The genetic diversity of *Mycobacterium avium* subsp. *hominissuis*

Inaugural-Dissertation

To obtain the academic degree

Doctor rerum naturalium (Dr. rer. nat.)

Submitted to the Department of Biology, Chemistry and Pharmacy

of the Freie Universität, Berlin

by

ANNESHA LAHIRI

from Kolkata (India)

February 2014

This work was accomplished between November 2010 and February 2014 at the Robert Koch Institute under the supervision of Dr. Astrid Lewin.

1st Reviewer: Prof. Dr. Rupert Mutzel
2nd Reviewer: Prof. Dr. Reinhard Burger

Date of Defence: 19.05.2014

Acknowledgement

I would like to thank my supervisor and guide, Dr. Astrid Lewin for her continuous support and patience that guided me past every obstacle during the entire duration of the project. I would also like to thank my supervisor, Prof. Dr. Rupert Mutzel and my research advisor, Prof. Dr. Lothar H. Wieler for their incredible support and advice throughout this work. I am grateful to the President of the Robert Koch-Institut, Prof. Dr. Reinhard Burger for giving me an opportunity to work in this institute.

Next, I would like to thank my parents for believing in me and encouraging me to fulfill my dreams.

My special thanks go to Rohan Pawar, Elisabeth Kamal, Robert Hauffe, Ralph Kunisch and Faisal Khattak for their guidance and assistance throughout the doctoral work. Finally I would like to express my gratitude to Dr. Esther-Maria Antáo who has been a constant and steadfast source of enthusiasm right from the very start of the project.

A.Table of Contents

B. Table of Figures 7

C. List of Tables 8

1. Introduction 9

1.1. The genus *Mycobacterium* 9

1.2. Shaping of the present day mycobacterial pathogen 11

1.3. Non tuberculosis mycobacteria 13

1.3.1. Ecology of Non tuberculosis mycobacteria 14

1.3.2. Commonly occurring Non tuberculosis mycobacteria 14

1.3.3. Identification of Non tuberculosis mycobacteria 16

1.3.4. Mycobacterium avium hominissuis – the most prominent of NTM 18

1.4. Genome of Mycobacterium avium hominissuis 19

1.4.1. Genome sequencing and insights into next generation sequencing 20

1.4.2. Understanding Mycobacterium avium hominissuis through genomic islands 21

1.4.3. Single nucleotide polymorphism (SNP) analysis in MAH 22

1.5. Rationale behind the project 23

1.6. Aims of the project 24

2. Materials and Methods 25

2.1. Collection of soil, dust, biofilm and water in Germany and in India 25

2.2. Isolation of NTM from soil, water, biofilms and dust from Germany and India 27

2.3. Media and growth conditions 27

2.4. List of bacterial strains used in this study 28

2.5. Molecular Biology techniques 29

2.5.1. Isolation of mycobacterial DNA 29

2.5.2. Identification of NTM strains 30

2.6. Genome sequencing of MAH strains from dust & child suffering from lymphadenitis.. 32
2.7. Softwares used for the identification of islands in MAH 34
2.8. SNP pattern recognition within MAH genomes 34
2.9. Additional materials used in the project 35
3. Results 37
3.1. Questioning the ecology of MAH in Germany and India 37
3.1.1. Inference from personal communications about Mycobacterium avium in Germany 37
3.1.2. Investigation of MAH in soil, dust, biofilms and water from Germany 39
3.1.3. Investigation of NTM and MAH in soil and dust from India 44
3.2. Sequencing of MAH genomes from Germany 50
3.2.1. Roche 454 sequencing platform for MAH genomes in Germany 50
3.2.2. Sequencing MAH genomes from strains in dust and a child with lymphadenitis 51
3.2.3. Raw reads and raw data processing 54
3.2.4. Denovo assembly of sequenced genome data 55
3.2.5. Processing of Ion torrent data 56
3.2.6. Submission of data into Genbank 58
3.3. Identification of a genome island in MAH with a flexible gene pool 59
3.3.1. Identification of regions specific to the MAH by Vista gateway 59
3.3.2. Identification of probable islands using Islandviewer 61
3.3.3. Identification of a 47 Kb genomic island in MAH 104 62
3.3.4. Identification of genomic islands in other sequenced MAH strains 68
3.4. SNP analysis of 3 genes in 36 MAH strains 79
3.4.1. Phylogenetic analysis of MAV_0846, MAV_0847 and MAV_0853 79
3.4.2. Phylogenetic analysis of the concatenated genes across thirty six strains 84

4. Discussion 85
4.1. Environmental MAH in Germany 85
4.2. Environmental NTM in India 88
4.3. Sequencing of two MAH genomes 93
4.4. Identification of a genome island in MAH 104 which represents a zone of diversity 95
4.5. SNP analysis of three MAH genes and their implications towards identifying plausible routes of infections. 97
5. Summary 99
6. Zusammenfassung 100
7. References 101
8. Appendix 114
8.1. List of abbreviations 114
8.2. Marker used during gel electrophoresis 116

B. Table of Figures

Figure 1: Estimated incidence of tuberculosis in the year 2012 10
Figure 2: Schematic depiction of the evolution of the mycobacterial pathogen..... 12
Figure 3: Collection points for environmental samples in Germany 25
Figure 4: States or counties in India where soil and dust was collected 26
Figure 5: Representative PCR results for detection of MAH in the environment 40
Figure 6: Representative gel for 16S r-RNA (complete) PCR for the Indian isolate 44
Figure 7: Representative gel for HSP65 gene specific PCR for the Indian isolates 45
Figure 8: Screen shot of NCBI BLAST analysis of complete 16S r-RNA sequencing..... 47
Figure 9: Illumina bio-analyzer profile of amplified DNA of MAH 27-1 52
Figure 10: Illumina bio-analyzer profile of amplified DNA product in MAH 2721 52
Figure 11: Bio-analyzer profile and the peak table of Ion torrent sequencing of MAH 27-1 53
Figure 12: Bio-analyzer profile and the peak table for ion torrent sequencing of MAH 2721..... 53
Figure 13: A representative screenshot of comparative genome analysis using VISTA browser. 60
Figure 14: A representative NCBI BLAST analysis of the Genome segment 3 63
Figure 15: Gene organization of the 47 Kb Genome Island in MAH 104..... 64
Figure 16: Duplication region in MAH 104 66
Figure 17: The GC curve of the island using the softwares geneious and the GC profile..... 67
Figure 18 :Gene organisation in MAH TH135 69
Figure 19: Gene organisation in MAH 27-1 71
Figure 20: Gene organisation of the genome island in the MAH strain 2721 72
Figure 21: Gene organisation of the island in the MAH isolate from deer MAH 10-4249..... 74
Figure 22: Gene organisation of the genome island in MAH 10-5606..... 76
Figure 23: The comparative analysis of flexible gene pool in MAH 78
Figure 24: Phylogenetic analysis of MAV_0846 (Carveol dehydrogenase). 81
Figure 25: Phylogenetic analysis of MAV_0853 - ddpX - D-alanyl-D-alanine dipeptidase..... 82
Figure 26: Phylogenetic analysis of MAV_0847 (TetR family transcriptional regulator) 83
Figure 27: Phylogenetic analysis of concatenated genes in thirty six MAH strains..... 84
Figure 28: Map of India with regions having performed NTM epidemiology studies..... 90
Figure 29: 100 bp plus DNA ladder used in gel electrophoresis 116

C. List of Tables

Table 1: MAH strains used in this study....28
Table 2: Essential components of a PCR reaction....31
Table 3: Essential components of a sequencing reaction....32
Table 4: Primers used for the identification of NTM species....32
Table 5: List of MAH strains sequenced by Roche/454 FLX Pyrosequencer....33
Table 6: List of additional chemicals used in the project....35
Table 7: List of instruments and materials during experimental proceedings....36
Table 8: NTM infections in hospitalized children in Germany, April 2003 to September 2005 ..38
Table 9: Mycobacterial species cultivated from surface water and drinking water in Germany ..39
Table 10: Environmental samples and isolation of MAH in Germany 42
Table 11: MAH from soil and dust identified in this study....43
Table 12: NTM identified from soil in India from this study....48
Table 13: Mycobacterial species isolated from Indian soil and dust....49
Table 14: Summary of the genome data obtained by Roche 454 sequencing platform 51
Table 15: The quality control (QC) statistical data of reads from MAH 27-1 and MAH 2721 54
Table 16: The processed data analysis from MAH 27-1 and MAH 2721 54
Table 17: Assembly statistics of contigs and scaffolds based on length 56
Table 18: Read length distribution of Ion torrent processed data....57
Table 19: Statistics of draft scaffolds generated from Illumina and Ion torrent sequencing data .58
Table 20: Seven regions specific to the MAH identified by comparative genomics....60
Table 21: Predicted islands in the seven regions specific to the MAH 62
Table 22: List of genes found within the genome island in MAH 104....65
Table 23: The genes found in the genome island in MAH TH 135....69
Table 24: The genes within the genome island in MAH 27-1....70
Table 25: The genome island specific genes in deer isolate MAH 10-4249 73
Table 26: The gene organisation in the MAH isolate from Pig (MAH 10 -5606)....77

1. Introduction

Emerging infectious diseases caused by opportunistic bacteria can pose a substantial threat to human population. Though there has been a long lasting understanding about the origin of bacterial pathogens and the molecular attributes of virulence, getting a perspective about the origin, evolution and ecology of bacteria is vital towards deciphering their mode of transmission and pathogenicity. Understanding the finer differences between pathogenicity and virulence is indispensable, as the former is a qualitative ability of a microbe to cause disease and the latter is a quantitative property referring to the degree of damage caused in the host by the bacteria [1]. A successful bacterial pathogen not only exploits its host to support its own survival but also regulates the host immune response to ensure its survival by striking a genetic balance between failed immunity and damaging hyper-immunity [2].

Functional and comparative genomics have been powerful tools, in this regard to understand the bacterial genome fluidity, genetic variability and assessing the most predominant strains in epidemiological settings [3]. The genomic era has brought forth an increased availability of DNA sequence information from multiple pathogenic and nonpathogenic variants of individual bacterial species thereby providing a rapid and unbiased means of uncovering the fundamental basis of pathogenicity and new approaches to combat infectious diseases [4].

1.1 The genus *Mycobacterium*

The unicellular, aerobic, gram-positive *Mycobacteria* are best defined by their acid-fast characteristics, mycolic acid rich cell walls, and high GC contents (61 to 71%) [5]. There are over 164 validated species and subspecies of slow growing and rapidly growing mycobacteria known today [6]. The genus harbors three major groups, the *Mycobacterium tuberculosis* complex (MTBC), the *Mycobacterium leprae* and all other non tuberculosis mycobacteria (NTM) [7]. *Mycobacterium leprae* causes leprosy, the MTBC causes tuberculosis (TB) and the NTM are opportunistic pathogens causing lung infections, lymph node infections, skin and joint infections, soft tissue infections and disseminated infections in immune-compromised hosts and Acquired Immuno Deficiency Syndrome (AIDS) patients. The MTBC currently comprises of seven species: *Mycobacterium tuberculosis* (MTb), *Mycobacterium africanum*, *Mycobacterium bovis*, *Mycobacterium microti*, *Mycobacterium pinnipedii Mycobacterium caprae*, and *Mycobacterium canettii*. [7]. MTb is the most

prominent member of this group and has been in the forefront of global health problems as it causes approximately 8.6 million cases of TB and 1.3 million deaths every year.[8]

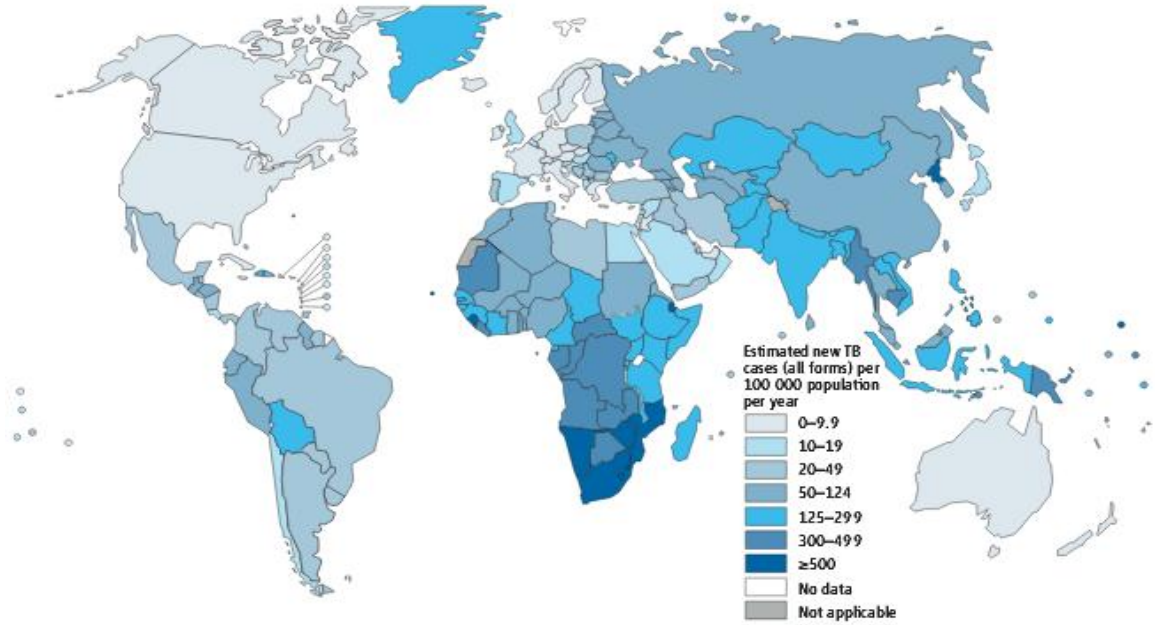


Figure 1: Estimated incidence of tuberculosis in the year 2012

Source: WHO global tuberculosis report 2013 [8]

The liability of disease due to mycobacteria can be measured in terms of incidence, prevalence and mortality. Incidence is defined as the number of new cases of the mycobacterial disease including the risk groups and relapse cases arising in a stipulated time period that corresponds to one year. Prevalence is defined by the number of people affected by the disease at a given time point. Mortality, on the other hand is the number of deaths caused by the disease in one year [9]. Efforts have been made by the World Health Organization (WHO) to introduce an international blue-print for TB control, a five-component directly observed treatment strategy (DOTS) which covers standard recording and reporting by national TB control programs (NTPs), diagnosis using sputum smear microscopy, a steady supply of first-line anti-TB drugs, short-course chemotherapy and political commitments [10]. Almost all countries have adopted this strategy and there has been a significant headway towards successful management of TB cases.

1.2 Shaping of the present day mycobacterial pathogen

Gene acquisition by horizontal gene transfer (HGT) and extensive genomic downsizing due to deletions have been dominant contributors to the unique biphasic genesis of MTb [11-13]. Comparative studies of the first available mycobacterial genomes apart from MTb (4.4 Mb) were those of *M. smegmatis* (7 Mb) and these revealed that MTb evolved by the loss of genetic material or deletion. On the other hand, additional comparative studies with the genomes of *M. marinum* (6.6Mb) [14] and *M. ulcerans* proposed that gene acquisition of foreign DNA through HGT was also a fundamental part of evolutionary history of MTb [13]. These examples suggest that both expansion and reduction of genomes have generated sufficient allelic diversity to allow the individualized evolution of mycobacterial species. HGT promotes rapid acquirement of novel functions by offering a spatial and temporal selective advantage to the species acquiring the genetic material [4]. Genomic deletions, on the other hand can be described as loss of genetic material. The loss of genetic material can be used as a comparative tool for assessing genetic variability compared to the reference strains that could account for transmission and pathogenesis associated genes [3].

Evidences show that the origin of the present day mycobacterial pathogen was strictly environmental as shown in Figure 2 (a) [15, 16]. Horizontal exchange in the progenitor clone played an important part in the evolution of MTb [17, 18]. A series of sequential selection of phenotypes such as protection from free living predators and capacity to transmit between hosts must have been selected thereafter for it to evolve as a terminally differentiated pathogen as represented in the Figure 2 (b) [1, 3, 19].

Cross species comparisons have been performed to assess these phenotypic trait acquisitions and many of the selected attributes share commonality to genomes of several NTM. Hence an increase in the number of recently available NTM genomes has provided us with the optimal reference point for revaluation of MTb evolution and questioning the determinants behind the virulence association of MTb [1].

(Table Contined on Next Page)

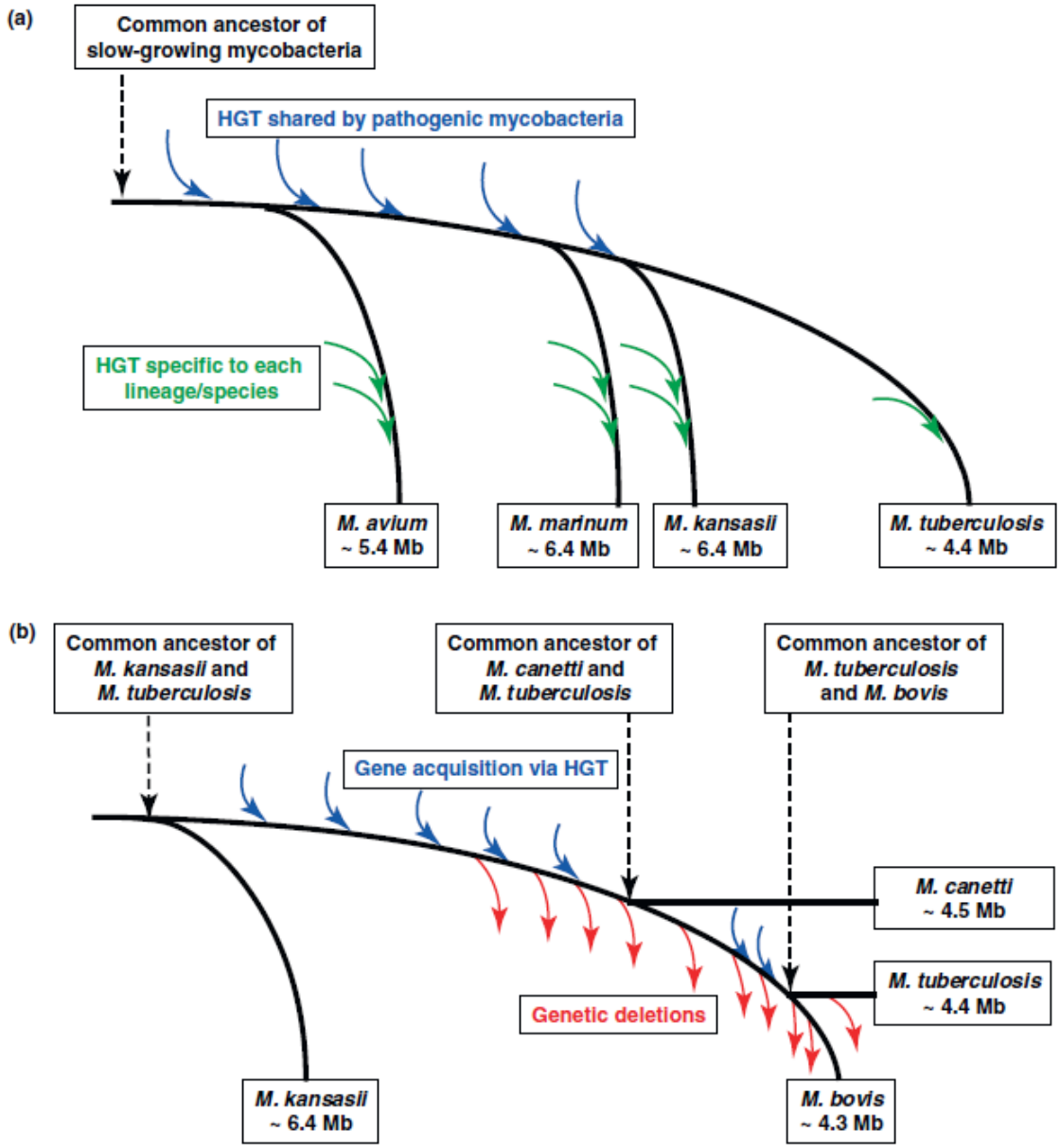


Figure 2: Schematic depiction of the evolution of the mycobacterial pathogen

The above Figure provides an overview about mycobacterial evolution [1].

(a) The common ancestor of slow-growing mycobacteria introduced genes through HGT and these genes were shared across all mycobacterial species represented by blue arrows. Specific genes in one or a few species are presented by the green arrows.

(b) MTb evolution progressed with common ancestors of M. kansasii and MTb. Alternating gene acquisitions shown by blue arrows followed by gene deletions (red arrows) resulted in the divergence of MTb and M. canetti strains.

1.3 Non tuberculosis mycobacteria

Those *Mycobacterium* species that are not members of the MTBC are often referred by the term ‘‘non tuberculosis mycobacteria’’ or ‘‘mycobacteria other than tuberculosis (MOTT)’’. Since these bacteria were different from the usual mycobacterial strains, they were also referred as ‘‘atypical,’’ originating from the misguided belief that they were uncommon strains of MTb [20].

