* sp. T103 and *M.* sp. TA5. The most abundant species present in the water samples included just one SG (*M. gordonae),* and further matched six species which were known to be PAH degraders; these were *M. chubuense*, *M.* sp. HE5, *M. vanbaalenii, M.* sp. K337W, *M.* sp. T103 *and M.* sp. MI21.2. The most abundant species in water included opportunistic pathogens such as *M. aurum*, *M. gordonae*, *M. mucogenicum*, *M. neoaurum* and *M. peregrinum*.

This study found that in terms of presence/absence of the same species detected in the *Mycobacterium* genus dataset, soil and water had similar profiles whereas the relative abundance of these species was very different. This is in concordance with one of the few studies which have compared mycobacteria communities across environmental materials; a Finnish study of lakes and forest soils using PCR-DGGE found that the soil and water environments identified different prevalent *Mycobacterium* species (Niva et al. 2006). In South India different isolate profiles for soil and water were observed, where the MAIS complex was predominant in water and *M. fortuitum* was prevalent in soil (Kamala et al. 1994). These findings were supported by a Malawian study which also found *M. fortuitum* to be prevalent in soil cultured isolates (Chilima et al. 2006). This study found that the relative abundance of *M. fortuitum* from all BLAST matches was negligible (0.18% and 0.74% in soil and water respectively), perhaps due to the relative ease of its cultivation compared to other more prevalent *Mycobacterium* species. In contrast the Ethiopian soil

samples were found to have a high prevalence of FG *M. neglectum* and *M. moriokaense* and SG *M. malmoense* and *M. colombiense*.

The results presented above show that *M. gordonae* was most prevalent in the water samples. This is in agreement with the water distribution systems survey in Paris which found *M. gordonae* to be the most prevalent isolate, as it comprised of 28.8% of all positive samples and was most abundant in surface waters (Le Dantec et al. 2002a). This is also in accordance with a survey in Seoul, Korea, which found *M. gordonae* prevalence to be the highest of all known *Mycobacterium* species in surface waters (Lee et al. 2008). *M. gordonae* also represented the highest number of isolates in a South African water distribution system (September et al. 2004). This suggests some concordance between cultivation and molecular methods and implicates *M. gordonae* as a widespread opportunistic pathogen. Other prevalent species in the Ethiopian waters included the FG *M. aurum* sp. ATCC 23070 and *M. neoaurum* and the SG *M. colombiense* and *M. riyadhense*. The agreement of these studies worldwide and the work presented in this thesis suggests that water communities select for specific mycobacteria.

A high diversity of BLAST matches to species and environmental isolates involved in PAH degradation were identified in both the soil and water samples. The *Mycobacterium* genus dataset for soil samples found that 20.4% of sequences were BLAST matched to species involved in biodegradation processes. A similar number of sequences were BLAST matched to PAH degrading species (24.2%) in water samples. The community profiles of the PAH degraders were very similar for soil and water samples, with 46 species found in both. Prevalent PAH degraders in soil included *M. flavescens*, which has been shown to utilise pyrene as the sole source of carbon and energy (Dean-Ross & Cerniglia 1996). *M. gilvum* was also prevalent in

soil samples; this has previously been isolated from PAH contaminated sites and has been characterised as a PAH degrading *Mycobacterium* (Hall et al. 2005; Leys et al. 2005). The PAH degrader *M. vanbaalenii* was prevalent in the water samples, and is commonly known to metabolise PAH compounds (Khan et al. 2002). The high diversity of species and environment isolates identified from Ethiopian soil and water samples presents great potential for bioremediation applications.

Few studies have identified SG in the environment as the group is often more difficult to detect compared to FG (Ghaemi et al. 2006; Kamala et al. 1994; Thorel et al. 2004), but it is important as many opportunistic pathogens belong to this group. In this study specific primers used to target a select group; the SG long helix 18 group (Leclerc et al. 2003) identified a wide diversity of SG.

Of the SG, the most prevalent species in soil was *M. malmoense* (24.6%); known as a potentially pathogenic species, it is the causative agent of cervical lymphadenitis and chronic pulmonary disease. *M. gordonae* (45.1%) was the most prevalent species in water; this is rarely causes disease and predominantly affects those that are immunocompromised. Members of the *M. avium* complex which are known opportunistic pathogens were relatively abundant in soil and water; comprising 28.2% and 16.2% of all SG sequence BLAST matches respectively. The prevalent species *M. simulans* and *M. riyadhense,* have been identified as causative agents for symptomatic pulmonary infections. The *M. tuberculosis* complex was identified as a prevalent group in water samples (3.1%). As members of the complex are indistinguishable in the 16S rRNA region, further work discussed in Chapter 6 has attempted to determine the prevalence of *M. bovis.* There are several SG of particular human health interest which cannot be discriminated using the *Mycobacterium* genus and the SG primers, such as *M. avium* subspecies, *M. ulcerans* and *M. marinum*, *M.*

gastri and M. *kansasii, M. colombiense and M. bouchedenohunse* and *M. angelicum* and *M. szulgai*. Therefore other gene targets are required in order to differentiate between these species.

The employment of the SG specific primer set has enabled a rare and comprehensive insight into the diversity of this potentially pathogenic group. Of all mycobacteria species, the *M. avium* complex has been most comprehensively researched in the environment owing to its ubiquitous nature and its status as an important opportunistic pathogen, particularly for those with HIV/AIDS (Horsburgh 1991). In this study matches to species belonging to the *M. avium* complex were relatively abundant. Members of the *M. avium* complex have been consistently isolated from soil and water environments, with higher numbers of isolates detected with higher concentrations of humic and fulvic acids (Kirschner et al. 1999). Interestingly, higher numbers of *M. avium* complex isolates were detected in the United States compared to Africa in a unique geographical study of water distribution systems (Von Reyn et al. 1993b). However similar numbers of isolates were obtained in Ugandan water samples compared to South-Eastern USA waters and isolates were found in higher numbers in Ugandan soil (Eaton et al. 1995). Exposure to the soil environment was shown to be a risk factor for *Mycobacterium avium- intracellulare* complex disease (Maekawa 2011). Therefore the prevalence of *M avium* complex that was detected in Ethiopian soil and water samples is significant as this environment poses a serious hazard as it could be a source of exposure and infection particularly to those with HIV/AIDS.

Mycobacterium species are also implicated in the reduction of BCG efficacy in lower latitude countries (Brandt et al. 2002; Flaherty et al. 2006; Weir et al. 2008a). Few studies have elucidated the *Mycobacterium* species which individuals are most

exposed to and subsequently may interfere with the efficacy normally obtained from the BCG vaccine. A skin test survey in Malawi found that overall individuals had greater exposure to *M. scrofulaceum*, *M. intracellulare* and *M. fortuitum* (Fine et al. 2001). In agreement, IFN-γ responses measured using the whole blood assay also found highest sensitivity to the MAIS complex (Black et al. 2001). Therefore it is of interest that the results presented in this chapter show a high relative abundance of the *M. avium* complex in the Ethiopian environment. However, an early skin test survey which took place in Ethiopian regions Hossana and Butajira found highest responses to the FG *M. chitae*, *M. dienhorferi*, *M. vaccae* and the SG *M. kansasii* (Lema & Stanford 1984). These species were detected in our study but were low in abundance. Sensitisation studies of Ethiopian wildlife found that *M. terrae* was the most commonly detected (Tschopp et al. 2010a; Tschopp et al. 2010b). *M. terrae* had a relatively low abundance in our survey, representing 0.04% of all BLAST matches. The species *M. nonchromogenicum* and *M. gordonae* were commonly isolated from Ethiopian cattle (Berg et al. 2009). Both species were detected in our study, but only *M. gordonae* was prevalent.

In summary, this study has provided a unique insight into the diversity of *Mycobacterium* species present within the Ethiopian environment. It has shown a wide diversity of both PAH-degrading FG and opportunistic pathogenic SG. This should offer a parallel to the studies that have shown that human, livestock and wildlife in Ethiopia have a wide ranging sensitisation to mycobacteria. Our findings have important implications for the clinical relevance of mycobacteria and should encourage more comprehensive correlations of the *Mycobacterium* species in the environment and exposure in these locations.

Chapter 5: Biogeographical trends determine the diversity and community composition of mycobacteria

5.1. Abstract

The biogeography of mycobacteria and specifically SG was elucidated on a countrywide scale across 42 villages in 9 different regions. Our findings demonstrate the importance of elevation and temperature for the community composition of mycobacteria in soil. For example an increased prevalence of sequences related to *M. neglectum* and reduced prevalence of those related to *M. insubricum* were observed at higher elevations. A non-linear relationship was observed between elevation and the diversity of mycobacteria, with a peak mid-elevation. The community composition of SG was correlated with several factors, however longitude was consistently influential for both composition and diversity. In particular sequences related to *M. malmoense* were highest in the east and *M. aemonae* highest in the west. In contrast latitude was the primary factor to explain the community composition of mycobacteria and specifically SG in water samples, for example the prevalence of *M. gordonae* was highest in the north and *M. colombiense* was highest in the south. A higher diversity of mycobacteria and particularly SG was observed in water samples from the south. These findings demonstrate the importance of bioclimatic and spatial factors for community structure of mycobacteria and may enable future predictive spatial mapping to identify areas with a higher risk of exposure and infection.

5.2. Introduction

Biogeography is the study of changes to community composition with variation in environmental, temporal and spatial factors. The biogeography of macroorganisms has been studied extensively and patterns such as the latitudinal species diversity gradient are consistently observed (Gaston 2000). The biogeography of the microbial world is less well understood. Reviews and research have recently sought to address this deficit (Fierer 2008; Fierer & Jackson 2006; Martiny et al. 2006; Prosser et al. 2007), perhaps prompted by improvements in molecular microbial detection methods which have heightened interest in this area.

There are many theories as to what determines the diversity and composition of organisms. The niche theory proposes that communities are differentially influenced by environmental heterogeneity and therefore thrive in different geographical and environmental locations (Tokeshi 1990). However, niche theory does not take into account the birth/death and dispersal of species. The neutral theory suggests that all species are ecologically equivalent where genetic variation is due to mutation and genetic drift and not selection (Chave 2004; McGill et al. 2006). As a result communities are structured from random processes and dispersal. It has been argued that the neutral model is too simplistic. It is likely that the interplay between both niche and neutral theory determine the composition of bacterial communities. Theories specific to bacterial communities must take into consideration their exceptional diversity; the high diversity is thought to be because bacteria have high dispersal rates due to their small size, high speciation rates due to the acquisition of genes from other organisms and their low extinction rates owing to their ability to withstand extreme environments (Fenchel 2003; Horner-Devine et al. 2004a; Ramette & Tiedje 2007).

There is an on-going debate as to whether bacterial communities display biogeographical trends, however mounting evidence suggest that they do. The subsequent question arises; what drives differences in beta-diversity? The Baas-Becking hypothesis suggests 'everything is everywhere, but the environment selects' (De Wit & Bouvier 2006). This proposes that bacteria are cosmopolitan, as these species are easily dispersed and that contemporary environmental conditions determine community composition. This cosmopolitan distribution was shown to be size dependent, organisms with <1mm showed this trend whereas those >10mm did not (Fenchel & Finlay 2003) therefore authors suggest that endemism is only prominent for macroorganisms (Fenchel & Finlay 2004). An alternative hypothesis is that differences are due to historical events, such as dispersal limitation, which created provinces with different assemblages (Martiny et al. 2006). Microbial endemism was observed in four soils collected from across the world which shared less than 5% of OTUs (Fulthorpe et al. 2008) and only 15% of OTUs were shared between 238 environmental assemblages (Nemergut et al. 2011). However, it is not possible to sequence everything; therefore it is difficult to determine whether OTUs are endemic. Overall, it is probable that both environmental variation and dispersal limitation are to some degree responsible for the observed differences in bacterial assemblages.

Spatial trends have been observed in microbial communities which were shown to be consistent with animal and plant communities (Green & Bohannan 2006). For example, the taxa-area relationship suggests that the greater the area sampled, the greater the number of species that will be detected, following a power-law relationship. A spatial study of salt marshes covering a scale of centimetres to metres found both ß-proteobacteria and plant communities exhibited a taxa-area relationship,

however the *z*-value (slope of the line) was lowest for ß-proteobacteria which suggests that this group are less dispersal limited than plants within a salt marsh (Horner-Devine et al. 2004b). ß-proteobacteria also exhibited a distance decay relationship, which explains that communities closer together in space will be more similar in composition than those that are further apart. A similar study on a larger scale found that Nitrosomonadales displayed a distance-decay relationship on a local scale but not on a continental scale, perhaps due to greater dispersal limitation at a local scale and the increased influence of environmental variables on a continental scale (Martiny et al. 2011).

Several environmental factors were shown to affect the richness, diversity and composition of microbial communities. The analyses of bacterial assemblages using the 16S rRNA gene for TRFLP and pyrosequencing found that the edaphic variable pH was the most influential factor for bacterial community composition (Fierer & Jackson 2006; Griffiths et al. 2011; Lauber et al. 2009; Nacke et al. 2011) and specifically negatively correlated with the relative abundance of Acidobacteria (Jones et al. 2009). A composite study of over 200 environmental samples demonstrated that salinity was the driving factor for differences in bacterial assemblages worldwide (Lozupone & Knight 2007). Sodium concentration and electrical conductivity salinity were correlated with archaeal community structure (Hollister et al. 2010). The environmental factors surface temperature and nutrient availability were revealed to correlate with changes in community structure in the marine environment (Gilbert et al. 2009). Consequently, environmental heterogeneity appears to be an important influence for microbial biogeography in addition to spatial factors.

The biogeography of *Mycobacterium* species is reviewed in section 1.6. Only one study has employed a multivariate approach to assess *Mycobacterium* species abundance (Jacobs et al. 2009) and no studies to date have used a multivariate approach to assess *Mycobacterium* species community composition and diversity. The majority of the research has employed cultivation techniques to assess *Mycobacterium* species abundance and composition, however cultivation is considered to only reveal a fraction of the diversity (Amann et al. 1995). Decontamination methods employed to isolate mycobacteria have been shown to bias cultivation toward *Mycobacterium* species that are more resistant to chemicals such as sodium hydroxide and oxalic acid (Le Dantec et al. 2002b). To address these limitations a comprehensive amplicon-based pyrosequencing approach and multivariate analysis were employed to elucidate the biogeography of mycobacteria composition and diversity across a wide range of environments. Ethiopia was chosen as a suitable location to determine the biogeography of mycobacteria in the soil and water environments due to its varying terrain and diverse climatic environments.

5.3. Aims

i) To compare the community composition of *Mycobacterium* species between the nine different climatic regions and soil/water.

ii) To use multivariate approaches to determine the correlations between bio-climatic variables and spatial factors with *Mycobacterium* species diversity and relative abundance.

iv) To determine if whether the SG opportunistic pathogens present any biogeographical trends such as differences with temperature, elevation, moisture, pH, latitude and longitude.

5.4.1. Bio-climatic & spatial characteristics

Villages were sampled from across Ethiopia, from the latitudes 4°42'-12°46'N and longitudes 34°15 – 37°52'E (Section 2.2.1). A total of 42 villages were sampled from 9 regions including villages from elevation transects in 2 regions (Section 2.2.2). In addition to GIS data collected at each soil and water sampling point, recordings were taken for elevation, temperature, moisture and pH (Section 2.2.4 & Table 5.1). Pairwise correlations of bio-climatic and spatial variables for the soil samples demonstrated significant relationships between elevation and soil temperature (CC = -0.9), moisture (CC = 0.5) and longitude (CC = 0.6) (Table 5.2). Soil temperature was also significantly correlated with moisture (CC = -0.6) and longitude (CC = -0.7). Equivalent pairwise correlations for water samples showed that elevation was significantly correlated to water temperature (CC = -0.9) and longitude (CC = 0.7) (Table 5.3). Temperature was also negatively associated with longitude (CC = -0.2). The environmental and spatial pairwise correlations for soil and water samples were in general agreement.

Table 5.1. Descriptive statistics of environmental variables for soil and water samples used in the biogeographical analysis.

Variable	No. of villages	Mean	Median	Std. Dev	Min	Max
Soil elevation (m)	35	2058.88	2068	924.68	365.2	3939.8
Soil temperature (°C)	35	23.65	22.99	5.89	13.17	36.63
Soil moisture (%)	35	12.24	9.94	7.045	1.39	28.17
Soil pH	35	5.27	5.32	0.56	4.07	6.38
Water elevation (m)	42	1902.62	1909	956.51	370	3958
Water temperature (°C)	42	23.93	23.47	6.77	11.65	39.15
Water pH	42	4.63	4.66	0.54	2.95	5.61

Table 5.2. Pairwise correlations between the environmental variables recorded for soil samples, for the 35 villages which were included in the analysis of soil samples. The asterisk represents significant variables (P<0.05) which are Bonferroni corrected.

	Elevation (m)	pH	Temperature (˚C)	Moisture (%)	Latitude (˚N)	Longitude (˚E)
Elevation (m)	1.000					
pH	0.232	1.000				
Temperature(˚C)	**-0.900***	-0.169	1.000			
Moisture (%)	**0.484***	0.263	**-0.614***	1.000		
Latitude (˚N)	0.325	-0.201	-0.177	-0.431	1.000	
Longitude (˚E)	**0.602***	0.013	**-0.696***	0.437	0.301	1.000

Table 5.3. Pairwise correlations between the environmental variables recorded for water samples, for the 42 villages which were included in the analysis of water samples. The asterisk represents significant variables (P<0.05) which are Bonferroni corrected.

	Elevation (m)	pH	Temperature (˚C)	Latitude (˚N)	Longitude (˚E)
Elevation (m)	1.000				
pH	0.232	1.000			
Temperature(˚C)	**-0.924***	-0.090	1.000		
Latitude (˚N)	0.379	0.299	-0.269	1.000	
Longitude (˚E)	**0.683***	0.109	**-0.718***	0.333	1.000

5.4.2. Outcome variable definitions

Several outcome variables were explored for the biogeographical analysis. For regression models, the outcome variables tested were the fraction of OTU richness, Phylogenetic Diversity (PD) metric and the Shannon diversity index (H'). The fraction of OTU richness was defined as the fraction of different OTUs per sample over the total number of OTUs in all samples. The PD metric was measured as the sum of branch length from a weighted phylogenetic tree of representative sequences belonging to each sample (Faith 1992). The H' index was calculated from the number of OTUs present within each sample. For the construction of heatmaps the relative abundance of individual *Mycobacterium* species was calculated from the number of BLAST matches of a species as a proportion of the total BLAST matches for all sequences in that sample. Beta-diversity Canonical Correspondence Analysis (CCA) and Pearson correlation tests were based on the relative abundance of an OTU,

and this was calculated from the number of sequences belonging to an OTU as a proportion of the total number of sequences for that sample. Principle Component Analysis (PCA) was carried out using both unweighted and weighted analysis. Unweighted PCAs depict the variation in sites based the presence/absence of OTUs whereas the weighted PCAs take into account the relative abundance of OTUs. The term community composition refers to the prevalence of OTUs. All variables were tested for normal distribution, if the variable was not normally distributed it was transformed.

5.5. Biogeographical variation in *Mycobacterium* genus communities

Community compositions were dissimilar in all village samples: Of 522 *Mycobacterium* genus soil OTUs, only 18 OTUs (3.5%) were present in all samples and 81 OTUs (15.5%) were present in only one sample. It was shown that 2, 20, and 22 OTUs were present exclusively in Butajira, Gambella and Turmi respectively. This suggests some microbial endemism, however in these regions only one sample was of sufficient quality to include in the analysis, so it may be unique to the sample as opposed to the region. Of 436 OTUs relating to the water samples, none were found in all samples and 42 OTUs (9.6%) were detected in only one water sample. Only 7 OTUs were found to be exclusive to one region (Turmi). Overall it is difficult to prove endemism due to insufficient sampling effort.

The relative abundance of *Mycobacterium* species in soil samples was plotted from low to high elevation (Figure 5.1), as significant differences were observed between OTU composition and elevation (section 5.7). The abundance of *M. neglectum*, *M. moriokaense*, *M. kansasii* and *M. bohemicum* tended to be higher in samples from higher elevations compared to lower elevations. In contrast *M. insubricum*, *M. flavescens* and *M. novocastrense* were more abundant at lower elevations. Species

that were consistently ubiquitous in all soil samples included *M. houstonense*, *M. gadium* and *M. fallax*. The relative abundance of BLAST matches to opportunistic pathogens in the *Mycobacterium* genus dataset was not significantly different between regions or between environmental substrates.

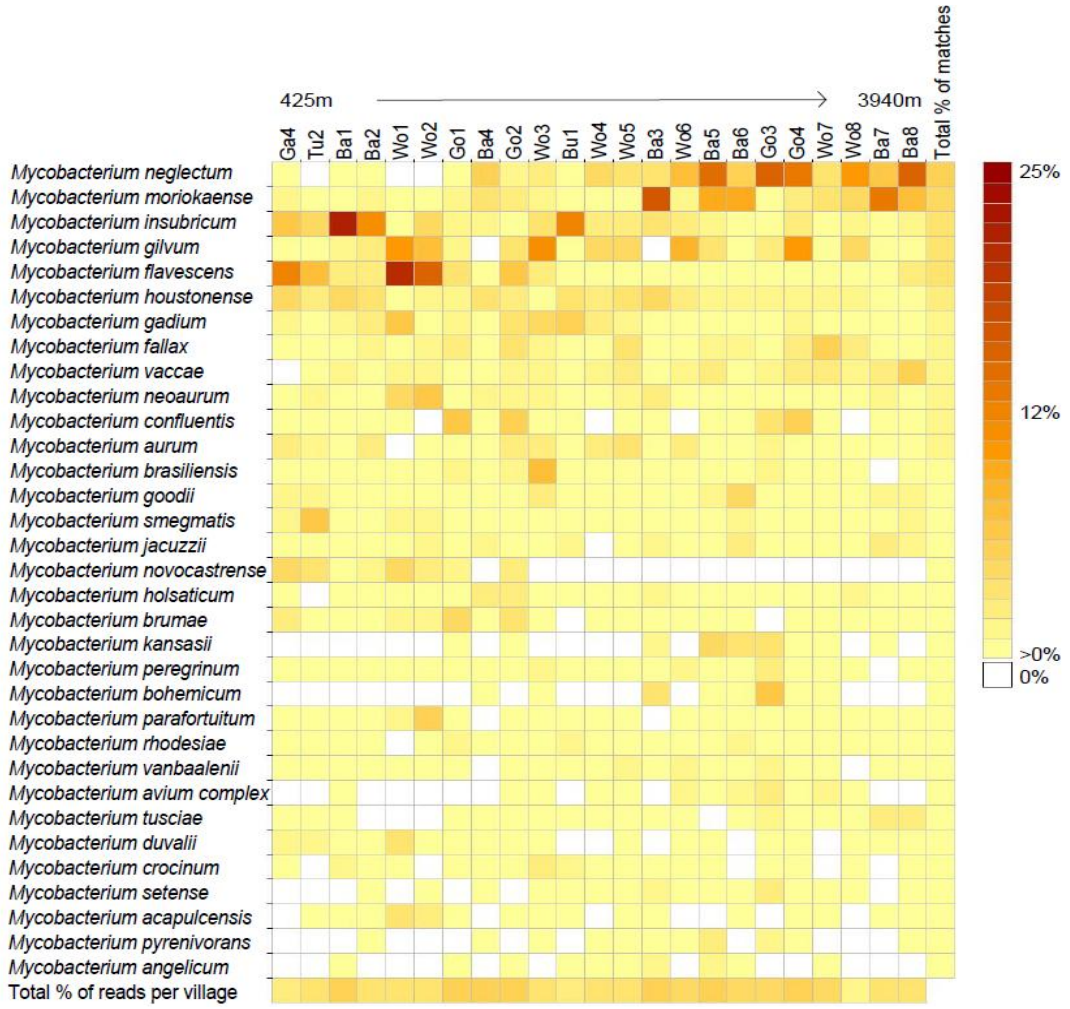


Figure 5.1. Heatmap of the relative abundance of *Mycobacterium* genus species in the soil samples as determined by the BLAST. *Mycobacterium* spp. includes the unidentified *Mycobacterium* species and the uncultured environmental species comprise 52.5% of all BLAST matches and are not included in the heatmap. The heatmap displays 90% (42.76%) of the remaining recognised abundant *Mycobacterium* species. Villages are ordered by elevation in metres.

Latitude was the only significant factor explaining variation in diversity and composition of *Mycobacterium* species in water samples (section 5.9). Differences in the numbers of *Mycobacterium* species were observed on a latitudinal gradient, and a

higher richness (i.e. numbers of different species) of *Mycobacterium* species was found in water samples collected from the south of Ethiopia compared to the north (Figure 5.2). Specifically, *M. gordonae* and *M. peregrinum* tended to be more relatively abundant in the north, whereas *M. neoaurum* appeared to be more prevalent in the South.

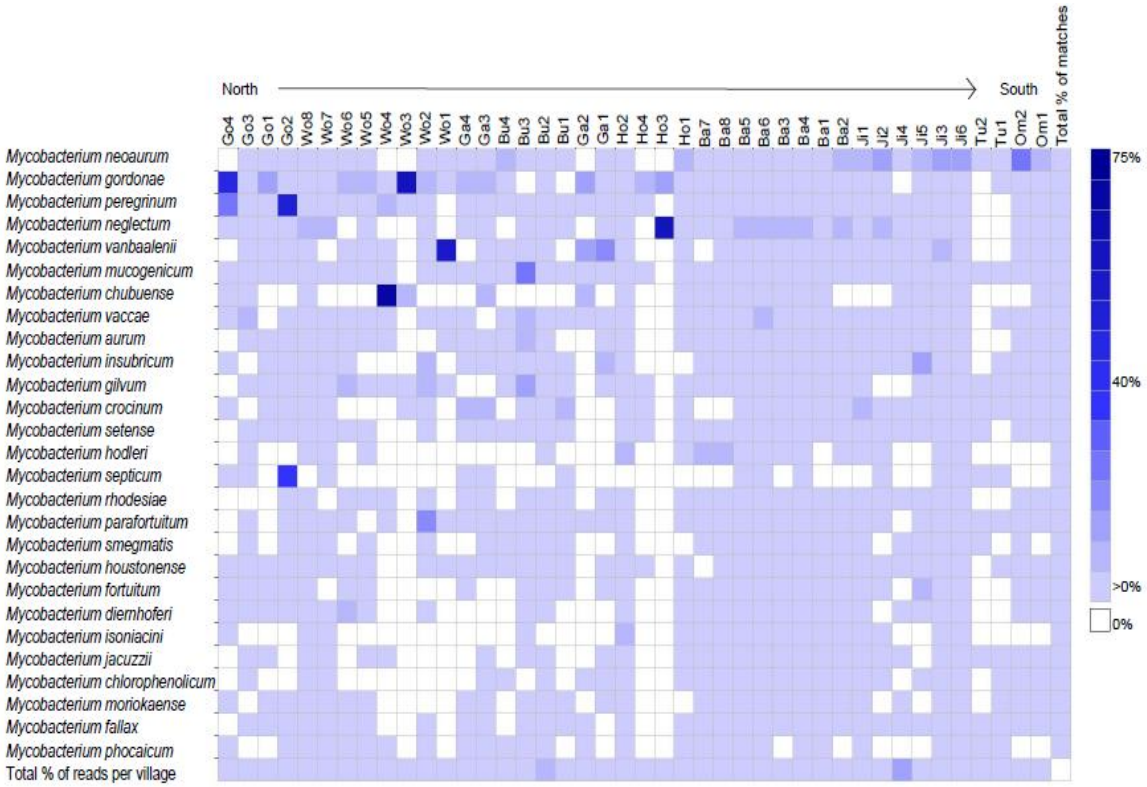


Figure 5.2. Heatmap of the relative abundance of *Mycobacterium* genus species in the water samples as determined by the BLAST. *Mycobacterium* spp. includes the unidentified *Mycobacterium* species and the uncultured environmental species comprise 63.21% of all BLAST matches and are not included in the heatmap. The heatmap displays 90% (33.11%) of the remaining recognised abundant *Mycobacterium* species, Villages are ordered by latitude.

Significant differences were also observed for additional environmental factors. For example OTUs BLAST matched to *M. smegmatis* ($r>0.7$, $P<0.04$) were significantly correlated with increased moisture in soil samples (Table 5.4). OTUs BLAST matched to *M.* sp. MI-5.1 P25 ($r>0.63$, $P<0.004$) and *M.* sp. CCBAU 25324 ($r>0.6$, $P<0.03$) were correlated with increased elevation of sites of water collection. The

species *M. gadium* (r = -0.6, *P* = 0.04) and *M. canariasense* (r = -0.6, *P* = 0.01) were correlated with decreased water temperature.

Table 5.4. OTU significance tests for the *Mycobacterium* genus dataset. Pairwise Pearson correlations coefficients and *P* values were calculated. Significant results after Bonferroni correction are shown ($P \leq 0.05$).

soil sample variables	OTU	Bonferroni corrected *P* value	R	Consensus Lineage
Elevation	87	**0.021**	-0.744	*Mycobacterium* sp.205
Temperature	493	**0.021**	0.744	*Mycobacterium* sp.G1368
Moisture	325	**0.036**	0.728	*Mycobacterium smegmatis*
Moisture	3	**0.005**	0.778	*Mycobacterium smegmatis*
Moisture	214	**0.041**	0.725	*Mycobacterium* sp.I5
Longitude	255	**0.005**	-0.781	*Mycobacterium* sp.GR-2009-164
Longitude	357	**0.076**	-0.705	*Mycobacterium insubricum*
water sample variables	**OTU**	**Bonferroni corrected *P* value**	**R**	**Consensus Lineage**
Elevation	275	**0.001**	0.667	*Mycobacterium* sp.MI-5.1P25
Elevation	233	**0.023**	0.579	*Mycobacterium isoniacini*
Elevation	245	**0.002**	0.645	*Mycobacterium* sp.CCBAU25324
Elevation	204	**0.016**	0.589	*Mycobacterium vaccae*
Elevation	77	**0.050**	0.557	*Mycobacterium* sp.A14N
Elevation	257	**0.027**	0.575	*Mycobacterium* sp.CCBAU25324
Elevation	4	**0.004**	0.626	*Mycobacterium hodleri*
Elevation	269	**0.003**	0.632	*Mycobacterium* sp.MI-5.1P25
Elevation	153	**0.012**	0.596	*Mycobacterium canariasense*
Temperature	275	**0.004**	-0.626	*Mycobacterium* sp.MI-5.1P25
Temperature	403	**0.051**	0.557	*Mycobacterium* sp.205
Temperature	371	**0.035**	-0.567	*Mycobacterium gadium*
Temperature	245	**0.041**	-0.563	*Mycobacterium* sp.CCBAU25324
Temperature	269	**0.037**	-0.566	*Mycobacterium* sp.MI-5.1P25
Temperature	153	**0.009**	-0.603	*Mycobacterium canariasense*
pH	364	**0.055**	-0.554	*Mycobacterium* sp.ATCC23070
Longitude	403	**0.003**	-0.635	*Mycobacterium* sp.205

5.6. Biogeographical variation in slow-growing mycobacteria communities

The SG dataset for soil samples had a total of 65 OTUs, of which 3 (4.6%) were found in all samples and 10 (15.4%) were detected in only one sample. Only two OTUs were found in exclusively in one region (Omorate). Similarly, the water

samples contained a total of 72 OTUs; only 1 OTU (1.4%) was found in all samples and 10 OTUs (13.9%) were found in one sample only. No OTUs were exclusively detected in one region, which suggests no microbial endemism. The variation of the relative abundance of BLAST matches to SG opportunistic pathogens was significantly higher between environmental substrates ($F(1,72) = 13.43$, $P < 0.001$), water samples (89.9%) had a higher mean prevalence of opportunistic pathogens compared to soil samples (76.2%). Regional variation in the relative abundance of SG opportunistic pathogens was not observed for soil ($F(7,24) = 1.6$, $P = 0.18$) or water samples ($F(8,33) = 2.1$, $P = 0.06$).

BLAST matches to all the sequences belonging to the SG soil dataset included 45 different species. Abundant species in all samples included *M. malmoense* (24.6%), *M. colombiense* (22.6%), *M. angelicum* (11.7%) and *M. riyadhense* (8.6%). The abundance of species was ordered along a longitudinal gradient, as this factor was shown to be significant for differences in OTU composition (section 5.8). The species *M. malmoense* and *M. colombiense* were more abundant in samples from the east of Ethiopia, and *M. aemonae, M. intracellulare* and *M. gastri* were more abundant in samples from the west (Figure 5.3). Several species were relatively abundant in the majority of samples including *M. asiaticum* and *M. simulans*.

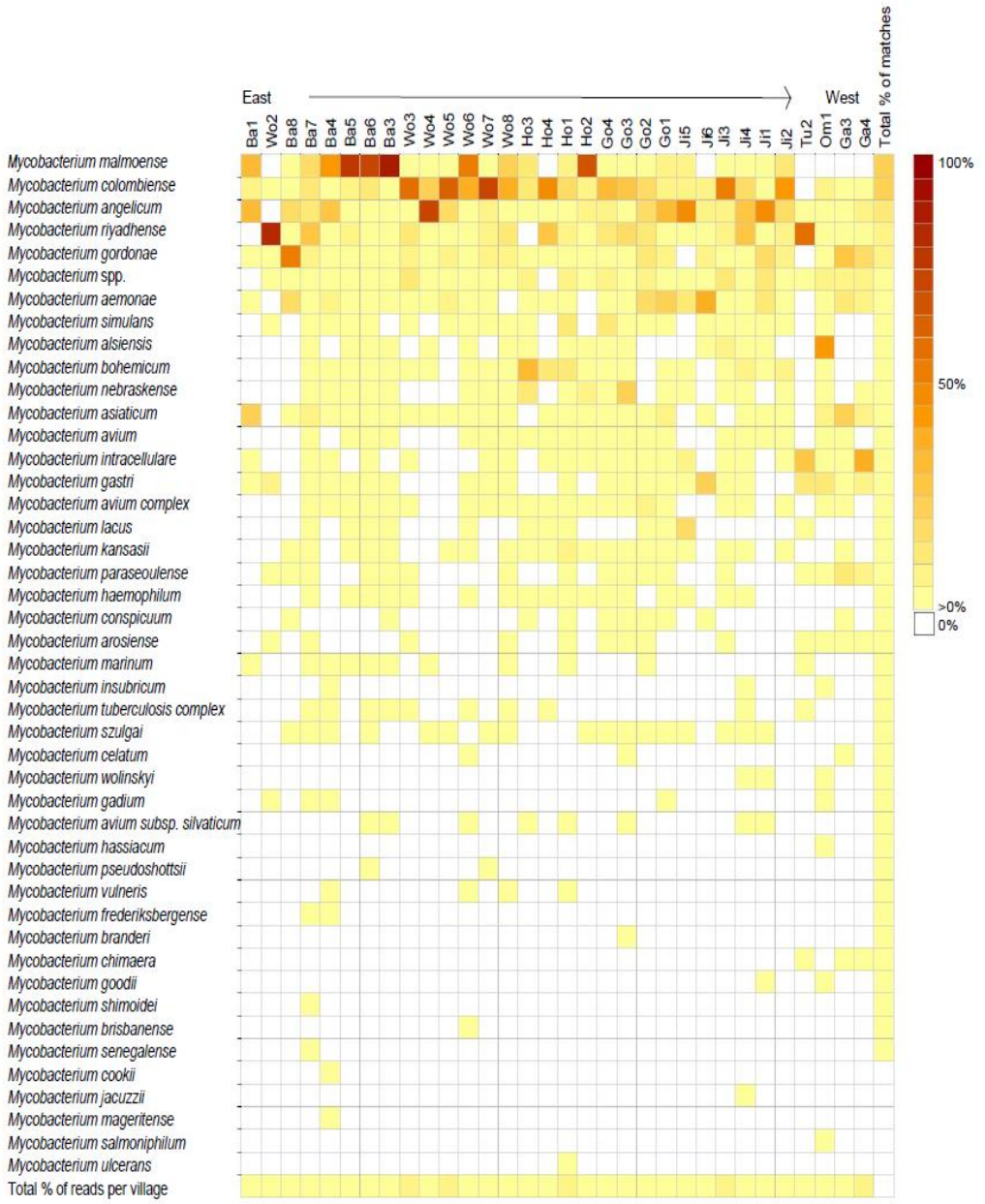


Figure 5.3. Heatmap of the relative abundance of SG in the soil samples as determined by BLAST. *Mycobacterium* spp. includes the unidentified *Mycobacterium* species and the environmental isolates. Villages are ordered by Longitude.

In the water samples there were 43 different SG identified from BLAST matches to individual sequences; the most abundant species included *M. gordonae* (45.1%), *M. colombiense* (12.8%), *M. riyadhense* (10.3%) and *M. malmoense* (5.6%). Variation in the relative abundance of species was observed along a latitudinal gradient (Figure 5.4), in agreement with compositional analysis which found latitude was a significant environmental factor (section 5.9); the highly abundant *M. gordonae* appears to be

most abundant in the northern parts of Ethiopia, whereas *M. colombiense* and *M. malmoense* are more abundant in the south. Central Ethiopia (Hossana and Butajira) had a higher relative abundance of *M. tuberculosis* complex and *M. asiaticum*.