NTM are opportunistic pathogens that cause lymphadenitis, lung infections, skin and soft tissue infections in immune-compromised hosts [21]. Infections from NTM were strictly pulmonary and nosocomial to some extent or restricted to cervical lymph nodes and skin prior to the AIDS epidemic. The emergence of the AIDS epidemic throughout the world brought a drastic change in the NTM depiction with increased infection rates in 25 to 50% of AIDS patients in the developed countries like United States and Europe [22, 23]. Today, NTM infections have become more prevalent in non HIV immune-compromised populations. Immune competent hosts, though uncommon, are also affected by these bacteria through trauma or surgical wounds.

NTM disease is found in high incidence in the developed countries like the United States and Europe, where the incidence of tuberculosis is low [24-26]. These pathogens are rarely found as cause of disease in the developing world where the burden due to the incidence of TB is still high. The rareness of NTM in these countries is not due to the absence of the organisms in these environments but rather, insufficient awareness and reporting along with lack of advanced diagnostic techniques [27]. Vaccination strategies like the Bacillus Calmette–Guérin (BCG) vaccination has been known to demonstrate protection against leprosy [28] . Its protective effect has been shown to reduce the incidence of leprosy. Likewise, it is unknown whether the BCG vaccination offers protective advantages to NTM infections in the developing TB endemic countries where BCG vaccination is still a trend for preventing TB infections.

1.3.1 Ecology of Non tuberculosis mycobacteria

NTM are opportunistic free-living saprophytes. Before the AIDS epidemic, NTM isolated from clinical sources were often considered as contaminants and discarded. But the AIDS epidemic brought about a radical change in the NTM profiles and they became prominent causative agents of disease in immune-compromised hosts. NTM prevalence varies according to different ecological and clinical settings. NTM are ubiquitous and are found in soil, dust [29], rivers, lakes oceans, drinking water distribution systems, in bath and showerheads and other manmade sites [30-32]. Different soil types like the presence of peat lands, varying levels of precipitation and concentrations of different metals ions like Fe, Al, Cu, and Cr can make a difference to the NTM concentrations. Furthermore the capacity of these mycobacteria to make biofilms [33, 34] and their resistance to chlorides [35] play a noteworthy part in their distribution in the environment [36]. Drinking-water distribution systems are the most common sources of NTM. Since they are recovered both from natural and constructed sources, the route of NTM infections is considered environmental unlike MTb, where transmission from one infected patient to another is a common phenomenon.

The ecology of NTM also varies with respect to TB endemic and TB non endemic settings. At present, the NTM to TB ratio is rather high in areas of low TB incidence. Developing countries with high TB prevalence have lesser recognized NTM infections. The lack of standardized and accepted criteria which defines the NTM respiratory diseases makes their reportage tougher in these countries.

1.3.2 Commonly occurring Non tuberculosis mycobacteria

NTM harbors more than 65 different kinds of environmental and opportunistic bacteria. They are traditionally classified as slow growing mycobacteria (SGM) and rapid growing mycobacteria (RGM) based on their differential growth rates [37]. RGM grows visible colonies within 3-7 days while SGM takes relatively longer (more than 7 days) [38]. *Mycobacterium avium* complex (MAC), *M. kansasii*, *M. gordonae* and *M. xenopi* are the commonly occurring SGM in clinical specimens while *M. chelonae*, *M. fortuitum,* and *M. abscessus* are the most prevalent RGM. A recent study by the NTM-Network European

Trials Group (NET) provided a snapshot of NTM species distribution in thirty countries across six continents. MAC followed by *M. gordonae* and *M. xenopi* were among the more predominant NTM in most countries [39].

MAC, the most prominent of the SGM includes several species and subspecies like the *M. avium* and *M. intracellulare*. While *M. intracellulare* commonly causes infections in immune-compromised hosts, *M. avium* comprises of four major subspecies of avian as well as mammalian nonobligatory pathogens. These are *M. avium avium* (MAA), *M. avium silvaticum* (MAS), *M. avium hominissuis* (MAH), and *M. avium paratuberculosis* (MAP) [40, 41] . While MAA and MAS cause Tb like diseases in birds [42], MAP is better known for causing Johne's disease in ruminants and probably Crohn's disease in humans [43].MAH is an important intracellular human pathogen affecting immune-compromised populations like older patients and children [44].

Infections due to *M. kansasii* were more predominant before the AIDS epidemic, however it still remains the second most important NTM after MAC, causing infections mostly in UK and Western Europe [45-47]. They are slow growing NTM that prefer a temperature range of 32°C to 42°C for their growth [20]. The ecological niche of this bacterium has been documented as water [45]. These bacteria are rarely recovered from soil. People with preexisting pulmonary disease, cancer, alcoholism and cystic fibrosis pose a risk to these infections.

M. gordonae is a nonpathogenic commensal that is the least pathogenic of the NTM. This avirulent bacterium is ubiquitous and is commonly isolated from soil and water [48]. It is mostly regarded as a contaminant. Cutaneous infections caused by *M. gordonae* are rare though there have been reports of their nosocomial transmission [48, 49].

Mycobacterium xenopi is a waterborne SGM which prefers growth at 45°C [50] and its growth preference differentiates it from the MAC. It is a common causal agent of mycobacterial pulmonary infections along with MAC and *M. kansasii*. Tap water and heated water distribution systems are common sources of their isolation but the incidence of their infections varies significantly with different geographical areas. The clinical presentation of the lung infection caused by *M. xenopi* mimics the disease caused by MTb and MAC [51]. More pulmonary *M. xenopi* infections are reported from Western Europe

[52]. Patients with underlying lung disease run a risk to these infections [53]. Extra pulmonary cases are rare and occur in immune-competent patients. However, knowledge on the pathogenesis underlying *M. xenopi* infections is limited.

Infections in humans by RGM are nosocomial and are frequently caused by *M. fortuitum*, *M. chelonae*, and *M. abscessus* [54]. Everyone is exposed to these bacteria because these bacteria are distributed in the environment in relatively large numbers. Pulmonary infections [55], skin and soft tissue infections [56], surgical wound infections and catheter-associated infections are the common manifestations of RGM disease. Additional sources of infections involve contaminated exposures to medical equipment like bronchoscopes [54], and surgical implants like prosthetic heart valves [57]. *M. fortuitum* manifests in the form of infections after cardiac bypass surgery while *M. chelonae* and *M. abscessus* cause skin and soft tissue infections. *M. abscessus* is responsible for lung infections in humans. [58]. *M. fortuitum* or *M. abscessus* associated risk factors involve pulmonary infections caused by previous mycobacterial disease, cystic fibrosis and bronchiectasis. Kidney transplantation or chronic renal failures are risk factors promoting infections with *M. chelonae* [58].

1.3.3 Identification of Non tuberculosis mycobacteria

Mycobacterial identification in the clinical laboratory still remains a fastidious and time-consuming procedure [59]. Identification of mycobacteria to species and subspecies level is vital to evaluate the clinical significance of a positive culture and to facilitate the correct effective antibiotic therapy. Assembling them in groups creates difficulty in deciphering the speciation of the isolate and also leaves out the clinically important NTM. NTM can be identified based on different phenotypic properties like differential growth rate and pigmentation, chemotaxonomic testing like High pressure liquid chromatography (HPLC) and different genotypic methods like 16S-rRNA typing.

Growth rates and pigmentation studies are conventional methods of identifying mycobacteria. These methods are not only time consuming but are also associated with significant delays in diagnosis. Differential growth rates help with the preliminary broad

classification of NTM into RGM and SGM. Pigmentation and smooth colonies help with the segregation of NTM from MTb which has rough non pigmented colonies.

Chemotaxonomic testing with HPLC serves as a rapid and reliable tool for mycobacterial identification. Primary mycobacterial cultures can be directly used for analysis. However the recognition of newly identified NTM subspecies is difficult with HPLC. There have been reports about difficulty faced during the differentiating of *M. abscessus* and *M. chelonae* with HPLC [60].

To overcome the disadvantages of conventional and chemotaxonomic typing, genotypic and molecular methods of identification have been an essential tool to isolate and identify mycobacteria [61]. Sequence-based methodologies have replaced routine clinical laboratory methodologies as these are more specific and easier to standardize. Molecular identification not only facilitates an improved accuracy in identification but also improves the turnaround time [62]. The three molecular targets that have proven effective in typing mycobacteria are the 16S r-RNA gene [59, 63] and the Heat shock protein 65 (HSP65) gene analysis [64-66] and the 16-23S Internal transcribed spacer (ITS) [67, 68]. The typing of the 16S r-RNA gene is an alternative to phenotypic identification and makes significant contributions towards discovery of new species [69]. The strains are characterized by amplification, followed by probe hybridization, restriction length polymorphism analysis or sequencing. However limitation in determining phylogenetic relationships and subspecies identification has posed as limitations of 16S rRNA. Hence alternative approaches like HSP65 typing and 16-23S ITS or a combinational approach is usually used to type mycobacterial isolates. The HSP65 gene has highly conserved primary structures and high ubiquity thereby making it a useful phylogenetic marker. Since it is found in all bacterial species as a single copy [70] this gene is not easily transferred from one bacterium to another which makes it all the more suitable for phylogenetic understanding of closely related species [71]. The 16S-23S ITS has a small genetic locus that is well flanked by conserved regions of the r-RNA operon, containing both conserved and highly variable signatures. These sequences have shown high level of spacer sequence variation in mycobacteria. This potential target is hence useful to derive additional phylogenetic information [72, 73].

1.3.4 Mycobacterium avium hominissuis – the most prominent of NTM

MAC is the most frequent etiological agent causing non tuberculous mycobacteriosis in developed nations like Europe, North America, South America and Australia [39]. MAH, an important member of the *M. avium* and MAC causing diseases in humans was suggested as separate subspecies to distinguish *M. avium* found in humans and pigs from those isolated from birds [74]. The medical relevance of MAH has become increasingly important with the rising population of immune-compromised patients due to longer life expectancy and AIDS epidemic [75]. Immune-compromised subjects with underlying pulmonary disease , children and patients with cystic fibrosis run a potential risk to be affected by MAH [76-78]. The accelerated rise in the prevalence of MAH infections has stimulated the initiation of studies with regard to rapid recovery, identification, ecology and genomics of these bacteria. Questions have been raised concerning the portal of entry and routes of MAH infections. The routes of transmission still remain ambiguous. The possibility that the natural environment could play a defining role to support this hypothesis has been validated by the ubiquitous nature of this bacteria being isolated from soil, water, dust and aerosols [79]. The high incidence of MAH infections correlated with high numbers of these bacteria in drinking water [77] and dust from potting soil (especially peat) [80] also points towards environment as a source of human infections.

MAH strains are genetically diverse and elucidate different geographical and host-dependent variations in their genetic diversity [81]. Numerous studies have attempted to create the connection between MAH infections and its putative reservoirs but a clear picture is not available yet [82].

Comparative genomics has provided genomic signatures that define members of a particular species and help to differentiate different clinical isolates, distinct host-associated variants and ecotypes (adapted to a specific habitat). Means of effectively discriminating among the closely related species and subspecies in MAC is vital towards better diagnostics and our understanding of epidemiology of these bacteria. The recent available methods of diagnosis use a five target multiplex PCR. This comprises of a 16S r-RNA gene target that identifies mycobacteria of all species, a chromosomal target called DT1 which is specific to MAA and *Mycobacterium intracellulare* and three insertion

elements namely the IS1311, IS900 and IS901. These targets help with the identification of MAH [83]. Other methods use a multiplex PCR based on IS900, IS901, IS1245 and the DNAJ gene. This method allows the detection of MAP, MAH and MAA/MAS in a single tube [84].

The characteristic features of MAH include the possession of multiple copies of IS1245, a variable 16S-23S ITS sequence and a more flexible and wider temperature range (24°C - 45°C). The absence of IS901 insertion sequence in the MAH makes it a defining feature to distinguish MAH from MAA and MAS [74]. Another insertion element, the IS900 is a defining feature to distinguish between MAH and MAP. Colony morphology variation is a characteristic feature of MAH isolates and three colony variants are observed in MAH. (i) a smooth, opaque, and dome type morphology (ii) a smooth, transparent, and flat type morphology and (iii) a rough type variant. Smooth transparent or smooth opaque types or a mixture of the two are observed most of the time in clinical isolates and AIDS patient isolates. Very little is known about genetics and regulation of colony variation [85].

1.4 Genome of Mycobacterium avium hominissuis

MAH is frequently isolated from soil, dust, water, children suffering from lymphadenitis [86] and immune-compromised older patients with cystic fibrosis or preexisting lung infections [87]. However only two completely annotated genomes, representative of this bacteria are available in the genome database. MAH 104, was sequenced at The Institute for Genomic Research (TIGR), back in the 1980s from an adult patient with AIDS in Southern California [88]. It has a genome size of 5475491 bp. It is predicted to encode 5120 protein genes, 46 t-RNA genes and 3 r-RNA genes. Annotation studies of this genome have resulted in the identification of genes unique to *M. avium* and other genes that share homology to other virulent mycobacteria. Evaluation of the whole-genome DNA microarray and PCR in 43 clinical isolates of *M. avium* related to strain 104, have shown a polymorphism rate of 13.5% between isolates. An eight fold-greater strain-to-strain variability was observed on a genomic level when compared to MTb isolates [89]. These observations with regard to genomic heterogeneity raises questions about strain 104 being a representative of virulent MAH isolates. The genetic analysis of the characteristics of MAH genome is still open.

The very recent introduction of a new MAH strain TH135 isolated from HIV-negative patient in Japan with pulmonary MAH disease has facilitated comparative analysis between the reference strain from an HIV positive patient and a second strain from an HIV negative patient [90]. It is plausible that the MAH repertoire and genetic profile of NTM in the east like Asia, east Africa and Japan may be well distinct from the profiles observed in the temperate west due to different geoclimatic conditions. The genome of TH135 is much smaller (4951217 bp) in comparison the MAH 104 genome. Since genomes of strains isolated from the environment and other clinical subjects like children suffering from lymphadenitis are still missing, the exact picture about the routes of infection of these bacteria continues to be uncertain. Introduction of more MAH genomes from varied niches would help to deliver better insights into this bacterium.

1.4.1 Genome sequencing and insights into next generation sequencing

Genome sequencing has seen a fundamental shift from automated Sanger sequence to next generation sequencing (NGS) which offers a relative advantage of large amount of data in a very short period of time [91, 92]. NGS requires the sequencing of every base several times as multiple observations of every base result in reliable base calling. Base calling accuracy is measured by Phred quality score (Q score) and this is used to assess the accuracy of sequencing platform. It indicates the probability that a given base is called incorrectly by the sequencer. Reliable base calling is expressed in terms of coverage metrics which is described as the number of times target sequences in a genome have been sequenced. The coverage of the sequence is directly proportional to the depth of the sequencing. Higher the coverage of the sequences, the greater is the associated depth of the genome sequencing. The ability of sequencing whole genomes has facilitated large scale comparative and evolutionary studies to understand different species better. The most commonly used sequencing platforms are Roche/454 FLX Pyrosequencer, Illumina Genome Analyzer and the Ion semiconductor sequencer. The Roche/454 FLX Pyrosequencer offers longer reads that helps with the mapping of repetitive regions as is the case with mycobacterial genomes. Reads are strings of newly acquired nucleotide bases that may vary in size depending on the kind of sequencing platform used. However the costs of reagents for this sequencing platform are very high. Longer reads are

beneficial towards construction de novo assemblies. Illumina Genome Analyzer on the other hand is the most widely used sequencing platform that offers more but shorter reads and is used for variant discovery by whole genome sequencing. The Ion semiconductor sequencer is a low cost rapid sequencer that can be used as a bench top machine. It offers shorter reads and a higher error rate in comparison to Roche/454 FLX Pyrosequencer [93].Single reads and paired end reads also have a big role in deciphering genomes. A paired end read is a simple modification to the standard single-read DNA library preparation facilitating the reading of both the forward and reverse template strands of each sequence during one paired-end read. Paired end reads enable precise alignment of reads by providing long range positional information.

Repeats in genomes arise when extra copies of biological sequences are introduced or inserted into genomes and these have always presented a technical challenge for alignment and assembly of genomes. Repeats generate ambiguities while assembling genomes and cause errors while interpreting results. Ignoring repeats is not a viable option as important biological informations are missed [94, 95]. The large number of repeats in the mycobacterial genomes makes it compulsory for the use of two or more different sequencing platforms to decipher the genome. Usually a combination of Roche/454 FLX Pyrosequencer and Illumina Genome Analyzer or Illumina Genome Analyzer and the Ion semiconductor sequencer are used to acquire both long and short reads that makes deciphering the genome an easier task.

1.4.2 Understanding Mycobacterium avium hominissuis through genomic islands

Bacterial genomes are dynamic entities comprising of core gene pools and flexible or adaptable gene pools. The core gene pool shares conserved genes within different isolates of the same species and contains specific information pertaining to specific cellular functions whereas the flexible gene pool encompasses genes that confer beneficial advantages to the bacteria under certain circumstances [96]. The expansion and reduction of the flexible gene pools can be attributed to clonal divergence, clonal selection and gene exchange. Genome plasticity through gene deletion and gene acquisition builds flexible gene pools and supports adaptive evolution in bacteria [97]. The flexible genome comprises mostly of mobile genetic elements such as plasmids, phages and genomic

islands (GI). GI are mobile genetic entities or clusters of horizontally transferred genes that can vary in size from 10 kb to 200 kb and contribute to rapid evolution of bacteria and confer survival advantages to the bacteria. The GI can be further classified into pathogenicity, fitness, resistance and symbiotic islands based on the functionality of the mobile regions. Genomic islands can be identified in bacteria based on differential GC content, codon usage, 16-20 bp of direct repeats and transfer ribonucleic acid (t-RNA) gene screening [96, 98].

Members of *M. avium* are bound to be phenotypically and genotypically diverse based on their diverse host specificities and ecological niches. Certain subspecies like the MAA cause severe diseases in domestic and wild birds while the MAP is specific to cattle. MAH has human and pig specific relevance in hosts. The genetic distinction between these groups coupled with the identification of islands and long sequence polymorphisms (LSPs) help in drawing a broader picture of the evolution of these clusters and the closely related members of the individual species. This is essential for structuring of distinct genomic profiles of these bacteria. Several islands have been identified in *M. avium* but only a few of them have been designated specific functions. One of the most recent findings was the identification of an island in *M. avium* specific for macrophage and amoeba infections [99]. Another of its kind was the identification of a 38 Kb pathogenicity island in MAP that encodes cell surface proteins expressed in hosts [100]. Numerous LSPs have also been identified within the *M. avium* but the functions associated with these polymorphisms have not been labeled [89]. More research towards exploring the flexible gene pools in MAH would not only help us understand the pathogenicity profile of the bacterium but also the host specific as well as environmental specific preferences of these bacteria.

1.4.3 Single nucleotide polymorphism (SNP) analysis in MAH

Practices for ascertaining as well as discriminating between different strains of MAH are crucial for the defining the possible routes of infection. Defining meaningful boundaries between subspecies in bacteria is complicated, yet this grouping is necessary for strain classification. Since genomes and phylo-geographic correlates to the genomes of different

strains are beginning to be mapped out, understanding the genetic variability and epidemiological dynamics have become relatively simpler.

MAH are natural inhabitants of the environment and genetic variability in them can result in different host colonization and different infection targets. Phylogenetic grouping of MAH isolates is necessary as genetic relationship between clinical pathogenic strains, environmental strains and reference strains can provide perceptions into the epidemiological surveillance of these bacteria. Though several grouping techniques like the multi locus sequence typing (MLST), fluorescent amplified fragment length polymorphism (FAFLP) and ribotyping are available, most of these techniques are costly and time-consuming. The SNP analysis is a simple, reproducible phylotyping method that is based on the principle that mutations that occur within the genomes can be used to define and understand bacteria better. The SNP are advantageous in phylogenetics and strain group/lineages identification. Since SNPs have an important role in defining disease severity, identification of SNPs in MAH genomes from varied environmental samples like soil, water and dust along with clinical subjects like immune-compromised older patients and children with lymphadenitis can help with establishing correlation between patient specific SNPs and environment specific SNPs. This can help us track the route of MAH infection.

1.5 Rationale behind the project

Since NTM repertoire under high and low Tb epidemic settings is expected to be different, the first goal of the study was to explore the NTM ecology and prevalence in a developed country like Germany and a developing country like India. MAH is a prominent NTM in Germany. Hence special importance was paid towards genome sequencing and SNP analysis of MAH in Germany. Introduction of two new European MAH genomes into the database to facilitate better comparative genomics was an important aspect of the study. The presence of genome islands in MAH was examined for deeper understanding of flexible gene pools in these bacteria. SNP polymorphism study, to deliver a better overview of the mutations within the strains was an important part of the study.

1.6 Aims of the project

a. Questioning the ecology of NTM and MAH in India and Germany
b. Sequencing MAH genomes from strains in dust and a child suffering from lymphadenitis
c. Identification of a genome island in MAH representing a flexible gene pool
d. SNP pattern recognition in MAH

2. Materials and Methods

The materials and methods comprises of sample collection methods implemented in Germany and India. NTM and MAH isolation techniques, softwares for genome island identification, genome sequencing and assembling methods and SNP analysis are also explained in this section.

2.1 Collection of soil, dust, biofilm and water in Germany and in India

Environmental NTM and MAH were investigated by the collection of soil, water, biofilm and dust. Soil and dust were collected in 50 ml sterile disposable conical falcon tubes (Carl Roth, Karlsruhe, Germany). Biofilms were collected by the use of four cotton swabs per sample. 90 ml – 500 ml of water from rivers, lakes and household taps were collected in sterile flasks. The contents of the water were centrifuged and then further processed. Sterile filters were also used alternatively to process the water samples and the contents of the filters were further treated. Soil and dust were filled up to half of the total content of the falcon tubes by using disposable sterile nitril gloves (Carl Roth, Karlsruhe, Germany) to avoid any contamination. House dust was collected from the contents of vacuum cleaners in households. Soil samples were taken either from close human environments (e.g. potting soil, garden soil, sand from playing grounds) or from areas involving lesser close contact with humans (e.g. forest, road side).

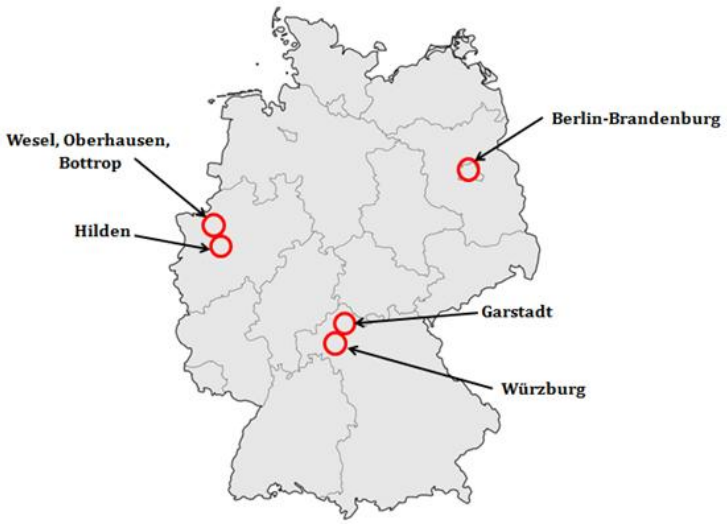


Figure 3: Collection points for environmental samples in Germany

The ecological samples in Germany were collected from Berlin, Brandenburg, Bavaria and North Rhine – Westphalia. Collection points in Germany are illustrated in Figure 3.

Environmental samples (soil and dust) in India were collected from the states of Andhra Pradesh, Odisha, West Bengal and Uttar Pradesh for isolation of NTM. The samples were collected from the state capitals like Hyderabad, Bhubaneswar, Kolkata and Delhi (Figure 4). Dust samples from rooms and the soil samples from sources close to human environments (potting soil, garden soil) and soil from farm land and outdoor areas such as road sides were collected in sterile falcon tubes and stored at 4° C until they were further processed.

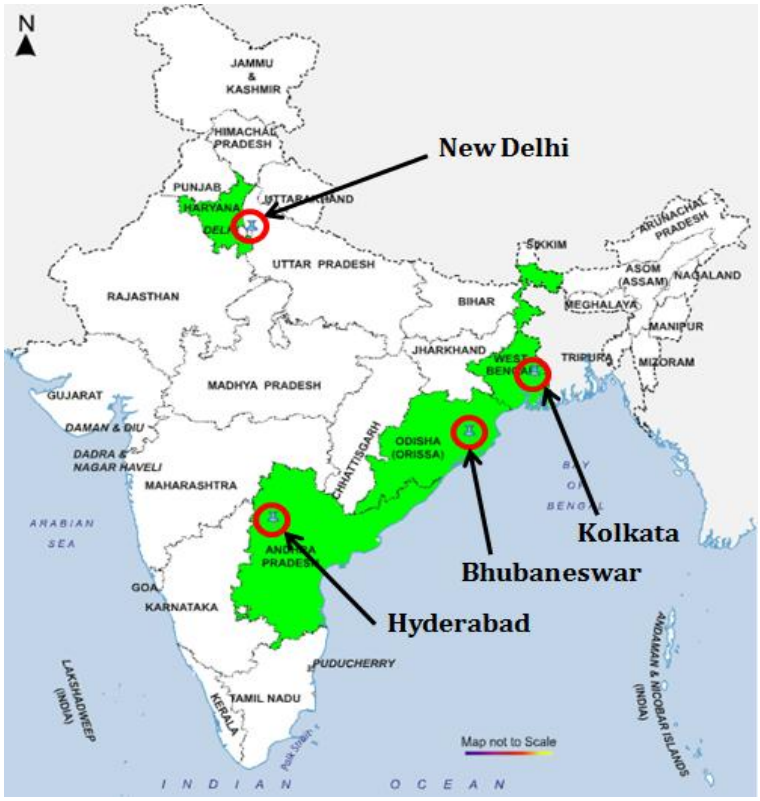


Figure 4: States or counties in India where soil and dust was collected

The above Figure is a representation of the states in India where the environmental samples were collected. The regions marked in green refer to the sample collection states whereas the areas marked in red color denote the states capitals and its surrounding areas where the samples were collected. The Figure above specifies Hyderabad, Bhubaneswar and Kolkata on the eastern coastal front of India as sample collection points. New Delhi, was also selected as the fourth collection point as it is the capital of India.

2.2 Isolation of NTM from soil, water, biofilms and dust from Germany and India

Two grams of soil or dust and swabs with biofilms from the sampling collection were transferred to new sterile facon tubes and fifteen ml of tryptone soya broth (TSB) (Oxoid, Hampshire, UK) was added to it. The contents of water were either centrifuged at 8600 g or passed through a sterile filter (0.22 µm Steritop-GP polyethersulfone (PES) Express Plus Membrane (Millipore Corporation, Billerica, MA, USA) before treatment with TSB. This was followed by vigorous vortexing. The assortment was shaken at 37°C for 5 hours to allow spore germination. Centrifugation at 500 g for 5 minutes at 4 °C was carried out to remove large particles. Five ml of the supernatant was then transferred to sterile centrifugation tubes and the mixture was treated with 2.5 ml of 1.5 M NaOH (Carl Roth, Karlsruhe, Germany), 2.5 ml of 0.3% Malachite green solution (Applichem Biochemica, Darmstadt, Germany) and 2.5 ml of 2.5 mg/ml Cycloheximide (Merck, Darmstadt, Germany). This was followed by 30 minutes of incubation at room temperature. Neutralization was carried out thereafter by addition of 2.5 ml of 1.5 M HCl (Carl Roth, Karlsruhe,Germany), followed by a centrifugation at 8600 g for 15 min at 4°C. The pellets were washed with sterile water and resuspended in 400 µl of sterile dH_2O. The suspension and its 1:10 dilution was then plated on Middlebrook 7H11 Agar (BD Biosciences, Heidelberg, Germany) with 10% ADC (2 g glucose, 5 g BSA, 0.85 g NaCl in 100 ml dH_2O) containing Cycloheximide (0.5 mg/ml). The agar plates were incubated at 28°C, 37°C and 42°C for three weeks until colonies were screened for NTM.

Sensitivity of the isolation method was tested by adding defined amounts of MAH 104 [88] to sterile soil, dust and sterilized tap water und quantifying the minimal number of bacteria required to recover MAH.

2.3 Media and growth conditions

Middlebrook 7H9 broth (BD Biosciences, Heidelberg, Germany), supplemented with either 10% ADC (2 g glucose, 5 g BSA, 0.85 g NaCl in 100 ml dH_2O) or 10% OADC (BD Biosciences, Heidelberg, Germany) and 0.05% Tween 80 (Roth, Karlsruhe, Germany) was used for the growth of mycobacterial strains in liquid media. Bacterial cultivation in solid media was made by using Middlebrook 7H11 agar (BD Biosciences, Heidelberg, Germany),

supplemented with 10% ADC or OADC and 0.5% Glycerol (Roth, Karlsruhe, Germany) at 37°C.

2.4 List of bacterial strains used in this study

Table 1: MAH strains used in this study

Strain	Origin / Description	Reference / Source
MAH 104	HIV Patient	NRCM*, Borstel, Germany
MAH 2721	Child with lymphadenitis	NRCM*, Borstel, Germany
MAH 10091/06	Child with lymphadenitis	NRCM*, Borstel, Germany
MAH 10203/06	Child with lymphadenitis	NRCM*, Borstel, Germany
MAH 4557/08	Child with lymphadenitis	NRCM*, Borstel, Germany
MAH 4023/08	Child with lymphadenitis	NRCM*, Borstel, Germany
MAH 3646/08	Child with lymphadenitis	NRCM*, Borstel, Germany
MAH 3449/08	Child with lymphadenitis	NRCM*, Borstel, Germany
MAH 3269/08	Child with lymphadenitis	NRCM*, Borstel, Germany
MAH 2630/08	Child with lymphadenitis	NRCM*, Borstel, Germany
MAH 2014/08	Child with lymphadenitis	NRCM*, Borstel, Germany
MAH 772/08	Child with lymphadenitis	NRC M*, Borstel, Germany
MAH 709/08	Child with lymphadenitis	NRCM*, Borstel, Germany
MAH 528/08	Child with lymphadenitis	NRCM*, Borstel, Germany
MAH 589/08	Child with lymphadenitis	NRCM*, Borstel, Germany
MAH 7673/04	Child with lymphadenitis	NRCM*, Borstel, Germany
MAH 11082/03	Child with lymphadenitis	NRCM*, Borstel, Germany
MAH 1620/04	Child with lymphadenitis	NRCM*, Borstel, Germany
MAH 883	Child with lymphadenitis	Univ of Dusseldorf- , Germany
MAH 9060/06	Adult with lung infection, isolated once	NRCM*, Borstel, Germany
MAH 10058/06	Adult with lung infection, isolated repeatedly	NRCM*, Borstel, Germany
MAH 9268/06	Adult with lung infection, isolated once	NRCM*, Borstel, Germany
MAH 8933/06	Adult with lung infection, isolated once	NRCM*, Borstel, Germany
MAH 9036/04	Child with lymphadenitis	NRCM*, Borstel, Germany
MAH 128	Soil (environment)	FLI^{+}, Jena, Germany
MAH 2514	Water (environment)	Univ of Düsseldorf^{-} , Germany
MAH 3044	Water (environment)	Univ of Düsseldorf^{-} , Germany

** National Reference Center for Mycobacteria, Borstel, Germany.*

$^{+}$ Friedrich Löffler Institute, Jena, Germany

$^{-}$ University of Düsseldorf, Germany

2.5 Molecular Biology techniques

The molecular biology techniques explained in this section constitute of DNA isolation techniques for mycobacterial strains followed by description of the different typing methods for NTM and MAH strain identification from Germany and India respectively.

2.5.1 Isolation of mycobacterial DNA

Mycobacterial DNA was extracted by the boiled lysate method and phenol-chloroform extraction method. The boiled lysate method is a quick method to obtain crude DNA extracts. DNA extraction by phenol-chloroform was performed for obtaining DNA of higher purity. The two isolation techniques are explained in detail below.

Boiled lysate method for crude DNA extraction

The mycobacterial strains grown on Middlebrook 7H11 agar petriplates (BD Biosciences, Heidelberg, Germany), supplemented with 10% ADC or OADC and 0.5% Glycerol (Roth, Karlsruhe, Germany) at 37°C were checked for single isolated colonies. A tooth pick was used to pick up small amount of colony material. The colony materials were mixed well to 50 µl of water in PCR tubes (Greiner Bio One, Frickenhausen, Germany). The bacterial concoction was heat killed at 96°C for 30 minutes. The resultant heat killed bacteria contained crude DNA extracts and were stored at 4°C for further use.

Phenol chloroform DNA extraction

The mycobacterial cultures were allowed to grow in Middlebrook 7H9 broth (BD Biosciences, Heidelberg, Germany), supplemented with 10% ADC (2 g glucose, 5 g BSA, 0.85 g NaCl in 100 ml dH_2O). When an optical density between 1.5 to 2.5 ($OD_{600\ nm}$) was reached, 3-5 ml of the cultures was centrifuged to form pellets at 6000 g at 4°C for 10 minutes. After discarding the supernatant, the pellets were re-suspended in 400 µl of TE-8 buffer (0,01 M Tris-HCl, 0,001 M EDTA, pH 8) followed by heat killing of the mycobacteria at 80°C for 30 mins. 5 µl of lysozyme (150 mg ml^{-1}) was then added to the suspension after cooling the heat killed bacteria to room temperature. This concoction was incubated at 37°C overnight followed by the addition of 70 µl of 10% Sodium dodecyl sulfate (SDS) and 5 µl of Proteinase K (20 mg ml^{-1}) (Sigma-Aldrich, Steinheim, Germany) to the lysate. Following two hours incubation at 65°C, 100 µl of 5 M NaCl and

100 µl of Cetyl-trimethyl-ammonium-bromide (CTAB: 10% CTAB in 0.7 M NaCl: SiGMA-ALDRICH, Taufkirchen, Germany) was added to the processed samples. Incubation at 65°C for 10 minutes was then carried out. DNA was extracted by a series of purification steps using chloroform iso-amyl alcohol, phenol/chloroform/iso-amyl alcohol (Roth, Karlsruhe, Germany), and a final a round of chloroform/iso-amyl alcohol extraction. 60 µl of isopropanol (Roth, Karlsruhe, Germany) was added for every 100 µl volume of supernatant and mixed well. This was incubated overnight at -20°C followed by a final washing step with 70% ethanol. The supernatant was discarded after centrifugation and the pellets were dried for two hours at 37°C. They were re-suspended in water and stored at 4°C for further use. The DNA was quantified by using the NanoDrop ND-1000 (Thermo Scientific, Epsom, UK).

2.5.2 Identification of NTM strains

The NTM strains were confirmed by performing Polymerase chain reaction (PCR) by using the DreamTaq kit (Fermentas St. Leon-Rot, Germany). PCR was carried out according to the recommendations of the manufacturers of the kit. The primers (Table 4) were purchased from Eurofins MWG operon, Ebersberg, Germany. The primers 16S r-RNA gene (complete), 16S r-RNA gene (partial) and HSP65 gene were used for typing of the NTM strains from India. The primers that follow later on were used for MAH identification. Isolates were classified as MAH under the condition that (i) the genus-specific PCR with primers MT1/MT2 was positive, (ii) the MAC-specific PCR with primers DnaJ-1 and DnaJ-2 was positive (iii) the *M. avium* specific PCR with primers MYCAV-R and MYCGEN-F was positive, (iv) IS1245 was present (primers IS1245-1 and 1245-2), (v) IS901 was absent (primers IS901-1 and IS902-2), and (vi) the isolates grew well at 42°C in primary culture and sub-cultures. The PCRs for the DnaJ gene, IS1245 and IS901 were performed as triplex-PCR as described by Moravkova and colleagues [84]. The size of the resulting DNA fragments was analyzed by agarose gel electrophoresis. The respective PCR-products were mixed with 6x DNA loading dye (Fermentas, St. Leon-Rot, Germany) and run on 1% to 2% agarose gels depending on the size on the fragment. A corresponding DNA size marker, namely Generuler™ 100 bp Plus DNA ladder (Fermentas, St. Leon-Rot, Germany) was included. Band visibility was attained by incubation of the agarose gel for twenty minutes in GelRed ™staining

solution (Biotrend, Köln, Germany). The gel was photographed by the Gel documentation system (Bio-Rad, München, Germany).

Mycobacterial isolates from India were identified by PCR and sequencing of the complete and partial 16S r-RNA gene and the HSP65 gene [101]. The primer sequences for the complete and partial 16S r-RNA gene were kindly provided to us by Dr. Elvira Richter (NRC for Mycobacteria, Borstel, Germany). The PCR products were run in 1% agarose gels together with Generuler™ 100 bp Plus DNA ladder and the gels were eluted following a successful run. Sequencing reactions were carried out by using the Prism Big Dye Terminator 3.1 FS Terminator Cycle Sequencing Ready Reaction Kit from PE Applied Biosystems 2500 (Darmstadt, Germany) along with an Applied Biosystems 3500xl Dx Genetic Analyser. The sequencing output was analyzed by the Lasergene program (version10.01; DNASTAR, Inc.) for the sequence assembly. Identification of NTM species was performed by comparing their sequences by the Basic Local Alignment Search Tool (BLAST) (http://BLAST.ncbi.nlm.nih.gov/BLAST.cgi). Similarity searches were also performed using SEPSITEST BLAST for the 16S r-RNA gene. RIDOM database was also used for the identification of the different mycobacterial strains. The reaction mix for the PCR (DreamTaq^{TM} DNA Polymerase Kit and dNTPs, Fermentas, St. Leon-Rot, Germany) is made up of the following constituents.

Table 2: Essential components of a PCR reaction

Master mix	**µl**
10 x Buffer	5
dNTPs, 2 mM	5
$MgCl_2$, 25 mM	1,6
Primer (Forward) 10 pmol/µl	1
Primer (Reverse) 10 pmol/µl	1
TaqPol (5u/µl)	0,25
dH_2O	21,15
Mycobacterial DNA template	15

Table 3: Essential components of a sequencing reaction

Master mix	µl
Mycobacterial DNA template	6
Primer 10 pmol/µl	0.5
BigDye3.1	2
5 x Buffer	1
dH_2O	0.5
End volume	10

Table 4: Primers used for the identification of NTM species

Target gene /target sequence	P	A	Oligonucleotide sequence (5′ - 3′)
16S r-RNA gene (complete)	1030	57° C	TGGAGAGTTTGATCCTGGCTCAG
			TGCACACAGGCCACAAGGGA
16S r-RNA gene (partial)	550	57° C	CGTGCTTAACACATGCAAGTC
			TTTCACGAACAACGCGACAA
HSP65 gene	300	60° C	AAGAAGTGGGGTGCCCCC
			CTTGGTCTCGACCTCCTTG
MT1/MT2	500	60° C	TTCCTGACCAGCGAGCTGCCG
			CCCCAGTACTCCCAGCTGTGC
DnaJ-1/DnaJ-2	140	58° C	GACTTCTACAAGGAGCTGGG
			GAGACCGCCTTGAATCGTTC
MYCGEN-F/ MYCAV-R	180	62° C	AGAGTTTGATCCTGGCTCAG
			ACCAGAAGACATGCGTCTTG
IS901-F/IS901-R	577	58° C	GGATTGCTAACCACGTGGTG
			GCGAGTTGCTTGATGAGCG
IS1245-F/IS1245-R	385	58° C	GAGTTGACCGCGTTCATCG
			CGTCGAGGAAGACATACGG

A – Annealing temperature in ° C

P – Product size in bp

2.6 Genome sequencing of MAH strains from dust & child suffering from lymphadenitis.

Genome sequencing of eight MAH strains (Table 5) from varying niches both environmental and clinical was carried out using the Roche/454 FLX Pyrosequencer at the bioinformatics core facility at the Robert Koch Institute. The strains sequenced are shown in the Table 5. Two out of the eight strains namely the MAH 27-1 and the MAH 2721 were re-sequenced

using the Illumina sequencing platform (Illumina GAIIx) and the Ion torrent sequencing platform. These strains were MAH isolated from dust and a child suffering from lymphadenitis. The DNA from the strains was isolated using the standard phenol-chloroform isolation method. The samples were sent for sequencing to Genotypic Technology Limited in India after DNA isolation and quantification. Following a successful library preparation and sequencing, the genome data was obtained in the form of raw reads.

Table 5: List of MAH strains sequenced by Roche/454 FLX Pyrosequencer

MAH strains	Environmental source
MAH 128	Soil
MAH 22-1	Soil
MAH 27-1	Dust
MAH 3044	Water
MAH 9036/04	Adult
MAH 11082/03	Child
MAH 2721	Child
MAH 10058/06	Adult

The raw reads were assembled in two ways. The first approach was aimed towards attaining high quality bases and longer contigs. Hence the low quality bases were trimmed or removed. The raw reads are first checked for quality by using a software SeqQC_v2.2. The Illumina reads were assembled using the software velvet-v1.2.2. Following a successful generation of Illumina contigs, the Ion torrent reads were mapped to the Illumina contigs to generate longer contigs also called super contigs or scaffolds. Scaffolding was performed by using the software SSPACE-V2.03. Gap closing was performed by using the software GapCloser-v1.124.

The scaffolds were ordered and those above 200 bp in size were selected for NCBI submissions. Rapid Annotation using Subsystem Technology (RAST) and ARTEMIS genome browser was used to facilitate better visualization of sequence features in the new genomes. This was initiated by concatenating the scaffolds with the help of reference guided assembly followed by its submissions onto the RAST server. The resultant was uploaded in the Artemis genome browser to gain further insights into the genome.

The second approach was aimed at reconfirming the results attained from the first approach. The raw data from both the sequencers was assembled using CLC genome work bench as it not only supports sequence data from all the major NGS platforms, such as Ion Torrent, Roche 454, and Illumina Genome Analyzer but also supports *de novo* assembly of hybrid data. Edena V3, a short read *de novo* assembler was also used to reconfirm the assembly.

2.7 Softwares used for the identification of islands in MAH

Vista gateway (http://pipeline.lbl.gov/cgi-bin/gateway2) was used as a comparative genomics tool to identify regions that were specific to the MAH but were absent in other mycobacterial species and subspecies like the MAP, MAA and the *M. intracellulare*. The MAH 104 was used as a reference strain.

Island viewer (http://www.pathogenomics.sfu.ca/islandviewer/query.php), software enabling computational visualization of genomic islands, was used to analyze the islands in MAH. Screening of flexible gene pools based on the presence of flanking t-RNAs, existence of duplication regions, and GC content of the island was accomplished to identify genome islands within MAH. The GC content of the islands was checked using the Geneious R 6.1.6 (Biomatters limited) and GC profile, which is a web based tool to analyze the variation in the GC content of the genomic sequences. Analysis of island and similarity searches was performed by BLAST.