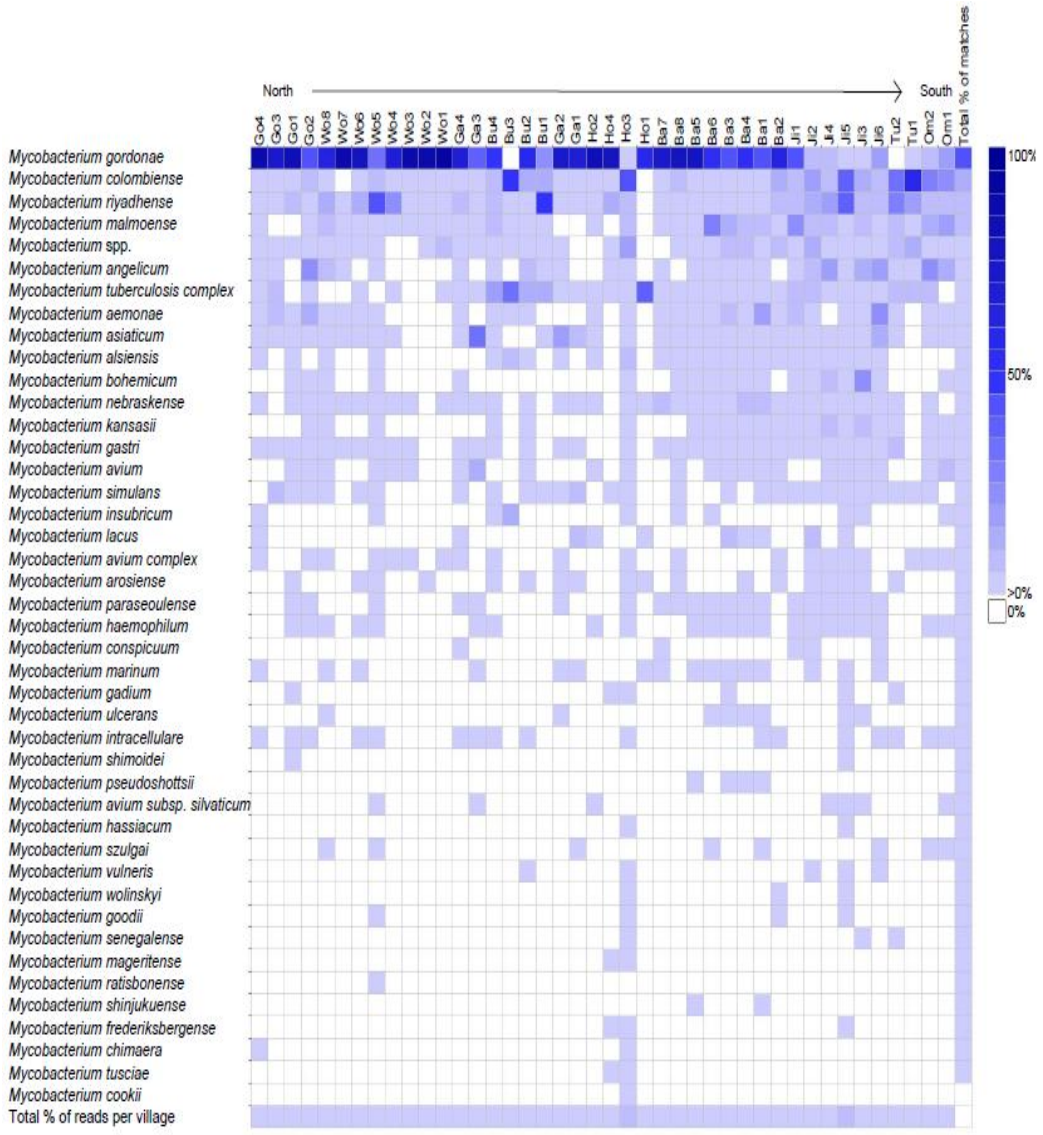


Figure 5.4. Heatmap of the relative abundance of SG in the water samples as determined by the BLAST. *Mycobacterium* spp. includes the unidentified *Mycobacterium* species and the environmental isolates. Villages are ordered by latitude.

Pathogenic SG associated with opportunistic infections in humans and animals were identified in the environmental samples, including *M. gordonae*, *M. kansasii*, *M. malmoense*, *M. avium* complex, *M. haemophilum*, *M. marinum* and *M. ulcerans*. The

relative abundance of OTUs and the relative abundance of BLAST matches to individual sequences were correlated with environmental and spatial factors (Figure 5.3, 5.4 & Table 5.5), highlighting the biogeographical component of exposure and infection risk. A pearson correlation analysis of opportunistic pathogens showed an OTU related to *M. riyadhense* to be significantly associated with increased soil temperature ($r = 0.6$, $P<0.05$). OTUs matched to *M. colombiense* were significantly more abundant in northern regions ($r = 0.6$, $P = 0.04$) and *M. asiaticum* was more prevalent in the western regions ($r = -0.6$, $P = 0.04$) (Table 5.5). Pearson correlations of the OTUs related to *M. gordonae* in water samples were associated with increased elevation ($r = 0.6$, $P<0.01$), decreased water temperature ($r = -0.7$, $P<0.005$) and were more abundant in western ($r = -0.6$, $P<0.01$) and northern regions ($r >0.6$, $P <0.001$) (Table 5.5), suggesting that *M. gordonae* is more prevalent in colder climates. In contrast, both *M. colombiense* ($r < -0.5$, $P <0.03$) and *M. malmoense* ($r = -0.5$, $P<0.05$) were found to be more abundant in southern compared to northern regions, and the former was also associated with acidic pH water conditions ($r < -0.6$, $P<.0.005$). An OTU related to *M. ulcerans* was associated with water samples collected from western areas ($r = -0.5$, $P = 0.05$). However, this OTU is also closely related to *M. marinum*, as *M. marinum* and *M. ulcerans* are identical in this region of the 16S rDNA gene, the species are synonymous in this analysis. This demonstrates that in some cases the 16S rRNA gene is not sufficient to resolve to the species level and additional genes should be employed to differentiate between these species.

Table 5.5. OTU significance tests for the SG. Pairwise Pearson correlations coefficients and *P* values were calculated. Significant results after Bonferroni correction are shown ($P \leq 0.05$).

Soil samples Variable	OTU	Bonferroni corrected *P* value	r	Consensus Lineage
Temperature	20	**0.049**	0.559	*Mycobacterium riyadhense*
Latitude	64	**0.036**	0.570	*Mycobacterium colombiense*
Longitude	58	**0.037**	-0.569	*Mycobacterium asiaticum*
Water samples Variables	**OTU**	**Bonferroni corrected *P* value**	**r**	**Consensus Lineage**
Elevation	58	**0.007**	0.560	*Mycobacterium gordonae*
Temperature	58	**0.000**	-0.652	*Mycobacterium gordonae*
pH	26	**0.000**	-0.632	*Mycobacterium colombiense*
pH	2	**0.000**	-0.637	*Mycobacterium colombiense*
Latitude	63	**0.001**	0.609	*Mycobacterium gordonae*
Latitude	26	**0.029**	-0.516	*Mycobacterium colombiense*
Latitude	0	**0.000**	0.673	*Mycobacterium gordonae*
Latitude	2	**0.011**	-0.547	*Mycobacterium colombiense*
Latitude	13	**0.000**	0.811	*Mycobacterium gordonae*
Latitude	11	**0.001**	0.623	*Mycobacterium gordonae*
Latitude	56	**0.047**	-0.500	*Mycobacterium malmoense*
Latitude	55	**0.008**	0.555	*Mycobacterium gordonae*
Longitude	9	**0.007**	-0.562	*Mycobacterium gordonae*
Longitude	71	**0.052**	-0.496	*Mycobacterium ulcerans*

5.7. Elevation and temperature explain *Mycobacterium* community biogeography in soils

The variation in soil *Mycobacterium* OTU community composition was tested for associations with elevation, soil temperature, pH, moisture, latitude and longitude. The Mantel tests demonstrated that, after correction of the *P* values, temperature and elevation were significant in explaining the variation in community composition in soils (Table 5.6). However, the Partial Mantel tests revealed that if elevation or temperature were controlled for, the other variable was no longer significant (Table 5.6), which indicates that these variables are closely correlated to each other. In concordance with this result the CCA of the *Mycobacterium* genus revealed elevation ($P = 0.001$) and temperature ($P = 0.001$) to be the only significant factors explaining

differences in soil community composition; these were negatively correlated with each other as indicated by the arrows pointing in opposite directions (Figure 5.5A). Soil community compositions in lower elevation/high temperature villages (elevation range: 1267-1406m) of the Bale (Ba1-2) and Woldiya (Wo1-2) transects (regions separated by approximately 600 km) clustered together (92 shared OTUs), as opposed to clustering with the respective higher elevation/low temperature villages (range 3518-3940m) within the same region Bale (Ba7-8) (64 shared OTUs) and Woldiya (Wo7-8) (43 shared OTUs). This supports the analyses above that climatic conditions can have a stronger influence on mycobacteria community structures than geographical proximity. Although the CCA of the randomly resampled dataset (1027 sequences per sample) did not show any environmental factors to be significant, the separation of high elevation/low temperature villages from low elevation/high temperature villages was still observed (Figure 5.5B). Similarly, the complimentary PCoA unweighted analysis of phylogenetic differences between samples displayed a gradient with elevation (Figure 5.6A). The gradient along the P1 axis component explained 22.1% of the variance from low to high elevations. This gradient was also observed for temperature (Figure 5.7A). The weighted PCoA analysis demonstrated a scattered correlation with elevation and temperature (Figure 5.6B & 5.7.B), which implies that these variables are equally likely to affect the abundance of OTUs. Consistent result shown by gradients with elevation and temperature were observed from the unweighted PcoA of the resampled data with the minimum number of sequences (1027) per sample (Fig 5.6C & 5.7C). Overall, complementary analyses suggest that elevation and temperature are significant for *Mycobacterium* species composition in soil samples. The fact that this trend was consistently observed throughout different analyses suggests that the elevation/temperature trend is real.

Table 5.6. Mantel and Partial Mantel tests for the *Mycobacterium* genus soil dataset comprising of 23 soil samples and the SG soil dataset comprising of 32 soil samples. Asterisks represent the relationships that were significant at the $P \leq 0.05$ level. *P* values are Bonferroni corrected (P value multiplied by the number of tests: 6)

Environmental variables		*Mycobacterium* genus dataset			SG dataset		
Effect of:	Controlling for :	R	*P* value	Corrected *P* value	R	*P* value	Corrected *P* value
Temperature	-	0.401	**0.001***	**0.006***	-0.021	0.608	3.648
Temperature	Elevation	0.192	**0.027***	**0.162**	-0.089	0.876	5.256
Temperature	pH	0.407	**0.001***	**0.006***	-0.022	0.589	3.534
Temperature	Moisture	0.403	**0.001***	**0.006***	-0.036	0.661	3.996
Temperature	Longitude	0.389	**0.001***	**0.006***	-0.115	0.960	5.760
Temperature	Latitude	0.397	**0.001***	**0.006***	-0.022	0.591	3.546
Elevation	-	0.392	**0.005***	**0.030***	0.086	0.209	1.254
Elevation	Temperature	0.17	0.118	0.708	0.121	0.123	0.738
Elevation	pH	0.395	**0.005***	**0.030***	0.089	0.196	1.176
Elevation	Moisture	0.393	**0.002***	**0.012***	0.077	0.174	1.044
Elevation	Longitude	0.398	**0.001***	**0.006***	-0.036	0.656	3.936
Elevation	Latitude	0.388	**0.010***	0.060	0.083	0.200	1.200
pH	-	0.072	0.259	1.554	-0.041	0.668	4.008
pH	Temperature	0.104	0.206	0.236	-0.042	0.671	4.026
pH	Elevation	0.089	0.221	1.326	-0.048	0.684	4.104
pH	Moisture	0.076	0.243	1.458	-0.045	0.671	4. 026
pH	Longitude	0.081	0.239	1.434	-0.058	0.755	4.530
pH	Latitude	0.076	0.273	1.638	-0.043	0.987	5.922
Moisture	-	0.081	0.215	1.290	0.040	0.291	1.746
Moisture	Temperature	-0.093	0.783	4.698	0.049	0.258	1.548
Moisture	Elevation	-0.088	0.769	4.614	0.009	0.425	2.550
Moisture	pH	0.084	0.242	1.452	0.044	0.288	1.728
Moisture	Longitude	0.034	0.37	2.220	-0.045	0.721	4.326
Moisture	Latitude	0.017	0.388	2.328	0.039	0.331	1.986
Longitude	-	0.114	0.179	1.074	0.199	**0.011***	0.066
Longitude	Temperature	-0.046	0.545	3.270	0.228	**0.004***	**0.024***
Longitude	Elevation	-0.137	0.901	5.406	0.184	**0.004***	**0.024***
Longitude	pH	0.120	0.194	1.164	0.203	**0.005***	**0.030***
Longitude	Moisture	0.087	0.263	1.578	0.200	**0.004***	**0.024***
Longitude	Latitude	0.094	0.227	1.362	0.197	**0.010***	0.060
Latitude	-	0.118	**0.022***	0.132	0.030	0.279	1.674
Latitude	Temperature	0.168	**0.024***	0.144	0.030	0.281	1.686
Latitude	Elevation	0.167	**0.017***	0.102	0.019	0.342	2.052
Latitude	pH	0.180	**0.012***	0.072	0.032	0.248	1.488
Latitude	Moisture	0.160	**0.039***	0.234	0.102	0.409	2.545
Latitude	Longitude	0.167	**0.015***	0.090	0.028	0.317	1.902

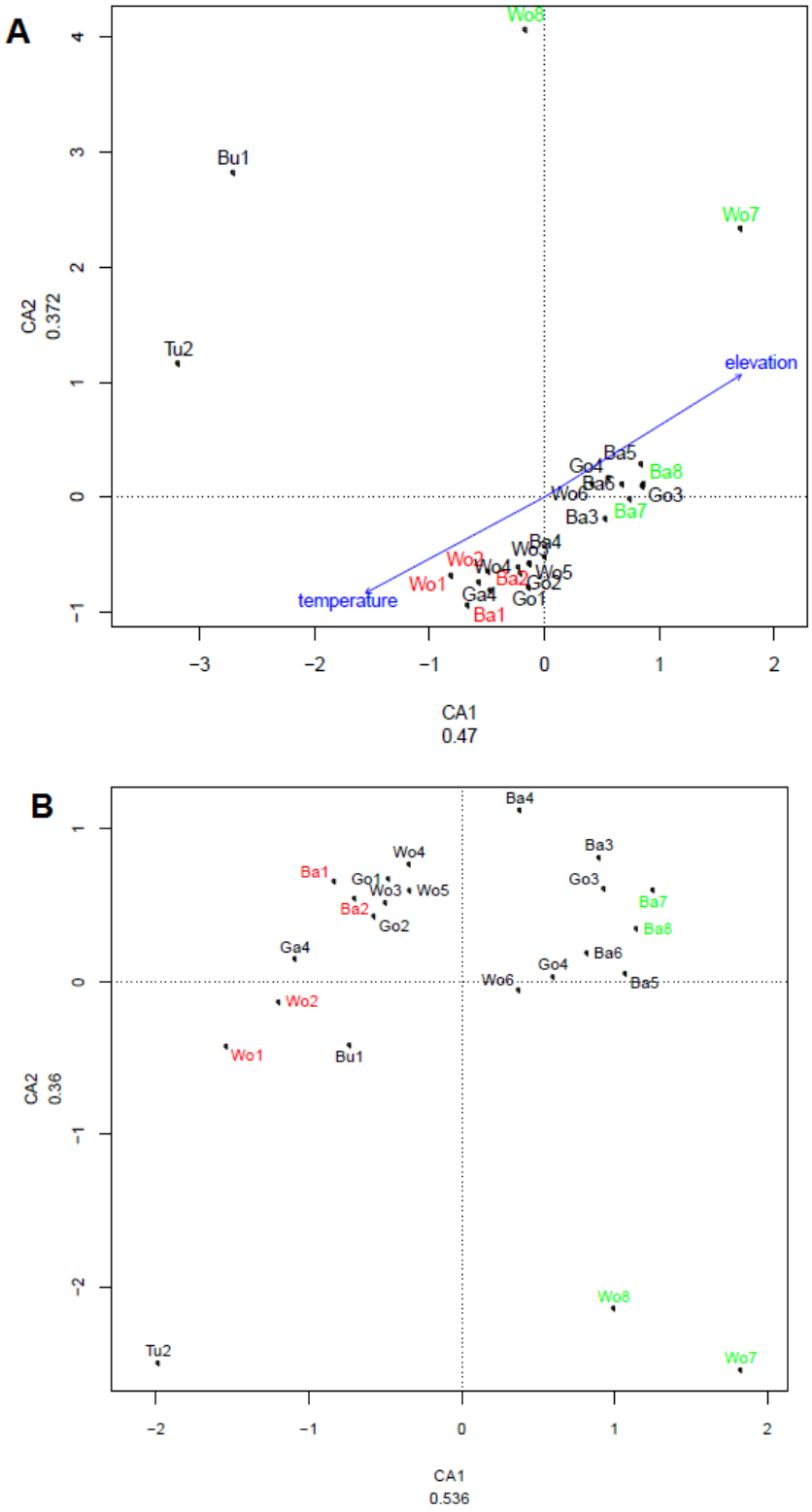


Figure 5.5. CCA plots of the *Mycobacterium* genus soil dataset comprising of 23 samples. Villages are denoted by the abbreviation of village name and number. **(A)** CCA plot of all sequences. The variation explained by temperature (r2 = 0.73, *P* = 0.001) and elevation (r2 = 0.56, *P* = 0.001) are shown by the arrows. (**B**) CCA plots of a random resample of 1027 sequences per sample. The samples highlighted in red are low elevation villages and samples highlighted in green are high elevation villages.

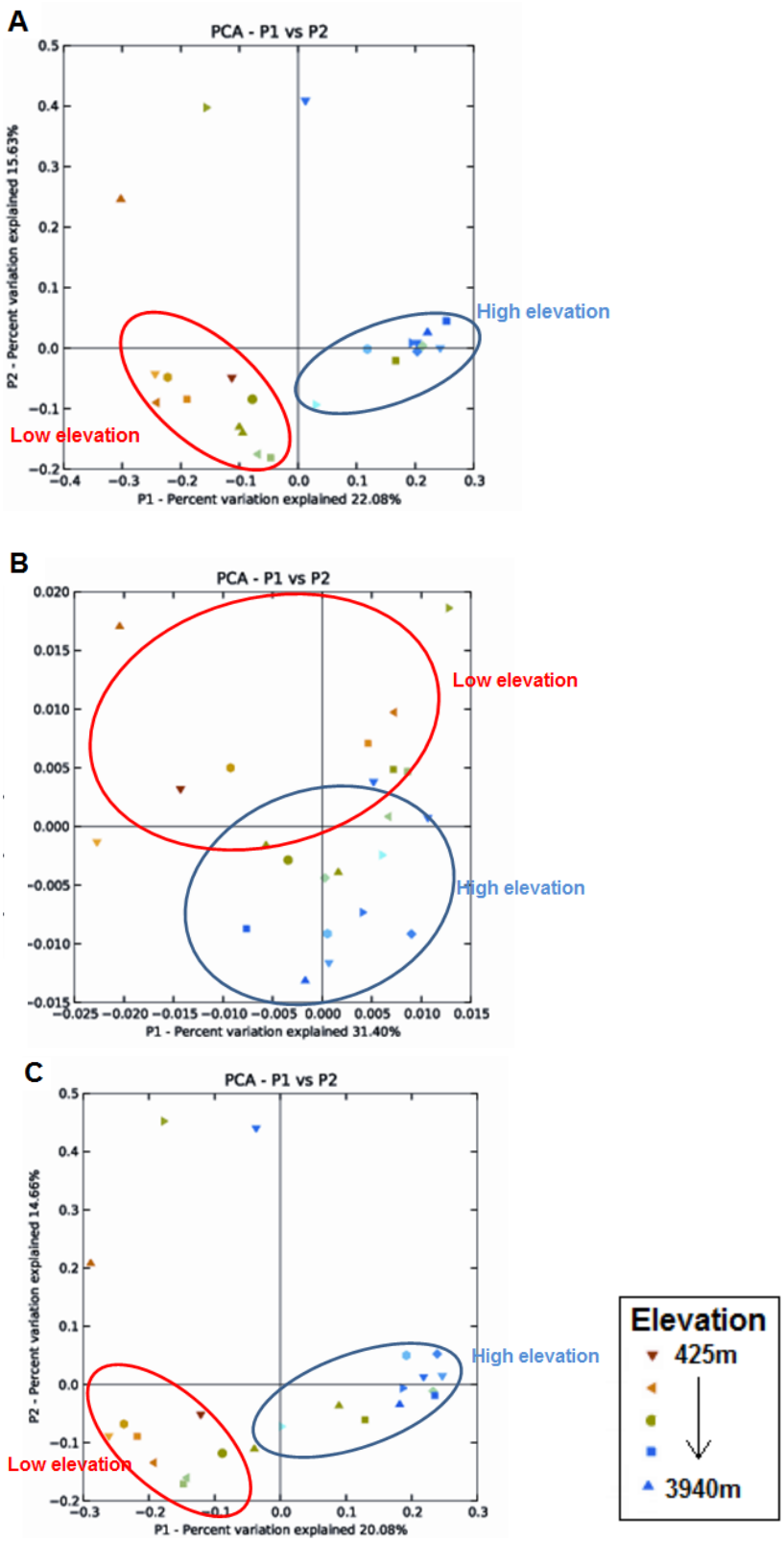


Figure 5.6. PCoA plots of the *Mycobacterium* genus soil dataset comprising of 23 soils. Sample points are coloured by **elevation**. The gradient from red to blue represents low elevations to high elevations. **(A)** Unweighted analysis of all sequences **(B)** Weighted analysis of all sequences **(C)** Unweighted analysis of a random resample (1027 sequences for each sample)

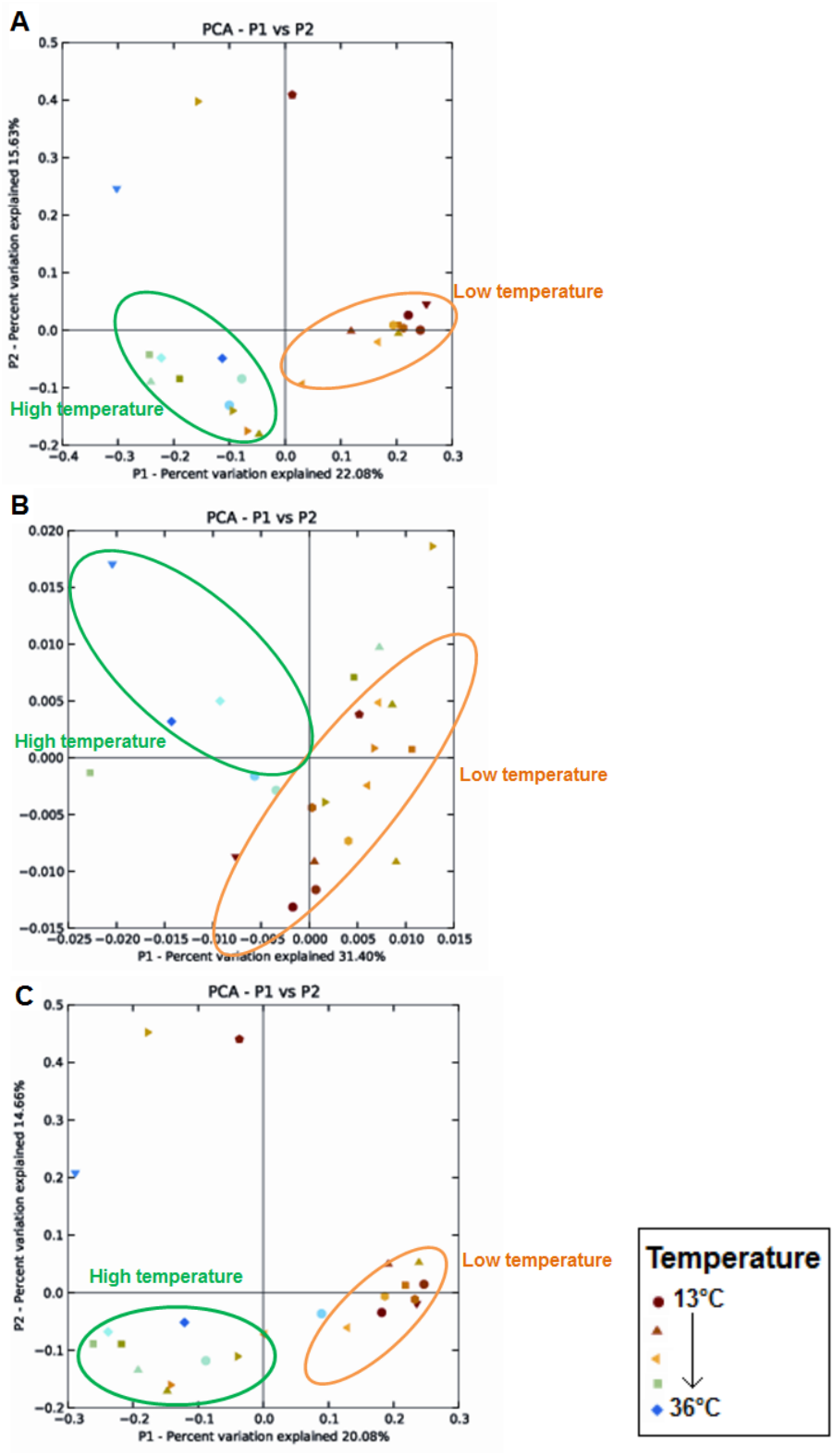


Figure 5.7. PCoA plots of the *Mycobacterium* genus soil dataset comprising of 23 soils. Sample points are coloured by **temperature**. The gradient from red to blue represents low to high temperatures. **(A)** Unweighted analysis of all sequences **(B)** Weighted analysis of all sequences **(C)** Unweighted analysis of a random resample (1027 sequences for each sample)

Three diversity and richness metrics were tested for correlations to bioclimatic and spatial variables to give an indication of changes in community structure in environmental samples. These metrics were the fraction of OTU richness, the phylogenetic diversity metric (PD) and the Shannon index (H'). Initially the variables were tested to determine whether there were significant differences between regions and between environmental sources compared to within regions.

The variation of the Shannon index was significantly higher between environmental sources compared to within environmental sources ($F(1,63) = 15.5$, $P = 0.0002$), soil samples had a higher mean (4.1) compared to water (3.4). Significantly greater variation was observed between regions for water samples ($F(8,33) = 3.3$, $P = 0.01$) but not soil samples ($F(5,17) = 0.9$, $P = 0.51$). The variation of the fraction of OTU richness was not significantly different between environmental sources ($F(1,63) = 2.3$, $P = 0.14$). However, there was a greater variation between regions for water samples compared to within regions ($F(8,33) = 5.25$, $P<0.005$) but this was not observed for soil samples. There was a greater variation between environmental sources for the PD metric compared to within sources; ($F(1,63) = 115.0$, $P<0.005$). The mean was higher for soil (1.8) compared to water (1.0). Greater variation was observed between regions compared to within regions for soil ($F(5,17) = 3.23$, $P = 0.03$) and water ($F(8,33) = 2.9$, $P = 0.01$).

Moisture was the only significant variable explaining differences in the Shannon diversity index using the generalised linear model (CC = -0.01, $P = 0.047$), however it only explained a small percentage of the variation (8%) (Table 5.7). Linear and quadratic regression models were compared for bioclimatic factors and the three outcome measures, and the model with the lowest Akakie's Information Criteria (AIC) value was chosen. A non-linear relationship was observed between elevation

and each of the three measures for richness and diversity of *Mycobacterium* species in soil over a range of 425m - 3940m, with peaks at mid-elevations (c. 2000-3000m) (Fraction of OTU richness R^2 = 0.4; phylogenetic diversity R^2 = 0.3; Shannon diversity R^2 = 0.2) (Figure 5.8A-C). For example, the Bale transect showed a peak in the fraction of OTU richness at mid elevation villages Ba3-6 (mean 0.4), compared to the higher elevation villages of Ba7-Ba8 and lower elevation villages Ba1-2 (mean 0.3). A similar trend was observed for the Woldiya transect; mid elevation villages Wo3-6 displayed a peak in the fraction of OTU richness (mean 0.39) compared to the higher elevation villages (Wo7-8) and lower elevation villages (Wo1-2) (mean 0.35). Temperature also displayed a non-linear relationship with the three metrics, however the models explained a lower degree of variance (R^2 values) for the fraction of OTU richness, the PD metric and the Shannon diversity index (0.2, 0.2 and 0.1 respectively) (Figure 5.9A-C). In summary, the diversity and richness measures are highest at the intermediate elevation and temperature points, perhaps because mycobacteria can more readily proliferate in less extreme environments. However, it is important to note that this is marginally significant.

Table 5.7. GLM for the Shannon diversity index (H') for the *Mycobacterium* genus soil dataset of 23 soil samples (pseudo R^2 = 0.08).

Variables	Range	CC (95% CI)	*P* value
Elevation (m)	365.5-3939.8	-0.0001 (-0.0002-0.0001)	0.536
Temperature (°C)	13.17-36.63	0.004 (-0.015-0.023)	0.703
Moisture (%)	1.39-28.17	-0.012 (-0.023--0.0002)	**0.047***
pH	4.07-6.38	-0.007 (-0.0278-0.265)	0.962
Latitude (°N)	5.030583-12.76961	0.014 (-0.028-0.056)	0.508
Longitude (°E)	34.263-39.87669	-0.054 (-0.120-0.013)	0.114
Elevation^2 (m)		-1.57E-08 (-5.03E-08 - 1.88E-08)	0.372
Temperature^2 (°C)		0.00006 (-0.0003-0.0004)	0.746
Moisture^2 (%)		-0.0004(-0.001-0.000)	**0.047***

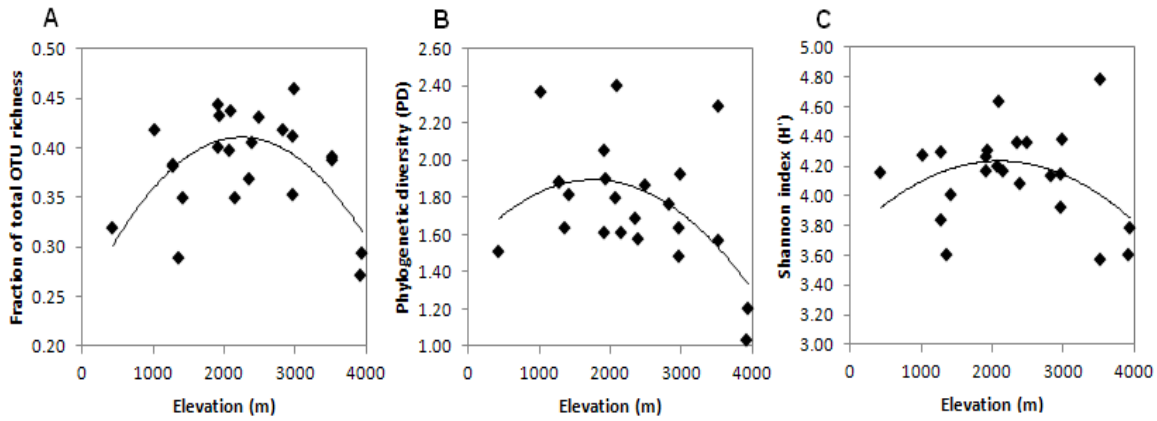


Figure 5.8. Quadratic relationships between the diversity of *Mycobacterium* genus soil dataset (23 samples) and Elevation (**A**) The fraction of OTU richness, data points represent the number of different OTUs per sample divided by the total number of different OTUs for all samples (R^2 value = 0.41). (**B**) Phylogenetic Diversity (PD) metric takes into account the fraction of total branch length for each sample (R^2 value = 0.27). **(C)** Shannon index takes into account the relative abundance of each OTU (R^2 = 0.15). Models were chosen based on the lowest AIC values.

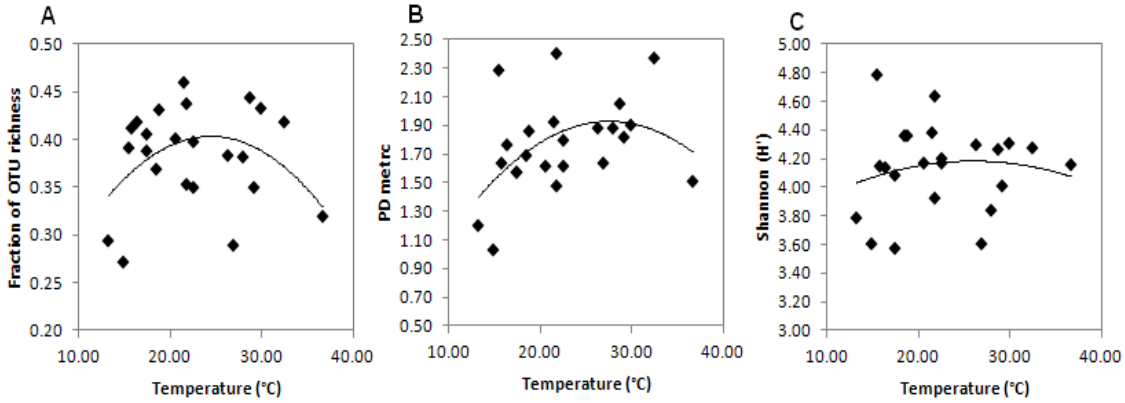


Figure 5.9. Quadratic relationships between the diversity of *Mycobacterium* genus soil dataset (23 samples) and Temperature **(A)** The fraction of OTU richness, data points represent the number of different OTUs per sample divided by the total number of different OTUs for all samples (R^2 value = 0.16). (**B**) Phylogenetic Diversity (PD) metric takes into account the fraction of total branch length for each sample (R^2 value = 0.23). **(C)** Shannon index takes into account the relative abundance of each OTU (R^2 = 0.02). Models were chosen based on the lowest AIC values.

5.8. Combination of bio-climatic and spatial factors were responsible for slow-growing mycobacteria community in soils

SG community composition in soil samples was correlated to bioclimatic factors using Mantel and Partial Mantel tests. Only longitude was found to be significant and after Bonferroni correction and was only significant after controlling for

environmental factors (Table. 5.6). CCA analyses of only the SG composition in soil samples demonstrated the significance of longitude ($P = 0.002$), moisture ($P = 0.004$) and temperature ($P = 0.012$) (Figure 5.10A). This suggests that a multifactorial process determines the structure of SG communities in soil. However as these factors have been shown to be correlated with each other (Table 5.2) it is difficult to determine the relative roles of these factors. The CCA of the randomly resampled dataset of 514 sequences for each sample also demonstrated the significance of longitude, moisture and temperature (Figure 5.10B). Longitude was consistently the most significant factor explaining the composition of SG, in agreement with the Mantel tests. PCoA of the phylogenetic similarities of SG failed to show any discrete patterns with bioclimatic factors. However a weak trend was observed with longitude, with some clustering of samples from the west (Figure 5.11) and particularly in the unweighted analysis (Figure 5.11 A&C). Overall longitude was the variable most strongly correlated with the composition of SG, perhaps suggesting that spatial factors are important for this group or that underlying environmental factors correlated with longitude influence this community.

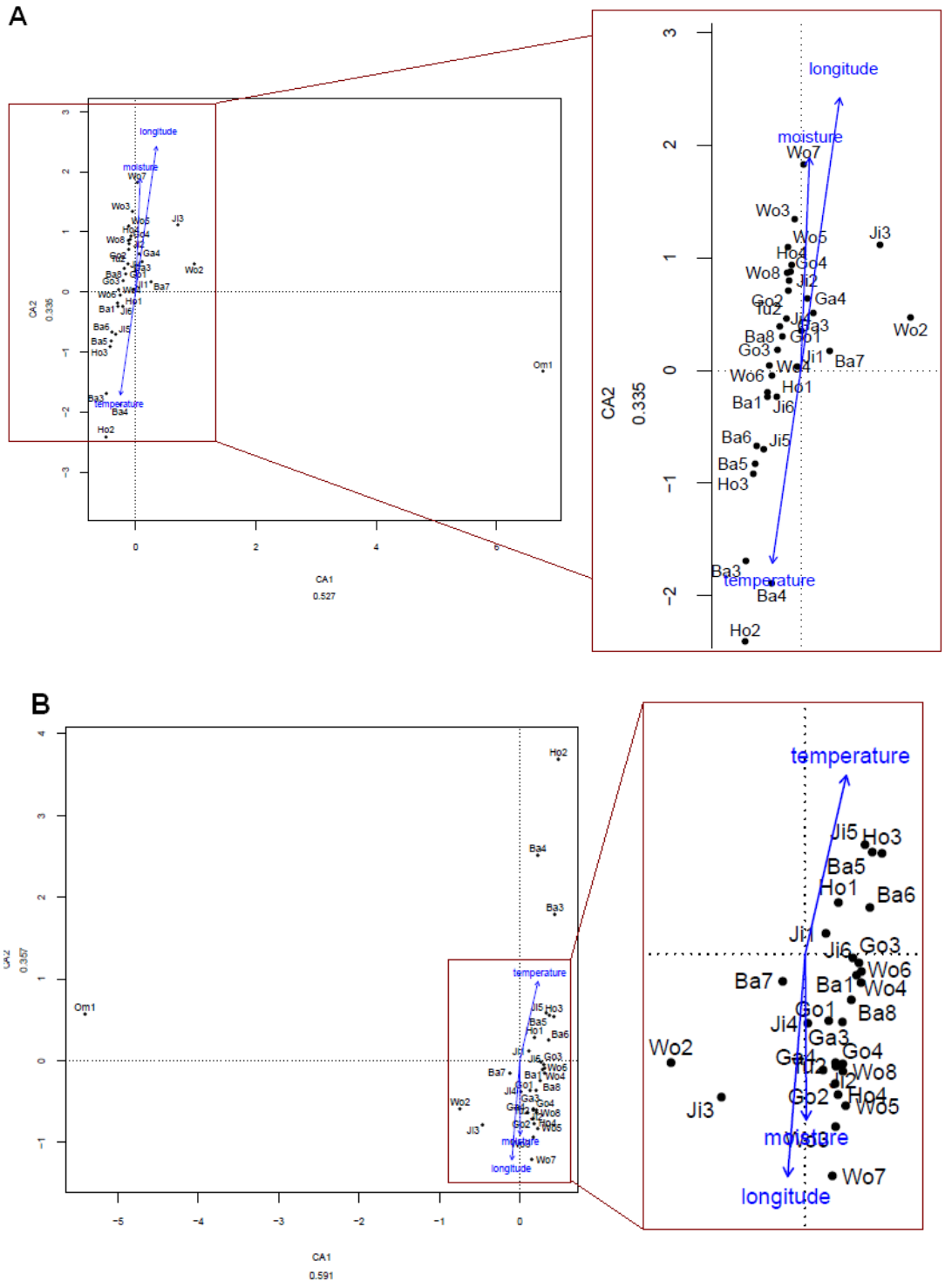


Figure 5.10. CCA plots of the SG soil dataset comprising of 32 samples. Villages are denoted by the abbreviation of village name and number. **(A)** CCA plot of all sequences. The variation explained by longitude (R^2 = 0.59, *P* = 0.001), moisture (R^2 = 0.35, *P* = 0.007) and temperature (R^2 = 0.30, *P* = 0.011) are shown by the arrows. (**B**) CCA plots of a random resample of 514 sequences per sample, the variation explained by longitude (R^2 = 0.36, *P* = 0.002), moisture (R^2 = 0.20, *P* = 0.050) and temperature (R^2 = 0.24, *P* = 0.011).