2.8 SNP pattern recognition within MAH genomes

A whole genome SNP analysis was performed with the help of Roche 454 sequencing of eight MAH genomes. Following the SNP analysis in eight genomes, candidate genes were selected based on the availability of complete gene sequence, made available by Roche 454 sequencing platform, presence of seven or more SNPs in every gene and virulence associations of the genes. In house perl scripts were used to identify the permutation combination of all genes in the eight strains that showed interesting SNPs or relatedness like SNP similarities between MAH strains from dust and MAH strains from children with lymphadenitis. The initial SNP analysis was performed in eight genomes and the results were rechecked in thirty six MAH strains. The SNP pattern recognition was carried out by using the software Geneious. A PCR purification kit (Zymoclean gel DNA recovery kit) was used

to ensure the purity of the template during the sequencing run. The purification was carried out according to the recommendations of the manufacturer. Once the sequences were retrieved, a lasergene core suit 10 (DNASTAR, Inc., Madison, WI, USA) was used to order the sequences. The sequences were imported to Geneious for further alignment and to view the SNPs present within the genes. The phylogenetic analysis of the sequences was performed using the software MEGA 5.2 [102]. MUSCLE [103, 104] and Geneious was used for aligning and concatenation of the gene sequences. The phylogenetic trees were estimated by Maximum Likelihood method (ML). ML uses a variety of substitution models but the Tamuna-Nei model was rendered the best for phylogenetic analysis.

2.9 Additional materials used in the project

Table 6: List of additional chemicals used in the project

Chemicals	Manufacturing company, location
Boric Acid	Roth, Karlsruhe, Germany
Tris	Roth, Karlsruhe, Germany
Tween 80	Sigma-Aldrich, Steinheim, Germany
Ethelenediaminetetraacetic acid (EDTA)	Merck, Darmstadt, Germany
Chloroform	Roth, Karlsruhe, Germany
DEPC-H_2O	Roth, Karlsruhe, Germany
Ethanol	Roth, Karlsruhe, Germany
Glycerin	Roth, Karlsruhe, Germany
Sodium chloride (NaCl)	Roth, Karlsruhe, Germany
ADC	2 g Glucose, 5 g BSA, 0,85 g NaCl, ad 100 ml d H2O
BBLTM-Middlebrook OADC	BD, Heidelberg, Germany
CTAB solution	4,1 g NaCl, 80 ml Aqua dest., 10 g CTAB and 100 ml d H_2O
ELGA Ultra water	Deionized by reverse osmosis and ion exchange
Middlebrook-Agar plates	6.3 g Middlebrook 7H11 Agar, 15 ml 10% Glycerin, add 270 ml dH2O, 30 ml ADC or OADC
Middlebrook liquid media	4.2 g of Middlebrook 7H9 Broth powder in 812ml of purified water containing 0,45ml Tween. 90 ml Middlebrook ADC or OADC is added for enrichment
1x TBE Electrophoresis buffer	90 mM Tris, 90 mM Boric acid, 20 mM, EDTA, pH 8,0

Table 7: List of instruments and materials during experimental proceedings

Instruments	Manufacturing company
Gel documentation system	Bio-Rad, München, Germany (Software: QuantityOne - 4.6.2)
Electrophoresis tanks	Agagel Mini - Biometra, Göttingen, Germany PeqLab 40-1214, PEQLAB Biotechnologie GmbH, Erlangen, Germany
Magnetic Rotater	MAG RCT - IKA, Staufen, Germany MR 3001 - Heidolph Instruments, Schwabach, Germany
pH-Meter pH 211	Hanna Instruments, Kehl am Rhein, Germany
Power Supply	ST 606, Gibco BRL, Eggenstein, Germany EPS 301Pharmacia Biotech, Freiburg, Germany
Thermocycler	T Gradient, Biometra, Göttingen, Germany GeneAmp PCR System 9700, Applied Biosystems, Darmstadt, Germany
Thermomixer comfort	Eppendorf, Hamburg, Germany
Vortexer VF 2	IKA, Staufen, Germany
Weighing balance	Sartorius Stedim Biotech, Göttingen, Germany
Centrifuge	Centrifuge 5415, Eppendorf, Hamburg, Germany Centrifuge Sigma 3K30, Osterode am Harz, Germany
Eppendorf-Reaction tubes (2 ml)	Eppendorf, Hamburg, Germany
Petriplates	Roth, Karlsruhe, Germany
ThinSeal™- cling films	ThinSeal™ cling films

3. Results

Since the environment is a documented source of NTM infections, the first part of the study covers ecological investigations of NTM in a TB endemic country setting such as India compared to a country with low TB incidence such as Germany. Special importance was paid towards understanding the ecology of MAH in Germany (Section 3.1).

The second section of the study incorporates genome sequencing of two MAH strains from dust and a child suffering from lymphadenitis (Section 3.2).

The third part of the study includes identification of genome islands in MAH (Section 3.3). Further to that, one genome island was chosen in specific, and the diversity of the chosen island was explored in both complete and draft MAH strains available in the database (Section 3.3).

The final part of the study integrates SNP analysis of three specific genes within the MAH genomes in order to define plausible infection routes for MAH (Section 3.4).

3.1 Questioning the ecology of MAH in Germany and India

This section elucidates the results from the study of ecological niche of MAH in Germany and NTM in India.

3.1.1 Inference from personal communications about Mycobacterium avium in Germany

Personal communication with Dr. Walter Haas at the Robert Koch Institute provided first insights into the epidemiology and incidence of NTM infections in Germany (Table 8). The study of 102 hospitalized children (less than 15 years of age) with NTM disease by Reuss and colleagues identified *M. avium* as the most frequently isolated species but on rare occasions, *M. kansasii*, *M. celatum*, *M. malmoense*, *M. intracellulare*, *M. chelonae*, *M. heidelbergense*, *M. interjectum* and *M. marinum* were also identified.

(Table Contined on Next Page)

Table 8: NTM infections in hospitalized children in Germany, April 2003 to September 2005

Species	N	(%) of positive samples
M. avium	80	78
M. kansasii	5	5
M. avium and *M. intracellulare* co-infection	4	4
M. celatum	3	3
M. malmoense	3	3
M. intracellulare	2	2
Co-infection with MTb (n=95)	2	2
M. chelonae	1	1
M. heidelbergense	1	1
M. interjectum	1	1
M. marinum	1	1
M. *avium*, M. *intracellulare* and *M. marinum* co-infection	1	1

N - Number of samples testing positive

Since water has often been made responsible for the transmission of NTM including *M. avium* [76, 105], drawing an idea about the abundance of *M. avium* in water from Germany was indispensable. Personal communication with Dr. Roland Schulze–Röbbecke assisted with the identification of most prevalent NTM species in both surface and drinking water samples from Germany (Table 9). Mycobacterial species found in both surface and drinking water samples from Germany comprised of *M. chelonae, M. gordonae, M. peregrinum. M. avium, M. intracellulare, M. lentiflavum* and *M. neglectum* in drinking water and *M. fallax, M. fortuitum, M. hiberniae* and *M. nonchromogenicum* in surface water rather than drinking water. Table 9 clearly shows that *M. avium* represented only 3% of the mycobacteria identified. Since *M. avium* was not abundant in water sources and yet was the most predominantly isolated species from paediatric subjects with NTM disease in Germany, it was henceforth hypothesized that this bacterium could have alternate reservoirs in Germany. Soil, water, biofilms and dust were thereafter collected from different locations in Germany to examine the MAH reservoirs.

Table 9: Mycobacterial species cultivated from surface water and drinking water in Germany

Type (number) of water samples tested	Mycobacteria identified	N	(% of samples tested positive)
Surface water (56)	*M. gordonae*	20	36
	M. peregrinum	8	14
	M. chelonae	6	11
	M. hiberniae	6	11
	M. mucogenicum	4	7
	M. nonchromogenicum	3	5
	M. fallax	2	4
	M. fortuitum	2	4
	M. porcinum	1	2
	M. smegmatis	1	2
	M. kansasii	1	2
	M. malmoense	1	2
Drinking water (79)	*M. gordonae*	35	44
	M. chelonae	31	39
	M. kansasii	8	10
	M. neglectum	8	10
	M. peregrinum	6	8
	M. lentiflavum	5	6
	M. intracellulare	3	4
	M. avium	**2**	**3**
	M. xenopi	1	1
	M. mucogenicum	1	1

N - Number of samples tested positive

3.1.2 Investigation of MAH in soil, dust, biofilms and water from Germany

127 environmental samples (42 samples of soil, 30 samples of dust, 36 samples of water and 19 samples of biofilm) were collected from different locations in Germany. The soil, dust, water and biofilm were collected in random as the primary aim of the study was to explore the alternate reservoirs of MAH in Germany. In order to identify MAH in soil, dust, water and biofilms, the colonies obtained following the incubation of isolation plates at 28°C, 37°C and 42°C for three weeks were exposed to genus-specific, *M. avium*-specific and MAH-specific PCR.

The genus specific PCR with primers MT1 and MT2 yielded a product size of 500 bp. The isolates identified positive with the genus-specific PCR were exposed to *M. avium*-specific PCR with MYCAV-R and MYCGEN-F primers. This reaction yielded a product size of 180 bp. The MAH-specific PCR was a multiplex PCR that was conducted using a combination of three primer pairs for DnaJ, IS1245 and IS901. Product sizes of 140 bp and 385 bp for the DnaJ gene and the IS1245 insertion element and absence of the product for the IS901 insertion element (577 bp) confirmed the presence of MAH strain (Figure 5).

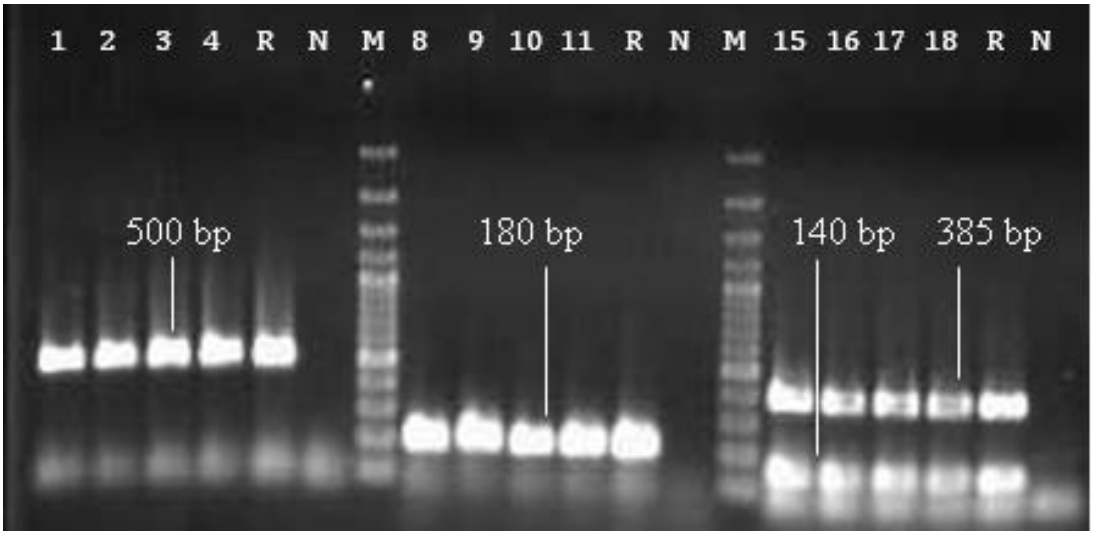


Figure 5: Representative PCR results for detection of MAH in the environment

PCR was performed with the DNA from different environmental isolates and the PCR products were analyzed by electrophoresis in a 2 % agarose gel. Lane 1-4: PCR amplification with MT1 and MT2 primers, lane 8-11: PCR amplification with MYCAV-R and MYCGEN-F primers and lanes 15-18: PCR amplification with primer for DnaJ, IS1245 and IS901.M: 100 bp marker (sizes of the bands are indicated in the appendix in the section 8.2), R: Reference strain MAH 104, N: Negative control (water).

127 environmental samples from soil, dust, water and biofilms from different places in Germany were screened for the identification of MAH. Water (42 samples) was primarily collected from lakes, rivers, taps and hand pumps. Biofilms (19 samples) were collected from kitchen sinks, bath tubs and traps. Dust samples (30 samples) comprised of contents of vacuum cleaners in households. Soil samples (42 samples) were collected either from close human environments (30 samples) (e.g. potting soil, garden soil, sand from playing

grounds) or (12 samples) from areas involving lesser contact to humans (e.g. forests). 18 samples (14%) from this study were confirmed to contain MAH. A higher incidence of MAH was found in the dust (33%) in comparison to soil (19%) (Table 10). Studies of water samples and biofilms did not yield any MAH strains. 4 water samples were identified with MAC but were tested negative for MAH. Since no MAH was identified in water, the results coincide with those of Dr. Roland Schulze–Röbbecke where the *M. avium* positive samples were as low as 3%. These results indicate that soil and dust rather than water could be possible reservoirs for MAH infections in Germany. No MAH was recovered from soils in forests while potting soils, garden soils and soils from playgrounds allowed MAH recovery.

Interestingly, all the positive samples came from environments close to human contact. MAH were isolated from 37.5 % of potting soil samples, 12.5 % of garden soil samples and 16.6 % of soil samples from play grounds. 38.5% of dust from vacuum cleaner bags accounted for MAH recovery. Further details of MAH recovery in Germany have been provided in the Tables 10 and 11. The high recovery of MAH from dust and soil, points towards dust and soil being possible sources of MAH infections in Germany.

The sensitivity testing of our isolation method revealed similar levels of sensitivity in our water and soil samples allowing recovery of MAH if at least 1,000 MAH per ml or per g were present in the samples. The method was one log less sensitive with dust samples. The detection of MAH samples from dust required around 10,000 MAH per g of dust to ensure recovery of mycobacteria.

(Table Contined on Next Page)

Table 10: Environmental samples and isolation of MAH in Germany

Samples Types	Origin of samples	Number of samples	Number of samples with MAH	% of samples with MAH
All samples		**127**	**18**	**14.1**
Dust		**30**	**10**	**33.3**
	Vacuum cleaner bag	26	10	38.5
	Surface floor	1	0	
	Surface furniture	1	0	
	Filter hair dryer	1	0	
	Filter respirator	1	0	
Soil		**42**	**8**	**19.0**
	flower pots	16	6	37.5
	Garden	8	1	12.5
	play ground	6	1	17
	urban space	4	0	
	Forest	3	0	
	Countryside	1	0	
	bird cage	1	0	
Water		**36**	**0**	**0**
	Lake	15	0	
	tap	8	0	
	Fountain	7	0	
	River	3	0	
	rain/puddle	3	0	
Biofilm		**19**	**0**	**0**
	sanitation facility	8	0	
	well walling	5	0	
	filter unit	3	0	
	river/lake	2	0	
	Aquarium	1	0	

Table 11: MAH from soil and dust identified in this study

MAH strains identified	Environmental source	City
MAH 6-1	Soil from flower pot	Berlin
MAH 14-1	Soil from playground	Berlin
MAH 22-1	Garden soil	Berlin
MAH 27-1	Dust from vacuum cleaner	Berlin
MAH 57-3	Dust from vacuum cleaner	Berlin
MAH 61-1	Dust from vacuum cleaner	Berlin
MAH 63-1	Dust from vacuum cleaner	Berlin
MAH 82-7	Dust from vacuum cleaner	Wesel
MAH 83-1	Dust from vacuum cleaner	Hilden
MAH 88-1	Dust from vacuum cleaner	Garstadt
MAH 89-1	Dust from vacuum cleaner	Würzburg
MAH 96-2	Soil from flower pot	Berlin
MAH 101-6	Dust from vacuum cleaner	Berlin
MAH 106	Soil from flower pot	Berlin
MAH 108	Dust from vacuum cleaner	Berlin
MAH 118	Soil from flower pot	Berlin
MAH 104-1	Soil from flower pot	Berlin
MAH 149-2	Soil from flower pot	Berlin

3.1.3 Investigation of NTM and MAH in soil and dust from India

Soil samples (89 samples) and dust samples (17 samples) from India were analysed for the occurrence of MAH. The analysis involved performing genus-specific PCR with MT1 and MT2 primers to identify the mycobacteria in the Indian environment. Following positive mycobacterial identification, 16S r-RNA specific and HSP65 gene specific primers as mentioned in Table 4 were used and the PCR results were checked on 1% agarose gels as shown in Figure 6 and 7.

The bands were eluted and sequencing was performed thereafter. The sequences were checked for quality followed by BLAST analysis to identify the different mycobacterial strains.

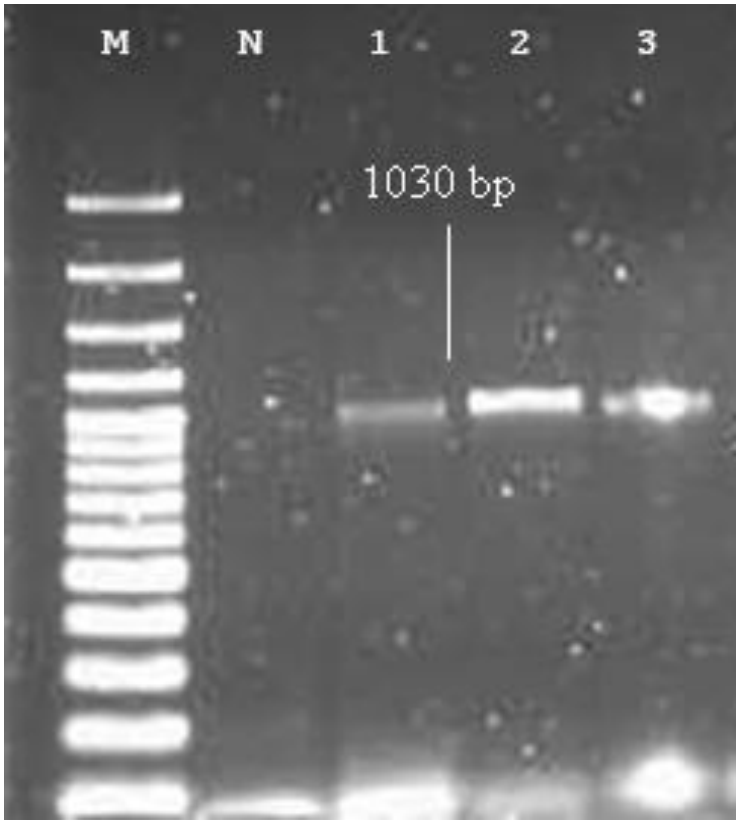


Figure 6: Representative gel for 16S r-RNA (complete) PCR for the Indian isolate

DNA was analyzed by electrophoresis in a 1 % agarose gel as shown in the Figure 6.Lanes 1-3: PCR amplification of positive mycobacterial samples with 16S r-RNA (complete) primers M: 100 bp marker (sizes of the bands are indicated in the appendices), Lane 2: N: Negative control (water)

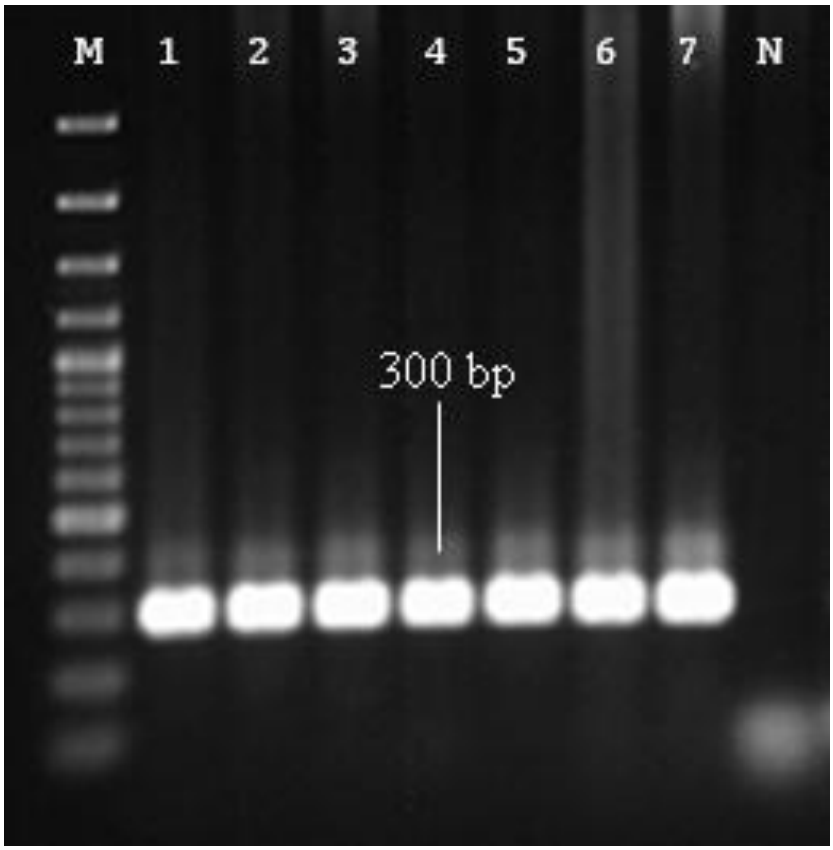


Figure 7: Representative gel for HSP65 gene specific PCR for the Indian isolates

DNA was analyzed by electrophoresis in a 1 % agarose gel and the representative gel result is provided in Figure 7. Lane 1-7: PCR amplification of positive mycobacterial isolates with HSP65 primers, M: 100 bp marker (sizes of the bands are indicated in the appendices), N: Negative control (water).

A demonstrative idea of the analysis is provided below. Following the sequencing of complete 16S r-RNA and the HSP65 gene, the partial 16S r-RNA sequence was extracted from the complete 16S r-RNA sequence. The complete 16S r-RNA sequence has a size of 1030 bp and the partial 16S r-RNA has a size of 550 bp and is an integral part of the complete sequence. The partial 16S r-RNA sequencing was performed on multiple occasions when the sequencing with complete 16S r-RNA was improper or lesser homology (> 90%) was observed during the BLAST analysis with complete 16S r-RNA gene in more than one species of mycobacteria. The partial 16S r-RNA sequence was used to ensure the correct typing of the mycobacterial strains and it was found to be more specific towards correct identification of mycobacterial strains.

The partial 16S r-RNA is highlighted in red colour in the illustrative example. All the three gene sequences (the complete 16S r-RNA, partial 16S r-RNA and the HSP65) were checked by NCBI BLAST for 99-100% homology and the mycobacterial species was confirmed thereafter if the results from all the three sequences were identical. Only those mycobacterial isolates that provided identical results with all the three typing techniques (99-100% homology) were considered as correctly typed and included in the strain collection list. Additional 16S r-RNA analysis was also performed by using the Sepsitest BLAST and RIDOM.

Complete sequence of 16S r-RNA gene for one of the Indian isolates:

> Complete sequence of 16S r-RNA gene

TGCTTACA**CATGCAAGTCGAACGGAAAGGCCCTTCGGGGTACTCGAGTGGCGAACGGGTGAGTAACACGTGG**
GTGATCTGCCCTGCACTTTGGGATAAGCCTGGGAAACTGGGTCTAATACCGAATATGACCACGCGCTTCATG
GTGTGTGGTGGAAAGCTTTTGCGGTGTGGGATGGGCCCGCGGCCTATCAGCTTGTTGGTGGGGTAATGGCCT
ACCAAGGCGACGACGGGTAGCCGGCCTGAGAGGGTGACCGGCCACACTGGGACTGAGATACGGCCCAGACTC
CTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCGACGCCGCGTGAGGGATG
ACGGCCTTCGGGTTGTAAACCTCTTTCAATAGGGACGAAGCGCAAGTGACGGTACCTATAGAAGAAGGACCG
GCCAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGTCCGAGCGTTGTCCGGAATTACTGGGCGTAAAGAG
CTCGTAGGTGGTTTGTCGCGTTGTTCGTGAAAACTCACAGCTTAACTGTGGGCGTGCGGGCGATACGGGCAG
ACTAGAGTACTGCAGGGGAGACTGGAATTCCTGGTGTAGCGGTGGAATGCGCAGATATCAGGAGGAACACCG
GTGGCGAAGGCGGGTCTCTGGGCAGTAACTGACGCTGAGGAGCGAAAGCGTGGGGAGCGAACAGGATTAGAT
ACCCTGGTAGTCCACGCCGTAAACGGTGGGTACTAGGTGTGGGTTTCCTTCCTTGGGATCCGTGCCGTAGCT
AACGCATTAAGTACCCCGCCTGGGGAGTACGGCCGCAAGGCTAAAACTCAAAGGAATTGACGGGGGCCCGCA
CAAGCGGCGGAGCATGTGGATTAATTCGATGCAACGCGAAGAACCTTACCTGGGTTGACATGCACAGGACGA
CTGCAGAGATG

Partial sequence of 16S r-RNA gene for one of the Indian isolates:

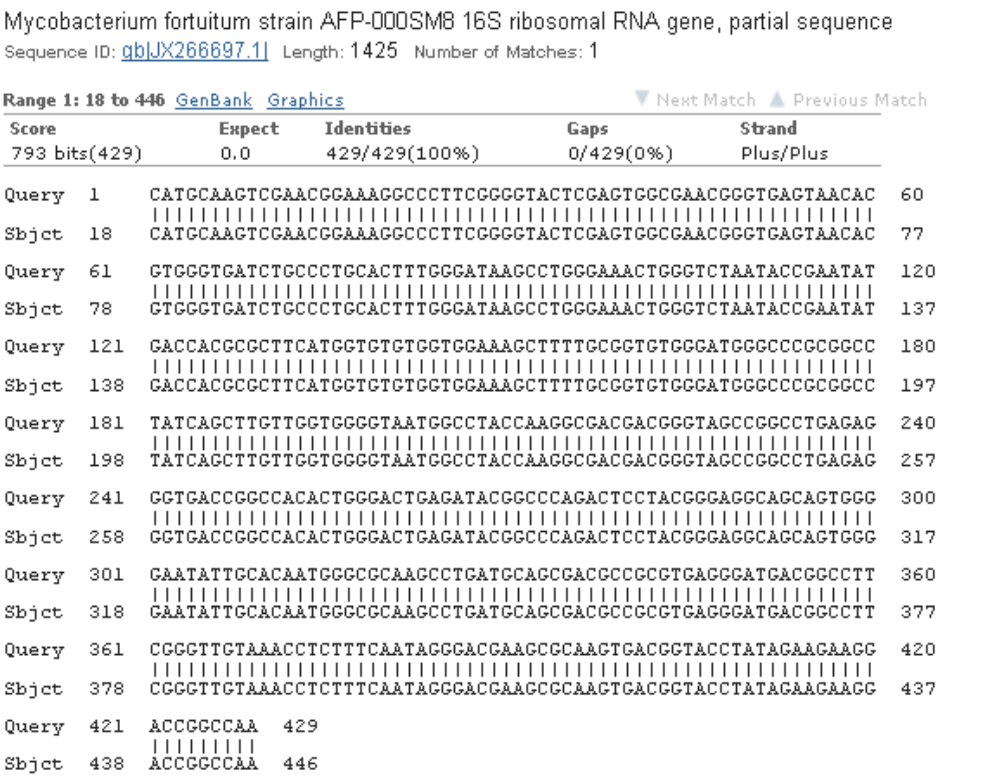
> Partial sequence of 16S r-RNA gene

CATGCAAGTCGAACGGAAAGGCCCTTCGGGGTACTCGAGTGGCGAACGGGTGAGTAACACGTGGGTGATCTG
CCCTGCACTTTGGGATAAGCCTGGGAAACTGGGTCTAATACCGAATATGACCACGCGCTTCATGGTGTGTGG
TGGAAAGCTTTTGCGGTGTGGGATGGGCCCGCGGCCTATCAGCTTGTTGGTGGGGTAATGGCCTACCAAGGC
GACGACGGGTAGCCGGCCTGAGAGGGTGACCGGCCACACTGGGACTGAGATACGGCCCAGACTCCTACGGGA
GGCAGCAGTGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCGACGCCGCGTGAGGGATGACGGCCTT
CGGGTTGTAAACCTCTTTCAATAGGGACGAAGCGCAAGTGACGGTACCTATAGAAGAAGGACCGGCCAA

Sequence of HSP gene for one of the Indian isolates:

> Sequence of HSP gene

TTAAGAAGTGGGGTGCCCCCACGATCACCAACGATGGTGTGTCCATCGCCAAGGAGATCGAGCTGGAGGACC
CGTACGAGAAGATCGGCGCTGAGCTCGTCAAAGAGGTCGCCAAGAAGACCGACGACGTCGCGGGCGACGGCA
CCACCACCGCCACCGTTCTGGCACAGGCCCTGGTTCGTGAAGGTCTGCGCAACGTCGCTGCCGGCGCCAACC
CGCTCGGCCTGAAGCGCGGCATCGAGAAGGCCGTCGAGAAGGTCACCGAGACGCTGCTGAAGAGCGCCAAGG
AGGTCGAGACCAAGAGA



Mycobacterium fortuitum strain AFP-000SM8 16S ribosomal RNA gene, partial sequence

Sequence ID: gb|JX266697.1| Length: 1425 Number of Matches: 1

Range 1: 18 to 446 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
793 bits(429)	0.0	429/429(100%)	0/429(0%)	Plus/Plus

```
Query  1    CATGCAAGTCGAACGGAAAGGCCCTTCGGGGTACTCGAGTGGCGAACGGGTGAGTAACAC  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  18   CATGCAAGTCGAACGGAAAGGCCCTTCGGGGTACTCGAGTGGCGAACGGGTGAGTAACAC  77

Query  61   GTGGGTGATCTGCCCTGCACTTTGGGATAAGCCTGGGAAACTGGGTCTAATACCGAATAT  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  78   GTGGGTGATCTGCCCTGCACTTTGGGATAAGCCTGGGAAACTGGGTCTAATACCGAATAT  137

Query  121  GACCACGCGCTTCATGGTGTGTGGTGGAAAGCTTTTGCGGTGTGGGATGGGCCCGCGGCC  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  138  GACCACGCGCTTCATGGTGTGTGGTGGAAAGCTTTTGCGGTGTGGGATGGGCCCGCGGCC  197

Query  181  TATCAGCTTGTTGGTGGGGTAATGGCCTACCAAGGCGACGACGGGTAGCCGGCCTGAGAG  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  198  TATCAGCTTGTTGGTGGGGTAATGGCCTACCAAGGCGACGACGGGTAGCCGGCCTGAGAG  257

Query  241  GGTGACCGGCCACACTGGGACTGAGATACGGCCCAGACTCCTACGGGAGGCAGCAGTGGG  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  258  GGTGACCGGCCACACTGGGACTGAGATACGGCCCAGACTCCTACGGGAGGCAGCAGTGGG  317

Query  301  GAATATTGCACAATGGGCGCAAGCCTGATGCAGCGACGCCGCGTGAGGGATGACGGCCTT  360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  318  GAATATTGCACAATGGGCGCAAGCCTGATGCAGCGACGCCGCGTGAGGGATGACGGCCTT  377

Query  361  CGGGTTGTAAACCTCTTTCAATAGGGACGAAGCGCAAGTGACGGTACCTATAGAAGAAGG  420
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  378  CGGGTTGTAAACCTCTTTCAATAGGGACGAAGCGCAAGTGACGGTACCTATAGAAGAAGG  437

Query  421  ACCGGCCAA  429
            |||||||||
Sbjct  438  ACCGGCCAA  446
```

Figure 8: Screen shot of NCBI BLAST analysis of complete 16S r-RNA sequencing.

Figure 8 shows a BLAST analysis with the sequences from a partial 16S r-RNA gene of an Indian NTM isolate. A 100% homology was observed in this case. The NTM isolate was typed as Mycobacterium fortuitum following a similar BLAST analysis with the sequences from complete 16S r-RNA gene and HSP65 gene.

Table 12 provides an overview of the different kinds of NTM identified from environmental sampling in India. No MAH was identified from our soil or dust studies in India. Hence the commonly occurring NTM in India were explored.

Table 12: NTM identified from soil in India from this study

Strains	Species/Subspecies	State/County
Ind011	*Mycobacterium IWGMT* 90203	Andhra Pradesh
Ind018-01	*Mycobacterium celatum*	Andhra Pradesh
Ind018-03	*Mycobacterium palustre*	Andhra Pradesh
Ind022-06	*Mycobacterium palustre*	Andhra Pradesh
Ind032-01	*Mycobacterium terrae* complex	Andhra Pradesh
Ind032-02	*Mycobacterium fortuitum*	Andhra Pradesh
Ind032-03	*Mycobacterium intracellulare*	Andhra Pradesh
Ind034	*Mycobacterium intracellulare*	Andhra Pradesh
Ind041-02	*Mycobacterium fortuitum*	Andhra Pradesh
Ind044-01	*Mycobacterium intermedium*	Andhra Pradesh
Ind045-02	*Mycobacterium terrae* complex	Andhra Pradesh
Ind047-01	*Mycobacterium intracellulare*	Andhra Pradesh
Ind053	*Mycobacterium parascrofulaceum*	Odisha
Ind054-03	*Mycobacterium terrae* complex	West Bengal
Ind058-01	*Mycobacterium scrofulaceum*	Odisha
Ind059	*Mycobacterium gordonae*	Odisha
Ind061	*Mycobacterium terrae* complex	New Delhi
Ind064	*Mycobacterium triplex*	Odisha
Ind066	*Mycobacterium simiae*	New Delhi
Ind072-01	*Mycobacterium fortuitum*	Odisha
Ind073-1	*Mycobacterium fortuitum*	Odisha
Ind073-04	*Mycobacterium terrae* complex	Odisha
Ind074-01	*Mycobacterium parascrofulaceum*	Odisha
Ind074-02	*Mycobacterium fortuitum*	Odisha
Ind077-03	*Mycobacterium asiaticum*	West Bengal
Ind077-07	*Mycobacterium intracellulare*	West Bengal
Ind081	*Mycobacterium asiaticum*	New Delhi
Ind082	*Mycobacterium asiaticum*	New Delhi
Ind083-01	*Mycobacterium asiaticum*	New Delhi
Ind083-03	*Mycobacterium intracellulare*	New Delhi
Ind084-02	*Mycobacterium fortuitum*	New Delhi
Ind090-01	*Mycobacterium intermedium*	New Delhi
Ind094-04	*Mycobacterium intermedium*	Odisha
Ind094-05	*Mycobacterium terrae* complex	Odisha
Ind095-01	*Mycobacterium fortuitum*	New Delhi
Ind095-02	*Mycobacterium parascrofulaceum*	New Delhi
Ind095-03	*Mycobacterium asiaticum*	New Delhi
Ind098-03	*Mycobacterium asiaticum*	New Delhi

Strains	Species/Subspecies	State/County
Ind101-02	*Mycobacterium intermedium*	Odisha
Ind101-04	*Mycobacterium asiaticum*	Odisha
Ind103-01	*Mycobacterium terrae* complex	Odisha
Ind103-02	*Mycobacterium scrofulaceum*	Odisha

42 samples (40%) out of 106 collected samples from India were identified as positive for NTM but 11 samples were either not revived or could not be identified to the species level. The source of all the bacteria identified from India was soil. 31 samples (29%) of the isolations were identified to the species level with a total of 42 different NTM strains isolated from soil. Table 12 and Table 13 clearly show that *M. terrae, M. fortuitum* and *M. asiaticum* were found as high as 17% in all the samples that were tested positive for mycobacteria in Indian soil and dust. *M. asiaticum* was more predominant in New Delhi. *M. intracellulare, M. scrofulaceum, M. parascrofulaceum* and *M. intermedium* were also amongst the commonly isolated species. NTM like *M. triplex, M. celatum, M. simiae, M. palustre and M. gordonae* were rare in the soil samples.

Table 13: Mycobacterial species isolated from Indian soil and dust

Species	Number of samples testing positive N=42	% of positive samples
M. fortuitum	7	17
M. terrae	7	17
M. asiaticum	7	17
M. intracellulare	5	12
M. scrofulaceum and M. parascrofulaceum	5	12
M. intermedium	4	10
M. palustre	2	5
M. gordonae	1	2
M. triplex	1	2
M. simiae	1	2
M. celatum	1	2
M. IWGMT 90203	1	2

3.2 Sequencing of MAH genomes from Germany

Subsequent to the ecology studies of MAH in Germany and NTM in India, the second principal objectives of the study involved introduction of European MAH reference genomes into the database. Hence strains isolated from our environmental MAH studies in Germany (Table 11) along with strains obtained from NRCM and FLI, Jena (Table 1) were selected for sequencing. Eight strains covering environmental (soil, dust and water) and clinical niches (adult patients with lung infections and children suffering from lymphadenitis) were selected for sequencing. The selection comprised of two MAH isolates from children suffering from lymphadenitis (MAH 11082/03 and MAH 2721), two strains from adult patients with lung infection (MAH 10058/06 and MAH 9036/04), two strains of soil origin (MAH 22-1 and MAH 128), one strain from water (MAH 3044) and one strain from dust (MAH 27-1).

3.2.1 Roche 454 sequencing platform for MAH genomes in Germany

The list of MAH strains sequenced by Roche platform had low and variable coverage. Particulars of the sequencing run are shown in the Table 14.

The raw genome data from Roche sequencer was obtained in the form of Sff files. Sff files contain the sequence data from a sequencing run. These files were obtained in the binary format and were converted into a text format such as the fasta format. Neubler assembler was used for assembling the contigs from raw data. This task was performed with assistance from the Bioinformatics core facility at the Robert Koch Institute. The sequencing with Roche 454 platform provided first insights into the size of the MAH genomes. Reference guided assemblies with the MAH 104 genome revealed heterogeneity and differences in the genomes. Several different SNPs and gaps were identified. However the low coverage of the genomes was an impending factor in the analysis.

Mycobacterial genomes are relatively big and have several repeats and duplications. Hybrids assemblies are essential for completion of mycobacterial genomes. Since the coverage with the Roche genome data was significantly lower, these low coverage genomes were unfit for draft genome submissions to the NCBI database. Hence two strains, namely the MAH 27-1 and MAH 2721 were selected, for re-sequencing by

Illumina and Ion torrent sequencing platform, with the intention that these genomes would be submitted to the NCBI database.

Table 14: Summary of the genome data obtained by Roche 454 sequencing platform

MAH strains	Environmental source	Number of runs	Reads	Nucleotides	Coverage
MAH 128	Soil	2	57,236 165,41	29,909,224 86,539,515	21.17
MAH 22-1	Soil	2	93,142 246,712	49,179,235 130,209,925	32.76
MAH 27-1	Dust	1	112,153	56,964,112	10.40
MAH 3044	Water	2	21,186 81,789	10,872,838 42,809,511	9.80
MAH 9036/04	Adult	1	31,512	15,905,201	2.90
MAH 11082/03	Child	2	106,541 106,941	55,436,951 56,441,815	20.43
MAH 2721	Child	3	37,183 41,769 36,748	20,526,844 23,319,351 20,645,597	11.78
MAH 10058/06	Adult	3	46,463	25,010,127	14.37

3.2.2 Sequencing MAH genomes from strains in dust and a child with lymphadenitis

The MAH 27-1, identified in dust from this study and MAH 2721, a strain isolated from a child suffering from lymphadenitis, obtained from National Reference Centre for Mycobacteria, were selected for further genome sequencing with Illumina and Ion torrent sequencing platforms. Following a successful library preparation of extracted genomic DNA, the samples were rendered ready for the sequencer.

The Illumina library preparation was paired end whereas the Ion torrent sequencing had single end library preparation. The bio-analyzer profile for Illumina and Ion torrent sequencing are shown in the Figures 9-12.

Peak	Size (bp)	Concentration (pg/µl)	Molarity (pmol/l)
1	35	125.00	5,411.3
2	305	502.46	2,499.1
3	312	125.88	612.1
4	328	545.30	2518.0
5	10,38	75.00	10.9

Region (From bp)	Region (to bp)	Average Size (bp)	Concentration (pg/µl)	Molarity (pmol/l)
200	1000	336	1,318.31	6,241.2

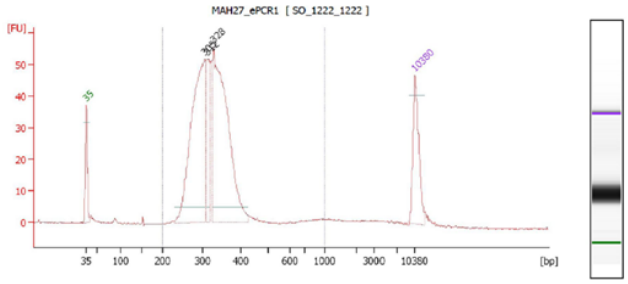


Figure 9: Illumina bio-analyzer profile of amplified DNA of MAH 27-1

The bio-analyzer profile for MAH 27-1 is shown above. 3 peaks were identified and noise of 0.2 was observed. The library was considered suitable for further sequencing.

Peak	Size(bp)	Concentration (pg/µl)	Molarity(pmol/l)
1	35	125.00	5,411.3
2	308	482.84	2,373.4
3	328	119.67	552.6
4	340	608.09	2710.7
5	10,38	75.00	10.9

Region (From bp)	Region (to bp)	Average Size (bp)	Concentration (pg/µl)	Molarity (pmol/l)
200	1000	346	1,449.71	6,604.0

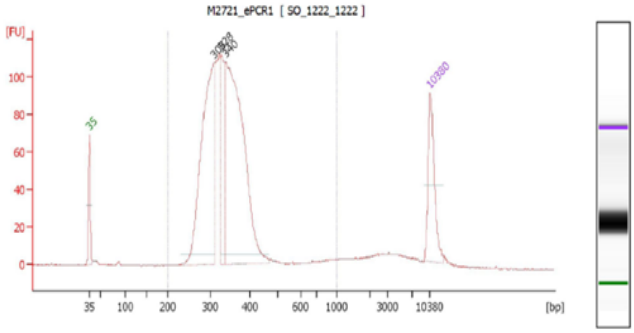


Figure 10: Illumina bio-analyzer profile of amplified DNA product in MAH 2721

The bio-analyzer profile for MAH 2721 is shown above. The library was considered suitable for further sequencing.