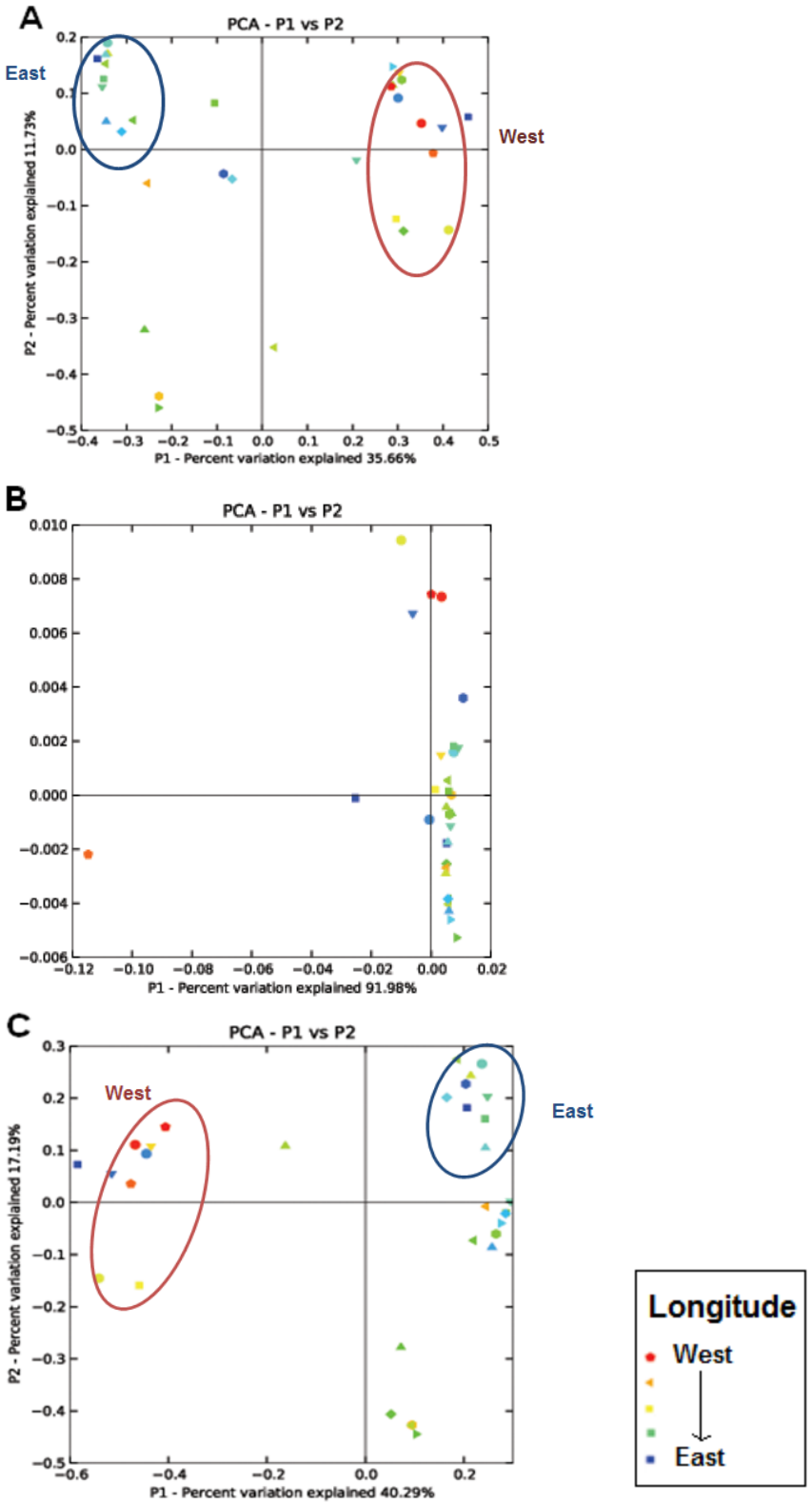


Figure 5.11. PCoA plots of the SG soil dataset comprising of 32 soils. Sample points are coloured by **longitude**, the gradient from red to blue represent west to east. **(A)** Unweighted analysis of all sequences **(B)** Weighted analysis of all sequences **(C)** Unweighted analysis of a random resample (514 sequences for each sample).

The diversity and richness metrics were also correlated with SG present in soil and water samples. Longitude was the only factor correlated to the differences in the Shannon diversity index (CC = -0.079, *P* = 0.007) (Table 5.8). Therefore longitude was shown to be significant to explain differences in both the composition and diversity (H') of SG. The PD metric and the fraction of OTU richness were not significantly correlated with bioclimatic and spatial factors. The diversity and richness estimates were also tested to evaluate if there were any significant differences between administrative regions or between environmental substrates.

The variation of the Shannon index for the SG was not significantly greater between environmental substrates compared to within substrates ($F(1,72) = 0.03$, $P=0.87$). The mean highest Shannon estimates was in soil samples (1.8) compared to water samples (1.9). High variation was also observed between regions compared to within regions for soil ($F(7,24) = 2.6$, $P = 0.04$) and water ($F(8,33) = 4.8$, $P<0.005$). No significant differences were observed for the fraction of OTUs and the PD metric between regions or between environmental sources.

Table 5.8. GLM for the Shannon diversity estimate (H') for the SG soil dataset of 32 soil samples. (Pseudo $R^2 = 0.19$)

Variables	Range	CC (95% CI)	*P* value
Elevation (m)	365.2-3939.8	-0.00003 (-0.00018-0.00012)	0.686
Temperature(°C)	13.17- 36.63	0.008 (-0.013-0.028)	0.451
Moisture (%)	1.39-28.17	-0.003 (-0.018-0.012)	0.727
pH	4.07-6.38	0.101(-0.069-0.272)	0.243
Latitude (°N)	4.705306-12.76961	-0.002 (-0.038-0.035)	0.922
Longitude (°S)	34.263-39.85836	-0.079 (-0.136--0.022)	**0.007***

5.9. Latitudinal gradient observed for community compositions and diversity estimates in water

The *Mycobacterium* species community composition in water samples demonstrated significant variation with latitude (r = 0.2, P = 0.005) even when other bioclimatic variables were controlled for (Table 5.9). In concordance with this finding, the CCAs of differences in OTU composition demonstrated that latitude (P = 0.043) was the only significant factor explaining variation (Figure 5.12A) and this remained the case after data was randomly resampled to 385 sequences per sample (Figure 5.12B). Differences in phylogenetic relatedness between villages illustrated latitudinal gradient across principal component 1, which explained 21.98% of the variance (Figure 5.13A). This gradient was also observed in the unweighted PCoA of the randomly resampled data (385 sequences per sample), explaining 17.27% of the variance (Figure 5.13C). While weighted analysis did not demonstrate a latitudinal gradient, the southern latitude villages did cluster together (Figure 5.13B). Spatial variation was the only significant factor for differences in the Shannon diversity index for water samples, as samples from the south were more diverse (CC = -0.1, P = 0.022) (Table 5.10); however this did not explain much of the variation R^2 = 0.1 (Figure 5.14A). The additional diversity estimates also displayed a linear relationship, however only weak correlations were found (Figure 5.14B&C). Latitude was consistently the only factor to explain the variation in *Mycobacterium* species composition and diversity in water samples, perhaps suggesting mechanisms of dispersal limitation and no apparent influence of environmental factors.

Table 5.9. Mantel and Partial Mantel tests for the *Mycobacterium* genus water dataset and the SG water dataset both comprising of 42 water samples. Asterisks represent the relationships that were significant at the $P \leq 0.05$ level. *P* values are Bonferroni corrected (*P* value multiplied by the number of tests: 5).

Environmental variables		*Mycobacterium* genus dataset			SG dataset		
Effect of:	Controlling for :	r	*P* value	Corrected P value	R	*P* value	Corrected *P* value
Temperature	-	0.051	0.274	1.370	-0.069	0.813	4.065
Temperature	elevation	0.020	0.375	1.875	-0.012	0.944	4.720
Temperature	pH	0.046	0.280	1.400	-0.08	0.837	4.185
Temperature	longitude	0.027	0.360	1.800	-0.121	0.97	4.850
Temperature	latitude	0.049	0.260	1.300	-0.073	0.823	4.115
Elevation	-	0.057	0.250	1.250	0.045	0.271	1.355
Elevation	temperature	0.032	0.334	1.670	0.113	0.089	0.445
Elevation	pH	0.041	0.298	1.490	0.017	0.385	1.925
Elevation	longitude	0.020	0.393	1.965	-0.029	0.645	3.225
Elevation	latitude	0.033	0.333	1.665	0.023	0.348	1.740
pH	-	0.120	0.143	0.715	0.207	**0.021***	0.105
pH	temperature	0.118	0.132	0.660	0.211	**0.020***	0.100
pH	elevation	0.113	0.142	0.710	0.203	**0.020***	0.100
pH	Longitude	0.115	0.145	0.725	0.201	**0.026***	0.130
pH	Latitude	0.094	0.173	0.865	0.186	**0.037***	0.185
Longitude	-	0.068	0.188	0.940	0.11	0.065	0.325
Longitude	Temperature	0.052	0.278	1.390	0.148	0.029*	0.145
Longitude	Elevation	0.041	0.244	1.220	0.104	0.038*	0.190
Longitude	pH	0.059	0.216	1.080	0.097	0.094	0.470
Longitude	Latitude	0.051	0.249	1.245	0.095	0.111	0.555
Latitude	-	0.204	**0.001***	**0.005***	0.188	**0.001***	**0.005***
Latitude	Temperature	0.204	**0.003***	**0.015***	0.189	**0.006***	**0.030***
Latitude	Elevation	0.199	**0.001***	**0.005***	0.184	**0.006***	**0.030***
Latitude	pH	0.190	**0.005***	**0.025***	0.163	**0.009***	**0.045***
Latitude	Longitude	0.119	**0.004***	**0.020***	0.18	**0.004***	**0.020***

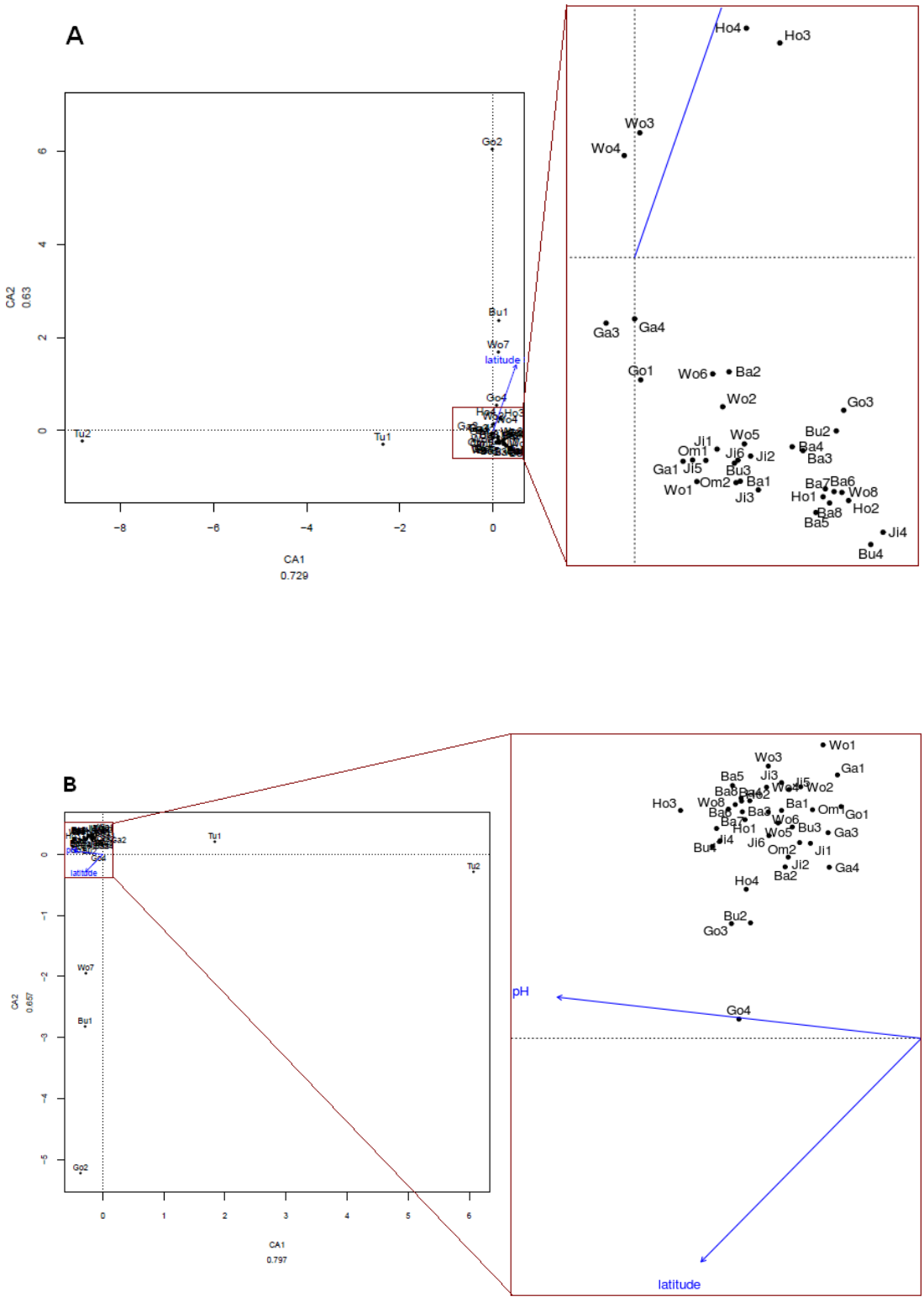


Figure 5.12. CCA plots of the *Mycobacterium* genus water dataset comprising of 42 water samples. Villages are denoted by the abbreviation of village name and number. **(A)** CCA plot of all sequences, the variation was explained by latitude (R^2 = 0.18, P = 0.001) is shown by the arrow. (**B**) CCA plot of a random resample (385 sequences per sample), the variation was explained by pH (R^2 = 0.19, P = 0.036) and latitude (R^2 = 0.13, P = 0.039).

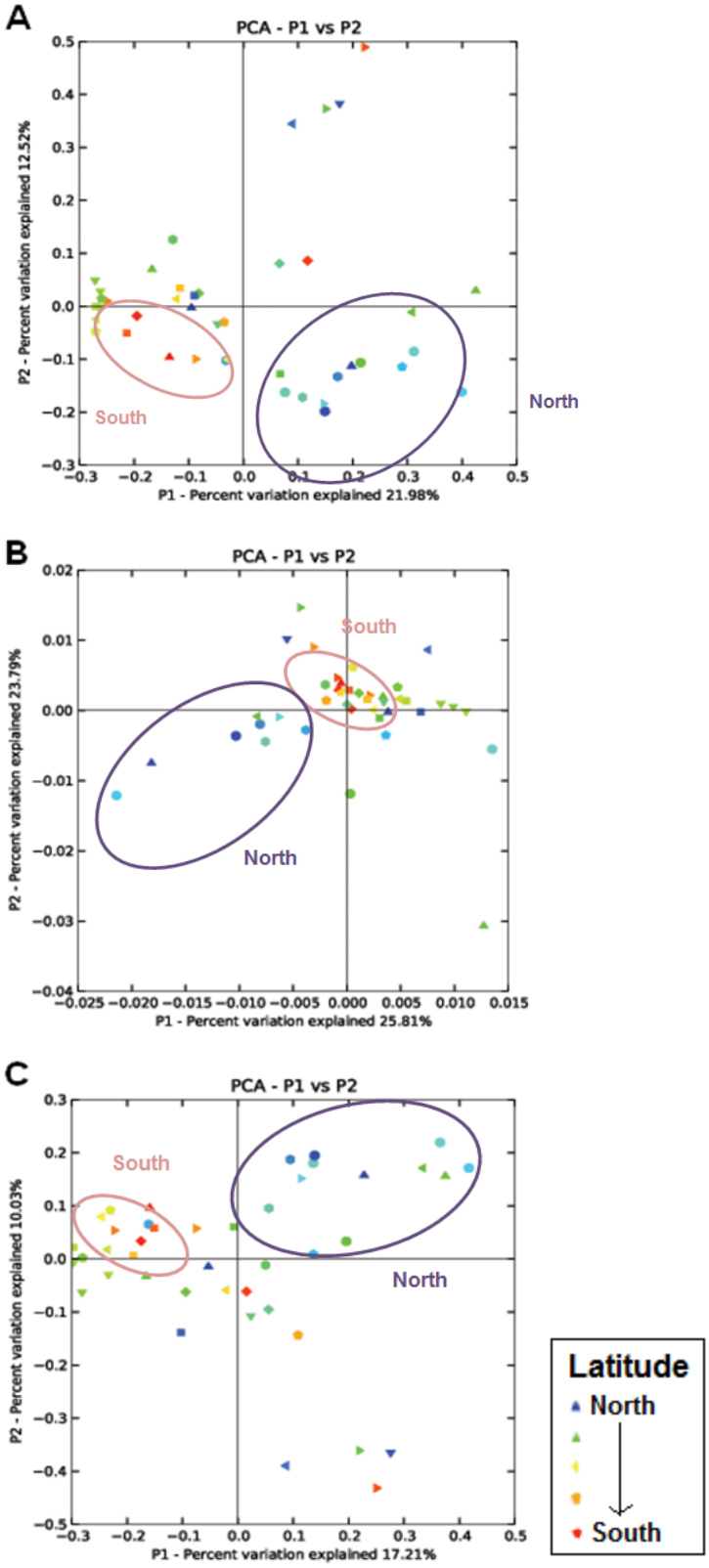


Figure 5.13. PCoA plots of the *Mycobacterium* genus water dataset comprising of 42 water samples. Sample points are coloured by **latitude**, the gradient from red to blue represent south to north latitudes **(A)** Unweighted analysis of all sequences **(B)** Weighted analysis of all sequences **(C)** Unweighted analysis of a random resample (385 sequences per sample).

Table 5.10. Univariate analysis of variables associated with the outcome variable Shannon diversity (H') of *Mycobacterium* genus species in 42 water samples (Pseudo R^2 = 0.13). Asterisks represent the relationships that were significant at the *P* <0.05 level. CC = correlation coefficient. CI = Confidence interval.

Variables	Range	CC (95% CI)	*P* value
Elevation (m)	370-3958	0.0001 (-0.0002-0.0003)	0.461
Temperature(˚C)	11.65-39.13	-0.028 (-0.060-0.005)	0.092
pH	2.95-5.61	0.210 (-0.222-0.642)	0.341
Latitude (˚N)	4.705306-12.76961	-0.121 (-0.225--0.018)	**0.022***
Longitude (˚E)	34.263-39.87669	0.055 (-0.095-0.205)	0.469

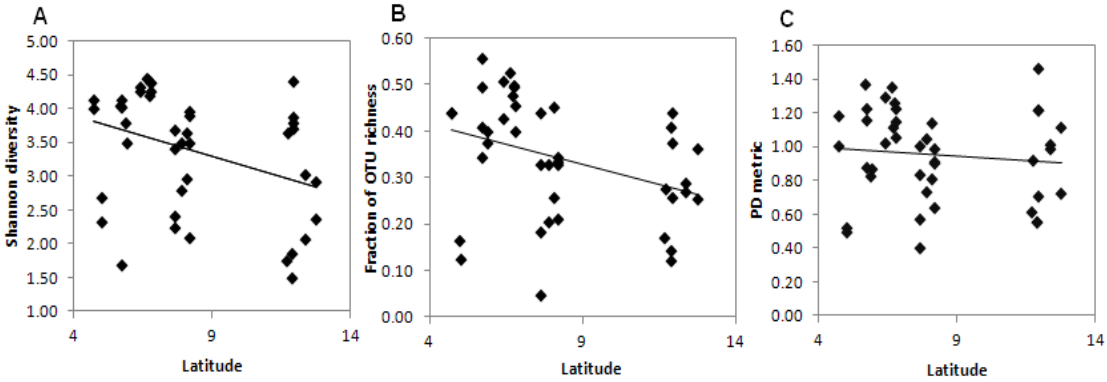


Figure 5.14. Linear relationships between diversity of the *Mycobacterium* genus water dataset (42 samples) (**A**) The Shannon diversity estimate (H') (R^2 value = 0.128). (**B**) The fraction of OTU richness (R^2 value = 0.127). **(C)** Phylogenetic Diversity (PD) metric (R^2 = 0.012). Models were chosen based on the lowest AIC values.

The SG depicted similar biogeographical trends, as latitude (*P* = 0.005) was the only factor to remain significant after Bonferroni correction and controlling for other factors (Table 5.9). In agreement, the CCA illustrated that latitude was the sole significant factor (*P* = 0.003), as samples from the northern regions Woldiya and Gonder clustered away from the southern villages (Figure 5.15A). This was consistent with the CCA of the random resampled data (950 sequences per sample), however other factors such as elevation, longitude and pH became significant (Figure 5.15B). The phylogenetic dissimilarities as depicted in the PCoA demonstrated a latitudinal gradient for the weighted analysis (Figure 5.16B), but this was not observed in the unweighted analysis of all sequences and the randomly resampled

data (Figure 5.16A&C). This suggests that latitude is particularly important in explaining the variation of abundant SG. Multivariate analysis revealed that both latitude and longitude remained significant after controlling for the other bioclimatic factors for differences in the Shannon diversity estimate (Table 5.11).

The diversity of the SG (R^2 = 0.3, P = 0.0001) in water sources displayed a monotonic decrease with latitude, from southern to northern regions as inferred by the Shannon estimate (H') (Figure 5.17A). However, no strong correlation was observed between latitude and the other diversity metrics, which were the fraction of OTU richness and the PD metric (Figure 5.17B&C). In concordance with the *Mycobacterium* genus dataset the diversity and community composition of SG was influenced mainly by latitude. The analysis suggests that the southern areas have a higher diversity of SG, perhaps due to the environmental conditions specific to the south or perhaps due to historical events causing these spatial differences.

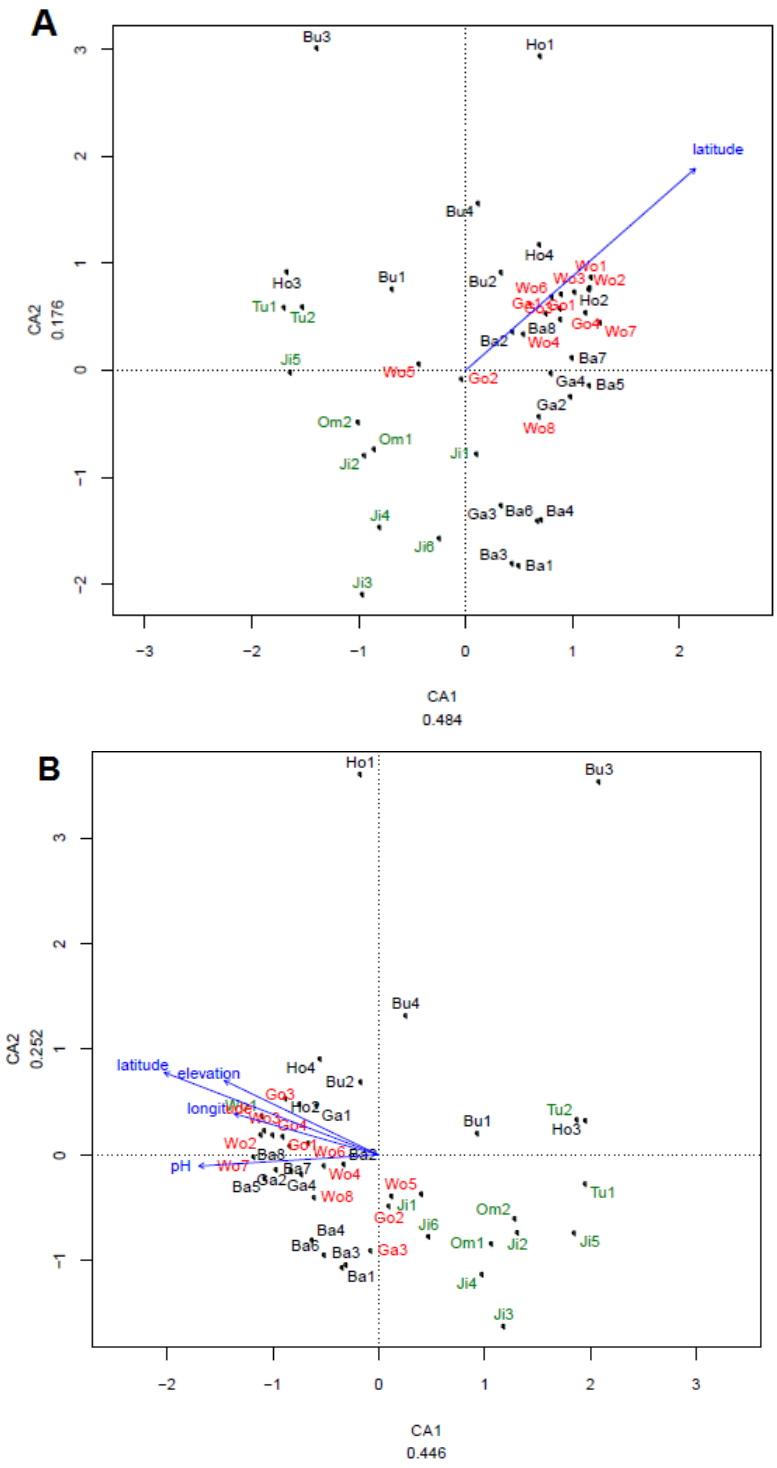


Figure 5.15. CCA plots of the SG water dataset comprising of 42 water samples. Villages are denoted by the abbreviation of village name and number. **(A)** CCA plot of all sequences, the variation was explained by latitude (r^2 = 0.37, *P* = 0.003) is shown by the arrow. (**B**) CCA plot of a random resample (950 sequences per sample), the variation was explained latitude (r^2 = 0.35, *P* = 0.001), elevation (r^2 = 0.19, *P* = 0.006), pH (r^2 = 0.21, *P* = 0.011) and longitude (r^2 = 0.15, *P* = 0.048). Villages highlighted in green and red are southern and northern latitudes respectively.

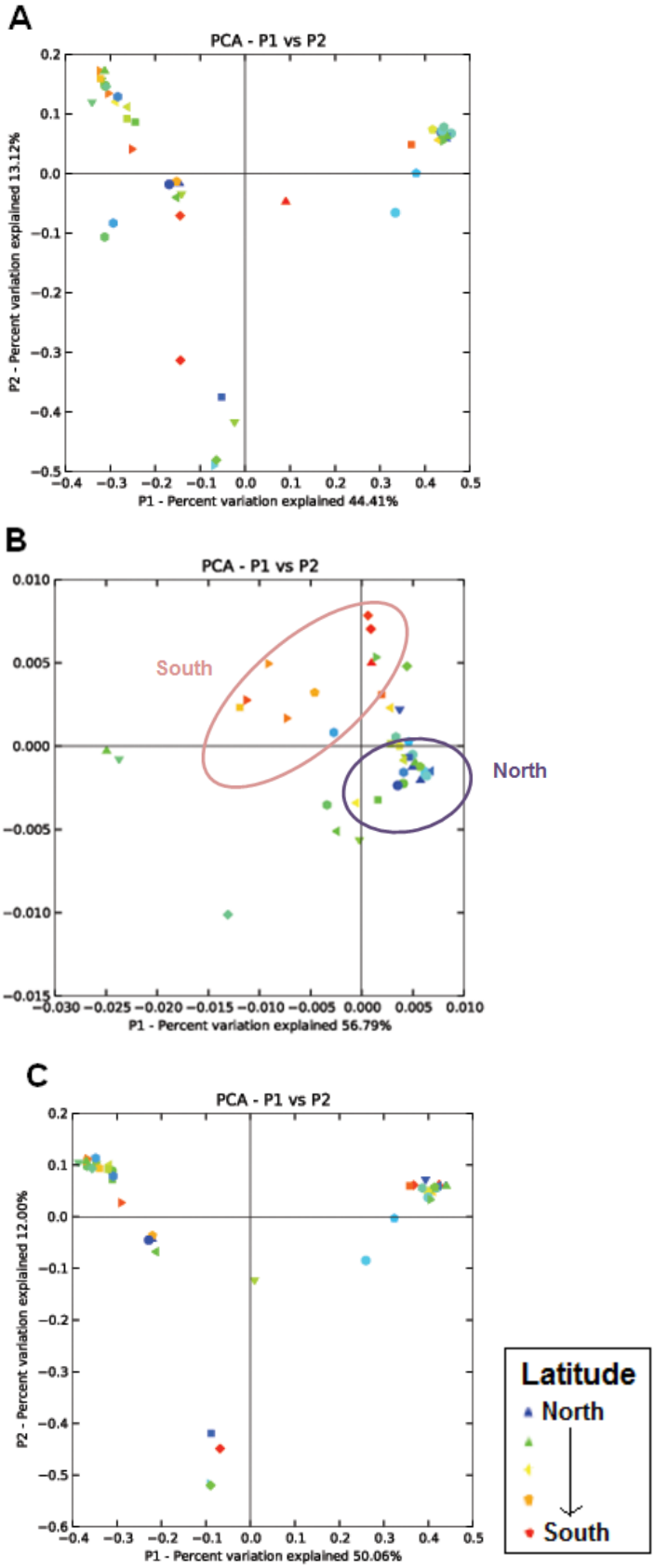


Figure 5.16. PCoA plots of the SG water dataset comprising of 42 water samples. Sample points are coloured by **latitude**, the gradient from red to blue represent south to north latitudes **(A)** Unweighted analysis of all sequences **(B)** Weighted analysis of all sequences **(C)** Unweighted analysis of a random resample (950 sequences for each sample).

Table 5.11. GLM for the Shannon diversity for the SG water dataset (42 samples) **(A)** Univariate analysis **(B)** Multivariate model of associated variables, pseudo R^2 = 0.385. Asterisks represent the relationships that were significant at the *P* <0.05 level. CC = correlation coefficient. CI = Confidence interval.

A

Variables	Range	CC (95% CI)	*P* value
Elevation (m)	370-3958	-0.0002 (-0.0003--0.0001)	**0.000***
Temperature(˚C)	11.65-39.13	0.017 (0.003-0.032)	**0.021***
pH	2.95-5.61	-0.115 (-0.316-0.087)	0.264
Latitude (˚N)	4.705306-12.76961	-0.091 (-0.129--0.053)	**0.000***
Longitude (˚E)	34.263-39.87669	-0.095 (-0.154--0.037)	**0.001***

B

Variables	Range	CC (95% CI)	*P* value
Latitude (˚N)	4.705306-12.76961	-0.079 (-0.115--0.043)	**0.000***
Longitude (˚E)	34.263-39.87669	-0.055 (-0.098--0.012)	**0.011***

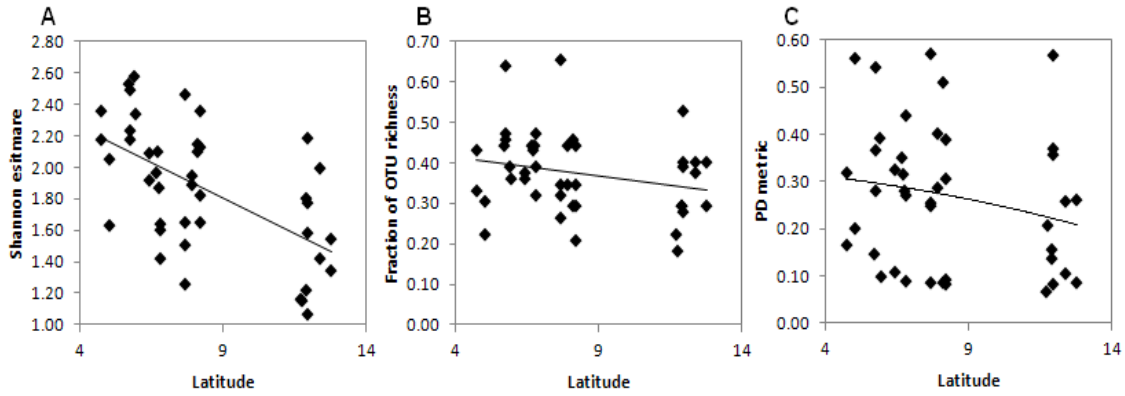


Figure 5.17. Linear relationships between diversity of the SG water dataset (42 samples) (**A**) The Shannon diversity estimate (H') (R^2 value = 0.338). (**B**) The fraction of OTU richness, data points represent the number of different OTUs per sample divided by the total number of different OTUs for all samples (R^2 value = 0.057). **(C)** Phylogenetic Diversity (PD) metric takes into account the fraction of total branch length for each sample (R^2 = 0.045). Models were chosen based on the lowest AIC values.

5.10. The effect of different spatial scales on the variation in species diversity

Variation in diversity between and within zones, regions and kebeles were compared to assess whether the proportion of variance was approximately the same between different sampling scales. The zones (<1000km) were the largest sampling unit, followed by regions (<100km) and small district kebeles (<10km). The one-way ANOVA results for the PD metric are shown here as it is an exemplar diversity metric. Overall, for the *Mycobacterium* genus soil and water datasets it was shown that regardless of sampling scale the variation between sampling units was significantly higher than within sampling units (Table 5.12), which suggests spatial proximity determines similarities in diversity at all spatial scales tested. As the F values are similar for both soils and waters, it suggests that the variance does not change markedly between sampling units. Therefore the analysis presented in this chapter will not be biased by comparisons made on the regional level. The variation in the number of OTUs for the SG datasets was not significant for all sampling units, which suggests that in this case spatial variation was not correlated with diversity (Table 5.14).

Table 5.12. One-way ANOVA of the PD metric for the different units of spatial scale; zones, regions and kebeles.

***Mycobacterium* genus soils**	Df	F	*P* value
Zones	4,18	4.21	**0.00***
Regions	5,17	3.23	**0.03***
Kebeles	12,10	3.66	**0.02***
***Mycobacterium* genus waters**	Df	F	*P* value
Zones	4,37	2.67	**0.05***
Regions	8,33	2.91	**0.01***
Kebeles	20,21	3.26	**0.00***
SG soils	Df	F	*P* value
Zones	4,27	0.40	0.81
Regions	7,24	1.44	0.24
Kebeles	17,14	1.58	0.20
SG waters	Df	F	*P* value
Zones	4,41	0.69	0.61
Regions	8,41	0.48	0.86
Kebeles	20,41	0.73	0.75

5.11. Discussion

This study has determined the biogeography of mycobacteria diversity and community composition in Ethiopian environmental samples collected from a wide range of climatic zones. Multivariate techniques revealed that community composition and diversity of mycobacteria in soil samples was correlated with lower temperatures and higher elevations. In contrast the composition of SG in soils appeared to be influenced by multiple bioclimatic factors, and longitude in particular explained the variation in diversity. The structure of assemblages in water samples was significantly associated with latitude; southern areas of Ethiopia consistently had higher diversity estimates for mycobacteria and specifically the SG. This biogeographical analysis has revealed that the organisation of mycobacteria communities is not random but influenced by both spatial factors and environmental factors.

The comprehensive pyrosequencing approach has for the first time enabled a biogeographical analysis of diversity and composition of mycobacteria as determined by OTUs and phylogenetic dissimilarities. The additional targeted pyrosequencing run specifically for SG enabled a rare insight into the group's community structure and diversity. It was of interest to assess whether the SG group presented any separate trends compared to the *Mycobacterium* genus, as this group includes many pathogenic species. Complementary multivariate approaches facilitated the determination of the relative roles of environmental variables. Sampling soil and water environments on a countrywide scale across a range of climatic regions provided a more complete analysis of the distribution of mycobacteria. However these findings demonstrated that many bioclimatic factors were correlated with each other and this limits the discrimination of their relative roles in shaping the

mycobacteria community structure. The chemical and physical characteristics of environmental samples, for example total organic carbon content and concentration of dissolved oxygen were not collected in this study. These characteristics have been found to be correlated with *Mycobacterium* species culturable counts and load as determined by quantative PCR (Iivanainen et al. 1999; Jacobs et al. 2009). Therefore it would be of interest to elucidate their roles in shaping the composition and diversity of mycobacteria. It is plausible that the spatial factors which this study found to be significant for distribution and diversity were confounding factors for other variables that were not measured.

Soil temperature/elevation was found to be principal factors explaining the *Mycobacterium* genus community composition. It has previously been demonstrated that different *Mycobacterium* species have different heat susceptibilities to temperature in suspension; for example *M. kansasii* was less resistant to heat compared to *M. fortuitum*, *M. intracellulare*, *M. marinum* and *M. xenopi* (Schulze-Röbbecke & Buchholtz 1992). This supports our findings of different compositions at different temperatures as individual species have varied optimums to temperature. As discussed, research has found correlations with temperature and total culturable counts of mycobacteria. The diversity of mycobacteria appeared to show a non-linear relationship with temperature, with the peak in diversity between 20°C-30°C. In contrast the latitudinal gradient observed for macroorganisms suggests that species diversity is highest in tropical temperatures (Gaston 2000). Therefore this suggests that there are different ecological variables which change the diversity of microorganisms and macroorganisms.