Peak	Size (bp)	Concentration (pg/µl)	Molarity (pmol/l)
1	35	125.00	5.411.3
2	308	537.31	2,645.9
3	10,38	75.00	10. 9

Region (From bp)	Region (to bp)	Average Size (bp)	Concentration (pg/µl)	Molarity (pmol/l)
200	1000	307	579.51	2,871.9

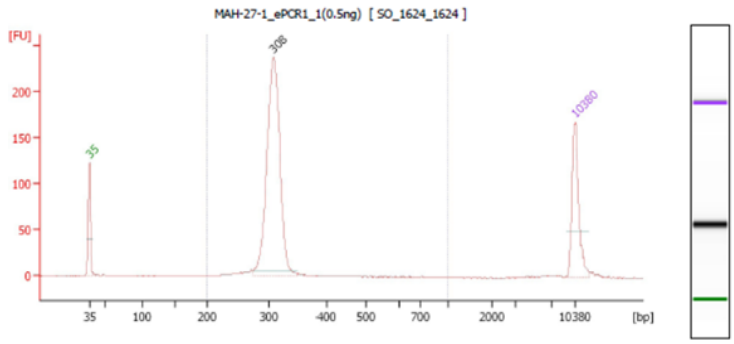


Figure 11: Bio-analyzer profile and the peak table of Ion torrent sequencing of MAH 27-1

The suitable bio-analyzer profile representing a suitable library for MAH 27-1 is shown above.

Peak	Size(bp)	Concentration (pg/µl)	Molarity(pmol/l)
1	35	125.00	5.411.3
2	304	552.47	2,753.9
3	10.38	75.00	10.9

Region (From bp)	Region (to bp)	Average Size (bp)	Concentration (pg/µl)	Molarity (pmol/l)
200	1000	305	567.59	2829.5

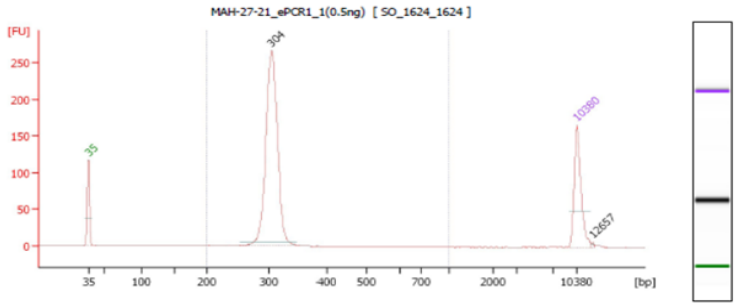


Figure 12: Bio-analyzer profile and the peak table for ion torrent sequencing of MAH 2721

The bio-analyzer profile for MAH 2721 is shown above.

3.2.3 Raw reads and raw data processing

The sequencers provide genome data in the form of raw reads. The raw reads are assembled together to form longer reads also called contigs. The contigs are assembled together to form super contigs or scaffolds. The scaffolds are synonymous to longer contigs. The Illumina paired end raw reads and the Ion torrent single end reads were quality checked using SeqQC v2.2. Quality check promotes removal of low quality bases. Processing of Illumina raw reads was done by screening for adapters. After the adapters were removed, the sequences were ready for assembly. Genome data generated by Illumina platform have been tabulated in the Tables 15 and 16.

Table 15: The quality control (QC) statistical data of reads from MAH 27-1 and MAH 2721

	Illumina MAH 2721	Illumina MAH 27-1	Ion torrent MAH 2721	Ion torrent MAH 27-1
Total number of raw read data	5,209,665	4,709,169	2,303,226	2,141,108
Total processed data	4,469,735	3,839,544	2,056,790	1,916,516

Table 16: The processed data analysis from MAH 27-1 and MAH 2721

	Illumina MAH 2721	Illumina MAH 27-1	Ion torrent MAH 2721	Ion torrent MAH 27-1
Fastq file size	1.09 GB	931.42 MB	710.12 MB	663.08 MB
Maximum Read Length	100 bp	100 bp	299 bp	289 bp
Minimum Read Length	50 bp	50 bp	50 bp	50 bp
Mean Read Length	84 bp	80 bp	170 bp	170 bp
Total Number (million) of Reads	**4469735 (4.47)**	**3839544 (3.84)**	**2056790 (2.06)**	**1916516 (1.92)**
Total Number of HQ Reads 1*	4469735 (4.47)	3839544 (3.84)	2056789 (2.06)	1916513 (1.92)
Percentage of HQ Reads	100%	100%	100%	100%

1 >70% of bases in a read with >17 phred score*

2 bases with >17 phred score*

The Illumina sequencing resulted in 5,209,665 bp for MAH 2721 and 4,709,169 bp for MAH 27-1. The Ion torrent sequencing resulted in 2,303,226 bp for MAH 2721 and 2,141,108 bp for MAH 27-1.The results from Illumina sequencing provided more genome information in comparison to the Ion torrent sequencing. Hence the data from Illumina sequencing was considered for primary assembly followed by Ion torrent data for additional mapping and contig extension. The low quality bases were removed before assembly to ensure a good assembly and high coverage. More statistical information is provided in the Table 16.

3.2.4 Denovo assembly of sequenced genome data

The generation of more genome data with Illumina platform resulted in a denovo assembly of Illumina data. Assembly was performed by using the software velvet-v1.2.20. Various parameters like total number of contigs generated, maximum contig length and total contig length were taken into consideration while making the assembly. Scaffolding was performed by using the software SSPACE-V2.03. Gap closing was performed by using the software GapCloser-v1.124. Due to the nature of paired end library data from the Illumina sequencing, significant decrease in the number of contigs could be made. 3226 contigs in MAH 27-1 strain was reduced to 2789 scaffolds whereas 3035 contigs were reduced to 2601 scaffolds in MAH 2721 strain in the initial round of scaffolding. The Ion torrent data from the genomes was used thereafter to ensure a better assembly and longer contigs. More information is provided in the section 3.2.5.

Table 17 provides an overview of the assembly statistics obtained from the assembly of Illumina data. Assembling the contigs to scaffolds should result in considerable reduction in the number to scaffolds to ensure a good assembly. The results shown in Table 17 display that contigs for both MAH 2721 and MAH 27-1 were reduced to lesser number of scaffolds after the assembly.

Maximum, minimum and average contig lengths in Table 17 provide an idea about the amount of genome data that was being assembled to form longer contigs or scaffolds. The number of Non-ATGC characters in a genome data delivers an impression about the number of Ns (unidentified bases) in the genome data. N50 value is the length of the smallest contig that represents a collection of few largest contigs within the genome

whose combined length forms 50% of the assembly. Higher N50 values denote good assembly and good contigs. The N50 value of the scaffolds was higher than that of the contigs (N50 value of 1801 to 2314 in MAH 2721 and N50 value of 1584 to 2063 in MAH 27-1) in the assembly performed.

Table 17: Assembly statistics of contigs and scaffolds based on length

Description	**MAH 2721**		**MAH 27-1**	
	Contigs	Scaffolds	Contigs	Scaffolds
Contigs Generated	3035	2601	3226	2789
Maximum Contig Length [bp]	35974	51059	21996	37655
Minimum Contig Length [bp]	125	125	117	125
Average Contig Length [bp]	1312,6	1619,1	1163,9	1433,9
Total Contigs Length [bp]	3983877	4211274	3754708	3999157
Number of Non-ATGC Characters	8940	112257	11803	122416
% of Non-ATGC Characters	0,224	2,666	0,314	3,061
Contigs >= 100 bp	3035	2601	3226	2789
Contigs >= 200 bp	3013	2595	3193	2783
Contigs >= 500 bp	2423	2203	2434	2240
Contigs >= 1 Kbp	1379	1370	1275	1318
Contigs >= 10 Kbp	10	15	6	11
N50 value	1801	2314	1584	2063

3.2.5 Processing of Ion torrent data

Further scaffolding and scaffold extension of the denovo assembled Illumina data was performed for both the strains with the help of Ion-Torrent data. The processed reads from Ion torrent data were mapped to Illumina scaffolds thereby initiating scaffold extension. Since the Ion torrent reads were much larger in comparison to the Illumina reads, scaffold extension was easier. The read length distribution for the Ion torrent processed data is shown in the Table 18 below.

Table 18: Read length distribution of Ion torrent processed data

Percentage of reads	MAH 2721	MAH 27-1
Percentage of reads between 51-100 bp	15.6	15.46
Percentage of reads between 101-150 bp	18.69	19.19
Percentage of reads between 151-200 bp	19.75	20.94
Percentage of reads between 201-250 bp	45.93	44.37
Percentage of reads above 250 bp	0.01	0.01

Ion torrent data incorporation reduced the MAH 2721 scaffolds from 2601 to 2247 scaffolds. 2789 scaffolds in MAH 27-1 were reduced to 1201 scaffolds. Further information about the hybrid scaffolds both from Illumina and Ion torrent data are provided in the Table 19. Processing with Ion torrent data reduced the number of Non-ATGC characters significantly.

The number of Ns (unidentified bases) was reduced to 3 in MAH 2721 whereas no unidentified bases were found in the scaffolds from MAH 27-1. The N50 values were higher after incorporation of the Ion torrent genome data. A good assembly was observed in the case of MAH 2721 (N50 value of 4578) and MAH 27-1 (N50 value of 8521).

The final results from the genome assembly revealed that the draft genome size for MAH 2721 was approximately 5.2 Mb whereas the draft genome of MAH 27-1 was 4.8 Mb. The assembly statistics have been provided in the Table 19. A 257X coverage was observed for MAH 2721 and a 235X coverage was observed for MAH 27-1.

(Table Contined on Next Page)

Table 19: Statistics of draft scaffolds generated from Illumina and Ion torrent sequencing data

Percentage of reads	MAH 2721	MAH 27-1
Number of Sequences Generated	2247	1201
Maximum Scaffold Length	52816	46324
Minimum Scaffold Length	104	112
Average Scaffold Length	2323.5	4016.9
Total Sequence Length	5220874 (5.2Mb)	4824352(4.8Mb)
Total Number of Non-ATGC Characters	3	0
Percentage of Non-ATGC Characters	0	0
No. of Sequences >= 100 bp	2247	1201
No. of Sequences >= 200 bp	2196	977
No. of Sequences >= 500 bp	1675	929
No. of Sequences >= 1 Kbp	1241	795
N50 value	4578	8521
X coverage (approx)	257**	235**

3.2.6 Submission of data into Genbank

RAST and Artemis Genome Browser were used to generate additional information about the genomes. Prior to the submission of the genomes in Genbank, a prescreening was done to ensure that the scaffolds to be submitted were gapless and were greater than 200 bp in size. MAH 27-1 scaffolds were submitted in the WGS format with 977 contigs and an N50 value of 8,599 bp. The GC content of the genome was 69%. A genome size of 4,793,926 bp was submitted to the Genbank. The entire genome was predicted to encode 5006 coding sequences (CDSs), 3 rRNA-encoding genes and 46 t-RNA-encoding genes. The RAST server annotation pipeline showed that MAH 27-1 shared its highest similarity with MAH 104 compared to all mycobacteria whose complete genome sequences were already available. Furthermore, the genome MAH 27-1 was found to be smaller than the two available genomes as mentioned above. This whole genome shotgun project has been deposited in DDBJ/EMBL/GenBank under the accession no AWXK00000000. The version described is the first version with AWXK00000000.1 as the accession number.

MAH 2721, on the other hand had 2,195 contigs with an N50 value of 4,586 bp and an approximate coverage of 257 fold. The GC content of the genome was 68.6%. The genome has an approximate size of 5,211,349 bp. The entire genome was predicted to encode 6443 coding sequences (CDSs), 3 rRNA-encoding genes and 48 t-RNA-encoding genes. The genome of MAH 2721 is smaller to MAH strain 104 but is much larger to MAH strain TH135 thereby pointing out, the genomic heterogeneity of MAH strains. This whole genome shotgun project has been deposited in DDBJ/EMBL/GenBank under the accession no. AWXJ00000000. The version described is the first version and has the accession number AWXJ00000000.1.

3.3 Identification of a genome island in MAH with a flexible gene pool

Identification of regions specific to the MAH is vital towards understanding the genes that might promote MAH as a potential pathogen. Hence inter-species and intra-species comparison are essential. Since the majority of the comparative analysis was performed before the introduction of new MAH TH135 genome (introduced in August 2013), MAH 104 was the only complete MAH genome available for comparative analysis then. Hence an inter-species comparison of the MAH 104 genome with other members of the MAC was performed.

3.3.1 Identification of regions specific to the MAH by Vista gateway

The vista browser contains microbial genomes that are available in the form of precompiled alignments. The MAH 104 was selected as reference genome for further comparison with other genomes like the MAP K-10 and MAA ATCC 25291 and *M. intracellulare* ATCC 13950.

The comparative analysis resulted in the identification of seven specific regions that were explicitly found in MAH 104 but were absent in MAP K-10 and MAA ATCC 25291 and *M. intracellulare* ATCC 13950 genomes. A representative screenshot of these regions are shown below in the Figure 13. The regions identified by the comparative analysis are presented in the Table 20.

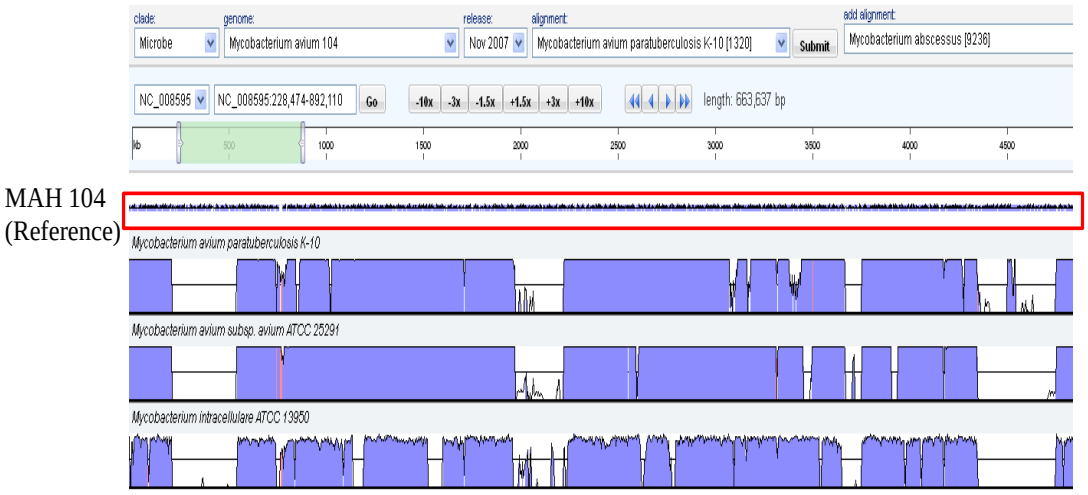


Figure 13: A representative screenshot of comparative genome analysis using VISTA browser

In the above Figure, the MAH 104 reference strain is marked in the red outline. The regions that are absent in the rest of the strains were considered as regions specific to the MAH 104.

It should be noted that the above Figure 13 provides a representative screen shot of the comparative genomics between the different genomes. All the seven regions identified in Table 22 have not been shown in the Figure 13 due to lack of image clarity while representation of the regions in one window. The screenshot image in Figure 13 is a zoomed version of the entire comparative genome analysis.

Table 20: Seven regions specific to the MAH identified by comparative genomics

Regions Identified	Genes	Start position – end position	Size of region
Region -1	**MAV_0253 - MAV_0298**	**254394 – 294226**	**39.83 Kb**
Region -2	MAV_0471 - MAV_0508	461330 – 493978	32.64 Kb
Region -3	**MAV_0779 - MAV_0841**	**746939 – 794035**	**47.09 Kb**
Region -4	**MAV_1458 - MAV_1506**	**1424505 – 1463494**	**38.98 Kb**
Region -5	MAV_1793 - MAV_2005	1788529 – 1987820	199.29 Kb
Region -6	MAV_2515 - MAV_2689	2548507 – 2724198	175.65 Kb
Region -7	MAV_3789 - MAV_3809	3916471 – 3939322	22.85 Kb

Since bacterial t-RNA genes are known to be insertion hotspot for GIs, these seven regions identified in the study as enumerated in Table 20 were checked for the presence of flanking t-RNA genes.

Three regions (Regions 1, 3 and 4) were identified with flanking t-RNA genes. Region 1 has 2 t-RNA genes, the t-RNA-serine and t-RNA-arginine at the 5′ end of the genome. Region 3 has t-RNA-lysine, t-RNA-glutamine, t-RNA-aspartate and t-RNA-phenylalanine at the 3′ end and region 4 has t-RNA-arginine in the 3′ end. These regions highlight the possibilities of GIs. To investigate the possibilities of GIs in these regions, the software islandviewer was used to explore their likelihood in these seven regions.

3.3.2 Identification of probable islands using Islandviewer

Island viewer uses a combination of 4 softwares to predict plausible islands in bacteria. The results from the seven regions identified above revealed that the regions 1, 2, 4, 5, 6 and 7 had probable islands (Table 21). Regions 1, 2, 4 and 7 contain one predicted island whereas regions 5 and 6 have seven and five predicted islands respectively. No islands were predicted in region 3.

Since regions 1 and 4 have one predicted island each along with flanking t-RNA genes and region 3 has no projected islands despite 4 flanking t-RNAs, further weightage was paid towards deciphering the genomic region 3.

(Table Contined on Next Page)

Table 21: Predicted islands in the 7 regions specific to the MAH

Regions Identified	Start position – end position of the region	Start position – end position of probable islands	Number of predicted islands
Region -1	**254394 – 294226**	**252,926 - 294,892**	**1**
Region -2	461330 – 493978	483,932 - 488,134	1
Region -3	**746939 – 794035**	**no island predicted**	**0**
Region -4	**1424505 – 1463494**	**1,456,328 - 1,463,365**	**1**
Region -5	1788529 – 1987820	1,801,703 - 1,805,910 1,809,699 - 1,816,301 1,823,660 - 1,831,906 1,878,666 - 1,884,543 1,919,399 - 1,927,530 1,956,188 - 1,961,917 1,977,459 - 1,990,545	7
Region -6	2548507 – 2724198	2,559,817 - 2,583,641 2,639,752 - 2,645,183 2,683,125 - 2,703,539 2,683,128 - 2,703,539 2,708,730 - 2,713,220	5
Region -7	3916471 – 3939322	3,919,180 - 3,939,322	1

3.3.3 Identification of a 47 Kb genomic island in MAH 104

Region 3 in MAH 104 comprises of 63 genes from MAV_0779 to MAV_0841 (Figure 14). 4 t-RNA genes namely t-RNA-lysine, t-RNA-glutamine, t-RNA-aspartate and t-RNA-phenylalanine follow after MAV_0841. 45 bp duplication regions are found at each end of the island as is shown in the Figure 16. The duplication region on the 3’ end of the genome is a part of the t-RNA lys gene (Mav_0842). 32 genes in this region are hypothetical by nature, 6 are conserved hypothetical genes and 17 are glycoprotein genes (GP genes). The segment also constitutes of phage element specific genes like the site-specific recombinase, phage integrase family protein, phage terminase protein and the phage tail protein. 4 other miscellaneous genes namely a PPE gene, the 30 KDa protein

gene, the 17 kDa surface antigen family protein gene and the excisionase family gene are also present in this island. No close homology was observed while performing similarity searches with complete genomes found on NCBI database. Similarity searches by NCBI BLAST analysis of draft genomes (30th January 2014) revealed that the region 3 shared close homology with three draft mycobacterial isolates. These were Mycobacterium sp. MAC_080597_8934 sharing 99% identity (94% query coverage), Mycobacterium sp. MAC_011194_8550 sharing 99% homology (69% query coverage) and *M. xenopi* RIVM700367 with 99% homology(53% query coverage) (Figure 14). A list of all genes found in the 47 Kb genome island in MAH 104 is presented in the Table 22. Figure 15 presents an idea about the gene organization in the 47 Kb island. Mycobacterium sp. MAC_080597_8934 and Mycobacterium sp. MAC_011194_8550 are recent introductions into the database (January, 2014).The NTM mentioned above have however not been typed to the species level. Hence it was difficult to perceive the MAC specific subspecies that shared homology to this genome island.

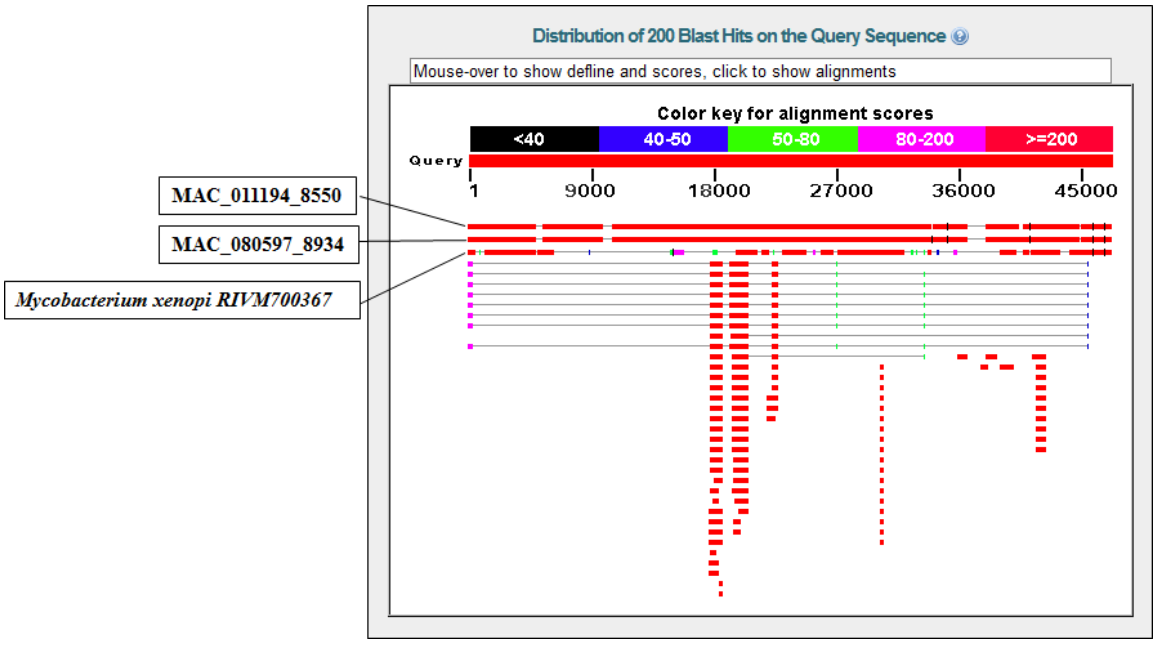


Figure 14: A representative NCBI BLAST Analysis of the Genome segment 3

The above BLAST results point towards a horizontal gene transfer event between M. xenopi and MAH or MAH and the MAC strains of Mycobacterium sp. MAC_080597_8934 and Mycobacterium sp. MAC_011194_8550.

The NCBI BLAST analysis results show that the genome segment shares homology with Mycobacterium sp. MAC_080597_8934, Mycobacterium sp. MAC_011194_8550 and *M. xenopi* RIVM700367.

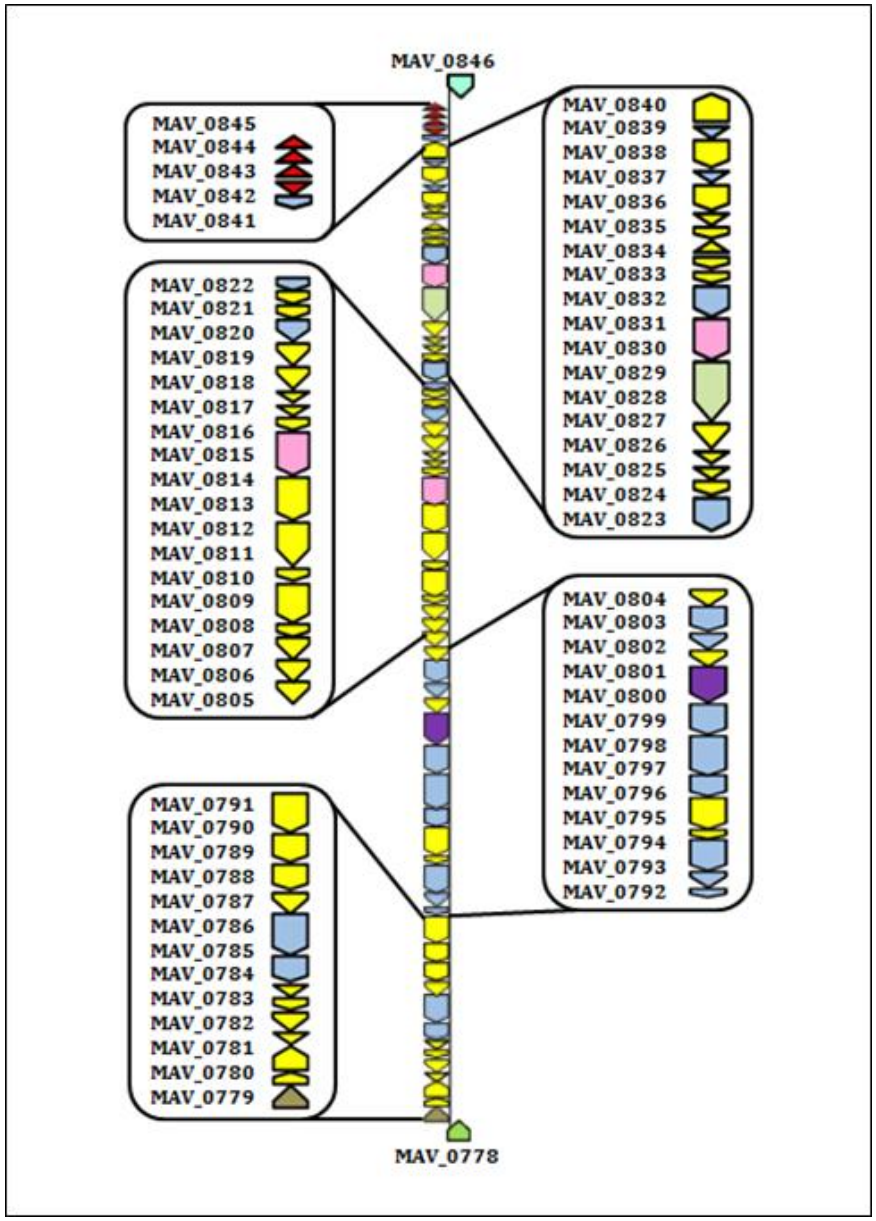


Figure 15: Gene organization of the 47 Kb Genome Island in MAH 104

The genes on either sides of the island marked in green color represent the flanking genes of the island. The hypothetical genes within the island are represented in yellow color. The genes marked in blue color represent glycoprotein genes within the island. The t-RNA genes are marked in red color.

Table 22: List of genes found within the genome island in MAH 104

Genes	Gene annotation	Gene	Gene annotation
MAV_0778	Methyltransferase	MAV_0812	hypothetical protein
MAV_0779	site-specific recombinase	MAV_0813	phage terminase
MAV_0780	hypothetical protein	MAV_0814	hypothetical protein
MAV_0781	hypothetical protein	MAV_0815	hypothetical protein
MAV_0782	hypothetical protein	MAV_0816	hypothetical protein
MAV_0783	hypothetical protein	MAV_0817	hypothetical protein
MAV_0784	hypothetical protein	MAV_0818	hypothetical protein
MAV_0785	hypothetical protein	MAV_0819	gp75 protein
MAV_0786	gp53 protein	MAV_0820	hypothetical protein
MAV_0787	gp50 protein	MAV_0821	hypothetical protein
MAV_0788	hypothetical protein	MAV_0822	gp79 protein
MAV_0789	hypothetical protein	MAV_0823	gp78 protein
MAV_0790	PPE family protein	MAV_0824	hypothetical protein
MAV_0791	hypothetical protein	MAV_0825	hypothetical protein
MAV_0792	gp34 protein	MAV_0826	hypothetical protein
MAV_0793	gp27 protein	MAV_0827	hypothetical protein
MAV_0794	gp28 protein	MAV_0828	17 kDa antigen protein
MAV_0795	hypothetical protein	MAV_0829	recT
MAV_0796	hypothetical protein	MAV_0830	gp60 protein
MAV_0797	gp37 protein	MAV_0831	hypothetical protein
MAV_0798	gp36 protein	MAV_0832	hypothetical protein
MAV_0799	gp23 protein	MAV_0833	hypothetical protein
MAV_0800	phage tail tape protein	MAV_0834	hypothetical protein
MAV_0801	hypothetical protein	MAV_0835	hypothetical protein
MAV_0802	gp32 protein	MAV_0836	hypothetical protein
MAV_0803	gp31 protein	MAV_0837	gp54 protein
MAV_0804	hypothetical protein	MAV_0838	hypothetical protein
MAV_0805	hypothetical protein	MAV_0839	gp54 protein
MAV_0806	hypothetical protein	MAV_0840	hypothetical protein
MAV_0807	hypothetical protein	MAV_0841	excisionase DNA protein
MAV_0808	hypothetical protein	MAV_0842	t-RNA-Lys
MAV_0809	hypothetical protein	MAV_0843	t-RNA-Glu
MAV_0810	hypothetical protein	MAV_0844	t-RNA-Asp
MAV_0811	hypothetical protein	MAV_0845	t-RNA-Phe
MAV_0812	hypothetical protein	MAV_0846	carveol dehydrogenase

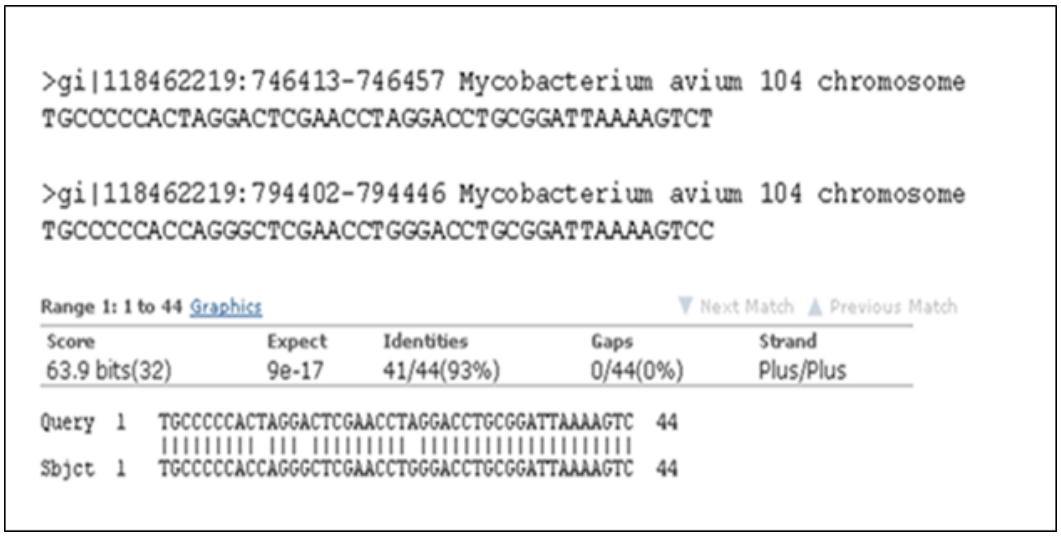


Figure 16: Duplication region in MAH 104

The Figure 16 shows the duplication regions that were found on either sides of the island. The duplication events on either sides of the genomic region point towards a possible genome island.

A genome segment is regarded as an island when it comprises of phage genes and transposable elements, has duplication regions on either side, has flanking t-RNA genes and also presents different GC content in comparison to the rest of the genome. Presence of phage genes, flanking t-RNA genes and duplication regions within the island have been shown in the Figures 15 and 16. The GC content of the island was analyzed thereafter. Since the island identified in our study was a result of HGT within mycobacterial species (Mycobacterium sp. MAC_080597_8934 sharing 99% identity (94% query coverage), Mycobacterium sp. MAC_011194_8550 sharing 99% homology (69% query coverage) and *M. xenopi* RIVM700367 with 99% homology (53% query coverage)), a huge difference in the GC content was not observed. A drop in the GC content was observed at the 3' end of the island. The GC content of the MAH 104 genome is 69% whereas the GC content of the island was 67%. Figure 17 displays the GC profile of the region.

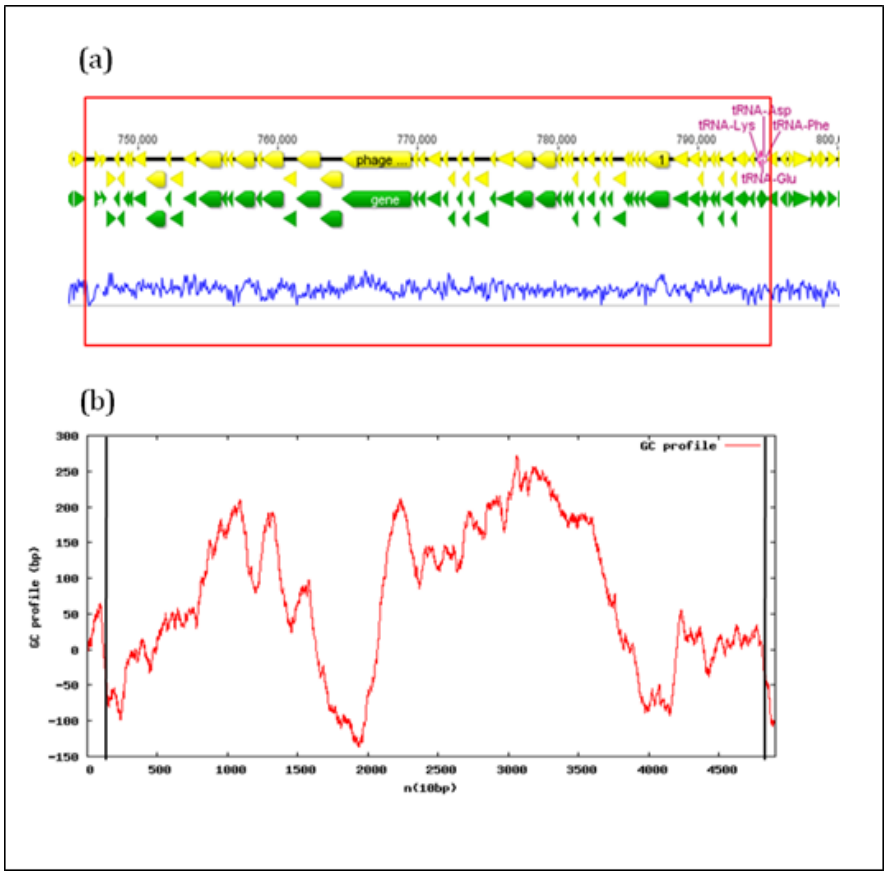


Figure 17: The GC curve of the island using the softwares geneious and the GC profile.

Figure 17(a) represents the GC curve obtained by the Geneious software. The blue lines in the Figure 17(a) represent the GC profile of the island. Figure17 (b) represents the GC curve by GC profile. The GC drop varies from one region of the island to another. A drop in the GC content is observed at the 3' end of the island.

Since this region displays phage genes, has duplication regions on either side, has flanking t-RNA genes and also presents a small drop in GC content followed by homology to *M. xenopi*, this region was considered as an island. It should be noted that most of the work related to the identification of the genome island was performed before the introduction of the two new draft genomes from MAC strains as mentioned above. Hence *M. xenopi* was the only homologous reference for the genomic region 3 available before January 2014. Also the lack of available species specifications from these MAC strains makes the analysis about the gene transfer tougher.

3.3.4 Identification of genomic islands in other sequenced MAH strains

Following the identification of the 47 Kb region in MAH 104, the next investigation required, looking into the genomes in other MAH strains in order to investigate if these strains contained the same genome islands as identified in MAH 104. The analysis was performed by identification of homologs of flanking genes namely the MAV_0778 (methyltransferase, putative, family protein) and MAV_0846 (carveol dehydrogenase). The genes flanking the island as mentioned above form a part of the core gene pool and are expected to be conserved in all MAH strains. The regions within the homologous expanses of MAV_0778 (methyltransferase, putative, family protein) and MAV_0846 (carveol dehydrogenase) were identified as islands in different strains of MAH. The homology of the islands was observed by performing BLAST analysis of whole island or parts of it and identifying the organism that showed maximum homology to these regions. Islands from two complete genomes namely MAH 104 and TH135 and four draft genomes namely MAH 27-1, MAH 2721, MAH 10-4249, and MAH 10-5606 were examined during this study. The comparative analysis provided insights into the flexible gene pool within this region in different MAH strains.

The newly sequenced MAH TH135 strain revealed a much smaller island of size 1.71 Kb instead of the 47 Kb island found in the genome MAH 104 (BLAST analysis results from 30th January 2014). The contents of the genome island and the regions flanking the island are shown in the Table 23. Two hypothetical genes namely the MAH_0684 and MAH_0685 were identified in the island. The structure of the flanking regions and that of the t-RNA remained consistent and were similar to the one observed in MAH 104. A diagrammatic representation of the island in MAH TH135 is provided in the Figure 18. The GC content of the island was found to be 50% in comparison to the whole MAH TH135 genome where the GC content is as high as 69.3%. The 1.71 Kb island did not share homology to complete genomes available during BLAST analysis. However analysis with draft genomes revealed that the two genes namely the MAH_0684 and MAH_0685 share close homology in the nucleotide level to the contig 000582 (99% identity, 86% query coverage) and contig 000583 (99% identity, 18% query coverage) of a draft genome of an MAH strain, MAH 10-4249 isolated from a deer in Colorado in USA.

Table 23: The genes found in the genome island in MAH TH 135

Genes	Gene annotation
MAH_0683	o-methyltransferase
MAH_0684	hypothetical protein
MAH_0685	hypothetical protein
MAH_t013	t-RNA-Lys
MAH_t014	t-RNA-Glu
MAH_t015	t-RNA-Asp
MAH_t016	t-RNA-Phe
MAH_0686	carveol dehydrogenase

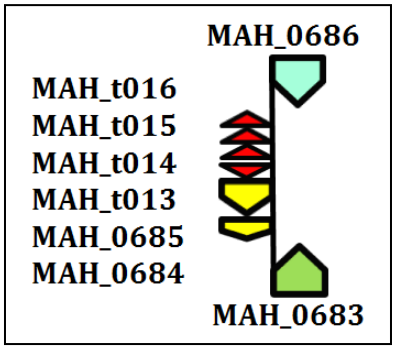


Figure 18 : Gene Organisation in MAH TH135

The Figure above illustrates the gene organisation in the genome island in MAH TH135. The island comprises of two hypothetical genes namely MAH_0684 and MAH_0685.

The island was examined in the draft genomes of MAH 27-1, which was sequenced during this study. The study exposed a completely dissimilar 14.2 Kb island in MAH 27-1 which was different from the ones observed in MAH 104 and MAH TH135. The island when exposed to BLAST analysis (30th January, 2014) with complete genomes exposed high levels of homology with a *M. intracellulare* isolate identified from a clinical subject in Korea, the MOTT-36Y (99% homology and 99% query coverage) and with *Mycobacterium yongonense* (*Mycobacterium* sp. *05-1390*) (99% homology and 99% query coverage), a slow growing mycobacterium related to the *M. intracellulare*. Homology studies with *M. intracellulare* MOTT 36Y exposed fourteen genes ranging from *W7S_03350* to *W7S_3410* followed by a t-RNA phenylalanine that were shared with MAH 27-1. The island is flanked by four t-RNAs after methyltransferase gene (*W7S_03345*) on one side (Figure 19). The GC content of the island was 62% whereas the

GC content of the MAH 27-1 genome is 69%. The island is represented by contigs 32 (60% query coverage, 100% identity) and contig 38 (40% query coverage 100% identity) in the WGS draft genomes of MAH 27-1 (Accession number AWXK01000000). Since the genome of MAH 27-1 is not annotated, the constituents of the island are shown in the Table 24 and the genes annotations provided represent the fourteen genes common to *M. intracellulare* clinical isolate, the MOTT-36Y and MAH 27-1.

The 14.2 Kb region identified in MAH 27-1 was also found in the contig 338 (99% query coverage 99% identity) of MAH 2721 implying that this region was common to both the MAH draft genomes sequenced in this study as shown in the Figure 20. The position of the four flanking t-RNA genes namely the t-RNA-lysine, t-RNA-glutamine, t-RNA-aspartate and t-RNA-phenylalanine present in the island of MAH 27-1 was reversed and was found at the start of the island and not the end of the island as is observed in the MAH 104 and MAH TH135.

Table 24: The genes within the genome island in MAH 27-1

Genes	Gene annotation
W7S_03345	methyltransferase, putative, family protein
W7S_t25859	t-RNA-Lys
W7S_t25861	t-RNA-Glu
W7S_t25863	t-RNA-Asp
W7S_t25865	t-RNA-Phe
W7S_03350	prophage integrase
W7S_03355	hypothetical protein
W7S_03360	hypothetical protein
W7S_03365	hypothetical protein
W7S_03370	regulatory protein
W7S_03375	hypothetical protein
W7S_03380	hypothetical protein
W7S_03385	hypothetical protein
W7S_03390	hydrolase
W7S_03395	MmpL family transport protein
W7S_03400	TetR family transcriptional regulator
W7S_03405	cytochrome P450 123B1 Cyp123B1
W7S_03410	TetR family transcriptional regulator
W7S_t25867	t-RNA-Phe
W7S_03415	carveol dehydrogenase

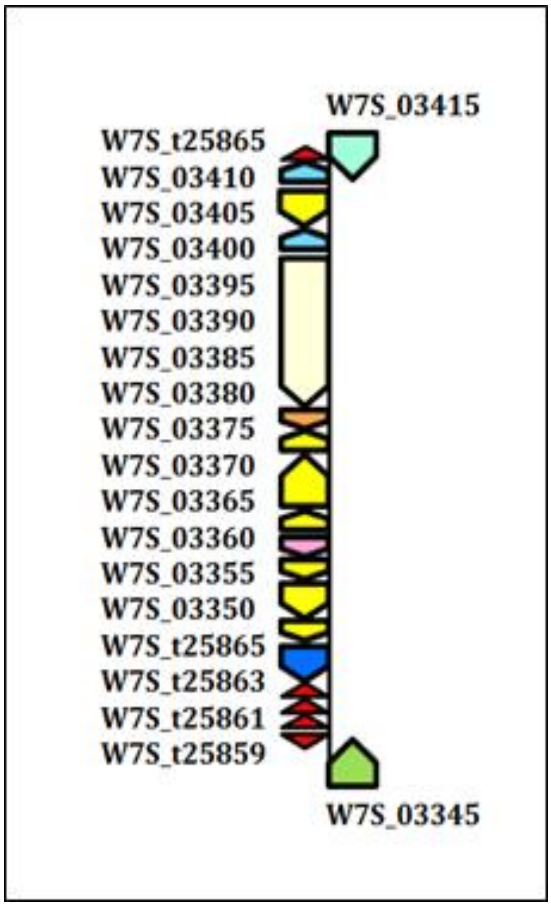


Figure 19: Gene Organisation in MAH 27-1

The above Figure presents the structure of the genome island in the MAH strain isolated from dust in this study, the MAH 27-1. The island identified in MAH 27-1 shares close homology to clinical M. intracellulare strain, the MOTT-36Y.