The relationship between elevation and diversity of microbial communities, unlike plant and animal communities is not well understood. Varying elevational trends

have been observed, for example a monotonic increase in microbial diversity from elevations 1820-4050 metres. This was compared to the benthic macro invertebrates in the same sampling sites, which displayed a non-linear relationship with elevation (Wang et al. 2011). In contrast, the diversity of Acidobacteria showed a significant monotonic decrease from elevations of 2460-3380 metres (Bryant et al. 2008). An additional study found no correlation between microbial diversity and elevation on a Peruvian montane gradient (200-3400 metres) (Fierer et al. 2011). However no study has researched the role of elevation for mycobacteria. Our results have shown that the richness and diversity of mycobacteria follows a non-linear relationship with elevation over a range of 425.4-3939.8 metres, and this was consistent with all three diversity measures employed. The non-linear relationship is similar to that normally observed in macroecology (Bhattarai & Vetaas 2003; Bryant et al. 2008; Richardson et al. 2000). The caveat of our analysis is that temperature was highly correlated with elevation.

Water samples from both the *Mycobacterium* genus and SG datasets had significantly higher diversity estimates with southern latitudes. Previous surveys of MAIS complex abundance also reported a geographical correlation; increased numbers of MAIS isolates were recovered in soils and waters from south-eastern United States compared to north-eastern parts (Brooks et al. 1984; Falkinham 1980). A recent study on a smaller scale found an increase in the mean estimated abundance as determined by quantitative PCR with increasing latitude in a coastal estuary environment (Jacobs et al. 2009). Therefore in agreement with previous studies our data suggests that mycobacteria community structure and diversity present geographical patterns. The results highlight that spatial factors influence mycobacteria composition and diversity in water samples, in concordance with

dispersal limitation theories (Fierer 2008; Martiny et al. 2006; Prosser et al. 2007). This is because dispersal limitation explains that microorganisms will be present in specific areas and therefore one would observe spatial trends.

Our findings have demonstrated that the composition of mycobacteria displays geographical and environmental trends, and therefore may explain the differential sensitisation to mycobacteria PPDs observed in different geographical locations (Black et al. 2001; Fine et al. 2001; Lema & Stanford 1984; Weir et al. 2003). The differences in sensitisation to mycobacteria observed between countries may be a result of the differences in the relative abundance and diversity of different species present in these environments. A direct comparison revealed the same *Mycobacterium* species that were detected in patient isolates were also detected in their local environment (Cayer et al. 2007; De Groote et al. 2006). The association between sensitisation to mycobacteria and geographical variation was observed in a survey where higher sensitisation was observed in southern parts of the Unites States compared to northern parts (Edwards et al. 1969). The observation of a higher diversity of mycobacteria particularly the SG in water samples collected from the southern parts of Ethiopia compared to the northern parts may have implications for a higher risk of exposure and infection in these areas. However, there is no accurate high resolution TB data to test this hypothesis.

There were a number of FG and SG that displayed trends with bioclimatic and spatial factors. It was of particular interest to examine the trends associated with polycyclic aromatic hydrocarbon (PAH) degraders and opportunistic pathogens. The FG PAH degraders *M.* sp. I5, *M. isoniacini* and *M. holderi* were identified in the Ethiopian environment; the former was correlated with increased moisture and the remaining two species were correlated with increased elevation. Therefore these results suggest

that these environmental conditions are likely to increase the probability of detection of PAH degraders, which could subsequently be identified for bioremediation purposes.

It is proposed that levels of exposure to mycobacteria play a role in the reduction of BCG efficacy in addition to determining risk of opportunistic disease. Research has demonstrated that differential pre-exposure to *Mycobacterium* species can affect the interferon-gamma response towards subsequent BCG vaccination, which is thought to be essential for a protective response against TB (Black et al. 2002; Weir et al. 2006). It is suggested that the abundance of mycobacteria increases at lower latitudes worldwide, and it is hypothesised that sensitisation to mycobacteria in environmental substrates is responsible for the lower BCG vaccine efficacy observed at lower latitudes (Fine 1995; Wilson et al. 1995).

Several SG opportunistic pathogens were correlated with bioclimatic factors. For example *M. bohemicum* and *M. kansasii*, which are capable of causing cervical lymphadenitis, were correlated with increased elevation. The species *M. ulcerans* and *M. novocastrense* were correlated with lowland areas and western areas respectively, and both are associated with cutaneous diseases. Many SG were associated with pulmonary TB; these included *M. malmoense*, *M. colombiense*, *M. riyadhense*, *M. asiaticum* and *M. gordonae*, and were correlated with both spatial and bioclimatic factors. These findings support the differences the composition of the causative agents of *Mycobacterium* disease in different countries. For example a comparison of 14 countries showed *M. avium*, *M. gordonae*, *M. kansasii* and *M. fortuitum* were more frequently isolated in Denmark, Germany, Czech republic and Iran respectively (Martin-Casabona et al. 2004). A similar review found that the species responsible for pulmonary disease were differentially more prevalent in different countries, for

example *M. kansasii* and *M. malmoense* were commonly detected in the UK and central USA, and *M. simiae* was found mostly in arid regions of USA, Cuba and Israel (Cook 2010). These differential trends in the causative agents of opportunistic disease are most likely to reflect differential exposure to mycobacteria.

In summary, the distribution of mycobacteria revealed significant correlations with latitude, temperature and elevation. Results would suggest a higher diversity of mycobacteria at lower latitudes, and mid-elevations (1500-2500m) and mid-temperatures (20ºC-30ºC). These findings have important implications for our understanding of the community composition of mycobacteria. Future work should aim to elucidate whether these trends are observed worldwide, and enable predictive spatial mapping based on these environmental and spatial correlates. This in turn will produce better correlations with BCG vaccine efficacy and the risk of opportunistic disease.

Chapter 6: Quantification of the *Mycobacterium* genus and the *M. tuberculosis* complex in the environment with implications for TB

6.1. Abstract

Mycobacteria are ubiquitous in the environment, however their abundance in the Ethiopian environment is not known. A quantitative PCR (qPCR) assay was employed to quantify abundance of mycobacteria in 42 soil samples and 42 water samples. Cell counts ranged from $9.47x10^5$ to $8.01x10^8$ cells per gram in soil samples and $3.35x10^1$ to $2.68x10^5$ cells per ml in water samples. Greater variation was observed between soil samples from different regions compared to within regions. A non-linear relationship was observed between mycobacteria abundance in soils and elevation. In contrast mycobacteria were significantly more abundant in eastern areas and areas with higher pH values in water samples. The abundance of *M. bovis* was quantified using a qPCR specific for the RD4 scar region. The species *M. bovis* was detected in 5/42 water samples ranging from $7.3x10^0$ cells per ml to $1.51x10^2$ cells per ml and detected in 1 soil sample ($2.94\ x10^3$ cells per gram). There was some concordance with the pyrosequencing BLAST results, where four samples positive for *M. bovis* also contained BLAST matches to the MTC. The detection of *M. bovis* in the water environment suggests a potential route of transmission. Overall, the incidence of extrapulmonary tuberculosis (2008/2009) was highest in Gonder. This correlated with the abundance of mycobacteria, which was also found to be highest in the region Gonder ($4.83x10^8$ cells per gram of soil). These results suggest abundance in the environment may be directly related to risk of disease.

6.2. Introduction

Environmental reservoirs are locations in which pathogens are able to thrive outside the host and via successful transmission routes can present a serious risk to animal and human health. Pathogens have been detected at a high abundance in waters (Brookes et al. 2004), soils (Santamaría & Toranzos 2003) and biofilms (Wingender & Flemming 2011). Successful environmental pathogens include *Campylobacter* sp. (Devane et al. 2005), *Salmonella Typhimurium* (Jyoti et al. 2011) and Enterotoxigenic *Escherichia coli* (Singh et al. 2010). A principal example is the life cycle of *Vibrio cholerae*; the species is known to proliferate in brackish and estuarine waters, and is particularly successful due to its survival in zooplankton (Vezzulli et al. 2010). Drinking from Lake Victoria and from local streams was found to increase the risk of cholera (Shapiro et al. 1999). Identification of environmental reservoirs and potential transmission routes of *Vibrio cholerae* has enabled successful control measures to be employed, such as the filtration of water through sari cloth (Huq et al. 2010).

The majority of *Mycobacterium* species survive in the environment, perhaps due to their tough cell wall as well as their ability to produce spores and proliferate within amoebae (Lamont et al. 2012; Thomas & McDonnell 2007). Selected species are known pathogens and therefore the transmission between the wildlife, livestock, humans and the environment is of particular concern. For example *M. leprae* was more likely to be detected in soil samples taken from leprosy endemic areas compared to areas where leprosy was not prevalent (Lavania et al. 2008). The species *M. ulcerans* has been commonly detected in riverine swamp areas (Fyfe et al. 2007; Portaels et al. 2008) and is of concern following its detection in environments subject to an outbreak of Buruli ulcer (Ross et al. 1997).

The species *M. bovis* is a member of the *M. tuberculosis* complex and is the etiologic agent of bovine TB. The burden of disease is of both clinical and economic importance. Although transmission is thought to occur through contact of infected wildlife, livestock and humans, the role of the environment as a reservoir of infection is the subject of on-going debate. The species *M. bovis* has been detected in the UK environment (Sweeney et al. 2007; Young et al. 2005), suggesting that an environmental reservoir does exist. Bovine TB in Ethiopia is of increasing concern; although the prevalence is relative low (4.7%) it is widespread throughout the country (Berg et al. 2009). Therefore it was of interest to assess whether *M. bovis* could be detected and quantified in the Ethiopian environment, as we anticipate significant environmental contamination due to shedding of *M. bovis*.

The obligate pathogen *M. tuberculosis* is the most pathogenic member of the MTC and is predominantly transmitted between humans. However it was of interest to assess whether the environment may play an important role. Previous work has shown that *M. tuberculosis* can remain pathogenic after 72 hours of exposure to sunlight (Caldwell 1925). In addition *M. tuberculosis* was found to survive on glass, sand and dust for a maximum of 300 days (Walther & Ewald 2004). Therefore although *M. tuberculosis* has not been naturally detected in the environment, these studies suggest *M. tuberculosis* can persist in the environment.

It was desirable to quantify the pathogen in the environment to assess whether it represents a significant reservoir. Cultivation methods reveal a fraction of the true abundance in the environment, whereas real-time quantitative PCR (qPCR) provides a more sensitive alternative. Unlike end-point PCR, qPCR measures the amount of target DNA during the early stages of the PCR reaction, thereby providing a more accurate estimate due to the fact that there are no limiting factors such as depleting

substrates and enzymes in the early stages of the PCR reaction (Wilhelm & Pingoud 2003). Non-specific qPCR chemistries include the employment of SYBR Green 1, which only emits light when bound to double-stranded DNA (Arya et al. 2005). The amount of DNA is therefore proportional to the amount of fluorescent signal. However this method is limited as it can bind non-specifically to all double-stranded DNA such as to primer dimers. The alternative TaqMan chemistry relies on a dual-labelled fluorogenic probe, with the reporter at the 5' end and the quencher at the 3' end of the probe (Heid et al. 1996). When the probe is bound to the sequence the reporter dye emission is suppressed by the quencher dye, however when the probe is cleaved due to strand displacement the reporter is separated from the quencher and fluorescent emission is detected. qPCR has been used extensively in microbial ecology to quantify the abundance of particular species or groups and to quantify genes involved in particular functions (Smith & Osborn 2009; Zhang & Fang 2006).

6.3. Aims

i) To quantify the *Mycobacterium* species abundance in Ethiopian soil and water environments and assess any correlations to bioclimatic and spatial factors.

ii) To quantify the abundance of *M. bovis* and the *M. tuberculosis* complex in Ethiopian soil and water samples to identify potential environmental reservoirs.

iii) To examine the incidence of smear-positive, smear-negative and extrapulmonary TB cases and relate to same locations of environmental sampling.

6.4.1. Optimisation and Specificity analysis of *Mycobacterium* genus assay

The previously designed primer set which targets the 16S-23S Internal Transcribed Space (ITS) and a part of the 23S gene of *Mycobacterium* species (Jacobs et al. 2009) was optimised on the ABI 7500 Fast Real-Time PCR System (Applied Biosystems Inc., CA, USA) as described in section 2.5. Three different primer/probe concentrations were evaluated for efficiency (Figure 6.1). The percentage efficiency of the reaction was calculated using the equation E = (10^(-1/slope)-1)*100. The efficiency was 108.5%, 99.1%, and 113.5% for the 5µM, 10µM and 20µM concentrations respectively. The probe and primer concentration of 10µM was chosen as this gave the efficiency closest to 100% and the R^2 value was the highest, suggesting that the primer set worked best at this concentration. The specificity of the primer set was confirmed using DNA extracted from *Mycobacterium* species, *Streptomyces* species and *Escherichia coli* (Table 6.1). As expected the primer set amplified the *Mycobacterium* species only. The limit of detection of the assay was reported at less than 1 cell/ml (Jacobs et al. 2009).

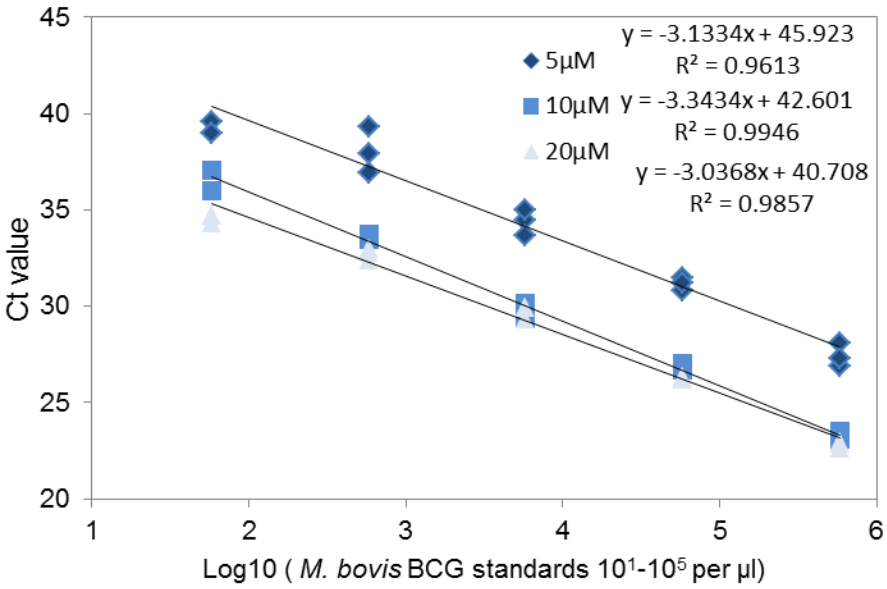


Figure 6.1. Optimisation of *Mycobacterium* genus primer/probe concentrations, tested at 5µM, 10µM and 20µM. Standards tested ranged from $5.82x10^{-1}$ per µl to $5.82x10^{5}$ per µl. All primer/probe concentrations were not detected below $5.82x10^{1}$ per µl.

Table 6.1. Reference strains used in this study to test the specificity of the *Mycobacterium* genus qPCR assay.

Species	Amplification
***Mycobacterium abscessus* DSM 44196T**	+
Mycobacterium aichense	+
***Mycobacterium aurum* ATCC 23366**	+
***Mycobacterium avium* subsp. *avium* DSM 44156T**	+
***Mycobacterium avium* subsp. *paratuberculosis* DSM 44133T**	+
Mycobacterium duvali	+
***Mycobacterium fortuitum* ATCC 14468**	+
Mycobacterium gilvum	+
***Mycobacterium kansasii* DSM 44162T**	+
***Mycobacterium intracellulare* DSM 43223T**	+
***Mycobacterium marinum* DSM 44344T**	+
Mycobacterium obuense	+
Mycobacterium phlei* ATCC 354**	+
Mycobacterium scrofulaceum* DSM 43992T	+
***Streptomyces griseus* DSM 40236T**	-
***Streptomyces bacillaris* DSM 40598T**	-
***Streptomyces coelicolor* DSM 40233T**	-
***Escherichia coli* ATCC 29842**	-

6.4.2. Quantification of the *Mycobacterium* genus in environmental samples

Real-time quantification of the entire *Mycobacterium* genus load was evaluated against the standard curves, which had seven dilutions from 5.8 $x10^5$ to 5.8$x10^{-1}$ cells per µl (Figure 6.2). Standards were detected at 5.8 $x10^1$ cells per µl and above. All soil samples were positive for the *Mycobacterium* genus apart from four soils, which were Ho4, Bu1, Bu2 and Bu4 (Figure 6.3). This is probably due to the observed inhibition in these samples as shown by the RD4-GFP qPCR assay (section 4.3.2). A retest of the *Mycobacterium* genus qPCR assay was carried out to ensure samples were negative. The *Mycobacterium* genus abundance in soil samples ranged from 9.5 $x10^5$ to 8.0 $x10^8$ cells per gram (Figure 6.3). From 42 water samples all were positive apart from seven; inhibition may have been responsible as six samples had delta Ct values above zero in the qPCR inhibition test which suggests there was some level of

inhibition, however the values were less than or equal to one which suggests negligible inhibition. No detection may also be due to very low levels present in these samples. For water samples the observed abundance of *Mycobacterium* species was lower and ranged from 3.4 x 10^1 to 2.7 x10^5 cells per ml. There were significantly greater of *Mycobacterium* genus cell counts between source types compared to within source types ($F(1,71) = 25.3$, $P<0.001$). The mean for soil samples (1.60x10^8 cells per gram) was higher compared to water samples (1.17x10^4 cells per ml). The difference between regions compared to within regions for soil samples was significantly greater ($F(8,29) = 4.2$, $P = 0.002$). The highest *Mycobacterium* genus load was in Gonder (mean 4.8x10^8 cells per gram) and the lowest abundance was in Omorate (mean 7.4x10^6 cells per gram). In contrast no significant differences were observed between regions for water samples. The highest *Mycobacterium* genus load were found in sample Ba1 (2.7x10^5 cells per ml) and the lowest in sample Bu1 (3.4x10^1 cells per ml).

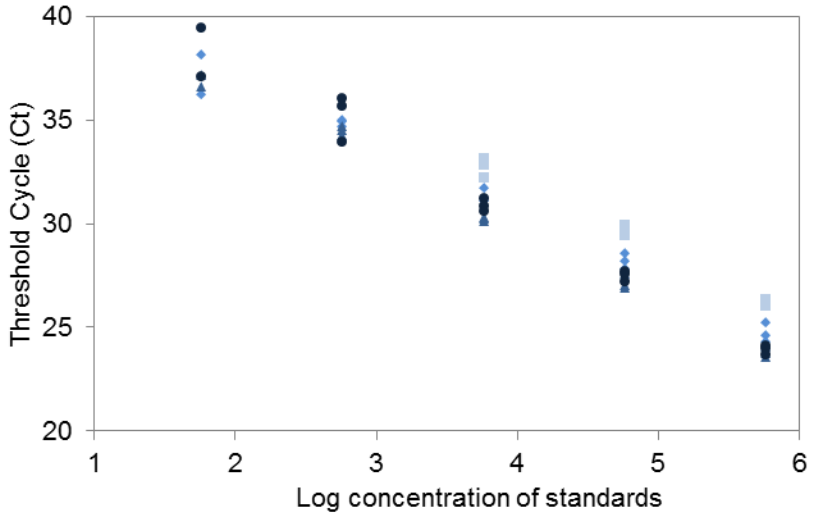


Figure 6.2. Standards for the *Mycobacterium* genus qPCR. Calibration standards generated curve lines R^2 = 0.982 to 0.990 with a –X slope of -3.199 to -3.650 and a Ct range of 23.52 to 39.41.

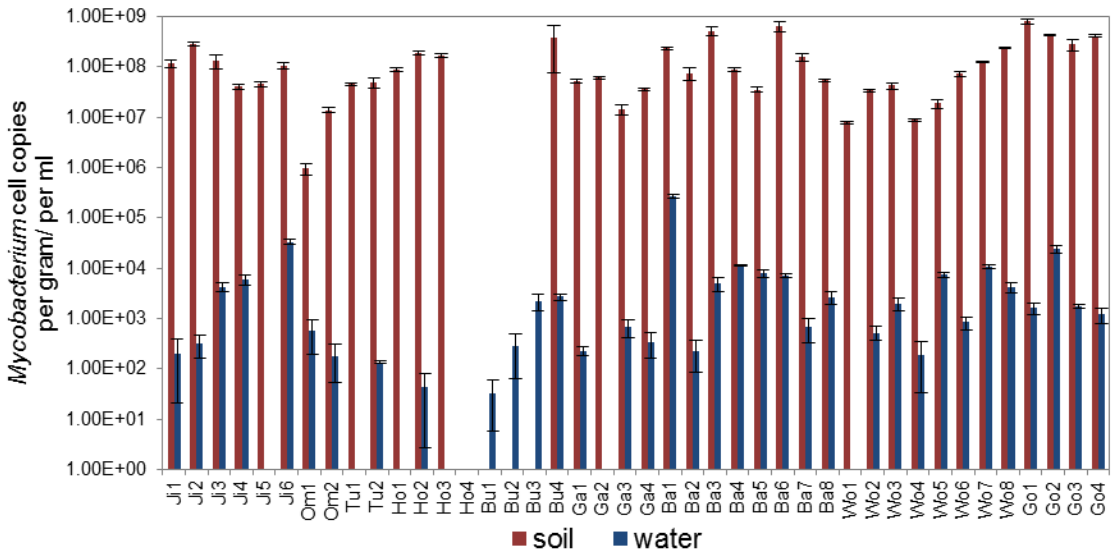


Figure 6.3. qPCR of the internal transcriber spacer and partial 23S region specific for the *Mycobacterium* genus. Error bars represent the standard deviation of the mean counts from three replicates per sample.

6.4.3. Bioclimatic correlations with *Mycobacterium* species abundance

The *Mycobacterium* genus abundance as determined by qPCR was correlated with the bioclimatic variables collected from the sampling sites using a generalised linear model (GLM) with a Gaussian distribution. The univariate analysis revealed the significance of increased elevation, decreased temperature and increased moisture for higher *Mycobacterium* genus abundance (Table 6.2A). The multivariate generalised linear model exemplified the significance of elevation (P = 0.002) and temperature (P = 0.044) after controlling for all other bioclimatic and spatial factors (Table 6.2B). The relationship between elevation, temperature, moisture and *Mycobacterium* genus abundance was evaluated further. Linear and quadratic models were compared and the model with the lowest Akaike's Information Criterion (AIC) was chosen and presented in this chapter. A non-linear relationship was observed between elevation and *Mycobacterium* genus abundance examined over the range of 425m-3940m, with peaks at mid-elevations (c. 2000-3000m) (abundance R^2 = 0.3) (Figure 6.4). For example, the Bale transect showed a peak in *Mycobacterium* genus load at mid elevation villages Ba3-6 (average 3.2×10^8 cells per gram of soil) compared to the

higher elevation villages of Ba7-Ba8 and lower elevation villages Ba1-2 (average $1.3x10^8$ cells per gram of soil). In Woldiya the higher elevation points Wo7-8 were higher on average ($1.8x10^8$ cells per gram of soil) compared to the mid-elevation villages Wo3-6 (average $3.6x10^7$ cells per gram of soil), however the lowest elevation points displayed the lowest *Mycobacterium* genus load counts (average $2.1x10^7$ cells per gram of soil).

A negative linear relationship was observed between temperature and *Mycobacterium* genus load, however the R^2 value for the linear relationship was low (R^2 = 0.1) (Figure 6.5). A weak positive linear relationship between the *Mycobacterium* genus abundance and moisture (R^2 = 0.1) was revealed (Figure 6.6). The correlations between soil bioclimatic variables and abundance suggest the potential to risk map areas of high *Mycobacterium* species exposure based on environmental factors.

Table 6.2. GLM for the Log *Mycobacterium* abundance of 38 soil samples **(A)** Univariate analysis **(B)** The multivariate model of associated variables, pseudo R^2 = 0.54. Asterisks represent the relationships that were significant at the *P* <0.05 level. CC = correlation coefficient. CI = Confidence interval.

A

Variables	Range	CC (95% CI)	*P* value
Elevation (m)	362.2-3939.8	0.0003 (0.0001-0.0005)	**0.002***
Temperature (°C)	13.17-38.21	-0.029 (-0.055--0.004)	**0.025***
Moisture (%)	1.39-28.17	0.026 (0.004-0.047)	**0.018***
pH	4.07-6.47	0.300 (-0.042-0.643)	0.085
Latitude (°N)	4.705306-12.76961	0.035 (-0.044-0.114)	0.388
Longitude (°E)	34.263-39.87669	0.070 (-0.026-0.167)	0.153
Elevation^2 (m)		5.27E-08 (1.19E-08-9.35E-08)	**0.011***
Temperature^2 (°C)		-0.001 (-0.001--0.0001)	**0.018***
Moisture^2 (%)		0.001 (0.0001-0.0016)	**0.029***

B

Variables	Range	CC (95% CI)	*P* value
Elevation (m)		0.002 (0.001-0.003)	**0.000***
Elevation^2 (m)		-3.00E-07(-4.04E-07 --1.29E-07)	**0.000***
Temperature (°C)		0.121 (0.069-0.173)	**0.000***
Moisture^2 (%)		0.001 (0.000-0.002)	**0.005***

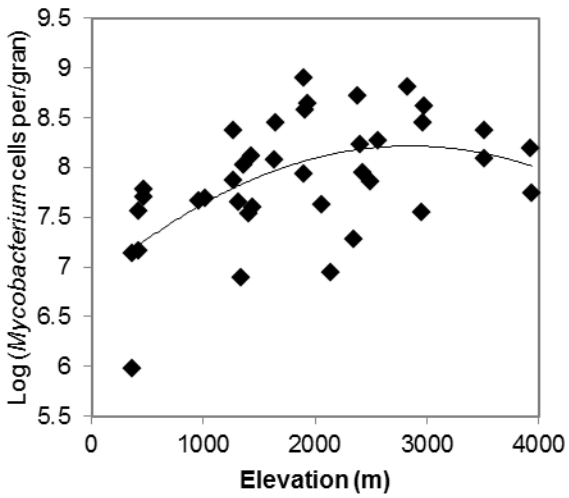


Figure 6.4. The relationship between *Mycobacterium* genus load in soil and elevation (R^2 = 0.317).

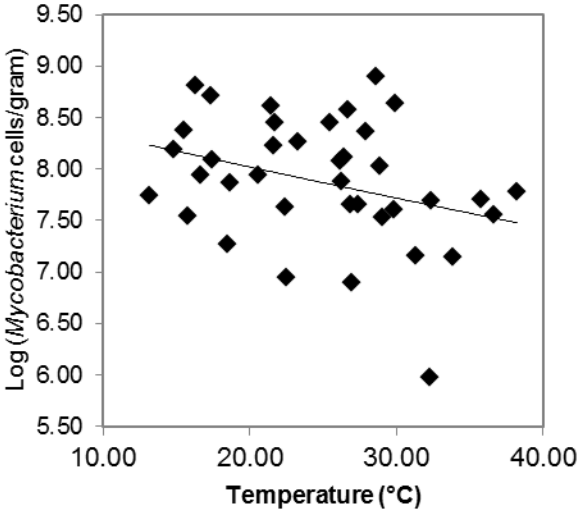


Figure 6.5. The relationship between *Mycobacterium* genus load in soil and temperature (R^2 = 0.102).

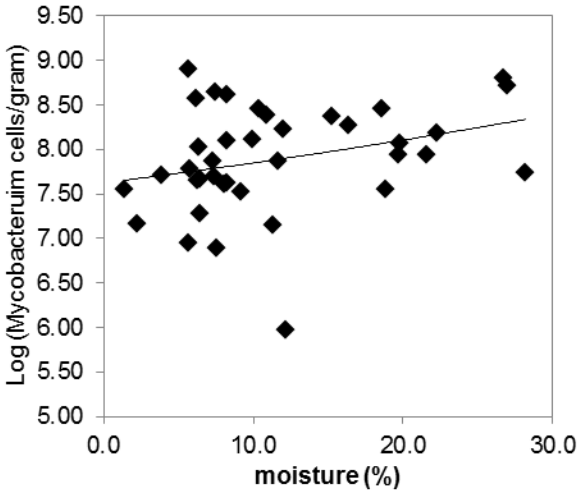


Figure 6.6. The relationship between *Mycobacterium* genus load in soil and moisture (R^2 = 0.09).

The relationship between bioclimatic factors, spatial variation and the *Mycobacterium* genus abundance in water samples was assessed using the same methods as described above. The GLM univariate analysis demonstrated the significance of lower temperatures, higher pH values and eastern areas for higher *Mycobacterium* genus abundance (Table 6.3A). After controlling for other bioclimatic and spatial variables the multivariate model demonstrated that only pH and longitude remained significant (Table 6.3B). These linear patterns were further demonstrated for pH, longitude and temperature (Figure 6.7, 6.8 & 6.9) as these had lowest AIC values. However, the R^2 values for these bioclimatic and spatial variables were small ($R^2 = 0.1$). Therefore this analysis suggests that both environmental and spatial factors contribute to the differences in the abundance of *Mycobacterium* species. The variables examined here represented those most likely to influence the composition of mycobacteria as determined by a literature review (section 1.6) however it is possible that other factors not measured may play a role.

Table 6.3. GLM for the Log *Mycobacterium* abundance of 35 water samples (A) Univariate analysis (B) The multivariate model of associated variables, pseudo $R^2 = 0.175$. Asterisks represent the relationships that were significant at the $P < 0.05$ level. CC = correlation coefficient. CI = Confidence interval.

A

Variables	Range	CC (95% CI)	*P* value
Elevation (m)	370-3958	0.0002 (0.0000-0.0004)	0.081
Temperature (°C)	11.65-36.05	-0.033 (-0.058--0.008)	**0.010***
pH	3.62-5.61	0.508 (0.097-0.919)	**0.015***
Latitude (°N)	4.705306-12.76961	0.030 (-0.060-0.119)	0.515
Longitude (°E)	34.263-39.85836	0.164 (0.041-0.287)	**0.009***

B

Variables	Range	CC (95% CI)	*P* value
pH	3.62-5.61	0.461 (0.087-0.834)	**0.016***
Longitude (°E)	34.263-39.85836	0.154 (0.029-0.279)	**0.016***

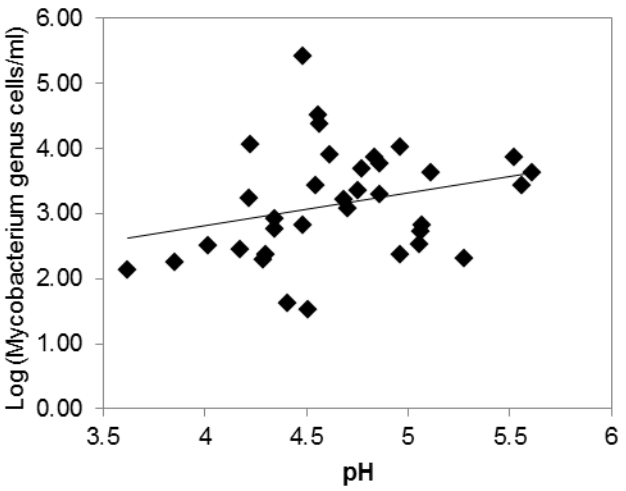


Figure 6.7. The relationship between *Mycobacterium* genus load in water and pH (R^2 = 0.077)

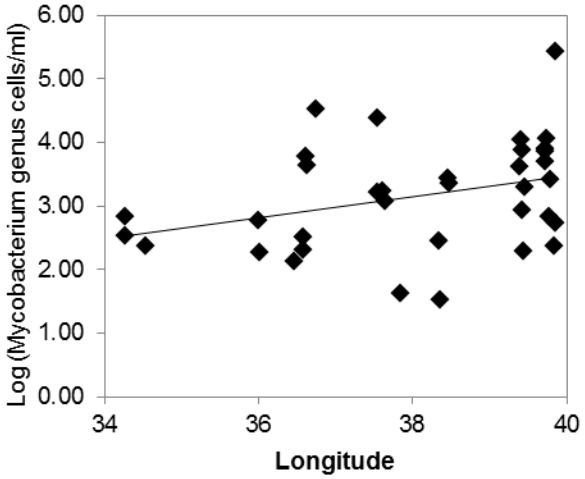


Figure 6.8. The relationship between *Mycobacterium* genus load in water and longitude (R^2 = 0.112).

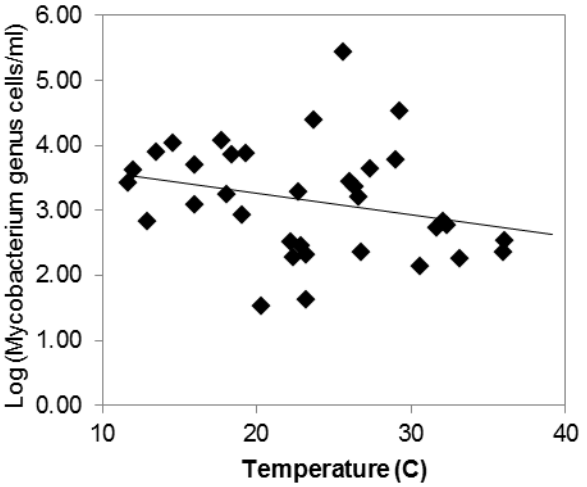


Figure 6.9. The relationship between *Mycobacterium* genus load in water and temperature (R^2 = 0.071).

6.5. Quantification of *M. bovis*

A previously designed qPCR assay was used to target the RD4 scar region specific for *M. bovis* (Sweeney et al. 2007). The limit of detection was reported as $8.5x10^4$ cells per gram of soil (Pontiroli et al. 2011). The standards for this set of experiments were *M. bovis* BCG, standards ranged from the range $8.5x10^5$ to $8.5x10^{-1}$ cells per µl (Figure 6.10). In total 7.1% (6/84) of environmental samples were positive in triplicate for *M. bovis*. One soil sample (2.4%) was positive, containing $2.9x10^3$ cells per gram, and five water samples (11.9%) were positive, containing between $7.3x10^0$ cells per ml to $1.5x10^2$ cells per ml (Figure 6.11). The *M. bovis* positive soil sample was collected from the region Bale, an area of high elevation (3940m) and afro-alpine vegetation (Figure 6.12A). The positive water sample from Woldiya was collected from water taps (Figure 6.12B), which suggests that *M. bovis* can survive in the water distribution system. Two out of four water samples from Butajira and Gambella tested positive for *M. bovis;* samples were collected from a wide variety of sources including wells, pumps and rivers (Figure 6.12C&D). This suggests that in these regions water may be a significant a reservoir of *M. bovis*.

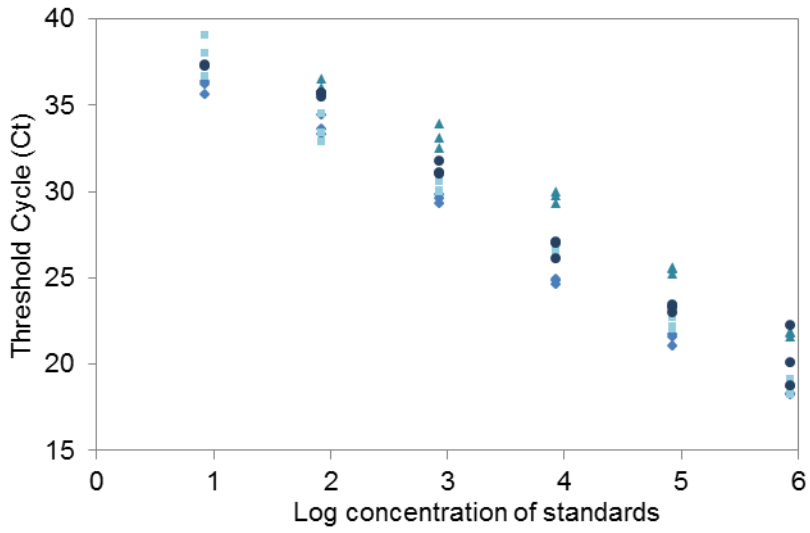


Figure 6.10 qPCR standards for *M. bovis* RD4. Calibration standards generated a curve line R^2 = 0.98-0.99 with a –X slope -3.66 to -3.82 and a Ct range of 18.2-38.0.

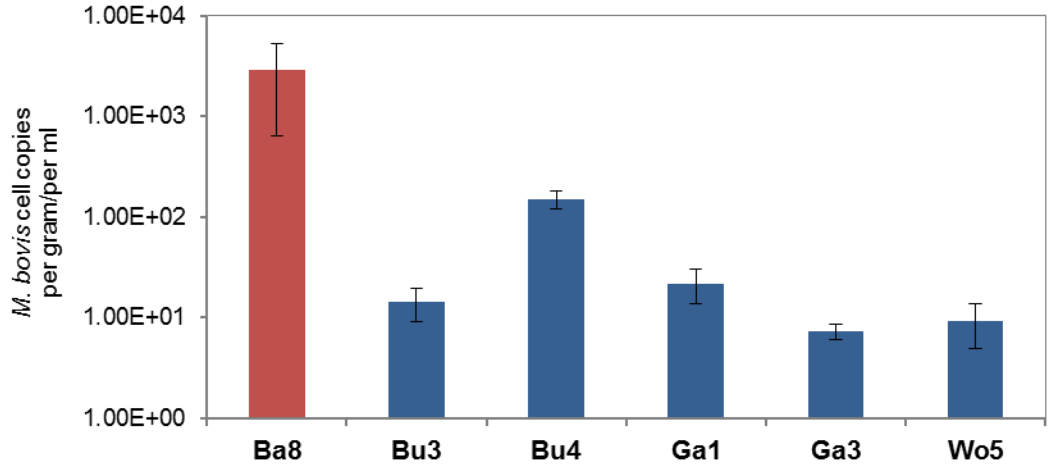


Figure 6.11. Mean numbers of *M. bovis* cell copies per gram of soil (red bar) or per ml of water (blue bar). Error bars represent the standard deviation of the mean counts from three replicates per sample.

Figure 6.12. Pictures of sample collection points from *M. bovis* positive villages **(A)** Ba8 **(B)**Wo5 **(C)** Bu4 **(D)** Ga1.

6.6. Quantification of *M. tuberculosis* and the MTC

For the quantification of *M. tuberculosis* and the MTC a previously designed multiplex PCR was employed (Reddington et al. 2011). Two different gene targets *lepA* and *wbbl1* were used to quantify the MTC and both *M. tuberculosis* and *M. canettii* respectively. The standards ranged from 0.73 genome equivalents per µl to $7.3x10^5$ genome equivalents per µl (Figure 6.13 & 6.14). The limit of detection was reported at 2.2 and 2.17 cell equivalents per reaction for the *M. tuberculosis/M. cannetti* probe and MTC probe respectively (Reddington et al. 2011). The multiplex qPCR found that only two water samples (Bu3 & Bu4) were positive for the MTC.

Average abundance values ranged from 1.83x 10^2 cells/ml to 1.81x10^3 cells/ml (Figure 6.15). This was in accordance with the *M. bovis* RD4 scar region qPCR, which found that Bu3 and Bu4 were positive for *M. bovis*. The abundance estimates for *M. bovis* were lower than the MTC abundance estimates; this suggests other members of MTC were present in these samples. However MTC was not detected in four samples that were positive for *M. bovis* in the RD4 scar region qPCR, perhaps as the MTC qPCR was less sensitive or that *M. bovis* was not present in that technical replicate. No samples (0/84) were positive for the *wbbl1* region specific for *M. tuberculosis* and *M. cannettii*, which suggests that the environment was not a reservoir of *M. tuberculosis.*

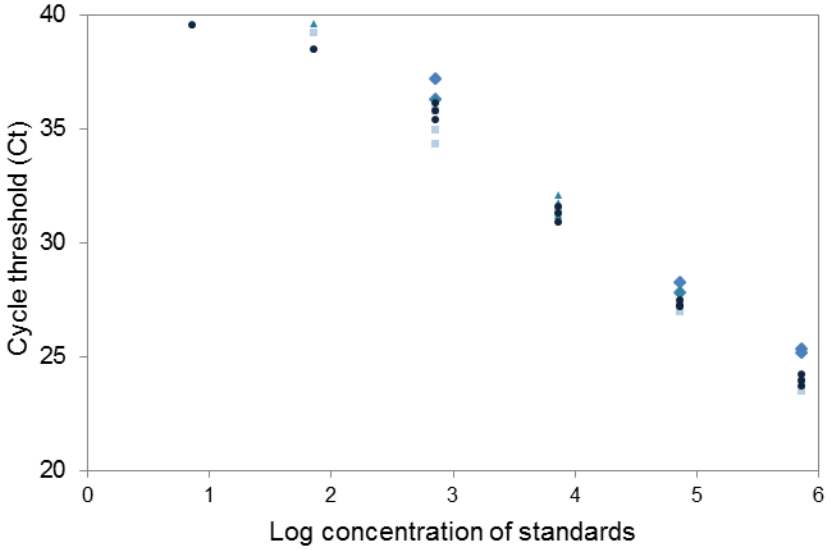


Figure 6.13 qPCR standards for MTC *lepA*. Calibration standards generated a curve line R^2 = 0.976-0.996 with a –X slope -3.500 to -3.922 and a Ct range of 23.520 -39.582.

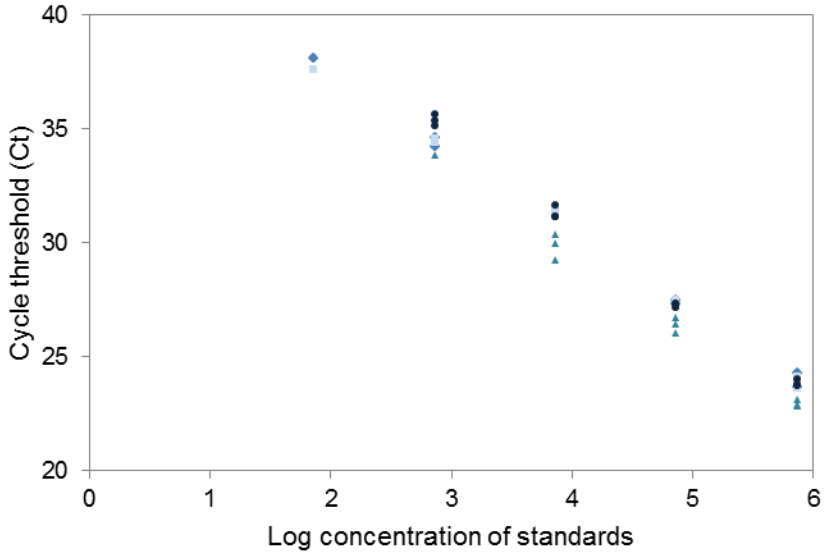


Figure 6.14. qPCR standards for *M. tuberculosis* and *M. canettii wbbl1*. Calibration standards generated a curve line R^2 = 0.991-0.999 with a –X slope -3.488 to -3.869 and a Ct range of 23.095 to 38.128.

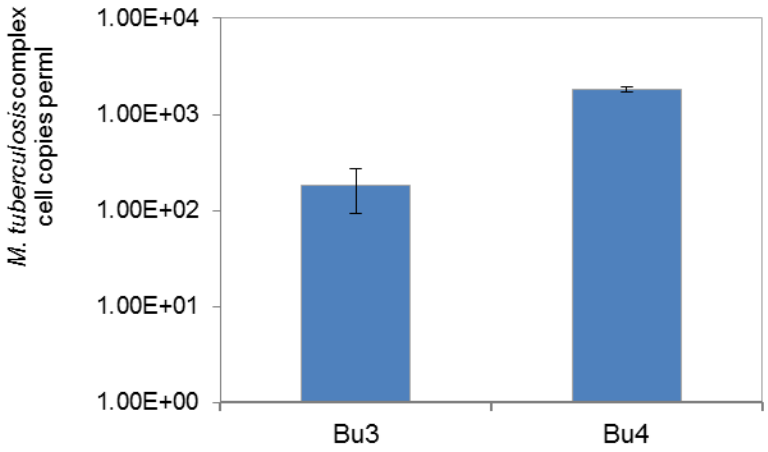


Figure 6.15. Mean numbers of *M. tuberculosis complex* cell copies per ml of water. Error bars represent the standard deviation of the mean counts from three replicates per sample.

6.7. Comparisons between qPCR and pyrosequencing BLAST results

Overall, 35/42 (83.3%) water samples and 11/23 (47.8%) soil samples were BLAST matched to *M. tuberculosis* complex. This indicates that MTC is widespread throughout the Ethiopian environment, especially in water sources (Figure 6.16). It was not possible to distinguish between members of the MTC using this approach, as the species within the complex are identical in their 16S rRNA gene. Therefore other gene targets were employed for qPCR to differentiate between MTC members.

Concordance was observed between the pyrosequencing BLAST results and the qPCR; for example, the four samples from Butajira and Gambella that were positive for *M. bovis* also contained sequences matched to the *M. tuberculosis* complex (Figure 6.16). However the two samples from Bale and Woldiya that were positive for *M. bovis* did not contain any BLAST matches for *M. tuberculosis* complex. Other samples that were positive for the *M. tuberculosis* complex but not the *M. bovis* qPCR may have contained other members of the *M. tuberculosis* complex. However the MTC was not detected in these samples using qPCR, perhaps because the pyrosequencing approach was more sensitive. Therefore pyrosequencing has a sensitivity of less than 1.83x 10^2 cells per ml.

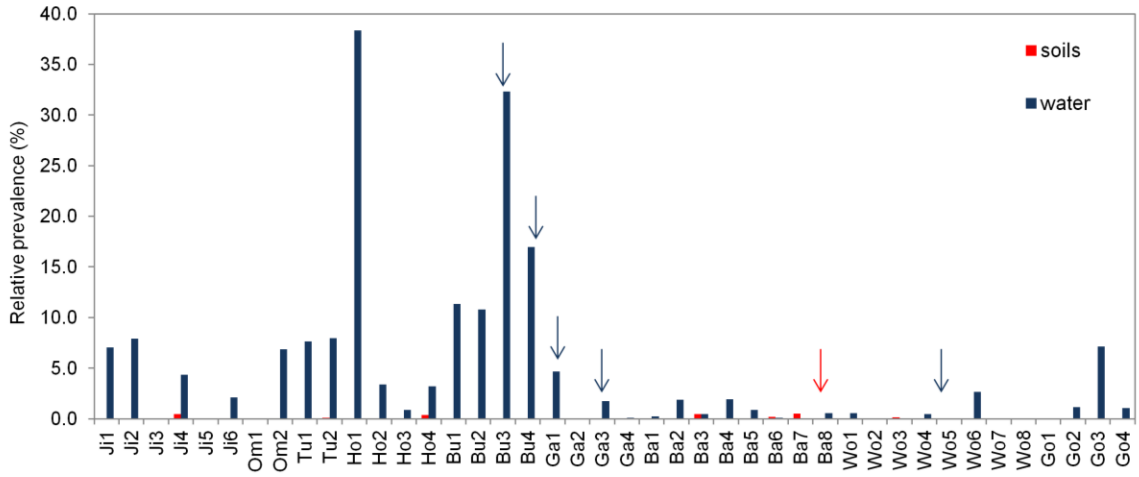


Figure 6.16. The prevalence of *M. tuberculosis* complex BLAST matches compared to all BLAST matched sequences. The arrows indicate where *M. bovis* was detected using TaqMan quantitative PCR**.** Arrows represent RD4 qPCR detection for soil (red) and water (blue).

6.8. Prevalence and Incidence of Tuberculosis in Ethiopia

Tuberculosis incidence data was collected from the Ministry of Health (MoH) and from health centres of villages in which environmental sampling took place. Data was recorded for the year 2008-2009 and was plotted on to ArcGIS maps provided by the Ministry of Agriculture and Rural Development (MARD). The incidence of

all new TB cases and smear negative cases were highest in Woldiya and Butajira (Figure 6.18 & 6.19). Whereas the incidence of smear positive cases was highest in southern regions in addition to Woldiya (Figure 6.20). Extrapulmonary TB cases were highest in Gonder followed by Jinka and Butajira (Figure 6.21). Overall, there was some agreement between the MoH and local health centre data particularly for smear positive and extrapulmonary TB cases. However there were comparative discrepancies between the regional and local records perhaps, due to the denominator population estimates which may be inaccurate for local health centre records. As the local health centre data is preliminary and the resolution of the MoH data is low, further TB incidence surveys should be carried out in these areas to provide more accurate records.

It was of interest that the highest incidence of extrapulmonary TB was in the region Gonder, as the abundance of mycobacteria in the environment was also highest in Gonder (mean 7.42×10^6 cells per gram). Therefore the environment may be a reservoir of infection and disease. The diversity of mycobacteria (no. of OTUs) was shown to be significantly higher in Bale (section 5.4), however no correlations were observed with the incidence of TB in Bale. The relative proportion of SG pathogenic species was significantly higher in Turmi (section 5.5). This was of interest as smear positive TB cases were higher in southern regions such as Turmi.

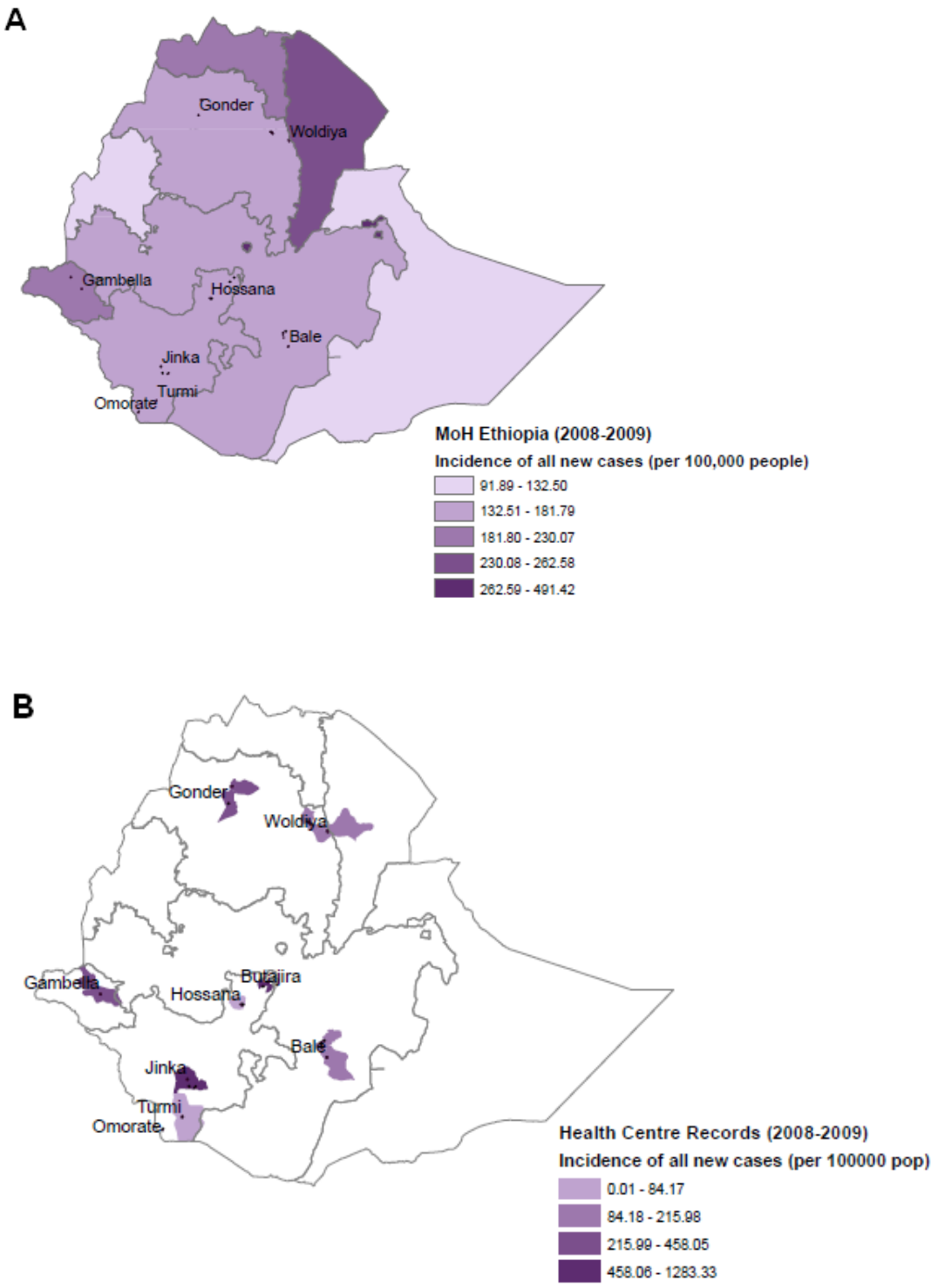


Figure 6.17. Incidence of all new cases between 2008 and 2009 **(A)** MoH **(B)** Health Centre Records.

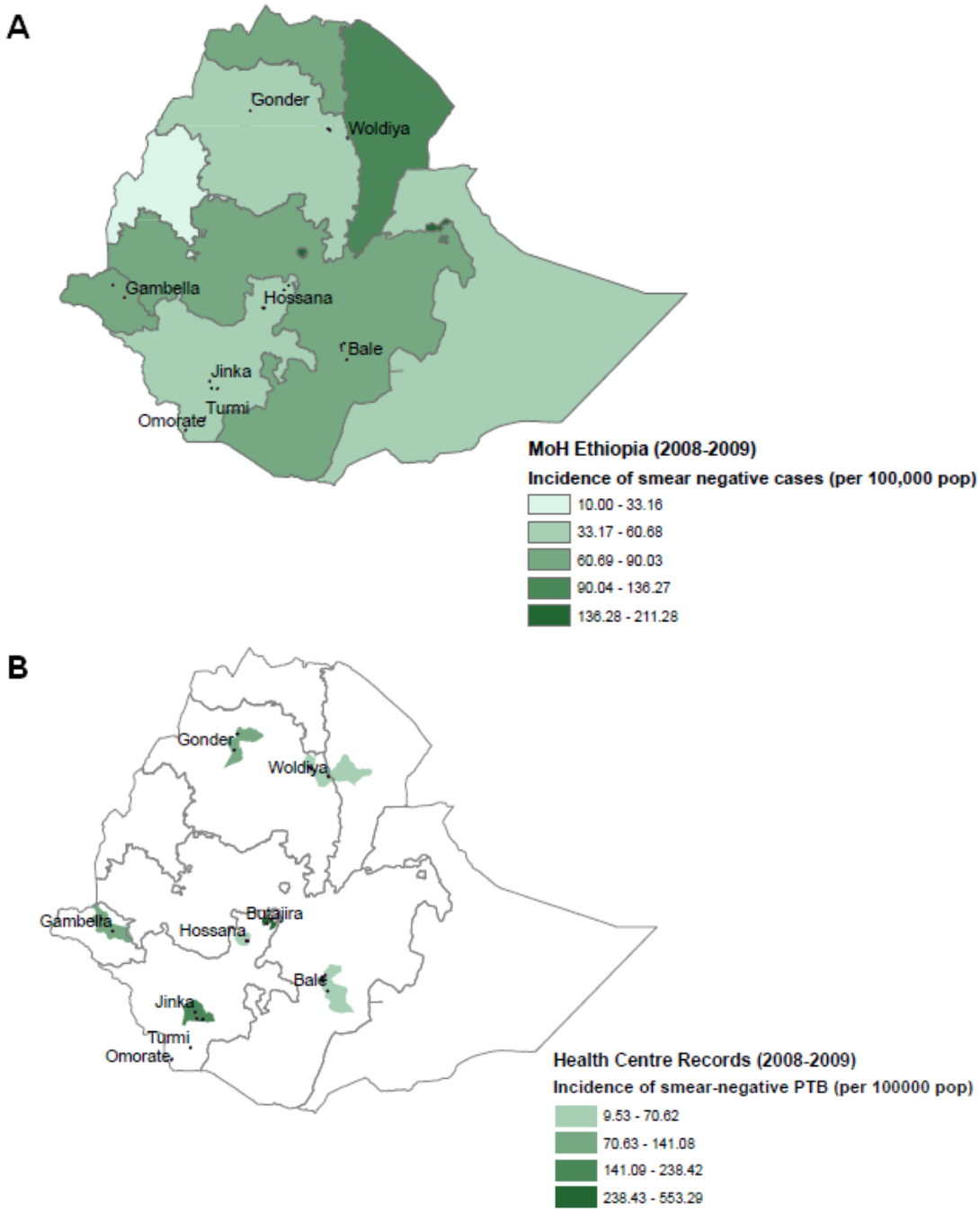


Figure 6.18. Incidence of smear negative cases between 2008 and 2009 **(A)** MoH **(B)** Health Centre Records.

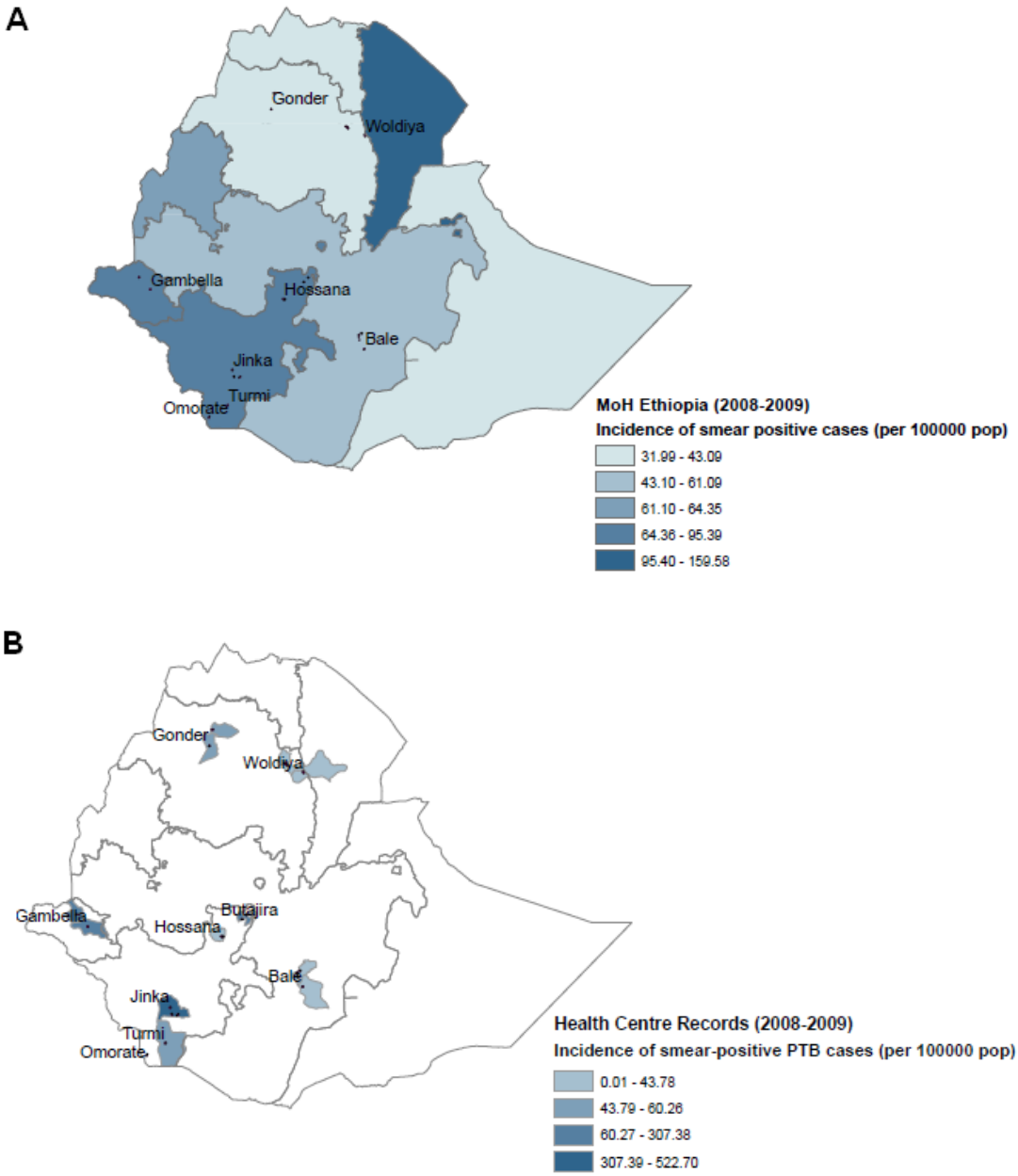


Figure 6.19. Incidence of smear positive cases between 2008 and 2009 **(A)** MoH **(B)** Health Centre Records.

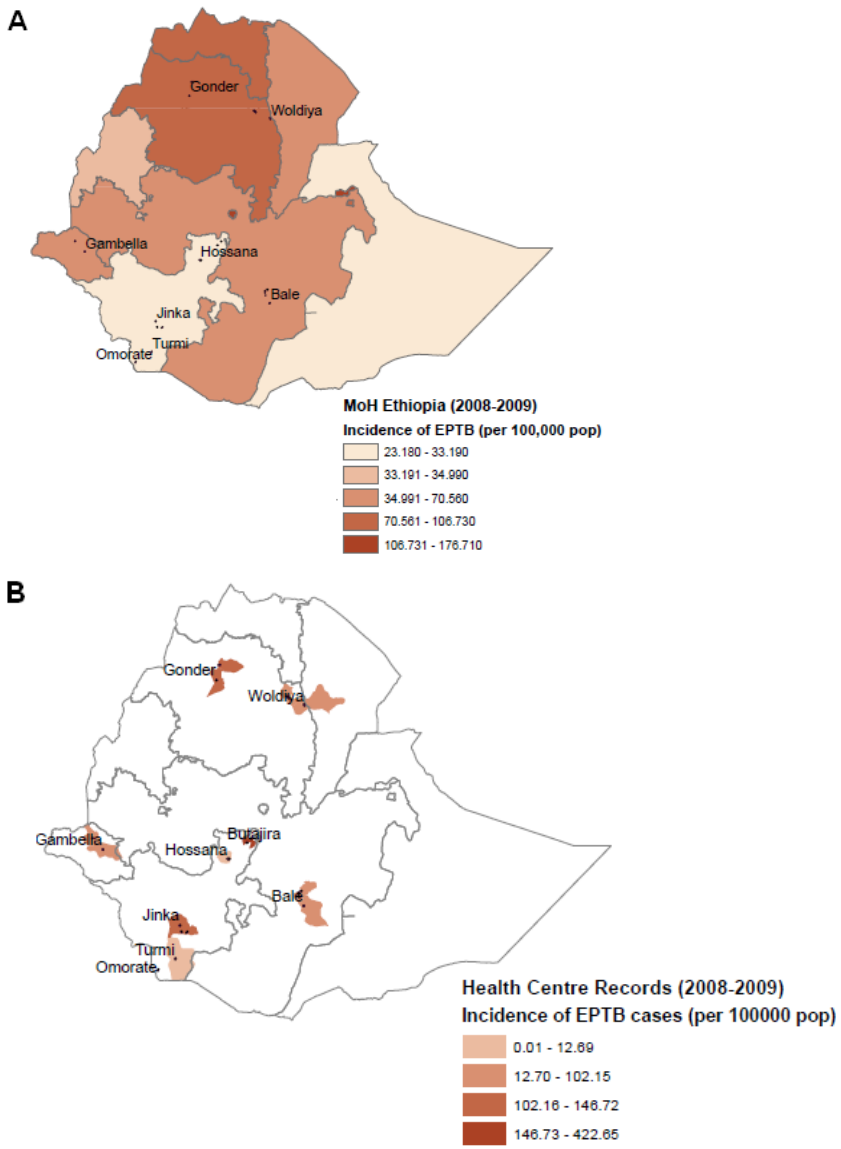


Figure 6.20. Incidence of extrapulmonary TB cases between 2008 and 2009 **(A)** MoH **(B)** Health Centre Records

6.9. Discussion

The quantification of the *Mycobacterium* genus in Ethiopian soil and water samples revealed that mycobacteria are ubiquitous and highly abundant in the environment. Bioclimatic and environmental factors such as elevation, temperature, pH and longitude were significant in explaining differences in the abundance of the *Mycobacterium* genus between regions. The abundance of the MTC complex and *M. bovis* was determined in order to complement the pyrosequencing BLAST results. This is the first study to successfully quantify *M. bovis* in the water environment and

to quantify *M. tuberculosis* complex in soil and water samples. This is of great clinical significance, as soil and water presents a hazardous reservoir for both human and animal infection.

The abundance estimates for *Mycobacterium* genus load varied for both soils and water, ranging from 10^5 to 10^8 cells per gram and 10^1 to 10^5 cells per ml respectively. The abundance values were higher for soil samples compared to water, perhaps because the high organic content of soil promotes the proliferation of *Mycobacterium* species. Total bacterial diversity in soil environments was estimated to be 1.5 $x10^{10}$ cells per gram of dry soil (Torsvik et al. 1990). Therefore this study suggests that *Mycobacterium* species represent approximately 1% of the total bacterial community in soils. Bacterial abundance in water was estimated to be $6x10^7$ cells per cm^{-3} (Torsvik 2002), and therefore *Mycobacterium* species also represent approximately 1% of the total bacterial community in water samples. Few studies have quantified the *Mycobacterium* load in the environment. Estimates similar to this study were detected in boreal soils, with up to 10^7 to^8 cells per gram of soil (Nieminen et al. 2006). Water samples taken from cooling towers were found to have 10^3 to 10^6 cells per litre (Adrados et al. 2011), whereas estimates from coastal estuary waters ranged from 0 to 10^3 per ml (Jacobs et al. 2009). House dust samples have been found to range from 10^4 to 10^6 cells per gram (Torvinen et al. 2010). It is important to note that the results from these studies are not strictly comparable as different targets were used and so efficiencies of qPCR assays will vary. However, these estimates are in agreement with our results, as soil samples had higher *Mycobacterium* species abundance estimates compared to water samples.

These results have demonstrated that bioclimatic and spatial factors may determine the abundance of *Mycobacterium* species in the environment. Specifically, a non-

linear relationship was observed between elevation and *Mycobacterium* genus abundance in soil samples, where mid-elevations had the highest cell counts abundance (R^2 = 0.3). This is in concordance with the non-linear relationship also observed between soil elevation and *Mycobacterium* species richness (R^2 = 0.4) and diversity (R^2 = 0.3) (Section 5.6). Only one published study has correlated *Mycobacterium* species abundance with environmental variables; high nitrogen and turbidity were positively correlated with abundance whereas dissolved oxygen was negatively correlated with abundance (Jacobs et al. 2009). In this study higher pH values and eastern areas were correlated with abundance in water samples but explained little variance. Therefore it may be the case that variables which were not measured, such as nitrogen content and turbidity, were influential factors in explaining *Mycobacterium* species abundance in this study.

The results of the pyrosequencing BLAST results demonstrated that MTC was very common in water samples (83.3% positive). Consistent results were observed for water samples Bu3 and Bu4, as these samples were positive for *M. bovis* and MTC using qPCR and contained BLAST matches to MTC using pyrosequencing methods. As different methods and different gene targets have demonstrated similar findings, this provides further evidence that *M. bovis* was present in these samples. Overall, *M. bovis* was detected in six samples, one of which was a soil sample. The species *M. bovis* was previously detected in soils collected from UK and Ireland, and the highest estimates ranged from 3.6 x 10^3 cells per gram of soil to 5.4x10^6 cells per gram (Sweeney et al. 2007; Young et al. 2005). These soils were collected from badger setts close to TB breakdown areas and demonstrated a potential transmission route between badgers and cattle. No studies thus far have detected and quantified *M. bovis* in the water environment. However, it can be proposed that water sources provide a

potential environmental reservoir, particularly in areas where bovine TB is prevalent. Studies have attempted to isolate *M. bovis* from water samples by cultivation in bovine TB areas, but were only able to cultivate other *Mycobacterium* species such as *M. terrae* (Michel et al. 2007).

The detection of *M. bovis* in the Ethiopian environment and particularly in water sources suggests that it may represent a significant reservoir for livestock, wildlife and humans. Where the infectious dose of *M. bovis* in cattle was shown to be as low as 1 colony forming unit via the intratracheal route (Dean et al. 2005). However the infectious dose is estimated to be as higher via the gastrointestinal route (10^6 cells) (Ashford et al. 2001). Individual cattle prevalence estimates for BTB in Ethiopia have ranged from 4.7% to 13.5% (Ameni et al. 2007; Berg et al. 2009; Fetene & Kebede 2009). Although estimates are relatively low, BTB is widespread throughout Ethiopia in the cattle population. The transmission of BTB from cattle to humans is thought to occur by drinking unpasteurised milk, consumption of raw meat and inhalation of aerosol droplets. It was estimated that in 3.1% of cases *M. bovis* is the causative agent of human TB (Ayele et al. 2004). Transmission of the MTC has also been found to occur from humans to cattle; potential routes include chewed tobacco given to cattle from farmers with TB (Ameni et al. 2011). This study proposes a novel transmission route between livestock and humans, as *M. bovis* positive water sources were shared drinking and bathing sources.

Although members of the MTC were detected using qPCR and pyrosequencing, the pathogenic species *M. tuberculosis* was not detected in the Ethiopian environment. It is acknowledged that the principal route of *M. tuberculosis* transmission is from person to person, and it is possible that transmission may occur via environmental reservoirs as this species has been detected in a variety of livestock and wildlife

species. For example, *M. tuberculosis* has been detected in cattle, mongooses and suricates (Alexander et al. 2002; Ameni et al. 2011; Ocepeck et al. 2005). However, as no *M. tuberculosis* was detected it suggests that the environment is not a significant reservoir. More extensive environmental sampling in TB endemic areas is required in order to confirm this finding.

The incidence of TB in Ethiopia is ranked 8th worldwide for the incidence of all forms of tuberculosis (WHO 2011b). The data collected from the Ministry of Health (MoH) and from local health centre records suggests that TB incidence was highest in the regions Woldiya and Butajira. Ethiopia is known to have of the highest rates of extrapulmonary TB worldwide (36% of all incident cases) (Berg et al. 2009) and this study found the highest incidence of extrapulmonary TB in Gonder. Therefore it was of interest that the highest abundance of *Mycobacterium* species as determined by qPCR was present in soil samples from Gonder. The incidence of extrapulmonary TB was of particular interest as it is has been suggested that the clinical presentation is due to the route of transmission, such as the ingestion and consumption of contaminated products. Cutaneous forms of TB are likely to occur via the infection of broken skin, perhaps via contact from soils or water. It is likely that opportunistic pathogenic *Mycobacterium* species detected in the environment presents a serious hazardous source of infection via these routes of exposure. Further work should aim to gain more comprehensive higher resolution TB incidence data, in order to make stronger correlations.

Chapter 7. Final Discussion

7.1. Overview

This project employed several molecular methods to determine the diversity and distribution of mycobacteria in ecologically diverse areas. Although pyrosequencing has been used extensively in microbial ecology, this is the first time it has been used to study the diversity of mycobacteria in the environment. This comprehensive approach has illustrated the high diversity of mycobacteria present within Ethiopian soil and water environments. Cultivation studies have shown a fraction of the true *Mycobacterium* species diversity. The collection of environmental samples on a countrywide scale and records of environmental variables enabled for the first time a multivariate biogeographical analysis of the *Mycobacterium* species community composition. Finally the quantification of the *Mycobacterium* genus and specifically of *M. bovis* suggests that the environment may be a significant reservoir of disease.

7.2. Diversity and abundance of mycobacteria

Pyrosequencing revealed an unprecedented diversity of both FG and SG compared with cultivation methods and with other molecular methods such as DGGE (Chapter 3). The composition of mycobacteria was varied among the Ethiopian villages, with over 200 different BLAST matches for the *Mycobacterium* genus and over 50 different BLAST matches for the SG (Chapter 4). Different profiles of *Mycobacterium* species were observed between different between environmental source types. A consistent trend revealed communities in soil encompassed more diversity of mycobacteria whereas water samples displayed a dominance of several species. This is in agreement with a cultivation study which found different profiles in different environmental source types and detected more isolates in soil samples compared to water samples (Kamala et al. 1994). In this project, soil and water

communities revealed the presence of similar species however the prevalence of these species were very different. This suggested that *Mycobacterium* species are readily exchanged across environments but that the differential properties of soil and water cause different species to proliferate in these environments. For example soil is known to contain high amounts of organic matter compared to water.

Abundant FG in soil samples included the polycyclic aromatic hydrocarbon (PAH) degraders *M. gilvum* and *M. flavescens*, the clinical isolates *M. insubricum* and *M. moriokaense* and the environmental species *M. neglectum*, *M.* sp. CNJ881 PL04 and *M.* sp. CNJ859 PL04 (Chapter 4). The ample presence of PAH degrading mycobacteria in the soil environment has been presented previously (Miller et al. 2004; Uyttebroek et al. 2006), however the presence of PAH degraders has not been shown in the Ethiopian environment before. The abundant SG in soil samples included the opportunistic pathogens *M. malmoense, M. colombiense* and *M. riyadhense*. In water samples abundant species included the environmental FG *M. aurum* ATCC 23070, *M.* sp. CCBAU 25324 and *M.* sp. MI21.2 and causative agents of bacteraemia *M. neoaurum* and *M. peregrinum*. SG communities in water samples were dominated by *M. gordonae* comprising 45.05% of all sequences. This species has been commonly detected in water distribution systems and swimming pools (Le Dantec et al. 2002a; Leoni et al. 1999; September et al. 2004). Other abundant SG in water samples were similar to those found in soil samples; *M. colombiense, M. riyadhense* and M. *malmoense.* The relatively abundant SG *M. colombiense* is part of the *Mycobacterium avium* complex and has been previously isolated from HIV positive individuals (Murcia et al. 2006). The species *M. riyadhense and M. malmoense* have been shown to cause pulmonary disease similar to that caused by *M. tuberculosis* (van Ingen et al. 2009b). The relatively high abundance of these

opportunistic pathogens in the environment suggests a reservoir of disease exists in the Ethiopian environment.

The environment as a potential reservoir of infection was further demonstrated by the high abundance of *Mycobacterium* species in soil samples which ranged from 9.47 $x10^5$ to 8.01 $x10^8$ cells per gram and in water samples abundance ranged from 3.35 x 10^1 to 2.68 $x10^5$ cells per ml (Chapter 6). The high abundance of *Mycobacterium* species suggests that these species can readily proliferate in the environment and pyrosequencing has shown that this includes many opportunistic pathogens. The primary route of infection for *M. tuberculosis* is via aerosols, where the infectious dose is thought to be <10 bacilli. However, the primary route from the environment is likely to be ingestion, where *M. avium* was shown to have an infectious dose of 10^4-10^7 via ingestion (Rusin et al. 1997). Therefore our abundance counts of mycobacteria suggest infection via ingestion.

The abundance of the etiologic agent of bovine TB *M. bovis* was determined using qPCR in all samples (Chapter 6). As the pyrosequencing of the 16S rRNA gene could not differentiate between members of the *M. tuberculosis* complex, a different target (RD4 scar region) was used to quantify the abundance of *M. bovis* (Sweeney et al. 2007). In the regions Gambella and Butajira *M. bovis* was detected from multiple water samples with estimates as high as $1.51x10^2$ cells per ml. This suggests that the water environment may be a source of infection for cattle and humans. This is the first time that *M. bovis* has been detected in the Ethiopian environment and also the first time *M. bovis* has been detected in water samples. This data suggests a common route of transmission for cattle, humans and wildlife as all share the same drinking and bathing sources.

7.3. Bioclimatic and spatial correlations

The composition, diversity and abundance of *Mycobacterium* species were significantly correlated with bioclimatic factors. This suggests that the differences in distribution of mycobacteria and specifically SG are determined by environmental heterogeneity in concordance with niche theory (De Wit & Bouvier 2006). However as spatial factors were shown to also be significant this suggests that historical events such as dispersal limitation are responsible for mycobacteria communities in specific regions (Martiny et al. 2006). It is likely that both environmental and spatial factors influence the diversity and distribution of *Mycobacterium* species.

The *Mycobacterium* species community composition in soil samples was shown to be significantly correlated with elevation and soil temperature. The elevation/temperature transects demonstrated different *Mycobacterium* species present at low elevation/high temperature sites compared high elevation/low temperature sites independent of geographical proximity. This suggests that the similar environmental conditions as determined by elevation/temperature are an important influence on the *Mycobacterium* species present. For example sequences matched to *M. neglectum*, *M. moriokaense*, *M. kansasii* and *M. bohemicum* were relatively more abundant at higher elevations/lower temperatures and *M. insubricum*, *M. flavescens* and *M. novocastrense* were relatively more abundant at lower elevations/higher temperatures. The changes in community composition were observed for the elevation/temperature transects of Bale and Woldiya, which suggests that elevation/temperature differences were more influential than spatial proximity.

The phylogenetic diversity, OTU diversity and richness displayed a significant non-linear relationship with elevation (Chapter 5). The abundance of *Mycobacterium* species as determined by qPCR also revealed a non-linear relationship with a peak mid-elevation over the range of 425.4-3939.8 metres (Chapter 6). The peak in diversity and abundance of *Mycobacterium* species at mid-elevations may be due to the moderate environmental conditions compared to the extreme conditions at high and low elevations/temperatures. This non-linear trend has been observed for plants and animal communities (Bhattarai & Vetaas 2003; Bryant et al. 2008; Richardson et al. 2000). However, this is the first time a non-linear trend has been observed for a bacterial community.

The SG in soil samples displayed different trends, a number of bioclimatic and environmental factors were responsible for the differences in community composition. Longitude was consistently a significant factor for distribution and diversity of SG (Chapter 5). For example sequences matched to *M. malmoense* and *M. colombiense* were more prevalent in the eastern regions of Ethiopia and *M. aemonae*, *M. intracellulare* and *M. gastri* were relatively more abundant in samples from the west. Spatial significance may be due to effective immigration where the profiles of SG are characteristically different in different regions. However, as longitude was correlated with increased elevation and lower temperatures (section 5.3.1), it is possible that it was a proxy for these factors. Spatial trends were also were observed for the community composition and diversity for all *Mycobacterium* species and specifically the SG in water samples. Differences in composition were observed with latitude and the diversity of mycobacteria was highest in southern regions of Ethiopia (Chapter 5). Specifically, *M. gordonae* and *M. peregrinum* were prevalent in the north and *M. neoaurum*, *M. colombiense* and *M. malmoense* were

prevalent in the South. These results suggest *Mycobacterium* species communities are determined at the regional level in water samples, perhaps due to limited dispersal of species on a countrywide scale.

7.4. Clinical implications of findings

The results have important implications on the prevalence of mycobacterial infection and disease in these areas. Overall, for all samples Bale had significantly higher richness estimates (no. of OTUs) for all *Mycobacterium* species (Chapter 5). The fraction of OTU richness for SG was highest in Jinka. The abundance of *Mycobacterium* species as determined by qPCR was significantly highest in Gonder (Chapter 6). Therefore the regions Bale, Jinka and Gonder are high risk areas for *Mycobacterium* species exposure and subsequently infection and disease. This correlates well with the MoH data and the records from local health centres which demonstrated that extrapulmonary TB cases were highest in Gonder and Jinka (Chapter 6). In addition the number of nontuberculous isolates in the cattle population in Ethiopia was shown to be highest in Jinka (Berg et al. 2009). It is plausible that many of the extrapulmonary TB cases are caused by *Mycobacterium* species in the environment. This is because as the route of transmission for extrapulmonary TB is likely to be through consumption of contaminated products or from contact with broken skin. However, *Mycobacterium* species can also cause pulmonary disease but differentiation from *Mycobacterium tuberculosi*s is problematic. Previous studies have shown a correlation between *Mycobacterium* species present in the potting soils and household plumbing and patients with mycobacterial disease (Falkinham 2011; De Groote et al. 2006). This provides further evidence between the presence of *Mycobacterium* species in the environment and presents a hazardous reservoir of infection.

The high diversity and abundance of *Mycobacterium* species in the Ethiopian environment sparks further interest as to the role of *Mycobacterium* species in blocking or masking BCG efficacy. The BCG coverage for Ethiopia was estimated to be 69% in 2010 (WHO 2011a), however the efficacy of BCG in Ethiopia is not known. Similar lower latitude countries with a similar demographic and climates have shown low BCG efficacy (Wilson et al. 1995). Due to the high rates of tuberculosis in Ethiopia and the reported high coverage of BCG it is likely that BCG provides little to no protection. It is conceivable that high exposure to *Mycobacterium* species in the Ethiopian environment may contribute to low BCG efficacy. Sensitisation studies in Malawi have shown a high sensitisation to *M. intracellulare* and *M. scrofulaceum* (Fine et al. 2001). Therefore it was interesting that this study found a high prevalence of the *Mycobacterium avium* complex in soils and water samples (28.2% and 16.2% of all SG BLAST matches respectively) (Chapter 4). Therefore it is likely that members of the *Mycobacterium avium* complex are prevalent in the environment and so individuals are likely to be most exposed to these species.

The livestock population of Ethiopia is one of the largest worldwide and conservative bovine tuberculosis prevalence is approximately 3% with highest estimates in Meskan Mareko (Central Ethiopia) (Tschopp et al. 2009). The regions Butajira and Gambella were of particular interest as *M. bovis* was detected in this study in two out of four of the village water samples (Chapter 6). This correlated with prevalence of BTB in Butajira which was shown to be relatively high (6.8%) (Tschopp et al. 2011, 2009). Therefore the water environment presents a route of transmission between livestock and between the human and wildlife populations in these areas.

7.5. Future Work

This study aimed to characterise the diversity and abundance of mycobacteria in the environment across nine different regions in Ethiopia. It would be of interest to assess the human sensitisation to mycobacteria in these areas. Using IFN-γ whole blood assays specific to certain *Mycobacterium* species it may be possible to correlate exposure to prevalent species in the environment. In addition BCG efficacy trials carried in two regions of Ethiopia with high and low exposure to mycobacteria would perhaps provide evidence to support this association. The determination of mycobacteria diversity and prevalence in additional geographical locations with known BCG efficacy estimates would provide additional evidence to support this hypothesis.

The spatial scale of this study has shown both environmental heterogeneity and geographical factors influence diversity, composition and abundance of mycobacteria. Further experiments should carry out replicates from the same geographical location for pyrosequencing as replicates from the same sampling site should demonstrate a high similarity in community structure. It would be of interest to assess the temporal changes of mycobacteria communities in the environment. Other gene markers such as *rpoB, gyrB* and 16S-23S internal transcriber spacer region should be used in future experiments to confirm findings related to diversity and prevalence of specific *Mycobacterium* species. Finally the diversity of *Mycobacterium* species RNA would provide insight into the diversity of the active microbial community. These methods would provide a better understanding of the diversity of mycobacteria in the environment.

References

Adekambi, T., Berger, P., Raoult, D., & Drancourt, M. (2006a). *rpoB* Gene Sequence-Based Characterization of Emerging Non-Tuberculous Mycobacteria with Descriptions of *Mycobacterium bolletii* sp. nov., *Mycobacterium phocaicum* sp. nov. and *Mycobacterium aubagnense* sp. nov. *International Journal of Systematic and Evolutionary Microbiology,* **56**, 133-43.

Adekambi, T., Salah, B. I., Khlif, M., Raoult, D., & Drancourt, M. (2006b). Survival of Environmental Mycobacteria in *Acanthamoeba polyphaga*. *Applied and Environmental Microbiology* **72,** 5974-5981.

Adrados, B., Julián, E., Codony, F., Torrents, E., Luquin, M., & Morató, J. (2011). Prevalence and Concentration of Non-Tuberculous Mycobacteria in Cooling Towers by Means of Quantitative PCR: a Prospective Study. *Current Microbiology*, **62**, 313-319.

Alexander, K. A., Pleydell, E., Williams, M. C., Lane, E. P., Nyange, J. F. C., & Michel, A. L. (2002). *Mycobacterium tuberculosis*: An Emerging Disease of Free-Ranging Wildlife. *Research in Veterinary Science*, **8**, 6-9.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, **215**, 403-410

Amann, R. I., Ludwig, W., & Schleifer, K. H. (1995). Phylogenetic Identification and *in-situ* Detection of Individual Microbial Cells Without Cultivation. *Microbiological Reviews*, **59**, 143-169.

Ameni, G., Aseffa, A., Engers, H., Young, D., Hewinson, G., & Vordermeier, M. (2006). Cattle Husbandry in Ethiopia is a Predominant Factor Affecting the Pathology of Bovine Tuberculosis and Gamma Interferon Responses to Mycobacterial Antigens. *Clinical and Vaccine Immunology*, **13**, 1030-1036.

Ameni, G., Aseffa, A., Engers, H., Young, D., Gordon, S. V., Hewinson, G., & Vordermeier, M. (2007). High Prevalence and Increased Severity of Pathology of Bovine Tuberculosis in Holsteins Compared to Zebu Breeds Under Field Cattle Husbandry in Central Ethiopia. *Clinical and Vaccine Immunology,* **14**, 1356-1361.

Ameni, G., Vordermeier, M., Firdessa, R., Aseffa, A., Hewinson, G., Gordon, S. V., & Berg, S. (2011). *Mycobacterium tuberculosis* Infection in Grazing Cattle in Central Ethiopia. *Veterinary Journal*, **188**, 359-361.

Anonymous. (1979). Trial of BCG Vaccines in South India for Tuberculosis Prevention: First Report. *Bulletin of the World Health Organization*, **57**, 819-827.

Ansorge, W. J. (2009). Next-Generation DNA Sequencing Techniques. *New Biotechnology*, **25**, 195-203.

Aranaz, A., Juan, L. D., Montero, N., Sanchez,C., Galka, M., Delso, C., Álvarez, J., Romero, B., Bezos, J., Vela, I A., Briones, V., Mateous, A., Dominguez, L (2004). Bovine Tuberculosis (*Mycobacterium bovis*) in Wildlife in Spain. *Journal of Clinical Microbiology*, **42**, 2602-2608.

Arya, M., Shergill, I. S., Williamson, M., Gommersall, L., Arya, N., & Patel, H. R. H. (2005). Basic Principles of Real-Time Quantitative PCR. *Expert Review of Molecular Diagnostics*, **5,** 209-219.

Ashford, D. A, Whitney, E., Raghunathan, P., & Cosivi, O. (2001). Epidemiology of Selected Mycobacteria That Infect Humans and Other Animals. *Revue Scientifique Et Technique*, **20,** 325-337.

Ayele, W. Y., Neill, S. D., Zinsstag, J., Weiss, M. G., & Pavlik, I. (2004). Bovine Tuberculosis: An Old Disease but a New Threat to Africa. *The International Journal of Tuberculosis and Lung Disease*, **8**, 924-937.

Barry, C. E., Lee, R. E., Mdluli, K., Sampson, A. E., Schroeder, B. G., Slayden, R. A., & Yuan, Y. (1998). Mycolic Acids: Structure, Biosynthesis and Physiological Functions. *Progress in Lipid Research*, **37**, 143-179.

Batley, J., & Edwards, D. (2009). Genome Sequence Data: Management, Storage, and Visualization. *BioTechniques*, **46**, 333- 335.

Beard, P. M., Daniels, M. J., Henderson, D., Pirie, A., Rudge, K., Buxton, D., Rhind, S., Grieg, A., Hutchins, M. R., McKendrink, I., Stevenson, K., Sharp, J. M (2001). Paratuberculosis Infection of Nonruminant Wildlife in Scotland. *Journal of Clinical Microbiology*, **39**, 1517-1521.

Behnke, A., Engel, M., Christen, R., Nebel, M., Klein, R. R., & Stoeck, T. (2011). Depicting More Accurate Pictures of Protistan Community Complexity Using Pyrosequencing of Hypervariable SSU rRNA Gene Regions. *Environmental Microbiology*, **13**, 340-349.

Benson-Mitchell, R., & Buchanan, G. (1996). Cervical Lymphadenopathy Secondary to Atypical Mycobacteria in Children. *The Journal of Laryngology and Otology*, **110**, 48-51.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Black, P. J (2008). Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry. *Nature*, **456**, 53-59

Berg, S, Firdessa, R., Habtamu, M., Gadisa, E., Mengistu, A., Yamuah, L., Ameni, G., et al. Vordermeier, M., Robertson, B. D, Smith, N. H., Engers, H., Young, D., Hewinson, G, Abraham Aseffa, A., Gordon, S. V (2009). The Burden of Mycobacterial Disease in Ethiopian Cattle: Implications for Public Health. *PloS one*, **4**, 1-8.

Bhattarai, K. R., & Vetaas, O. R. (2003). Variation in Plant Species Richness of Different Life Forms Along a Subtropical Elevation Gradient in the Himalayas, East Nepal. *Global Ecology and Biogeography*, **12**, 327-340.

Black, G. F., Dockrell, H. M., Crampin, a C., Floyd, S., Weir, R. E., Bliss, L., Sichali, L., Fine, P. E. M (2001). Patterns and Implications of Naturally Acquired Immune Responses to Environmental and Tuberculous Mycobacterial Antigens in Northern Malawi. *The Journal of Infectious Diseases*, **184**, 322-329.

Black, G. F., Weir, R. E., Floyd, S., Bliss, L., Warndorff, D. K., Crampin, A. C., Ngwira, B., Dockrell, H. M (2002). BCG-Induced Increase in Interferon-Gamma Response to Mycobacterial Antigens and Efficacy of BCG Vaccination in Malawi and the UK: Two Randomised Controlled Studies. *The Lancet*, **359**, 1393-1401.

Blackwell, J. M., Barton, H. C., White, J. K., Roach, T. I. A., Shaw, M.-anne, Whitehead, S. H., Mock, B. A., Baker, A. M. (1994). Genetic Regulation of Leishmanial and Mycobacterial Infections: the *Lsh/Ity/Bcg* Gene Story Continues. *Immunloogical Letters*, **43**, 99-107.

Bland, C. S., Ireland, J. M., Lozano, E., Alvarez, M. E., & Primm, T. P. (2005). Mycobacterial Ecology of the Rio Grande. *Applied and Environmental Microbiology*, **71**, 5719-5727.

Boddinghaus, B., Rogall, T., Flohr, T., Blöcker, H., & Böttger, E. C. (1990). Detection and Identification of Mycobacteria by Amplification of rRNA. *Journal of Clinical Microbiology*, **28**, 1751-1759.

Bohannan, B. J., & Hughes, J. (2003). New Approaches to Analyzing Microbial Biodiversity Data. *Current Opinion in Microbiology*, **6**, 282-287.

Bojalil, L. F., Cerbon, J., & Trujillo, A. (1962). Adansonian Classification of Mycobacteria. *Journal of General Microbiology*, **28**, 333-346.

Bonah, C. (2005). The "Experimental Stable" of the BCG Vaccine: Safety, Efficacy, Proof and Standards, 1921-1933. *Studies in History and Philosophy of Biological and Biomedical Sciences*, **36,** 696-721.

Brandt, L., Cunha, J. F., Olsen, A. W., Chilima, B. Z., Hirsch, P., Appelberg, R., & Andersen, P. (2002). Failure of the *Mycobacterium bovis* BCG Vaccine: Some Species of Environmental Mycobacteria Block Multiplication of BCG and Induction of Protective Immunity to Tuberculosis. *Infection and Immunity*, **70**, 672-678.

Brennan, P. J., & Nikaido, H. (1995). The Envelope of Mycobacteria. *Annual Review of Biochemistry*, **64**, 29-63.

Brewer, T. F. (2000). Preventing Tuberculosis with Bacillus Calmette-Guérin Vaccine: A Meta-Analysis of the Literature. *Clinical Infectious Diseases*, **31**, S64-S67.

Britton, Warwick J, & Lockwood, D. N. J. (2004). Leprosy. *The Lancet*, **363**, 1209-1219.

Brookes, J. D., Antenucci, J., Hipsey, M., Burch, M. D., Ashbolt, N. J., & Ferguson, C. (2004). Fate and Transport of Pathogens in Lakes and Reservoirs. *Environment International*, **30**, 741-759.

Brooks, R. W., Parker, B. C., Gruft, H., & Falkinham, J. O. (1984). Epidemiology of Infection by Nontuberculous Mycobacteria. *American Review of Respiratory Disease*, **130**, 630-633.

Brosch, R, Gordon, S. V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., Garnier, T., Cole, S. T (2002). A New Evolutionary Scenario for the *Mycobacterium tuberculosis* Complex. *Proceedings of the National Academy of Science of the USA*, **99**, 3684-3689.

Brosch, Roland, Gordon, S. V., Garnier, T., Eiglmeier, K., Frigui, W., Valenti, P., Dos Santos, S., Cole, S. T (2007). Genome Plasticity of BCG and Impact on Vaccine Efficacy. *Proceedings of the National Academy of Science of the USA*, **104**, 5596-5601.

Brown-Elliott, B. A., & Wallace, R. J. (2002). Clinical and Taxonomic Status of Pathogenic Nonpigmented or Late-Pigmenting Rapidly Growing Mycobacteria. *Clinical Microbiology Reviews*, **15,** 716-746.

Bryant, J. A., Lamanna, C., Morlon, H., Kerkhoff, A. J., Enquist, B. J., & Green, J. L. (2008). Microbes on Mountainsides: Contrasting Elevational Patterns of Bacterial and Plant Diversity. *Proceedings of the National Academy of Science of the USA*, **105**, 11505-11511.

Buddle, M. B., Wards, B. J., Aldwell, F. E., Collins, D. M., De Lisle, G. W. (2002) Influence of Sensitisation to Environmental Mycobacteria on Subsequent Vaccination Against Bovine Tuberculosis. *Vaccine*, **20,** 1126-1133

Butler, W. R., Jost, K. C., & Kilburn, J. O. (1991). Identification of Mycobacteria by High-Performance Liquid Chromatography. *Journal of Clinical Microbiology*, **29,** 2468-2472.

Caldwell, M. E. (1925). Viability of *Mycobacterium Tuberculosis* in a Semi-Arid Environment. *The Journal of Infectious Diseases*, **37**, 465-472.

Calvo-Bado, L.A, Oakley, B. B., Dowd, S. E., Green, L. E., Medley, G. F., Ul-Hassan, A., Bateman, V., Wellington, E. M. H (2011). Ovine Pedomics: The First Study of the Ovine Foot 16S rRNA-Based Microbiome. *The ISME Journal*, **5**, 1-12.

Caporaso, J. G., Bittinger, K., Bushman, F. D., DeSantis, T. Z., Andersen, G. L., & Knight, R. (2010a). PyNAST: A Flexible Tool for Aligning Sequences to a Template Alignment. *Bioinformatics*, **26**, 266-267.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Knight, R (2010b). QIIME Allows Analysis of High-Throughput Community Sequencing Data. *Nature Methods*, **7**, 335-336.

Cayer, M-P., Veillette, M., Pageau, P., Hamelin, R., Bergeron, M-J., Mériaux, A., Cormier, Y., Duchaine, C (2007). Identification of Mycobacteria in Peat Moss Processing Plants: Application of Molecular Biology Approaches. *Canadian Journal of Microbiology*, **99**, 92-99.

Cayrou, C., Turenne, C., Behr, M. A., & Drancourt, M. (2010). Genotyping of *Mycobacterium avium* Complex Organisms Using Multispacer Sequence Typing. *Microbiology*, **156,** 687-694.

Chan, O. C., Wolf, M., Hepperle, D., & Casper, P. (2002). Methanogenic Archaeal Community in the Sediment of an Artificially Partitioned Acidic Bog Lake. *FEMS Microbiology Ecology*, **42,** 119-129.

Chang, C., Wang, L., Liao, C., & Huang, S. P. (2002). Identification of Nontuberculous Mycobacteria Existing in Tap Water by PCR-Restriction Fragment Length Polymorphism. *Applied and Environmental Microbiology*, **68,** 3159-3161.

Chave, J. (2004). Neutral Theory and Community Ecology. *Ecology Letters*, **7**, 241-253.

Chaves, D., Sandoval, A., Rodríguez, L., García, J. C., Restrepo, S., & Zambrano, M. M. (2010). Comparative Analysis of Six *Mycobacterium tuberculosis* Complex Genomes. *Biomedica*, **30**, 23-31.

Chilima, B. Z., Clark, I. M., Floyd, S., Fine, P. E. M., & Hirsch, P. R. (2006). Distribution of Environmental Mycobacteria in Karonga District, Northern Malawi. *Applied and Environmental Microbiology*, **72**, 2343-2350.

Chimara, E., Ferrazoli, L., & Leão, S. C. (2004). *Mycobacterium tuberculosis* Complex Differentiation Using *gyrB*-Restriction Fragment Length Polymorphism Analysis. *Memórias do Instituto Oswaldo Cruz*, **99**, 745-748.

Cirillo, J D, Falkow, S., Tompkins, L. S., & Bermudez, L. E. (1997). Interaction of *Mycobacterium avium* With Environmental Amoebae Enhances Virulence. *Infection and Immunity*, **65**, 3759-3767.

Cohan, F. M. (2002). What Are Bacterial Species? *Annual Review of Microbiology*, **56**, 457-487.

Colditz, G. A., Brewer, T. F., Berkey, C. S., Burdick, E., Fineberg, H. V., & Mosteller, F. (1994). Vaccine in the Prevention of Tuberculosis Efficacy of BCG. *The Journal of the American Medical Association*, **271,** 698-702.

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, a S., Tiedje, J. M. (2009). The Ribosomal Database Project:

Improved Alignments and New Tools for rRNA Analysis. *Nucleic Acids Research*, **37,** D141-D145.

Collins, C. H. (1985). *Mycobacterium marinum* Infections in Man. *Journal of Hygeine*, **94**, 135-149.

Collins, F. M. (1989). Mycobacterial Disease, Immunosuppression, and Acquired Immunodeficiency Syndrome. *Clinical Microbiology Reviews*, **2**, 360-77.

Converse, P. J. (2000). Dual Infection: The Challenge of HIV/AIDS and Tuberculosis in Ethiopia. *Northeast African Studies*, **7**, 147-165.

Cook, J. L. (2010). Nontuberculous Mycobacteria: Opportunistic Environmental Pathogens for Predisposed Hosts. *British Medical Bulletin*, **96**, 45-59.

Coppenraet, E. S. B. V., Lindeboom, J. A., Prins, J. M., Peeters, M. F., Claas, E. C. J., & Kuijper, E. J. (2004). Real-Time PCR Assay Using Fine-Needle Aspirates and Tissue Biopsy Specimens for Rapid Diagnosis of Mycobacterial Lymphadenitis in Children. *Journal of Clinical Microbiology*, **42**, 2644-2650.

Corner, L. A. L. (2011). Natural Transmission of *Mycobacterium bovis* Infection in Captive Brushtail Possums (*Trichosurus vulpecula*). *New Zealand Veterinary Journal*, **50**, 37-41.

Cosivi, O., Grange, J. M., Daborn, C. J., Raviglione, M. C., Fujikura, T., Cousins, D., Robinson, R. A., Meslin, F. (1999). Zoonotic Tuberculosis due to *Mycobacterium bovis* in Developing Countries. *Emerging Infectious Diseases*, **4**, 59-70.

Covert, T. C., Rodgers, M. R., Reyes, a L., & Stelma, G. N. (1999). Occurrence of Nontuberculous Mycobacteria in Environmental Samples. *Applied and Environmental Microbiology*, **65**, 2492-2496.

Crampin, A. C., Glynn, J. R., & Fine, P. E. M. (2009). What has Karonga taught us? Tuberculosis studied over three decades. *International Journal of Tuberculosis and Lung Disease*, **13**, 153-164.

Crump, J. A., Van Ingen, J., Morrissey, A. B., Boeree, M. J., Mavura, D. R., Swai, B., Thielman, N. M., Van Soolingen, D. (2009). Invasive Disease Caused by Nontuberculous Mycobacteria, Tanzania. *Emerging Infectious Diseases*, **15**, 53-55.

Curtis, T P, Sloan, W. T., & Scannell, J. W. (2002). Estimating Prokaryotic Diversity and its Limits. *Proceedings of the National Academy of Science of the USA*, **99**, 10494-10499.

Dailloux, M., Laurain, C., Weber, M., & Hartermann, P. (1999). Water and Nontuberculous Mycobacteria. *Water Research*, **33**, 2219-2228.

Danelishvili, L., Wu, M., Stang, B., Harriff, M., Cirillo, S., Cirillo, J. D., Bildfell, R., Bermudez, L. E. (2007). Identification of *Mycobacterium avium* Pathogenicity

Island Important for Macrophage and Amoeba Infection. *Proceedings of the National Academy of Science of the USA*, **104**, 11038-11043.

De Groote, M. A., Pace, N. R., Fulton, K., & Falkinham, J. O. (2006). Relationships Between *Mycobacterium* Isolates From Patients With Pulmonary Mycobacterial Infection and Potting Soils. *Applied and Environmental Microbiology*, **72**, 7602-7606.

De Santis, T. Z., Hugenholtz, P., Keller, K., Brodie, E. L., Larsen, N., Piceno, Y. M., Phan, R., Andersen, G. L. (2006). NAST: A Multiple Sequence Alignment Server for Comparative Analysis of 16S rRNA Genes. *Nucleic Acids Research*, **34,** W394-W399.

De Wit, R., & Bouvier, T. (2006). "Everything is Everywhere, But, the Environment Selects"; What Did Baas Becking and Beijerinck Really Say? *Environmental Microbiology*, **8**, 755-758.

Dean, G. S., Rhodes, S. G., Coad, M., Whelan, A. O., Cockle, P. J., Clifford, D. J., Hewinson, R. G., Vordemeier, M. (2005). Minimum Infective Dose of *Mycobacterium bovis* in Cattle. *Infection and Immunity*, **73**, 6467-6471.

Dean-Ross, D., & Cerniglia, C. E. (1996). Degradation of Pyrene by *Mycobacterium flavescens*. *Applied Microbiology and Biotechnology*, **46**, 307-312.

Demangel, C., Garnier, T., Rosenkrands, I., & Cole, S. T. (2005). Differential Effects of Prior Exposure to Environmental Mycobacteria on Vaccination with *Mycobacterium bovis* BCG or a Recombinant BCG Strain Expressing RD1 Antigens. *Infection and Immunity*, **73**, 2190-2196.

Devane, M. L., Nicol, C., Ball, A., Klena, J. D., Scholes, P., Hudson, J. a, Baker, M. G., Savill, M. G. (2005). The Occurrence of *Campylobacter* Subtypes in Environmental Reservoirs and Potential Transmission Routes. *Journal of Applied Microbiology*, **98**, 980-990.

DiGiuseppe Champion, P. A., & Cox, J. S. (2007). Protein Secretion Systems in Mycobacteria. *Cellular Microbiology*, **9**, 1376-1384.

Dobner, P., Feldmann, K., Rifai, M., Löscher, T., & Rinder, H. (1996). Rapid Identification of Mycobacterial Species by PCR Amplification of Hypervariable 16S rRNA Gene Promoter Region. *Journal of Clinical Microbiology*, **34**, 866-869.

Donnelly, C. A, Woodroffe, R., Cox, D. R., Bourne, F. J., Cheeseman, C. L., Clifton-Hadley, R. S., Wei, G., Morrison, W. I. (2006). Positive and Negative Effects of Widespread Badger Culling on Tuberculosis in Cattle. *Nature*, **439**, 843-846.

Donoghue, H. D., Overend, E., & Stanford, J. L. (1997). A Longitudinal Study of Environmental Mycobacteria on a Farm in South-West England. *Journal of Applied Microbiology*, **82**, 57-67.

Dorman, S. E., & Holland, S. M. (1998). Mutation in the Signal-Transducing Chain of the Interferon-Gamma Receptor and Susceptibility to Mycobacterial Infection. *The Journal of Clinical Investigation*, **101**, 2364-2369.

Dumbrell, A. J., Nelson, M., Helgason, T., Dytham, C., & Fitter, A. H. (2010). Relative Roles of Niche and Neutral Processes in Structuring a Soil Microbial Community. *The ISME journal*, **4**, 337-345.

Eaton, T., Falkinham, J. O., Asiu, T. O., & Daniel, T. M. (1995). Isolation and Characteristics of *Mycobacterium avium* Complex From Water and Soil Samples in Uganda. *Tubercle and Lung Disease*, **76**, 570-574.

Eddyani, M., De Jonckheere, J. F., Durnez, L., Suykerbuyk, P., Leirs, H., & Portaels, F. (2008). Occurrence of Free-Living Amoebae in Communities of Low and High Endemicity for Buruli Ulcer in Southern Benin. *Applied and Environmental Microbiology*, **74**, 6547-6553.

Edgar, R. C. (2010). Search and Clustering Orders of Magnitude Faster Than BLAST. *Bioinformatics*, **26**, 2460-2461.

Edgar, R. C. (2011a). OTUPipe. Unpublished. Retrieved January 2012, from http://drive5.com/otupipe/

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011b). UCHIME Improves Sensitivity and Speed of Chimera Detection. *Bioinformatics*, **27**, 2194-2200.

Edwards, L. B., & Tolderlund, K. (1952). BCG Vaccine Studies. III. Preliminary Report on Effect of Sunlight and BCG Vaccine. *Bulletin of the World Health Organization*, **5**, 245-248.

Edwards, L. B., Acquaviva, F. A., Livesay, V. T., Cross, F. W., & Palmer, C. E. (1969). An Atlas of Sensitivity to Tuberculin, PPD-B, and Histoplasmin in the United States. *American Review of Respiratory Disease*, **99**, 1-65.

Edwards, U., Rogall, T., Blocker, H., Emde, M., & Bottger, Eric, C. (1989). Isolation and Direct Complete Nucleotide Determination of Entire Genes. Characterization of a Gene Coding for 16S Ribosomal RNA. *Nucleic acids research*, **17**, 7843-7853.

Ehrt, S., & Schnappinger, D. (2007). *Mycobacterium tuberculosis* Virulence: Lipids Inside and Out. *Nature Medicine*, **13**, 284-285.

Elias, D., Mengistu, G., Akuffo, H., & Britton, S. (2006). Are Intestinal Helminths Risk Factors For Developing Active Tuberculosis? *Tropical Medicine & International Health*, **11**, 551-558.

Engelbrektson, A., Kunin, V., Wrighton, K. C., Zvenigorodsky, N., Chen, F., Ochman, H., & Hugenholtz, P. (2010). Experimental Factors Affecting PCR-Based

Estimates of Microbial Species Richness and Evenness. *The ISME journal*, **4**, 642-647.

Euzeby, J. P. (1997). List of Prokaryotic Names with Standing in Nomenclature. *International Journal of Systematic Bacteriology*. **47**, 590-592

Faegri, A., Torsvik, V., & Goksoyr, J. (1977). Bacterial and Fungal Activities in Soil: Seperation of Bacteria and Fungi by Rapid Fractionated Centrifugation Technique. *Soil Biology and Biochemistry*, **9,** 105-112.

Faith, D. P. (1992). Conservation Evaluation and Phylogenetic Diversity. *Biological Conservation*, **61**, 1-10.

Falkinham, J. O. (1980). Epidemiology of Infection by Nontuberculous Mycobacteria .I. Geographic Distribution in the Eastern-United States. *American Review of Respiratory Disease*, **121**, 931-937.

Falkinham, J. O. (1996). Epidemiology of Infection by Nontuberculous Mycobacteria. *Clinical Microbiology Reviews*, **9**, 177-215.

Falkinham, J. O. (2009a). The Biology of Environmental Mycobacteria. *Environmental Microbiology Reports*, **1**, 477-487.

Falkinham, J. O. (2009b). Surrounded by Mycobacteria: Nontuberculous mycobacteria in the human environment. *Journal of Applied Microbiology*, **107**, 356-367.

Falkinham, J. O. (2011). Nontuberculous Mycobacteria from Household Plumbing of Patients with Nontuberculous Mycobacteria Disease. *Emerging Infectious Diseases*, **17**, 419-424.

Fenchel, T. (2003). Biogeography for Bacteria. *Science*, **301**, 925-926.

Fenchel, T., & Finlay, B. J. (2003). Is Microbial Diversity Fundamentally Different from Biodiversity of Larger Animals and Plants? *European Journal of Protistology*, **490,** 486-490.

Fenchel, T., & Finlay, B. J. (2004). The Ubiquity of Small Species: Patterns of Local and Global Diversity. *BioScience*, **54**, 777-784.

Ferris, M. J., Muyzer, G., & Ward, D. M. (1996). Denaturing Gradient Gel Electrophoresis Profiles of 16S rRNA-Defined Populations Inhabiting a Hot Spring Microbial Mat Community. *Applied and Environmental Microbiology*, **62**, 340-346.

Fetene, T., & Kebede, N. (2009). Bovine Tuberculosis of Cattle in Three Districts of Northwestern Ethiopia. *Tropical Animal Health and Production*, **41**, 273-277.

Fierer, N. (2008). Microbial Biogeography: Patterns in Microbial Diversity Across Space and Time. *In Accessing Uncultivated Microorganisms: from the Environment*

to Organisms and Genomes and Back. pp. 95-115. Edited by Zengler, K., ASM Press, Washington DC.

Fierer, N., & Jackson, R. B. (2006). The Diversity and Biogeography of Soil Bacterial Communities. *Proceedings of the National Academy of Science of the USA,* **103**, 626-631.

Fierer, N., Mccain, C. M., Meir, P., Zimmerman, M., Rapp, J. M., Silman, M. R., Knight, R., (2011). Microbes Do Not Follow the Elevational Diversity Patterns of Plants and Animals. *Ecology*, **92**, 797-804.

Fine, P. E. M. (1995). Variation in Protection by BCG: Implications of and for Heterologous Immunity. *The Lancet*, **346,** 1339-1345.

Fine, P. E. M., Carneiro, I. A. M., Milstien, J. B., & Clements, J. C. (1999). Issues Relating to the Use of BCG in Immunization Programmes A Discussion Document. *World Health Organization*, 1-45.

Fine, P. E. M., Floyd, S., Stanford, J. L., Nkhosa, P., Kasunga, A., Chaguluka, S., Warndorff, D. K., Ponnighaus, J. M. (2001). Environmental Mycobacteria in Northern Malawi: Implications for the Epidemiology of Tuberculosis and Leprosy. *Epidemiology and Infection*, **126**, 379-387.

Fine, P. E. M. (2001). BCG: The Challenge Continues. *Scandinavian Journal of Infectious Diseases*, **33**, 243-245.

Fischer, S. G., & Lerman, L. S. (1983). DNA Fragments Differing by Single Base-Pair Substitutions are Separated in Denaturing Gradient Gels: Correspondence with Melting Theory. *Proceedings of the National Academy of Science of the USA***, 80**, 1579-1583.

Flaherty, D. K., Vesosky, B., Beamer, G. L., Stromberg, P., & Turner, J. (2006). Exposure to *Mycobacterium avium* Can Modulate Established Immunity Against *Mycobacterium tuberculosis* Infection Generated by *Mycobacterium bovis* BCG Vaccination. *Journal of Leukocyte Biology*, **80,** 1262-1271.

Frostegard, A., Courtois, S., Ramisse, V., Bernillon, D., Gall, F. L., Nesme, X., Simonet, P., (1999). Quantification of Bias Related to the Extraction of DNA Directly from Soils *Applied and Environmental Microbiology*, **65**, 5409-5420.

Fulthorpe, R. R., Roesch, L. F. W., Riva, A., & Triplett, E. W. (2008). Distantly Sampled Soils Carry Few Species in Common. *The ISME journal*, **2**, 901-910.

Fyfe, J. A M., Lavender, C. J., Johnson, P. D. R., Globan, M., Sievers, A., Azuolas, J., & Stinear, T. P. (2007). Development and Application of Two Multiplex real-time PCR Assays for the Detection of *Mycobacterium ulcerans* in Clinical and Environmental Samples. *Applied and Environmental Microbiology*, **73**, 4733-4740.

Fyfe, J. A. M., Lavender, C. J., Handasyde, K. A, Legione, A. R., O'Brien, C. R., Stinear, T. P., Pidot, S. J., Johnson, P. D. R (2010). A Major Role for Mammals in the Ecology of *Mycobacterium ulcerans*. *PLoS neglected tropical diseases*, **4**, e791.

Galassi, L., Donato, R., Tortoli, E., Burrini, D., Santianni, D., & Dei, R. (2003). Nontuberculous Mycobacteria in Hospital Water Systems: Application of HPLC For Identification of Environmental Mycobacteria. *Journal of Water and Health*, **1**, 133-139.

Gaston, K. J. (2000). Global Patterns in Biodiversity. *Nature*, **405**, 220-227.

George, K. M., Chatterjee, D., Gunawardana, G., Welty, D., Haymann, J., Lee, R., & Small, P. L. C. (1999). Mycolactone: A Polyketide Toxin from *Mycobacterium ulcerans* Required for Virulence. *Science*, **283**, 854-857.

Ghaemi, E., Ghazisaidi, K., Koohsari, H., Khodabakhshi, B., & Mansoorian, A. (2006). Environmental Mycobacteria in Areas of High and Low Tuberculosis Prevalence in the Islamic Republic of Iran. *Eastern Mediterranean Health Journal*, **12**, 280-285.

Ghosh, J., Larsson, P., Singh, B., Pettersson, B. M. F., Islam, N. M., Sarkar, S. N., Dasgupta, S., Kirsebom, L. A. (2009). Sporulation in Mycobacteria. *Proceedings of the National Academy of Science of the USA*, **106**, 10781-10786.

Gilbert, J. A, Field, D., Swift, P., Newbold, L., Oliver, A., Smyth, T., Somerfield, P. J., Joint, I.. (2009). The Seasonal Structure of Microbial Communities in the Western English Channel. *Environmental Microbiology*, **11**, 3132-3139.

Gilles, A., Meglecz, E., Pech, N., Ferreira, S., Malausa, T., & Martin, J.-F. (2011). Accuracy and Quality Assessment of 454 GS-FLX Titanium Pyrosequencing. *BMC Genomics*, **12,** 1-11.

Giongo, A., Crabb, D. B., Davis-Richardson, A. G., Chauliac, D., Mobberley, J. M., Gano, K. A, Mukherjee, N., Triplett, E. W. (2010). PANGEA: Pipeline for Analysis of Next Generation Amplicons. *The ISME journal*, **4**, 852-861.

Glickman, S. E., Kilburn, J. O., Butler, W. R., & Ramos, L. S. (1994). Rapid Identification of Mycolic Acid Patterns of Mycobacteria by High-Performance Liquid Chromatography using Pattern Recognition Software and a *Mycobacterium* Library. *Journal of Clinical Microbiology*, **32**, 740-745.

Gordon, S. V., Bottai, D., Simeone, R., Stinear, T. P., & Brosch, R. (2009). Pathogenicity in the Tubercle Bacillus: Molecular and Evolutionary Determinants. *BioEssays*, **31**, 378-388.

Green, J. L., & Bohannan, B. J. M. (2006). Spatial Scaling of Microbial Biodiversity. *Trends in Ecology & Evolution*, **21**, 501-507.

Griffith, D. E, Aksamit, T., Brown-Elliott, B. A, Catanzaro, A., Daley, C., Gordin, F., Holland, S. M., & Winthrop. K. (2007). An Official ATS/IDSA Statement: Diagnosis, Treatment, and Prevention of Nontuberculous Mycobacterial Diseases. *American Journal of Respiratory and Critical Care Medicine*, **175**, 367-416.

Griffiths, R I, Whiteley, A. S., O'Donnell, A. G., & Bailey, M. J., (2000). Rapid Method for Coextraction of DNA and RNA From Natural Environments for Analysis of Ribosomal DNA and rRNA-Based Microbial Community Composition. *Applied and Environmental Microbiology*, **66**, 5488-5491.

Griffiths, R. I, Thomson, B. C., James, P., Bell, T., Bailey, M. J., & Whiteley, A. S. (2011). The Bacterial Biogeography of British Soils. *Environmental Microbiology*, **13**, 1642-1654.

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., Ciulla, D., Birren, B. W (2011). Chimeric 16S rRNA Sequence Formation and Detection in Sanger and 454-pyrosequenced PCR Amplicons. *Genome Research*, **21**, 494-504.

Hall, K., Miller, C. D., Sorensen, D. L., Anderson, A., & Sims, R. (2005). Development of a Catabolically Significant Genetic Probe for Polycyclic Aromatic Hydrocarbon-Degrading Mycobacteria in Soil. *Biodegradation*, **16,** 475-484.

Haritash, A. K., & Kaushik, C. P. (2009). Biodegradation Aspects of Polycyclic Aromatic Hydrocarbons (PAHs): A Review. *Journal of Hazardous Materials*, **169,** 1-15.

Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., & Xie, Z. (2008). Single-Molecule DNA Sequencing of a Viral Genome. *Science*, **320**, 106-109.

Heid, C. A, Stevens, J., Livak, K. J., & Williams, P. M. (1996). Real Time Quantitative PCR. *Genome Research,* **6**, 986-994.

Hert, D. G., Fredlake, C. P., & Barron, A. E. (2008). Advantages and Limitations of Next-Generation Sequencing Technologies: A Comparison of Electrophoresis and Non-Electrophoresis Methods. *Electrophoresis*, **29**, 4618-4626.

Heuer, H., Krsek, M., Baker, P., Smalla, K., & Wellington, E. M. (1997). Analysis of Actinomycete Communities by Specific Amplification of Genes Encoding 16S rRNA and Gel-Electrophoretic Separation in Denaturing Gradients. *Applied and Environmental Microbiology*, **63**, 3233-3241.

Ho, P., Zhang, L., Wei, X., & Seah, G. T. (2009). *Mycobacterium chelonae* sensitisation induces CD4(+)-mediated cytotoxicity against BCG. *European Journal of Immunology*, **39**, 1841-1849.

Ho, P., Wei, X., & Seah, G. T. (2010). Regulatory T cells Induced by *Mycobacterium chelonae* Sensitization Influence Murine Responses to Bacille Calmette-Guerin. *Journal of Leukocyte Biology*, **88**, 1073-1080.

Hollister, E. B., Engledow, A. S., Hammett, A. J. M., Provin, T. L., Wilkinson, H. H., & Gentry, T. J. (2010). Shifts in Microbial Community Structure Along an Ecological Gradient of Hypersaline Soils and Sediments. *The ISME journal*, **4**, 829-838.

Hong, S., Bunge, J., Leslin, C., Jeon, S., & Epstein, S. S. (2009). Polymerase Chain Reaction Primers Miss Half of rRNA Microbial Diversity. *The ISME journal*, **3**, 1365-1373.

Hori, M., Fukano, H., & Suzuki, Y. (2007). Uniform Amplification of Multiple DNAs By Emulsion PCR. *Biochemical and Biophysical Research Communications*, **352,** 323-328.

Horner-Devine, M. C., Carney, K. M., & Bohannan, B. J. M. (2004a). An Ecological Perspective on Bacterial Biodiversity. *Proceedings Of The Royal Society*, **271**, 113-122.

Horner-Devine, M. C., Lage, M., Hughes, J. B., & Bohannan, B. J. M. (2004b). A Taxa-Area Relationship for Bacteria. *Nature*, **432**, 750-753.

Horsburgh, R. (1991). *Mycobacterium avium* Complex Infection in the Acquired Immunodeficiency Syndrome. *The New England Journal of Medicine*, **324,** 1332-1338.

Horsburgh, C. R., Caldwell, M. B., & Simonds, R. J. (1993). Epdiemiology of Disseminated Nontuberculous Mycobacterial Disease in Children With Acquired Immunodeficiency Syndrome. *Pediatric Infectious Disease Journal*, **12**, 219-222.

Horsburgh, C. R. (1996). Epidemiology of *Mycobacterium avium* complex. In *Mycobacterium avium Complex Infection: Progress in Research and Treatment*.1st edn, pp. 1-22. Edited by J. Korvick & C. Benson. New York: Marcel Dekker.

Huq, A., Yunus, M., Sohel, S. S., Bhuiya, A., Emch, M., Luby, S. P., Russek-Cohen, E., Colwell, R. R. (2010). Simple Sari Cloth Filtration of Water Is Sustainable and Continues To Protect Villagers from Cholera in Matlab, Bangladesh. *MBio*, **1**, e00034-e00037.

Huse, S. M., Huber, J. A, Morrison, H. G., Sogin, M. L., & Welch, D. M. (2007). Accuracy and Quality of Massively Parallel DNA Pyrosequencing. *Genome Biology*, **8,** 1-9

Hussein, Z., Landt, O., Wirths, B., & Wellinghausen, N. (2009). Detection of Non-Tuberculous Mycobacteria in Hospital Water by Culture and Molecular Methods. *International Journal of Medical Microbiology*, **299**, 281-290.

IMCR (2006). Influence of Sex, Age & Nontuberculous Infection at Intake on the Efficacy of BCG: Re-Analysis of 15-Year Data From a Double-Blind Randomized Control Trial in South India. *The Indian Journal of Medical Research,* **123,** 119-124.

Iivanainen, E., Martikainen, P. J., & Katila, M. L. (1997). Comparison of Some Decontamination Methods and Growth Media for Isolation of Mycobacteria From Northern Brook Waters. *Journal of Applied Microbiology,* **82,** 121-127.

Iivanainen, E., Martikainen, P. J., Väänänen, P., & Katila, M. L. (1999). Environmental Factors Affecting the Occurrence of Mycobacteria in Brook Sediments. *Journal of Applied Microbiology,* **86**, 673-681.

Jacobs, J., Rhodes, M., Sturgis, B., & Wood, B. (2009). Influence of Environmental Gradients on the Abundance and Distribution of *Mycobacterium* spp. in a Coastal Lagoon Estuary. *Applied and Environmental Microbiology,* **75,** 7378-7384.

Joachimiak, M. P., Weisman, J. L., & May, B. C. (2006). JColorGrid: Software For the Visualization of Biological Measurements. *BMC Bioinformatics*, **7,** 1-5.

Johnson, P D R, & Lavender, C. J. (2009). Correlation Between Buruli Ulcer and Vector-Borne Notifiable Diseases *Emerging Infectious Diseases*, **15**, 614-615.

Jones, R. T., Robeson, M. S., Lauber, C. L., Hamady, M., Knight, R., & Fierer, N. (2009). A Comprehensive Survey of Soil Acidobacterial Diversity Using Pyrosequencing and Clone Library Analyses. *The ISME journal,* **3**, 442-453.

Jyoti, A., Vajpayee, P., Singh, G., Patel, C. B., Gupta, K. C., & Shanker, R. (2011). Identification of Environmental Reservoirs of Nontyphoidal Salmonellosis: Aptamer-Assisted Bioconcentration and Subsequent Detection of *Salmonella Typhimurium* by Quantitative Polymerase Chain Reaction. *Environmental Science & Technology,* **45,** 8996-9002.

Kamala, T., Paramasivan, C. N., Herbert, D., Venkatesan, P., & Prabhakar, R. (1994). Isolation and Identification of Environmental Mycobacteria in the *Mycobacterium bovis* BCG Trial Area of South India. *Applied and Environmental Microbiology*, **60**, 2180-2183.

Kana, B. D., Gordhan, B. G., Downing, K. J., Sung, N., Vostroktunova, G., Machowski, E. E., Tsenova, L., & Mizrahi, V. (2008). The Resuscitation-Promoting Factors of *Mycobacterium tuberculosis* Are Required For Virulence and Resuscitation From Dormancy But Are Collectively Dispensable For Growth *in vitro*. *Molecular Microbiology,* **67**, 672-684.

Kanagawa, T. (2003). Bias and Artifacts in Multitemplate Polymerase Chain Reactions (PCR). *Journal of Bioscience and Bioengineering*, **96**, 317-323.

Kanaly, R. A., & Harayama, S. (2000). Biodegradation of High-Molecular-Weight Polycyclic Aromatic Hydrocarbons by Bacteria. *Journal of Bacteriology*, **182**, 2059-2067.

Kankya, C., Muwonge, A., Olet, S., Munyeme, M., Biffa, D., Opuda-Asibo, J., Skjerve, E., & Oloya, J., (2010). Factors Associated With Pastoral Community Knowledge and Occurrence of Mycobacterial Infections in Human-Animal Interface Areas of Nakasongola and Mubende Districts, Uganda. *BMC Public Health*, **10**, 1-8

Kankya, C., Muwonge, A., Djønne, B., Munyeme, M., Opuda-Asibo, J., Skjerve, E., Oloya, J., & Johanson, T. B (2011). Isolation of Non-Tuberculous Mycobacteria From Pastoral Ecosystems of Uganda: Public Health Significance. *BMC Public Health*, **11**, 1-9

Kasai, H., Ezaki, T., & Harayama, S. (2000). Differentiation of Phylogenetically Related Slowly Growing Mycobacteria by Their *gyrB* Sequences. *Journal of Clinical Microbiology*, **38**, 301-308.

Kawai, M., Yamagishi, J., Yamaguchi, N., Tani, K., & Nasu, M. (2004). Bacterial Population Dynamics and Community Structure in a Pharmaceutical Manufacturing Water Supply System Determined by Real-Time PCR and PCR-Denaturing Gradient Gel Electrophoresis. *Journal of Applied Microbiology*, **97**, 1123-1131.

Kennedy, G. M., Morisaki, J. H., & Champion, P. A. D. (2012). Conserved Mechanisms of *Mycobacterium marinum* Pathogenesis Within the Environmental Amoeba *Acanthamoeba castellanii*. *Applied and Environmental Microbiology*, **78**, 2049-2052.

Khan, A. A., Kim, S.-jae, Paine, D. D., & Cerniglia, C. E. (2002). Classification of a Polycyclic Aromatic Hydrocarbon-Metabolizing Bacterium, *Mycobacterium* sp. strain PYR-1, as *Mycobacterium vanbaalenii* sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, **52**, 1997-2002.

Kim, Kijeong, Lee, H., Lee, M.-K., Lee, S.-A., Shim, T.-S., Lim, S. Y., Koh, W.-J., Kim, B-J. (2010). Development and Application of Multiprobe Real-Time PCR Method Targeting the *hsp65* Gene For Differentiation of *Mycobacterium* Species From Isolates and Sputum Specimens. *Journal of Clinical Microbiology*, **48**, 3073-3080.

Kircher, M., & Kelso, J. (2010). High-Throughput DNA Sequencing-Concepts and Limitations. *BioEssays*, **32**, 524-536.

Kirschner, R. (1992). Epidemiology of Infection by Nontuberculous Mycobacteria - *Mycobacterium avium*, *Mycobacterium intracellulare* and *Mycobacterium scrofulaceum* in Acid Brown Water Swamps of the South-Eastern United States and Their Association With Environmental Variables. *American Review of Respiratory Disease*, **145,** 271-275.

Kirschner, P., Springer, B., Vogel, U., Meier, A, Wrede, A, Kiekenbeck, M., Bange, F. C., & Botteger, E. C. (1993). Genotypic Identification of Mycobacteria by Nucleic Acid Sequence Determination: Report of a 2-Year Experience in a Clinical Laboratory. *Journal of Clinical Microbiology*, **31**, 2882-2889.

Kirschner, R., Parker, B., & Falkinham, J. O. (1999). Humic and Fulvic Acids Stimulate the Growth of *Mycobacterium avium*. *FEMS Microbiology Ecology*, **30**, 327-332.

Kleespies, M., Kroppenstedt, R. M., Rainey, F. A, Webb, L. E., & Stackebrandt, E. (1996). *Mycobacterium hodleri* sp. nov., a New Member of the Fast-Growing Mycobacteria Capable of Degrading Polycyclic Aromatic Hydrocarbons. *International Journal of Systematic Bacteriology*, **46**, 683-687.

Kopecky, J., Kyselkova, M., Omelka, M., Cermak, L., Novotna, J., Grundmann, G., Moënne-Loccoz, Y., & Sagova-Mareckova, M (2011). Environmental Mycobacteria Closely Related to the Pathogenic Species Evidenced in an Acidic Forest Wetland. *Soil Biology and Biochemistry*, **43**, 697-700.

Kox, L. F., van Leeuwen, J., Knijper, S., Jansen, H. M., & Kolk, A. H. (1995). PCR Assay Based on DNA Coding for 16S rRNA for Detection and Identification of Mycobacteria in Clinical Samples. *Journal of Clinical Microbiology*, **33**, 3225-3233.

Krsek, M., & Wellington, E. M. (1999). Comparison of Different Methods for the Isolation and Purification of Total Community DNA From Soil. *Journal of Microbiological Methods*, **39**, 1-16.

Kumar, P. S., Brooker, M. R., Dowd, S. E., & Camerlengo, T. (2011). Target Region Selection Is a Critical Determinant of Community Fingerprints Generated by 16S Pyrosequencing. *PLoS one*, **6,** 1-8

Kunin, V., Engelbrektson, A., Ochman, H., & Hugenholtz, P. (2010). Wrinkles in the Rare Biosphere: Pyrosequencing Errors Can Lead to Artificial Inflation of Diversity Estimates. *Environmental Microbiology*, **12**, 118-123.

Lalor, M. K., Floyd, S., Gorak-Stolinska, P., Weir, R. E., Blitz, R., Branson, K., Fine, P. E. M., & Dockrell. H. M. (2011). BCG vaccination: a role for vitamin D? *PLoS one*, **6**, 1-11

Lamont, E. A, Bannantine, J. P., Armién, A., Ariyakumar, D. S., & Sreevatsan, S. (2012). Identification and Characterization of a Spore-Like Morphotype in Chronically Starved *Mycobacterium avium* subsp. *paratuberculosis* Cultures. *PloS one*, **7**, 1-10

Lauber, C. L., Hamady, M., Knight, R., & Fierer, N. (2009). Pyrosequencing-Based Assessment of Soil pH as a Predictor of Soil Bacterial Community Structure at the Continental Scale. *Applied and Environmental Microbiology*, **75,** 5111-5120.

Lavania, M., Katoch, K., Mohan, V., & Kumar, A. (2008). Detection of Viable *Mycobacterium leprae* in Soil Samples: Insights into Possible Sources of Transmission of Leprosy. *Infection, Genetics and Evoution*, **8**, 627-631.

Le Dantec, C., Duguet, J., Montiel, A., Dumoutier, N., Dubrou, S., & Vincent, V. (2002a). Occurrence of Mycobacteria in Water Treatment Lines and in Water Distribution Systems. *Applied and Environmental Microbiology*, **68**, 5318-5325.

Le Dantec, C., Duguet, J., Montiel, A., Dumoutier, N., Dubrou, S., & Vincent, V. (2002b). Chlorine Disinfection of Atypical Mycobacteria Isolated from a Water Distribution System. *Applied and Environmental Microbiology*, **68**, 1025-1032.

Leclerc, M. C., Thomas, F., & Guégan, J. F. (2003). Evidence for Phylogenetic Inheritance in Pathogenicity of *Mycobacterium*. *Antonie van Leeuwenhoek*, **83**, 265-274.

Lee, E-S., Mok-Young, L., Han, S-H., & Jong, K. (2008). Occurence and Molecular Differentiation of Environmental Mycobacteria in Surface Waters. *Journal of Microbiology and Biotechnology*, **18**, 1207-1215.

Lema, E., & Stanford, J. (1984). Skin-Test Sensitisation by Tubercle Bacilli and by Other Mycobacteria in Ethiopian School-Children. *Tubercle*, **65**, 285-293.

Leoni, E., Legnani, P., Mucci, M. T., & Pirani, R. (1999). Prevalence of Mycobacteria in a Swimming Pool Environment. *Journal of Applied Microbiology*, **87**, 683-688.

Levy-Frébault, V. V., & Portaels, F. (1992). Proposed Minimal Standards for the Genus *Mycobacterium* and for Description of New Slowly Growing *Mycobacterium* Species. *International Journal of Systematic Bacteriology*, **42**, 315-323.

Leys, N. M., Ryngaert, A., Bastiaens, L., Wattiau, P., Top, E. M., Verstraete, W., & Springael, D. (2005). Occurrence and Community Composition of Fast-Growing *Mycobacterium* in Soils Contaminated with Polycyclic Aromatic Hydrocarbons. *FEMS Microbiology Ecology*, **51**, 375- 388

Li, W., & Godzik, A. (2006). Cd-hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences. *Bioinformatics*, **22**, 1658-1659.

Lim, S. Y., Kim, B.-J., Lee, M.-K., & Kim, K. (2008). Development of a Real-Time PCR-Based Method for Rapid Differential Identification of *Mycobacterium* Species. *Letters in Applied Microbiology*, **46,** 101-106.

Liu, J., Tran, V., Leung, A. S., Alexander, D. C., & Zhu, B. (2009). BCG Vaccines Their Mechanisms of Attenuation and Impact on Safety and Protective Efficacy. *Human Vaccines*, **5**, 70-78.

Lonnroth, K., Raviglione, M., & Dye, C. (2006). Global Epidemiology of Tuberculosis: Prospects for Control. *The Lancet*, **367**, 938-939.

Lozupone, C. A, & Knight, R. (2007). Global Patterns in Bacterial Diversity. *Proceedings of the National Academy of Science of the USA*, **104**, 11436-11440.

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., & Knight, R. (2011). UniFrac: An Effective Distance Metric for Microbial Community Comparison. *The ISME Journal*, **5,** 169-172.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, Buchner, A., & Schleifer, K-H. (2004). ARB: A Software Environment for Sequence Data. *Nucleic Acids Research*, **32**, 1363-1371.

Maekawa, K., Ito, Y., Hirai, T., Kubo, T., Imai, S., Tatsumi, S., Fuijita, K., Takakura, S., Niimi, A., Iinuma, Y., Ichyama, S., Togashi, K. & Mishima, M. (2011) Environmental Risk Factors for Pulmonary *Mycobacterium-avium-intracellulare* Complex Disease. *Chest*, **140**, 723-729.

Magurran, A. E. (2004). Measuring Biological Diversity 1st edn, pp. 1-256. Blackwell Science Ltd,.

Manning, E. J. B. & Collins, M. T. (2001). *Mycobacterium avium* subsp. *paratuberculosis:* Pathogen, Pathogenesis and Diagnosis. *Revue Scientifique et Technique*, **20**, 133-150.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L.A, Berka, J., & Rothberg, J. M. (2005). Genome Sequencing in Microfabricated High-Density Picolitre Reactors. *Nature*, **437**, 376-380.

Marras, T. (2002). Epidemiology of Human Pulmonary Infection With Nontuberculous Mycobacteria. *Clinics in Chest Medicine*, **23**, 553-567.

Martin-Casabona, N., Bahrmand, A. R., Bennedsen, J., Thomsen, V. O., Curcio, M., Fauville-Dufaux, M., Feldman, K., & Watt, B. (2004). Non-Tuberculous Mycobacteria: Patterns of Isolation. A Multi-Country Retrospective Survey. *The International Journal of Tuberculosis and Lung Disease*, **8**, 1186-1193.

Martin-Laurent, F., Philippot, L., Hallet, S., Chaussod, R., Soulas, G., & Catroux, G. (2001). DNA Extraction from Soils: Old Bias for New Microbial Diversity Analysis Methods. *Applied and Environmental Microbiology*, **67**, 2354-2359.

Martiny, J. B. H., Bohannan, B. J. M., Brown, J. H., Colwell, R. K., Fuhrman, J. A, Green, J. L., Horner-Devine, M. C., & Stayley, J. (2006) Microbial Biogeography: Putting Microorganisms on the Map. *Nature Reviews Microbiology*, **4,** 102-112.

Martiny, J. B. H., Eisen, J. A., Penn, K., Allison, S. D., & Horner-Devine, M. C. (2011). Drivers of Bacterial ß-Diversity Depend on Spatial Scale. *Proceedings of the National Academy of Science of the USA*, **108**, 7850-7854

McGill, B. J., Maurer, B. A, & Weiser, M. D. (2006). Empirical Evaluation of Neutral Theory. *Ecology*, **87**, 1411-1423.

Medie, F. M., Salah, B. I., Henrissat, B., Raoult, D., & Drancourt, M. (2011). *Mycobacterium tuberculosis* Complex Mycobacteria as Amoeba-Resistant Organisms. *PloS one*, **6**, 1-7

Mendum, T. A., Chilima, B. Z., & Hirsch, P. R. (2000). The PCR Amplification of Non-Tuberculous Mycobacterial 16S rRNA Sequences From Soil. *FEMS Microbiology Letters*, **185,** 189-192

Metzker, M. L. (2010). Sequencing technologies - The Next Generation. *Nature Review Genetics*, **11**, 31-46.

Michel, Anita L, de Klerk, L.-M., Gey van Pittius, N. C., Warren, R. M., & van Helden, P. D. (2007). Bovine Tuberculosis in African buffaloes: Observations Regarding *Mycobacterium bovis* Shedding into Water and Exposure to Environmental Mycobacteria. *BMC veterinary research*, **3**, 1-7

Miller, C. D., Hall, K., Liang, Y. N., Nieman, K., Sorensen, D., Issa, B., Anderson, A. J., & Sims, R. C. (2004). Isolation and Characterization of Polycyclic Aromatic Hydrocarbon-Degrading *Mycobacterium* Isolates From Soil. *Microbial ecology*, **48**, 230-238.

Miltner, E. C., & Bermudez, L. E. (2000). *Mycobacterium avium* Grown in *Acanthamoeba castellanii* is Protected From the Effects of Antimicrobials. *Antimicrobial Agents and Chemotherapy*, **44**, 1990-1994.

Morita, H., Usami, I., Torii, M., Nakamura, A., Kato, K., Itoh, M. (2005) Isolation on Nontuberculous Mycobacteria From Patients With Pneumoconiosis. *Journal of Infection and Chemotherapy*, **11**, 89-92

Mostowy, S., Cousins, D., Brinkman, J., Aranaz, A., & Behr, M. A. (2002). Genomic Deletions Suggest a Phylogeny for the *Mycobacterium tuberculosis* Complex. *The Journal of Infectious Diseases*, **186**, 74-80.

Murcia, M. I., Tortoli, E., Menendez, M. C., Palenque, E., & Garcia, M. J. (2006). *Mycobacterium colombiense* sp. nov., a Novel Member of the *Mycobacterium avium* Complex and Description of MAC-X as a New ITS Genetic Variant. *International Journal of Systematic and Evolutionary Microbiology*, **56**, 2049-2054.

Muyzer, G., de Waal, E. C., & Uitterlinden, A G. (1993). Profiling of Complex Microbial Populations by Denaturing Gradient Gel Electrophoresis Analysis of Polymerase Chain Reaction-Amplified Genes Coding for 16S rRNA. *Applied and Environmental Microbiology*, **59,** 695-700.

Muyzer, G., & Smalla, K. (1998). Application of Denaturing Gradient Gel Electrophoresis (DGGE) and Temperature Gradient Gel Electrophoresis (TGGE) in Microbial Ecology. *Antonie van Leeuwenhoek*, **73**, 127-141.

Myers, R. M., Fischer, S. G., Lerman, L. S., & Maniatis, T. (1985). Modification of the Melting Properties of Duplex DNA by Attachment of a GC-rich DNA Sequence as Determined by Denaturing Gradient Gel Electrophoresis. *Nucleic Acids Research*, **13**, 3131-3145.

Nacke, H., Thürmer, A., Wollherr, A., Will, C., Hodac, L., Herold, N., Schöning, I., & Daniel, R. (2011). Pyrosequencing-Based Assessment of Bacterial Community Structure Along Different Management Types in German Forest and Grassland Soils. *PloS one*, 6, 1-12

Narang, R., Narang, P., & Mendiratta, D. K. (2009). Isolation and Identification of Nontuberculous Mycobacteria From Water and Soil in Central India. *Indian Journal of Medical Microbiology*, **27**, 247-250.

Nemergut, D. R., Costello, E. K., Hamady, M., Lozupone, C., Jiang, L., Schmidt, S. K., Fierer, N., & Knight. R. (2011). Global Patterns in the Biogeography of Bacterial Taxa. *Environmental Microbiology*, **13,** 135-144.

Ngan, G. J. Y., Ng, L. M., Jureen, R., Lin, R. T. P., & Teo, J. W. P. (2011). Development of Multiplex PCR Assays Based on the 16S-23S rRNA Internal Transcribed Spacer for the Detection of Clinically Relevant Nontuberculous Mycobacteria. *Letters in Applied Microbiology*, **52**, 546-554.

Nielsen, S. S., & Toft, N. (2009). A Review of Prevalences of Paratuberculosis in Farmed Animals in Europe. *Preventive Veterinary Medicine*, **88,** 1-14.

Nieminen, T., Pakarinen, J., Tsitko, I., Salkinoja-Salonen, M., Breitenstein, A., Ali-Vehmas, T., & Neubauer, P. (2006). 16S rRNA Targeted Sandwich Hybridization Method for Direct Quantification of Mycobacteria in Soils. *Journal of Microbiological Methods*, **67**, 44-55.

Niva, M., Hernesmaa, A., Haahtela, K., Salkinoja-salonen, M., & Haukka, K. (2006). Actinobacterial Communities of Boreal Forest Soil and Lake Water are Rich in Mycobacteria. *Boreal Environment Research*, **11**, 45-53.

Norby, B., Fosgate, G. T., Manning, E. J. B., Collins, M. T., & Roussel, A. J. (2007). Environmental Mycobacteria in Soil and Water on Beef Ranches: Association Between Presence of Cultivable Mycobacteria and Soil and Water Physicochemical Characteristics. *Veterinary Microbiology*, **124**, 153-159.

Nunn, P., Williams, B., Floyd, K., Dye, C., Elzinga, G., & Raviglione, M. (2005). Tuberculosis Control in the Era of HIV. *Nature Reviews Immunology*, **5,** 819-826.

Nunn, P., Reid, A., & De Cock, K. M. (2007). Tuberculosis and HIV Infection: The Global Setting. *The Journal of Infectious Diseases*, **196**, S5-14.

Oakley, B. B., Carbonero, F., Van Der Gast, C. J., Hawkins, R. J. & Purdy, K. J. (2010) Evolutionary Divergence and Biogeography of Sympatric Niche-Differentiated Bacterial Populations. *The ISME Journal*, **4**, 488-497

Oakley, B. B., Carbonero, F., Dowd, S. E., Hawkins, R. J., & Purdy, K. J. (2012). Contrasting patterns of niche partitioning between two anaerobic terminal oxidizers of organic matter. *The ISME Journal*, **6**, 905-914.

Ocepeck, M., Pate, M., Manca, Z.-D., & Poljak, M. (2005). Transmission of *Mycobacterium tuberculosis* From Human to Cattle. *Journal of Clinical Microbiology*, **43**, 3555-3557.

Oettinger, T., Jorgensen, M., Laderfoged, A., Haslov, K., & Andersen, P. (1999). Development of the *Mycobacterium bovis* BCG vaccine: Review of the Historical and Biochemical Evidence for a Genealogical Tree. *Tubercle and Lung Disease*, **79**, 243-250.

O'Brien, D. J., Schmitt, S. M., Fierke, J. S., Hogle, S. A., Wintersen, S. R., Cooley, T. M., Moritz, W, E., & Kaneene, J. B. (2002). Epidemiology of *Mycobacterium bovis* Disease in White-Tailed Deer, Michigan, USA, 1995-2000. *Preventive Veterinary Medicine*, **54**, 47-63.

Palmer, C. E., & Long, M. W. (1966). Effects of Infection With Atypical Mycobacteria on BCG Vaccination and Tuberculosis. *American Review of Respiratory Disease*, **94**, 553-568.

Parashar, D., Das, R., Chauhan, D. S., Sharma, V. D., Lavania, M., Yadav, V. S., Chauhan, S. V. S., & Katoch, V. M. (2009). Identification of Environmental Mycobacteria Isolated From Agra, North India by Conventional & Molecular Approaches. *The Indian Journal of Medical Research*, **129,** 424-431.

Petric, I., Philippot, L., Abbate, C., Bispo, a, Chesnot, T., Hallin, S., Laval, K.,& Martin-Laurent, F. (2011). Inter-Laboratory Evaluation of the ISO Standard 11063 "Soil Quality - Method To Directly Extract DNA From Soil Samples". *Journal of Microbiological Methods*, **84**, 454-460.

Phillips, M. S., & Von Reyn, C. F. (2001). Nosocomial Infections Nontuberculous Mycobacteria. *Clinical Infectious Diseases*, **33**, 1363-1374.

Pickup, R., Rhodes, G., & Saunders, J. R. (2004). Extraction of Microbial DNA From Aquatic Sources in Freshwater. *Molecular Microbial Ecology Manual* 2nd ed, pp. 41-52. Kluwer Academic Publishers.

Piersimoni, C. (2009). Extrapulmonary Infections Associated with Nontuberculous Mycobacteria in Immunocompetent Persons. *Emerging Infectious Diseases*, **15**, 1351-1358.

Pinsky, B. A, & Banaei, N. (2008). Multiplex Real-Time PCR Assay for Rapid Identification of *Mycobacterium tuberculosis* Complex Members to the Species Level. *Journal of Clinical Microbiology*, **46**, 2241-2246.

Ponnighaus, J. M., Fine, P. E. M., Sterne, J. A. C., Wilson, R., Msosa, E., Gruer, P. J. K., Jenkins, P, A., & Bliss, L. (1992). Efficacy of BCG Vaccine Against Leprosy and Tuberculosis in Northern Malawi. *Tuberculosis*, **339**, 636-639.

Pontiroli, A., Travis, E. R., Sweeney, F. P., Porter, D., Gaze, W. H., Mason, S., Hibberd, V., & Wellington. E. M. H. (2011). Pathogen Quantitation in Complex

Matrices: A Multi-Operator Comparison of DNA Extraction Methods With a Novel Assessment of PCR Inhibition. *PloS one*, **6**, 1-11.

Portaels, F., Elsen, P., Guimaraes-Peres, A., Fonteyne, P. A., & Meyers, W. M. (1999). Insects in the Transmission of *Mycobacterium ulcerans* Infection. *The Lancet*, **353**, 986.

Portaels, F., Meyers, W. M., Ablordey, A., Castro, A. G., Chemlal, K., de Rijk, P., Elsen, P., et al. (2008). First Cultivation and Characterization of *Mycobacterium ulcerans* From the Environment. *PLoS Neglected Tropical Diseases*, **2,** 1-12

Pourahmad, F., Thompson, K. D., Adams, A, & Richards, R. H. (2009). Comparative Evaluation of Polymerase Chain Reaction-Restriction Enzyme Analysis (PRA) and Sequencing of Heat Shock Protein 65 (*hsp65*) Gene for Identification of Aquatic Mycobacteria. *Journal of Microbiological Methods*, **76**, 128-135.

Power, C. A., Wei, G., & Bretscher, P. A. (1998). Mycobacterial Dose Defines the Th1 / Th2 Nature of the Immune Response Independently of Whether Immunization Is Administered by the Intravenous, Subcutaneous, or Intradermal Route, *Infection and Immunity*, **66**, 5743-5750.

Primm, T. P., Lucero, C. A., & Falkinham, J. O. (2004). Health Impacts of Environmental Mycobacteria. *Clinical Microbiology Reviews*, **17**, 98-106.

Prosser, J. I., Bohannan, B. J. M., Curtis, T. P., Ellis, R. J., Firestone, M. K., Freckleton, R. P., Green, J. L., Green, L. E., Killham, K., Lennon, J. J., Osborn, M., Solan, M., van der Gast, J. & Young, J. P. W. (2007). The Role of Ecological Theory in Microbial Ecology. *Nature Reviews Microbiology*, **5**, 384-392.

Proud, A. (2006). Bovine TB Special Edition. *Government Veterinary Journal* (DEFRA), 16.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glöckner, F. O. (2007). SILVA: A Comprehensive Online Resource for Quality Checked and Aligned Ribosomal RNA Sequence Data Compatible With ARB. *Nucleic Acids Research*, **35**, 7188-7196.

Quince, C., Lanzen, A., Davenport, R. J., & Turnbaugh, P. J. (2011). Removing Noise from Pyrosequenced Amplicons. *BMC Bioinformatics*, **12**, 1-18

R Development Core Team. (2011). R: A Language and Environment for Statistical Computing. (Vienna, Austria).

Rahbar, M., Lamei, A., Babazadeh, H., & Yavari, S. A. (2010). Isolation of Rapid Growing Mycobacteria from Soil and Water in Iran. *Journal of Biotechnology*, **9,** 3618-3621.

Ramette, A., & Tiedje, J. M. (2007). Biogeography: An Emerging Cornerstone for Understanding Prokaryotic Diversity, Ecology, and Evolution. *Microbial Ecology*, **53**, 197-207.

Rastogi, N, Legrand, E., & Sola, C. (2001). The Mycobacteria: An Introduction to Nomenclature and Pathogenesis. *Revue Scientifique et Technique*, **20**, 21-54.

Reddington, K., O'Grady, J., Dorai-Raj, S., Maher, M., van Soolingen, D., & Barry, T. (2011). Novel Multiplex Real-Time PCR Diagnostic Assay for Identification and Differentiation of *Mycobacterium tuberculosis*, *Mycobacterium canettii*, and *Mycobacterium tuberculosis* Complex Strains. *Journal of Clinical Microbiology*, **49**, 651-657.

Regassa, A., Medhin, G., & Ameni, G. (2008). Bovine Tuberculosis is More Prevalent in Cattle Owned by Farmers With Active Tuberculosis in Central Ethiopia. *The Veterinary Journal*, **178**, 119-125.

Reich, J. M. & Johnson, R. E. (1992) *Mycobacterium avium* Complex Pulmonary Disease Presenting as an Isolated Lingular or Middle Lobe Pattern. *Chest*, **101**, 1605-1609.

Richardson, B. A., Richardson, M. J., Scatena, F. N., & Mcdowell, W. H. (2000). Effects of Nutrient Availability and Other Elevational Changes on Bromeliad Populations and Their Invertebrate Communities in a Humid Tropical Forest in Puerto Rico. *Journal of Tropical Ecology*, **16**, 167-188.

Richter, B. G., & Sexton, D. P. (2009). Managing and Analyzing Next-Generation Sequence Data. *PLoS Computational Biology*, 5, 1-4

Ringuet, H., Honore, S., Varnerot, A., Berche, P., Gaillard, J. L., & Vincent, V. (1999). *hsp65* Sequencing for Identification of Rapidly Growing Mycobacteria**.** *Journal of Clinical Microbiology***, 37,** 852-857

Robe, P. (2003). Extraction of DNA from Soil. *European Journal of Soil Biology*, **39**, 183-190.

Roche, P. W., Triccas, J. A, & Winter, N. (1995). BCG Vaccination Against Tuberculosis: Past Disappointments and Future Hopes. *Trends in Microbiology*, **3**, 397-401.

Rodrigues, L. C., Diwan, V. K., & Wheeler, J. G. (1993). Protective Effect of BCG Against Tuberculous Meningitis and Miliary Tuberculosis: A meta-analysis. *International Journal of Epidemiology*, **22**, 1154-1158.

Rogall, T, Wolters, J., Flohr, T., & Böttger, E. C. (1990). Towards a Phylogeny and Definition of Species at the Molecular Level Within the Genus *Mycobacterium*. *International Journal of Systematic Bacteriology*, **40,** 323-330.

Ronaghi, M, Karamohamed, S., Pettersson, B., Uhlén, M., & Nyrén, P. (1996). Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. *Analytical Biochemistry*, **242**, 84-89.

Ronaghi, Mostafa, Uhlén, M., Nyrén, P., Uhlen, M., & Nyren, P. (1998). A Sequencing Method Based on Real-Time Pyrophosphate. *Science*, **281**, 363-365.

Ronaghi, Mostafa. (2001). Pyrosequencing Sheds Light on DNA Sequencing. *Genome Research*, **11**, 3-11.

Ross, B. C., Johnson, P. D. R., Oppedisano, F., Marino, L., Sievers, A., & Stinear, T. (1997). Detection of *Mycobacterium ulcerans* in Environmental Samples During an Outbreak of Ulcerative Disease. *Microbiology*, **63**, 4135-4138.

Rossello-Mora, R., & Amann, R. (2001). The Species Concept for Prokaryotes. *FEMS microbiology reviews*, **25**, 39-67.

Roth, A., Fischer, M., Hamid, M. E., Michalke, S., Ludwig, W., & Mauch, H. (1998). Differentiation of Phylogenetically Related Slowly Growing Mycobacteria Based on 16S-23S rRNA Gene Internal Transcribed Spacer Sequences. *Journal of Clinical Microbiology*, **36**, 139-147.

Roth, A., Reischl, U., Streubel, A., Naumann, L., Kroppenstedt, R. M., Habicht, M., Fischer, M., & Mauch, H. (2000). Novel Diagnostic Algorithm for Identification of Mycobacteria using Genus-Specific Amplification of the 16S-23S rRNA Gene Spacer and Restriction Endonucleases. *Journal of Clinical Microbiology*, **38**, 1094-1104.

Rusin, P., Rose, J. B., Haas, C. N., & Gerba, C. P. (1997). Risk Assessment of Opportunistic Bacterial Pathogens in Drinking Water. *Reviews of Environmental Contamination and Toxicology*, **152**, 57-83.

Ruyon, E. H. (1959). Anonymous Mycobacteria in Pulmonary Disease. *The Medical Clinics of North America*, **43,** 273-290.

SAC. (2002). Assessment of Surveillancee and Control of Johne's Disease in Farm Animals in GB. *Veterinary Division of the Scottish Agricultural College*, 1-245.

Salah, B. I., Adekambi, T., Raoult, D., & Drancourt, M. (2008). *rpoB* Sequence-Based Identification of *Mycobacterium avium* Complex Species. *Microbiology*, **154**, 3715-3723.

Salah, B. I., Cayrou, C., Raoult, D., & Drancourt, M. (2009a*). Mycobacterium marseillense* sp. nov., *Mycobacterium timonense* sp. nov. and *Mycobacterium bouchedurhonense* sp. nov., members of the *Mycobacterium avium* complex. *International Journal of Systematic and Evolutionary Microbiology*, **59**, 2803-2808.

Salah, B. I., Ghigo, E., & Drancourt, M. (2009b). Free-Living Amoebae, A Training Field for Macrophage Resistance of Mycobacteria. *Clinical Microbiology and Infection*, **15**, 894-905.

Salah, B. I., & Drancourt, M. (2010). Surviving Within the Amoebal Exocyst: The *Mycobacterium avium* Complex Paradigm. *BMC Microbiology*, **10,** 1-8

Sanderson, J. D., Moss, M. T., Tizard, M. L., & Hermon-Taylor, J. (1992). *Mycobacterium paratuberculosis* DNA in Crohn's Disease Tissue. *Gut*, **33**, 890-896.

Santamaría, J., & Toranzos, G. A. (2003). Enteric Pathogens and Soil: A Short Review. *International Microbiology*, **6**, 5-9.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A, & Weber, C. F. (2009). Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*, **75,** 7537-7541.

Schloss, P. D., Gevers, D., & Westcott, S. L. (2011). Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. *PLoS one*, **6,** 1-14

Schneegurt, M. A, Dore, S. Y., & Kulpa, C. F. (2003). Direct Extraction of DNA From Soils for Studies in Microbial Ecology. *Current issues in molecular biology*, **5,** 1-8.

Schulze-Röbbecke, R., & Buchholtz, K. (1992). Heat Susceptibility of Aquatic Mycobacteria. *Applied and Environmental Microbiology*, **58**, 1869-1873.

September, S. M., Brozel, V. S., & Venter, S. N. (2004). Diversity of Nontuberculoid *Mycobacterium* Species in Biofilms of Urban and Semiurban Drinking Water Distribution Systems. *Applied and Environmental microbiology*, **70**, 7571-7573.

Shapiro, R. L., Otieno, M. R., Adcock, P. M., Phillips-Howard, P. A, Hawley, W. A, Kumar, L., Waiyaki, P., & Slutsker, L. (1999). Transmission of Epidemic *Vibrio cholerae* O1 in Rural Western Kenya Associated with Drinking Water from Lake Victoria: An Environmental Reservoir for Cholera? *The American Journal of Tropical Medicine and Hygiene*, **60**, 271-276.

Shinnick, T. M., & Good, R. C. (1994). Mycobacterial Taxonomy. *European Journal of Clinical Microbiology & Infectious Diseases*, **13**, 884-901.

Shitaye, J. E., Tsegaye, W., & Pavlik, I. (2007). Bovine Tuberculosis Infection in Animal and Human Populations in Ethiopia: A Review. *Veterinarni Medicina*, **8**, 317-332.

Singh, G., Vajpayee, P., Ram, S., & Shanker, R. (2010). Environmental Reservoirs for Enterotoxigenic *Escherichia coli* in South Asian Gangetic Riverine System. *Environmental Science & Technology*, **44**, 6475-6480.

Smith, C. J., & Osborn, A. M. (2009). Advantages and Limitations of Quantitative PCR (Q-PCR)-Based Approaches in Microbial Ecology. *FEMS Microbiology Ecology*, **67**, 6-20.

Smith, D., Wiegeshaus, E., & Balasubramanian, V. (2000). An Analysis of Some Hypotheses Related to the Chingelput Bacille Calmette-Guérin Trial. *Clinical Infectious Diseases*, **31**, S77-80.

Smith, N. H., Hewinson, R. G., Kremer, K., Brosch, R., & Gordon, S. V. (2009). Myths and Misconceptions: The Origin and Evolution of *Mycobacterium tuberculosis*. *Nature reviews Microbiology*, **7,** 537-544.

Sogin, M. L., Morrison, H. G., Huber, J. a, Mark Welch, D., Huse, S. M., Neal, P. R., Arrieta, J. M., & Herndl, G, J. (2006). Microbial Diversity in the Deep Sea and the Underexplored "Rare Biosphere". *Proceedings of the National Academy of Science of the USA*, **103**, 12115-12120.

Somerville, C. C., Knight, I. T., Straube, W. L., & Colwell, R. R. (1989). Simple, Rapid Method for Direct Isolation of Nucleic Acids from Aquatic Environments. *Applied and Environmental Microbiology*, **55**, 548-554.

Springer, B., Stockman, L., Teschner, K., Roberts, G. D., & Böttger, E. C. (1996). Two-Laboratory Collaborative Study on Identification of Mycobacteria: Molecular Versus Phenotypic Methods. *Journal of Clinical Microbiology*, **34**, 296-303.

Stackebrandt, E., & Goebel, B. M. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International Journal of Systematic Bacteriology*, **44**, 846-849.

Stahl, D. A, & Urbance, J. W. (1990). The Division Between Fast- and Slow-Growing Species Corresponds to Natural Relationships Among the Mycobacteria. *Journal of Bacteriology*, **172**, 116-124.

Staley, J. T. (2006). The Bacterial Species Dilemma and the Genomic-Phylogenetic Species Concept. *Philosophical Transactions of the Royal Society* **361**, 1899-1909.

Sterne, J. A. C., Rodrigues, L. C., & Guedes, I. N. (1998). Does the Efficacy of BCG Decline with the Time Since Vaccination? *International Journal of Tuberculosis and Lung Disease*, **2**, 200-207.

Stienstra, Y., van der Graaf, W. T., Meerman, G. J., The, T. H., de Leij, L. F., & van der Werf, T. S. (2001). Susceptibility to Development of *Mycobacterium ulcerans* Disease: Review of Possible Risk Factors. *Tropical Medicine & International Health*, **6**, 554-562.

Stinear, T., Ford, T., & Vincent, V. (2004). Analytical Methods for the Detection of Waterborne and Environmental Pathogenic Mycobacteria. *In Pathogenic mycobacteria in water: A guide to public health consequences, monitoring and management* pp55-73. London, UK: WHO IWA Publishing

Stingley, R. L., Brezna, B., Khan, A. A, & Cerniglia, C. E. (2004). Novel Organization of Genes in a Phthalate Degradation Operon of *Mycobacterium vanbaalenii* PYR-1. *Microbiology*, **150**, 3749-3761.

Sweeney, F P, Courtenay, O., Hibberd, V., Hewinson, R. G., Reilly, L. a, Gaze, W. H., & Wellington, E. M. H. (2007). Environmental Monitoring of *Mycobacterium bovis* in Badger Feces and Badger Sett Soil by Real-Time PCR, as Confirmed by Immunofluorescence, Immunocapture, and Cultivation. *Applied and Environmental Microbiology*, **73**, 7471-7473.

Tegbaru, B., Wolday, D., Messele, T., Legesse, M., Mekonnen, Y., Miedema, F., & van Baarle, D. (2006). Tuberculin Skin Test Conversion and Reactivity Rates Among Adults With and Without Human Immunodeficiency Virus in Urban Settings in Ethiopia. *Clinical and Vaccine Immunology*, **13**, 784-789.

Telenti, A., Marchesi, F., Balz, M., Bally, F., Böttger, E. C., & Bodmer, T. (1993). Rapid Identification of Mycobacteria to the Species Level by Polymerase Chain Reaction and Restriction Enzyme Analysis. *Journal of Clinical Microbiology*, **31**, 175-178.

Tell, L. A., Woods, L., & Cromie, R. L. (2001). Mycobacteriosis in Birds Mycobacterial Species Affecting Birds Epizootiology of Avian Mycobacteriosis. *Revue Scientifique et Technique* **20**, 180-203.

Tenant, R., & Bermudez, L. E. (2006). *Mycobacterium avium* Genes Upregulated Upon Infection of *Acanthamoeba castellanii* Demonstrate a Common Response to the Intracellular Environment. *Current microbiology*, **52,** 128-133.

Ter Braak, C. J. F. (1986) Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis. *Ecology*, **67**, 1167-1179

Thibert, L., & Lapierre, S. (1993). Routine Application of High-Performance Liquid Chromatography for Identification of Mycobacteria. *Journal of Clinical Microbiology*, **31**, 1759-1763.

Thomas, V., & McDonnell, G. (2007). Relationship Between Mycobacteria and Amoebae: Ecological and Epidemiological Concerns. *Letters in Applied Microbiology*, **45**, 349-357.

Thomson, R., Carter, R., Gilpin, C., Coulter, C., & Hargreaves, M. (2008). Comparison of Methods for Processing Drinking Water Samples for the Isolation of *Mycobacterium avium* and *Mycobacterium intracellulare*. *Applied and Environmental Microbiology*, **74,** 3094-3098.

Thorel, M. F., Huchzermeyer, H. F., & Michel, A. L. (2001). *Mycobacterium avium* and *Mycobacterium intracellulare* infection in mammals. *Revue Scientifique et Technique*, **20**, 204-218.

Thorel, M. F., Falkinham, J. O., & Moreau, R. G. (2004). Environmental Mycobacteria from Alpine and Subalpine Habitats. *FEMS Microbiology Ecology*, **49**, 343-347.

Tokeshi, M. (1990). Niche Apportionment or Random Assortment: Species Abundance Patterns Revisited. *Journal of Animal Ecology*, **59**, 1129-1146.

Torrens, J. K., Dawkins, P., Conway, S. P., & Moya, E. (1998). Non-tuberculous mycobacteria in cystic fibrosis. *Thorax*, **53**, 182-185.

Torsvik, V., Goksøyr, J., & Daae, F. L. (1990). High Diversity in DNA of Soil Bacteria. *Applied and Environmental Microbiology*, **56**, 782-787.

Torsvik, V. (2002). Prokaryotic Diversity: Magnitude, Dynamics, and Controlling Factors. *Science*, 296, 1064-1066.

Tortoli, Enrico, Bartoloni, A., Bottger, E. C., Emler, S., Garzelli, C., Magliano, E., Mantella, A., & Urbano, P. (2001). Burden of Unidentifiable Mycobacteria in a Reference Laboratory. *Journal of Clinical Microbiology*, **39**, 4058-4065.

Torvinen, E., Suomalainen, S., Lehtola, M. J., Miettinen, I. T., Zacheus, O., Paulin, L., Katila, M.-leena & Martikainen, P. (2004). Mycobacteria in Water and Loose Deposits of Drinking Water Distribution Systems in Finland. *Applied and Environmental Microbiology*, **70**, 1973-1981.

Torvinen, E., Meklin, T., Torkko, P., Suomalainen, S., Reiman, M., Katila, M.-L., Paulin, L., & Nevalainen, A (2006). Mycobacteria and Fungi in Moisture-Damaged Building Materials. *Applied and Environmental Microbiology*, **72**, 6822-6824.

Torvinen, E., Torkko, P., & Rintala, A. N. H. (2010). Real-Time PCR Detection of Environmental Mycobacteria in House Dust. *Journal of Microbiological Methods*, **82**, 78-84.

Towe, S., Wallisch, S., Bannert, A., Fischer, D., Hai, B., Haesler, F., Kleineidam, K., & Scloter, M. (2011). Improved Protocol for the Simultaneous Extraction and Column-Based Separation of DNA and RNA from Different Soils. *Journal of Microbiological Methods*, **84**, 406-412.

Traag, B. A, Driks, A., Stragier, P., Bitter, W., Broussard, G., Hatfull, G., Chu, F., & Losick, R. (2010). Do Mycobacteria Produce Endospores? *Proceedings of the National Academy of Sciences of the USA*, **107**, 878-881**.**

Trunz, B. B., Fine, P. E. M., & Dye, C. (2006). Effect of BCG Vaccination on Childhood Tuberculous Meningitis and Miliary Tuberculosis Worldwide: A Meta-Analysis and Assessment of Cost-Effectiveness. *The Lancet*, **367**, 1173-1180.

Tsai, Y. L., & Olson, B. H. (1991). Rapid Method for Direct Extraction of DNA from Soil and Sediments. *Applied and Environmental Microbiology*, **57**, 1070-1074.

Tschopp, R., Schelling, E., Hattendorf, J., Aseffa, A., & Zinsstag, J. (2009). Risk factors of Bovine Tuberculosis in Cattle in Rural Livestock Production Systems of Ethiopia. *Preventive Veterinary Medicine,* **89**, 205-211.

Tschopp, R., Aseffa, A., Schelling, E., Berg, S., Hailu, E., Gadisa, E., Habtamu, M., & Zinnstag, J. (2010a). Bovine Tuberculosis at the Wildlife-Livestock-Human Interface in Hamer Woreda, South Omo, Southern Ethiopia. *PloS one*, **5**, 1-7.

Tschopp, R., Berg, S., Argaw, K., Gadisa, E., Habtamu, M., Schelling, E., Young, D., & Zinnstag, J. (2010b). Bovine Tuberculosis in Ethiopian Wildlife. *Journal of Wildlife Diseases,* **46,** 753-762.

Tschopp, R., Bobosha, K., Aseffa, A., Schelling, E., Habtamu, M., Iwnetu, R., Hailu, E., & Zinnstag, J (2011). Bovine Tuberculosis at a Cattle-Small Ruminant-Human Interface in Meskan, Gurage Region, Central Ethiopia. *BMC Infectious Diseases,* **11**, 1-10.

Uyttebroek, M., Breugelmans, P., Janssen, M., Wattiau, P., Joffe, B., Karlson, U., Ortega-Calvo, J.-J., & Springael, D. (2006). Distribution of the *Mycobacterium* Community and Polycyclic Aromatic Hydrocarbons (PAHs) Among Different Size Fractions of a Long-Term PAH-Contaminated Soil. *Environmental Microbiology*, **8**, 836-847.

Vaerewijck, M. J. M., Huys, G., Palomino, J. C., Swings, J., & Portaels, F. (2005). Mycobacteria in Drinking Water Distribution Systems: Ecology and Significance for Human Health. *FEMS Microbiology Reviews*, **29**, 911-934.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., & Johnson, S. M. (2008). A High-Resolution, Nucleosome Position Map of *C. elegans* Reveals a Lack of Universal Sequence-Dictated Positioning. *Genome Research,* **18,** 1051-1063.

Vaneechoutte, M., De Beenhouwer, H., Claeys, G., Verschraegen, G., De Rouck, A, Paepe, N., Elaichouni, A, & Porteals, F. (1993). Identification of *Mycobacterium* Species by Using Amplified Ribosomal DNA Restriction Analysis. *Journal of Clinical Microbiology,* **31**, 2061-2065.

Van Ingen, J, Boeree, M. J., Dekhuijzen, P. N. R., & van Soolingen, D. (2009a). Environmental Sources of Rapid Growing Nontuberculous Mycobacteria Causing Disease in Humans. *Clinical Microbiology and Infection*, **15**, 888-893.

Van Ingen, Jakko, Al-Hajoj, S. A M., Boeree, M., Al-Rabiah, F., Enaimi, M., de Zwaan, R., Tortoli, E., & van Soolingen, D (2009b). *Mycobacterium riyadhense* sp. nov., a non-tuberculous species identified as *Mycobacterium tuberculosis* complex by a commercial line-probe assay. *International Journal of Systematic and Evolutionary Microbiology,* **59**, 1049-1053.

Velmurugan, K., Chen, B., Miller, J. L., Azogue, S., Gurses, S., Hsu, T., Glickman, M., (2007). *Mycobacterium tuberculosis nuoG* is a Virulence Gene that Inhibits Apoptosis of Infected Host Cells. *PLoS pathogens*, **3**, e110.

Vezzulli, L., Pruzzo, C., Huq, A., & Colwell, R. R. (2010). Environmental Reservoirs of *Vibrio cholerae* and Their Role in Cholera. *Environmental Microbiology Reports*, **2**, 27-33.

Von Reyn, C. F. V., Barber, T. W., Arbeit, R. D., Sox, C. H., Gerald, T., Connor, O., Brindle, R. J., & Magnusson, M. (1993a). Evidence of Previous Infection with *Mycobacterium intracellulare* Complex Among Healthy Subjects: International Study of Reactions Dominant Skin Test Mycobacterial. *Journal of Infectious Disease*, **168**, 1553-1558.

Von Reyn, C F, Waddell, R. D., Eaton, T., Arbeit, R. D., Maslow, J. N., Barber, T. W., Brindle, R. J., & Falkinham, J. O. (1993b). Isolation of *Mycobacterium avium* Complex From Water in the United States, Finland, Zaire, and Kenya. *Journal of Clinical Microbiology*, **31**, 3227-3230.

Von Wintzingerode, F., Göbel, U. B., & Stackebrandt, E. (1997). Determination of Microbial Diversity in Environmental Samples: Pitfalls of PCR-Based rRNA Analysis. *FEMS Microbiology Reviews*, **21**, 213-229.

WHO. (2011a). Global Tuberculosis Control: WHO report 2011. *WHO* **393**, 1-258

WHO. (2011b). Ethiopia Reported Immunization Coverage. Vaccine-preventable Diseases: Monitoring System 2011 Global Summary.

Wagner, D., & Young, L. S. (2004). Nontuberculous Mycobacterial Infections: A Clinical Review. *Infection*, **32**, 257-270.

Wallace, R. J., Glassroth, J., Griffith, D. E., Olivier, K. N., Cook, J. L., & Gordin, F. (1997). Diagnosis and Treatment of Disease Caused by Nontuberculous Mycobacteria. *American Journal of Respiratory and Critical Care Medicine*, **156**, S1-S26.

Wallace, R. J., Brown, B. A, & Griffith, D. E. (1998). Nosocomial Outbreaks/Pseudo-Outbreaks Caused by Nontuberculous Mycobacteria. *Annual Review of Microbiology*, **52**, 453-490.

Walther, B. A, & Ewald, P. W. (2004). Pathogen Survival in the External Environment and the Evolution of Virulence. *Biological Reviews*, **79**, 849-869.

Wang, J., Soininen, J., Zhang, Y., Wang, B., Yang, X., & Shen, J. (2011). Contrasting Patterns in Elevational Diversity Between Microorganisms and Macroorganisms. *Journal of Biogeography*, **38**, 595-603.

Ward, D. M., Ferris, M. J., Nold, S. C., & Bateson, M. M. (1998). A Natural View of Microbial Biodiversity within Hot Spring Cyanobacterial Mat Communities. *Microbiology and Molecular Biology Reviews*, **62**, 1353-1370.

Wards, B. J., de Lisle, G. W., & Collins, D. M. (2000). An *esat6* Knockout Mutant of *Mycobacterium bovis* Produced by Homologous Recombination Will Contribute

to the Development of a Live Tuberculosis Vaccine. *Tubercle and Lung Disease*, **80**, 185-189.

Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O., Krichevsky, M. I., Moore, L. H., & Truper, H. G. (1987). Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *International Journal of Systematic Bacteriology*, **37**, 463-464.

Weinbauer, M. G., Fritz, I., Wenderoth, D. F., & Höfle, M. G. (2002). Simultaneous Extraction from Bacterioplankton of Total RNA and DNA Suitable for Quantitative Structure and Function Analyses. *Applied and Environmental Microbiology*, **68**, 1082-1087.

Weir, R. E., Fine, P. E. M., Nazareth, B., Floyd, S., Black, G. F., King, E., Stanley, C., & Branson, K. (2003). Interferon- g and Skin Test Responses of Schoolchildren in Southeast England to Purified Protein Derivatives From *Mycobacterium tuberculosis* and Other Species of Mycobacteria. *Clinical and Experimental Immunology*, **134**, 285-294.

Weir, R. E., Black, G. F., Dockrell, H. M., Floyd, S., Fine, P. E. M., Chaguluka, S. D., Stenson, S., & Blackwell, J. M. (2004). Mycobacterial Purified Protein Derivatives Stimulate Innate Immunity: Malawians Show Enhanced Tumor Necrosis Factor Alpha, Interleukin-1B (IL-1B), and IL-10 Responses Compared to Those of Adolescents in the United Kingdom. *Infection and Immunity*, **72**, 1807-1811.

Weir, R. E., Black, G. F., Nazareth, B., Floyd, S., Stenson, S., Stanley, C., Branson, K., & Dockrell, H M. (2006). The Influence of Previous Exposure to Environmental Mycobacteria on the Interferon-Gamma Response to Bacille Calmette-Guérin Vaccination in Southern England and Northern Malawi. *Clinical and Experimental Immunology*, **146**, 390-399.

Weir, R. E., Fine, P. E. M., Floyd, S., Stenson, S., Stanley, C., Branson, K., Britton, W. J., & Dockrell, H M. (2008a). Comparison of IFN-g Responses to Mycobacterial Antigens as Markers of Response to BCG Vaccination. *Tuberculosis*, **88**, 31-38.

Weir, R. E., Gorak-Stolinska, P., Floyd, S., Lalor, M. K., Stenson, S., Branson, K., Blitz, R., & Dockrell, H M. (2008b). Persistence of the Immune Response Induced by BCG Vaccination. *BMC Infectious Diseases*, 8, 1-9.

Whan, L., Grant, I. R., & Rowe, M. T. (2006). Interaction Between *Mycobacterium avium* subsp. *paratuberculosis* and Environmental Protozoa. *BMC Microbiology*, **6**, 1-6.

Wickremasinghe, M., Ozerovitch, L. J., Davies, G., Wodehouse, T., Chadwick, M. V., Shah, P., & Wilson, R. (2005). Non-Tuberculous Mycobacteria in Patients with Bronchiectasis. *Thorax*, **60,** 1045-1051.

Wilhelm, J., & Pingoud, A. (2003). Real-Time Polymerase Chain Reaction. *ChemBioChem*, **4**, 1120-1128.

Willumsen, P., Karlson, U., Stackebrandt, E., & Kroppenstedt, R. M. (2001). *Mycobacterium frederiksbergense* sp. nov., a Novel Polycyclic Aromatic Hydrocarbon-Degrading *Mycobacterium* Species. *International Journal of Systematic and Evolutionary Microbiology*, **51**, 1715-22.

Wilson, M. E., Fineberg, H. V., & Colditz, G. A. (1995). Geographic Latitude and the Efficacy of Bacillus Calmette-Guérin Vaccine. *Clinical Infectious Diseases*, **20**, 982-991.

Wingender, J., & Flemming, H.-C. (2011). Biofilms in Drinking Water and Their Role as Reservoir for Pathogens. *International Journal of Hygiene and Environmental Health*, **214**, 417-423.

Wittes, R. C. (2000). Immunology of Bacille Calmette-Guérin and Related Topics. *Clinical Infectious Diseases*, **31**, S59-63.

Wolinsky, E. (1992). Mycobacterial Diseases Other Than Tuberculosis. *Clinical Infectious Diseases*, **15,** 1-10.

Xiong, L., Kong, F., Yang, Y., Cheng, J., & Gilbert, G. L. (2006). Use of PCR and Reverse Line Blot Hybridization Macroarray Based on 16S-23S rRNA Gene Internal Transcribed Spacer Sequences for Rapid Identification of 34 *Mycobacterium* species. *Journal of Clinical Microbiology*, **44**, 3544-3550.

Yassin, M. A., Takele, L., Gebresenbet, S., Girma, E., Lera, M., Lendebo, E., & Cuevas, L. E. (2004). HIV and Tuberculosis Coinfection in the Southern Region of Ethiopia: A Prospective Epidemiological Study. *Scandinavian Journal of Infectious Diseases*, **36,** 670-673.

Young, J. S., Gormley, E., & Wellington, E. M. H. (2005). Molecular Detection of *Mycobacterium bovis* and *Mycobacterium bovis* BCG (Pasteur) in Soil. *Applied and Environmental Microbiology*, **71**, 1946-1952.

Young, S. L., Slobbe, L., Wilson, R., Buddle, B. M., de Lisle, G. W., & Buchan, G. S. (2007). Environmental Strains of *Mycobacterium avium* Interfere with Immune Responses Associated with *Mycobacterium bovis* BCG vaccination. *Infection and Immunity*, **75**, 2833-2840.

Youssef, N., Sheik, C. S., Krumholz, L. R., Najar, F. Z., Roe, B. a, & Elshahed, M. S. (2009). Comparison of Species Richness Estimates Obtained Using Nearly Complete Fragments and Simulated Pyrosequencing-Generated Fragments in 16S rRNA Gene-Based Environmental Surveys. *Applied and Environmental Microbiology*, **75**, 5227-5236.

Zhang, T., & Fang, H. H. P. (2006). Applications of Real-Time Polymerase Chain Reaction for Quantification of Microorganisms in Environmental Samples. *Applied Microbiology and Biotechnology*, **70**, 281-289.

Zhou, J., Wu, L., Deng, Y., Zhi, X., Jiang, Y.-H., Tu, Q., Xie, J., & Yang, Y. (2011). Reproducibility and Quantitation of Amplicon Sequencing- Based Detection. *The ISME Journal,* **5**, 1-11.

Zwerling, A., Behr, M. A., Verma, A., Brewer, T. F., Menzies, D., & Pai, M. (2011). The BCG World Atlas: A Database of Global BCG Vaccination Policies and Practices. *PLoS medicine,* **8**, 1-8.

Appendix 1: QIIME commands

```
*** Quality control of individual fasta files for each sample (Do
for 42 samples x4 datasets samples)

split_libraries.py  -f split_mapping/fasta_qual/454Reads.1_mycoB.fna
-q split_mapping/fasta_qual/454Reads.1_mycoB.qual -o
demultiplexed_len_400/1_2myco_demultiplexed/ -b 0  -m
split_mapping/1_mycoB_mapping.txt -M 1  -n 1000000 -l 400 -k


split_libraries.py  -f split_mapping/fasta_qual/454Reads.1_2myco.fna
-q split_mapping/fasta_qual/454Reads.1_2myco.qual -o
demultiplexed_len_420/1_2myco_demultiplexed/ -b 0  -m
split_mapping/1.2myco_mapping.txt -M 1  -n 1000000 -l 420 -k

*** Reverse complement the slow-growing mycobacteria datasets

adjust_seq_orientation.py -i combined_2myco_seqs_len420_soil.fna
adjust_seq_orientation.py -i combined_2myco_seqs_len420_water.fna

**** Chimera check, denoise and cluster data into OTUs using OTUPipe

/usr/local/qiime-svn-2011-09-22/bin/pick_otus.py -i
combined_mycoB_seqs_len400_soil.fas -d --db_filepath gold.fa -m
usearch -o otu_pipe_mycoB_soil_len400_97/ --word_length 64 --minsize
2 -d

/usr/local/qiime-svn-2011-09-22/bin/pick_otus.py -i
combined_mycoB_seqs_len400_water.fas -d --db_filepath gold.fa -m
usearch -o otu_pipe_mycoB_water_len400_97/ --word_length 64 --
minsize 2 -d

/usr/local/qiime-svn-2011-09-22/bin/pick_otus.py -i
combined_2myco_seqs_len420_soil_rc.fna  --db_filepath gold.fa -m
usearch -o otu_pipe_2myco_soil_len420_97/ --word_length 64 --minsize
2

/usr/local/qiime-svn-2011-09-22/bin/pick_otus.py -i
combined_2myco_seqs_len420_water_rc.fna --db_filepath gold.fa -m
usearch -o otu_pipe_2myco_water_len420_97/ --word_length 64 --
minsize 2


**** pick representative sequences for OTUs (most abundant sequence
in the OTU)

pick_rep_set.py -i otu_pipe_mycoB_soil_len400_97.txt -f
combined_mycoB_seqs_len400_soil.fna -o rep_set.fna

pick_rep_set.py -i otu_pipe_mycoB_water_len400_97.txt -f
combined_mycoB_seqs_len400_water.fna

pick_rep_set.py -i otu_pipe_2myco_soil_len420_97.txt -f
combined_2myco_seqs_len420_soil_rc.fna -o rep_set.fna


pick_rep_set.py -i otu_pipe_2myco_water_len420_97.txt -f
combined_2myco_seqs_len420_water_rc.fna -o rep_set.fna
```

```
*** Taxonomic Assignment with local BLAST (SILVA_Myco database)

assign_taxonomy.py -i rep_set.fna -m blast -t
taxonomy_mapping_file.txt -r myco_1020.fas

assign_taxonomy.py -i combined_mycoB_seqs_len400_soil.fna -m blast -
t taxonomy_mapping_file.txt -r myco_1020.fas

assign_taxonomy.py -i combined_mycoB_seqs_len400_water.fna -m blast
-t taxonomy_mapping_file.txt -r myco_1020.fas

assign_taxonomy.py -i combined_2myco_seqs_len420_soil_rc.fna -m
blast -t taxonomy_mapping_file.txt -r myco_1020.fas


assign_taxonomy.py -i combined_2myco_seqs_len420_water_rc.fna -m
blast -t taxonomy_mapping_file.txt -r myco_1020.fas


**** Taxonomic Assignment with RDP

assign_taxonomy.py -i rep_set.fna -m rdp

assign_taxonomy.py -i combined_mycoB_seqs_len400_soil.fna -m rdp

assign_taxonomy.py -i combined_mycoB_seqs_len400_water.fna -m rdp

assign_taxonomy.py -i combined_2myco_seqs_len420_soil_rc.fna -m rdp

assign_taxonomy.py -i combined_2myco_seqs_len420_water_rc.fna -m rdp


*** Create PyNAST alignments of the representative sequences
align_seqs.py -i rep_set.fna -t core_set_aligned.fasta.imputed.txt


*** Alignment Filtering (Use:0.80 gap + 0.10 entropy)

filter_alignment.py -i pynast_aligned/rep_set_aligned.fna -o
pynast_aligned/ -e 0.10 -g 0.80


*** Phylogenetic Tree building
make_phylogeny.py -i pynast_aligned/rep_set_aligned_pfiltered.fasta
-o pynast_aligned/rep_set.tre


*** Making OTU Table

make_otu_table.py -i otu_pipe_mycoB_soil_len400_97.txt -t
blast_assigned_taxonomy/rep_set_tax_assignments.txt -o otu_table.txt

make_otu_table.py -i otu_pipe_mycoB_water_len400_97.txt -t
blast_assigned_taxonomy/rep_set_tax_assignments.txt -o otu_table.txt

make_otu_table.py -i otu_pipe_2myco_soil_len420_97.txt -t
blast_assigned_taxonomy/rep_set_tax_assignments.txt -o otu_table.txt
```

```
make_otu_table.py -i otu_pipe_2myco_water_len420_97.txt -t
blast_assigned_taxonomy/rep_set_tax_assignments.txt -o otu_table.txt

*** Number of of seqs assigned to OTUs
per_library_stats.py -i otu_table.txt


*** Beta-diversity PCoA plots

beta_diversity_through_plots.py -i otu_table.txt -m
combined_mapping.txt -o beta_div/ -f -t pynast_aligned/rep_set.tre -
-suppress_distance_histograms -e X


*** Taxa summary through plots - by Sort column

summarize_taxa_through_plots.py -i otu_table_sorted.txt -m
combined_mapping.txt -o taxa_plots_sorted/ -f -p
qiime_parameters.txt'


*** Alpha diversity

alpha_rarefaction.py -t pynast_aligned/rep_set.tre -m
combined_mapping.txt -i otu_table.txt -o alpha_div/ -n 20

alpha_diversity.py -i otu_table.txt -m PD_whole_tree -o
alpha_div.txt -t pynast_aligned/rep_set.tre


*** OTU significance tests- Pearson correlation

otu_category_significance.py -i otu_table.txt -m
mapping_ENVO_soil.txt -s correlation -c elevation -f 2 -o
otu_catsig_corr_elevation.txt
otu_category_significance.py -i otu_table.txt -m
mapping_ENVO_soil.txt -s correlation -c latitude -f 2 -o
otu_catsig_corr_latitude.txt
otu_category_significance.py -i otu_table.txt -m
mapping_ENVO_soil.txt -s correlation -c longitude -f 2 -o
otu_catsig_corr_longitude.txt
otu_category_significance.py -i otu_table.txt -m
mapping_ENVO_soil.txt -s correlation -c pH -f 2 -o
otu_catsig_corr_pH.txt
otu_category_significance.py -i otu_table.txt -m
mapping_ENVO_soil.txt -s correlation -c temp -f 2 -o
otu_catsig_corr_temp.txt
otu_category_significance.py -i otu_table.txt -m
mapping_ENVO_soil.txt -s correlation -c water_content_soil -f 2 -o
otu_catsig_corr_watercontent.txt

otu_category_significance.py -i otu_table.txt -m
mapping_ENVO_freshwater_habitat.txt -s correlation -c elevation -f 2
-o otu_catsig_corr_elevation.txt
otu_category_significance.py -i otu_table.txt -m
mapping_ENVO_freshwater_habitat.txt -s correlation -c latitude -f 2
-o otu_catsig_corr_latitude.txt
otu_category_significance.py -i otu_table.txt -m
mapping_ENVO_freshwater_habitat.txt -s correlation -c longitude -f 2
-o otu_catsig_corr_longitude.txt
```

```
otu_category_significance.py -i otu_table.txt -m
mapping_ENVO_freshwater_habitat.txt -s correlation -c pH -f 2 -o
otu_catsig_corr_pH.txt
otu_category_significance.py -i otu_table.txt -m
mapping_ENVO_freshwater_habitat.txt -s correlation -c temp -f 2 -o
otu_catsig_corr_temp.txt
```

Appendix 2: Abstracts for papers to be submitted

Biogeographical trends in Climate and Latitude determine the diversity and community composition of Mycobacteria

The biogeography of mycobacteria was elucidated on a countrywide scale across varied climatic regions. This is of particular interest as members of the *Mycobacterium* genus are opportunistic pathogens and are also hypothesized to be responsible for the reduction in Bacillus Calmette-Guerin (BCG) vaccine efficacy. With the objective of sampling from a country with a high prevalence of tuberculosis and from an area with varied terrain, soil and water samples were collected from forty-two villages across Ethiopia. An amplicon based pyrosequencing approach was employed to target the 16S rRNA gene and used to determine the composition of species belonging to the *Mycobacterium* genus and a separate pyrosequencing assay specifically for a group of slow-growing mycobacteria. Our findings demonstrate the importance of elevation and temperature for the community composition of mycobacteria in soil. A non-linear relationship was observed between elevation and the outcome variables *Mycobacterium* species richness, diversity and abundance. In contrast latitude was the primary factor to explain the composition and diversity of mycobacteria in water samples. This study has demonstrated significant spatial and environmental correlations with mycobacteria and specifically slow-growing mycobacteria distributions. These results have implications for the risk of infection and similar biogeographical surveys on a worldwide scale may provide improved correlations with BCG vaccine efficacy.

High diversity of mycobacteria and detection of *M. bovis* in Ethiopian soils and water

The species within the genus *Mycobacterium* are commonly detected in a variety of environments, many are important in soil degradative processes however some are opportunistic pathogens. Soil and water source samples were collected from a total of forty-two residential sites in nine different administrative regions and a total of eleven different climatic zones. To determine community composition, metagenomic DNA was extracted and amplicon pyrosequencing was employed to target the 16S rRNA gene with specific targets for the *Mycobacterium* genus and a selected group of slow-growing mycobacteria. A quantitative PCR (qPCR) assay targeting the RD4 scar region was used to determine the abundance of *M. bovis* in the environment. These approaches enabled a comprehensive characterization of the relative abundance and distribution of mycobacteria on a countrywide spatial scale. The study revealed a greater diversity of both fast-growing (222 species) and slow-growing mycobacteria (59 species) than previously reported and indicated a marked heterogeneity in community structure between different environmental sources. Prevalent species in soil were the fast growers *M. neglectum, M. moriokaense* and the slow growers *M. malmoense* and *M. colombiense*, in contrast to water which were the fast growers *M. aurum sp*. ATCC 23070, *M. neoaurum* and the slow-growers *M. gordonae* and *M. colombiense*. The species *M. bovis* was detected in six environmental samples where real-time qPCR abundance ranged from approximately 7 copies per ml to 3000 copies per gram. Results indicate ample opportunity for human exposure to mycobacteria with potential pathogenic and opportunistic species both in soil and water sources.