The investigation of the island in the draft genome of MAH 2721 revealed an island of size, at least 31.5 Kb. The island was identified in two contigs of the draft genome, namely the contig 338 and the contig 341. However the connecting link between the two contigs was not identified in this study as the reassembly of raw reads obtained from Illumina and Ion torrent data did not assist with the gap filling. The GC content of the island was hence not looked into. The region presented on the contig 338 in MAH 2721 resembles the 14.2 Kb region observed in MAH 27-1 from *W7S_03350* to *W7S_3410* followed by the t-RNA phenylalanine (*W7S_t25865*). The 14.2 Kb region in MAH 2721 gives way to a 7.2 Kb region that does not share homology greater than 50% to any other mycobacteria or known bacteria. The second region found on contig 341 covers a 10 Kb

region that shares homology to *Mycobacterium septicum* DSM 44393 (41% query coverage 76% identity) and gives way to the Carveol dehydrogenase gene which is one of the flanking regions of the island The pictoral representation of the island is provided in the Figure 20.

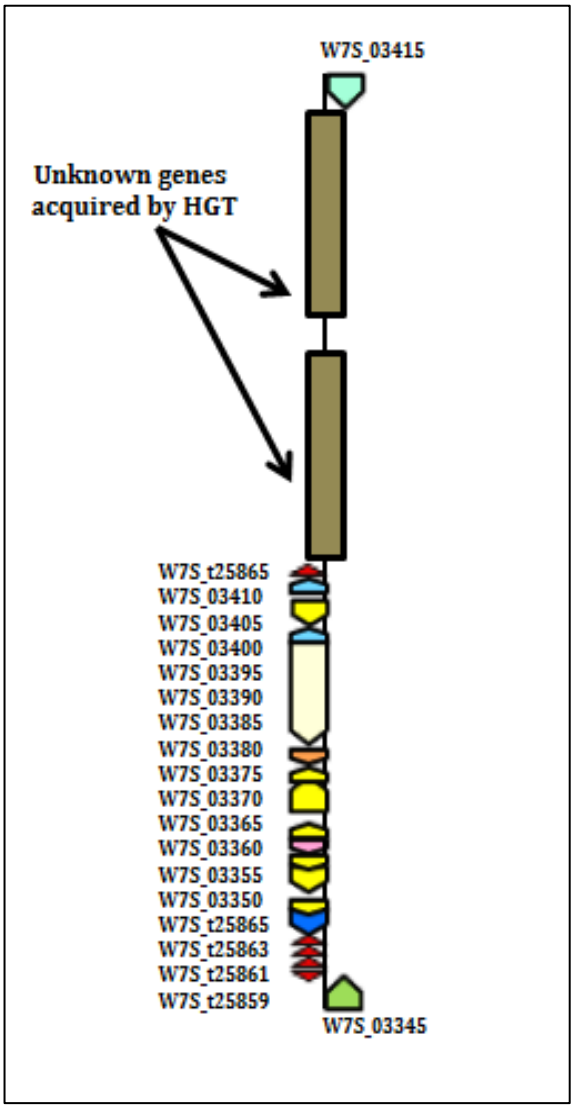


Figure 20: Gene organisation of the genome island in the MAH strain 2721

The above illustration represents the at least 31.5 Kb island in MAH 2721. The flanking genes present in the island remain consistent but the position of the t-RNA genes within the island are reversed when compared to MAH 104 and MAH TH135.

Identification of genes within the island of the deer isolate MAH 10-4249 was easier as it was an annotated draft genome. A genome island of approximately 19 Kb was identified in MAH 10-4249 on contig 582 and contig 583. The genome island provides a mix of

genes from MAH TH135 (two hypothetical genes) and several genes from *M. intracellulare* isolate the MOTT-36Y (91% query coverage and 82% identity). Homology searches with draft genomes of MAH 27-1 (Contigs 32 and contigs 38) and MAH 2721 (contig 338) also revealed an identity of 82%. The flanking region of S-adenosyl-L-methionine-dependent methyltransferase (O971_18005) was identified in contig 583 followed by a non-annotated region that shares close homology to MAH_0684. The rest of the island is found on contig 582 starting with a non-annotated region that shares close homology to MAH_ 0685 followed by the 4 t-RNAs, t-RNA-Lys, t-RNA-Glu, t-RNA-Asp, t-RNA-Phe (Table 25). The island ends with oxidoreductase which is a homolog for carveol dehydrogenase. A non-annotated t-RNA Phe gene was also identified before oxidoreductase (Figure 21).

Table 25: The genome island specific genes in deer isolate MAH 10-4249

Genes	Gene annotation	Contig
O971_18005	S-adenosyl-L-methionine-dependent methyltransferase	Contig 583
Not annotated	Shares homology to MAH_0684	Contig 583
Not annotated	Shares homology to MAH_0685	Contig 582
O971_18000	t-RNA-Lys	Contig 582
O971_17995	t-RNA-Glu	Contig 582
O971_17990	t-RNA-Asp	Contig 582
O971_17985	t-RNA-Phe	Contig 582
O971_17980	Integrase	Contig 582
O971_17975	cell division protein FtsK	Contig 582
O971_17970	plasmid replication	Contig 582
O971_17965	hypothetical protein	Contig 582
O971_17960	hypothetical protein	Contig 582
O971_17955	Regulator	Contig 582
O971_17950	TetR family transcriptional regulator	Contig 582
O971_17945	hypothetical protein	Contig 582
O971_17940	alpha/beta hydrolase	Contig 582
O971_17935	membrane protein; disrupted	Contig 582
O971_17930	TetR family transcriptional regulator	Contig 582
O971_17925	cytochrome P450	Contig 582
O971_17920	TetR family transcriptional regulator	Contig 582
O971_17915	hypothetical protein	Contig 582
Not annotated	t-RNA-Phe	Contig 582
O971_17910	Oxidoreductase	Contig 582

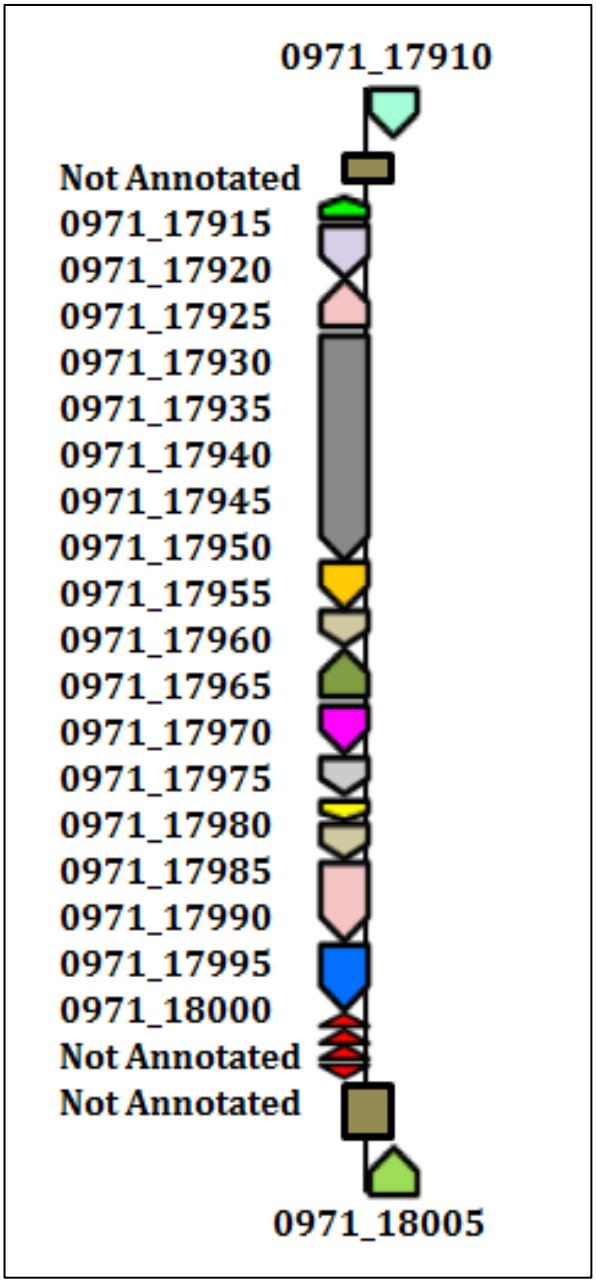


Figure 21: Gene organisation of the island in the MAH isolate from deer MAH 10-4249

The above gene map illustrates the genome island in the MAH isolate from deer. The genome island in MAH 10-4249 represents a mixture of a genome island both from MAH TH135 represented by the non-annotated region next to the flanking gene (0971_18005). A lot of genes in the island share homology to the genome island specific genes in the M. intracellulare MOTT-36Y.

The MAH isolate from pig (MAH 10-5606) contains an island of size 33.5 Kb on contigs 660 and 661. While 12.7 Kb of the island is found on contig 660, the rest of the island of approximately 20.7 Kb was identified on contig 661. The genes found on the island are represented in the Table 26. The flanking regions remain constant. Majority of the genes in the island are hypothetical genes. BLAST analysis (30th January 2014) with island found on contig 661 (20.7Kb) region revealed homology to 86% identity (75% query coverage) to clinical intracellulare strain *Mycobacterium intracellulare* MOTT-64 which is a frequently encountered clinical genotype in South Korea. BLAST analysis of the island from contig 660 revealed a region that did not share close homology to any complete genome available in the NCBI database. However analysis with draft genomes available in NCBI database revealed 99% identity (100% query coverage) with *Mycobacterium avium* strains namely, *Mycobacterium avium* 09-5983 (Contig 591) and *Mycobacterium avium* 05-4293 (contig 80) from USA. Since these strains were not subtyped they did not make any value added contribution to the analysis.

(Figure Contined on Next Page)

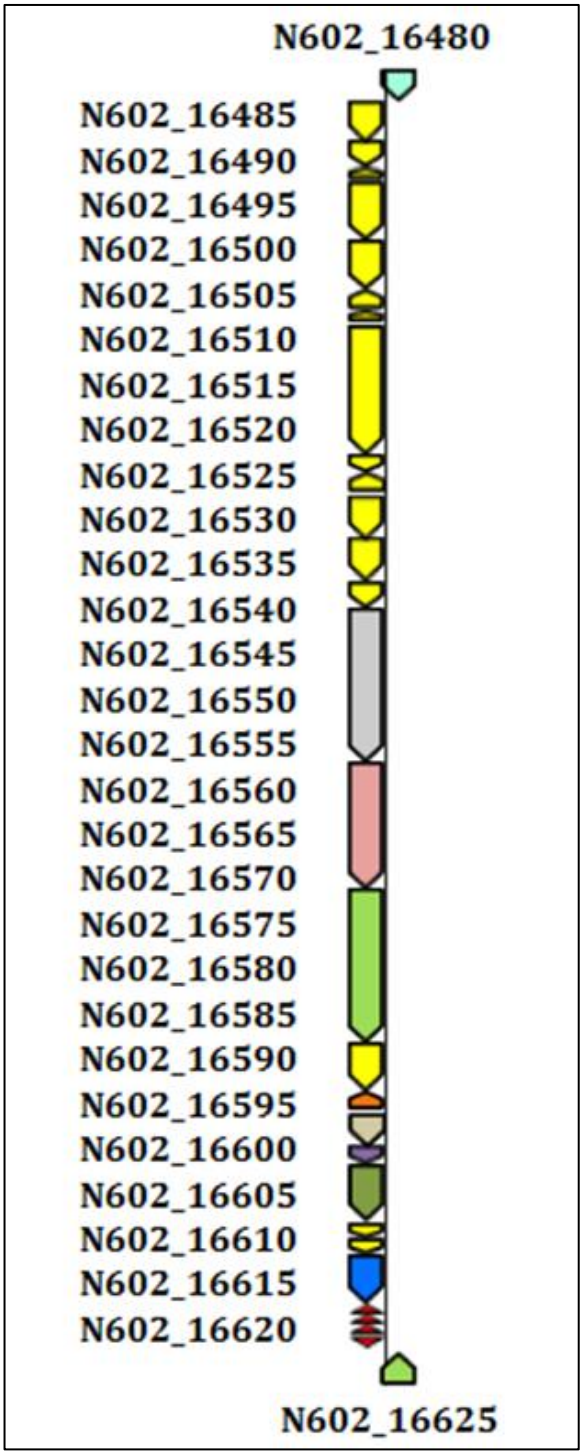


Figure 22: Gene organisation of the genome island in MAH 10-5606

The organization of genes in MAH 10-5606 is presented in the above diagram. A majority of the genes within the island are hypothetical in origin. The island shares 86% identity with Mycobacterium intracellulare MOTT-64.

Table 26: The gene organisation in the MAH isolate from Pig (MAH 10 -5606)

Genes	Gene annotation	Contig
N602_16625	S-adenosyl-L-methionine-dependent methyltransferase	Contig 661
N602_16620	t-RNA-Lys	Contig 661
N602_16615	t-RNA-Glu	Contig 661
N602_16610	t-RNA-Asp	Contig 661
N602_16605	t-RNA-Phe	Contig 661
N602_16600	Integrase	Contig 661
N602_16595	hypothetical protein	Contig 661
N602_16590	hypothetical protein	Contig 661
N602_16585	cell division protein FtsK; disrupted	Contig 661
N602_16580	plasmid replication, integration and excision activator	Contig 661
N602_16575	GntR family transcriptional regulator	Contig 661
N602_16570	DNA mismatch repair protein MutT	Contig 661
N602_16565	hypothetical protein	Contig 661
N602_16560	ATPase AAA; disrupted	Contig 661
N602_16555	hypothetical protein	Contig 661
N602_16550	helicase; disrupted	Contig 661
N602_16545	hypothetical protein	Contig 661
N602_16540	hypothetical protein	Contig 661
N602_16535	hypothetical protein	Contig 661
N602_16530	Transposase	Contig 660
N602_16525	hypothetical protein	Contig 660
N602_16520	hypothetical protein	Contig 660
N602_16515	hypothetical protein	Contig 660
N602_16510	hypothetical protein	Contig 660
N602_16505	hypothetical protein	Contig 660
N602_16500	hypothetical protein	Contig 660
N602_16495	hypothetical protein	Contig 660
N602_16490	hypothetical protein	Contig 660
N602_16485	hypothetical protein	Contig 660
N602_16480	Oxidoreductase	Contig 660

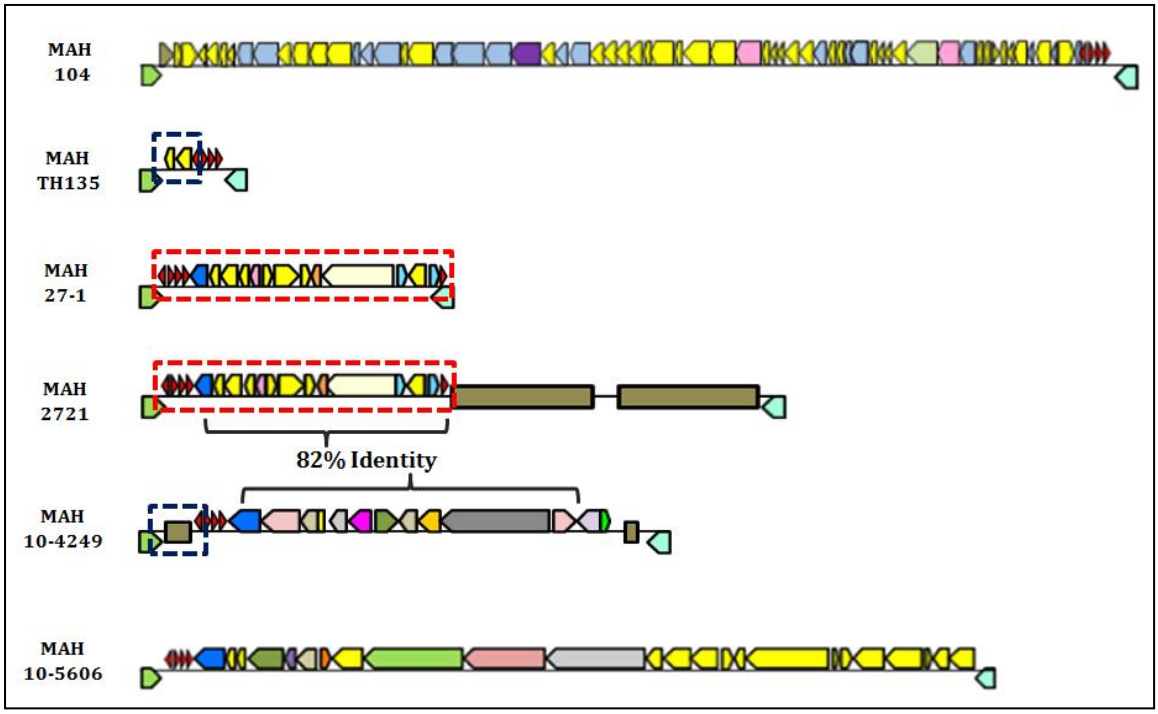


Figure 23: The comparative analysis of flexible gene pool in MAH

The above Figure provides an overview of the genome island (Region-3) in MAH. The homologous regions in the different islands are presented by hased boxes. The region of homology in MAH 27-1 and MAH 2721 is represented in the above Figure (red hashed box). The region of homology in MAH TH135 and MAH 10-4249 is shown in the black hashed box. The t-RNA genes are presented in red colour.

The above Figure 23 clearly shows that the region 3 presents a typical example of a flexible diverse gene pool. The comparative analysis reveals that the islands in MAH 104 and MAH 5606 are different from the rest of the islands (region 3). MAH TH135 shares homology with MAH 10-4249 as is represented in the Figure 23 (black hashed box). MAH 10-4249 shares 82% homology with MAH 27-1 and MAH 2721. MAH 27-1 and MAH 2721 shares 14 genes and 4 t-RNA genes apart from the flanking genes (represented by the red hashed lines) in the island. While the MAH 104 and MAH TH135 share the same gene organisational positions of the t-RNA genes (upstream of the island near the 3' end), the organisation of the t-RNA genes in the rest of the MAH strains stand close to the 5' end of the island.

3.4 SNP analysis of 3 genes in 36 MAH strains

In-house perl scripts were used to identify gene candidates with SNP patterns (from eight MAH strains) from the genome data generated by Roche sequencing. Different permutation combinations were applied to the individual genes to observe similarities and dissimilarities between different MAH groups originating from both environmental and clinical sources. Similarities between MAH strains from dust and MAH strains from children with lymphadenitis were observed but these results were necessary to be reproduced in larger MAH strain sets. Three genes were selected for SNP analysis based on identical SNP patterns in the MAH strains from dust and children with lymphadenitis and literature reviews of the virulence associations of these genes. Since the whole genome SNP analysis was performed with genome information of eight strains from Roche 454 sequencing platform (which had a low coverage as shown in Table 12), the SNP analysis was reproduced in thirty six MAH strains from varying environmental and clinical sources. 17 MAH strains (8 strains isolated from soil and 9 strains isolated from dust) were environmental in origin whereas 19 strains (14 strains isolated from children suffering from Lymphadenitis and 5 strains from adults with lung infections) were isolated from clinical subjects. All the three genes selected in the SNP analysis were surprisingly found in a region upstream of the genome island identified in our study in section 3.3.3. The genes chosen for analysis were MAV_0846 (carveol dehydrogenase), MAV_0847 (TetR family transcriptional regulator) and MAV_0853 (ddpX-D-alanyl-D-alanine dipeptidase).

3.4.1 Phylogenetic analysis of MAV_0846, MAV_0847 and MAV_0853

The phylogenetic analysis was performed to see if the grouping of the clinical MAH isolates was dependent on any of the environmental niche that these organisms were found in.

The phylogenetic analysis of MAV_0846 (Carveol dehydrogenase) gene in Figure 24 was performed from 36 MAH strains. The results revealed 21 strains in one big cluster comprising of 8 strains of children suffering from lymphadenitis, 2 strains from soil, 7 strains from dust and 4 strains from adults with lung infection. The strains in this cluster are closely related. The cluster contained a major bulk of clinical isolates (8 clinical isolates out of 14 MAH isolates in children suffering from Lymphadenitis and 4 out of 5 isolates from adults with lung infection) sharing close proximity to dust isolates (7 MAH isolates out of 9

isolates) implying that dust could have an important role to play in MAH infections in Germany. MAH isolates from soil were divided amongst several other sub branches.

Phylogenetic analysis in MAV_0853 identified that the environmental and clinical isolates are unevenly distributed across different branches as is shown in the Figure 25. However an interesting feature of the phylogenetic analysis was a big cluster of 17 strains comprising of 7 MAH isolates from dust, 5 from children suffering from lymphadenitis, 4 strains from adults with lung infection and 1 isolate from soil. A very similar cluster of 21 MAH strains was observed in the phylogenetic analysis of MAV_0846 (Carveol dehydrogenase). 17 strains belonging to the cluster identified in the gene MAV_0857 were found to be grouped together in the phylogenetic analysis of MAV_0846.

The results from the phylogenetic analysis of MAV_0847 (TetR family transcriptional regulator) showed that the environmental and clinical isolates are unevenly mixed. However majority of the MAH strains isolated from children suffering from lymphadenitis (8 MAH strains out of 14 MAH strains) were clustered together implicating that these strains are closely related to one another (Figure 26).

(Figure Contined on Next Page)

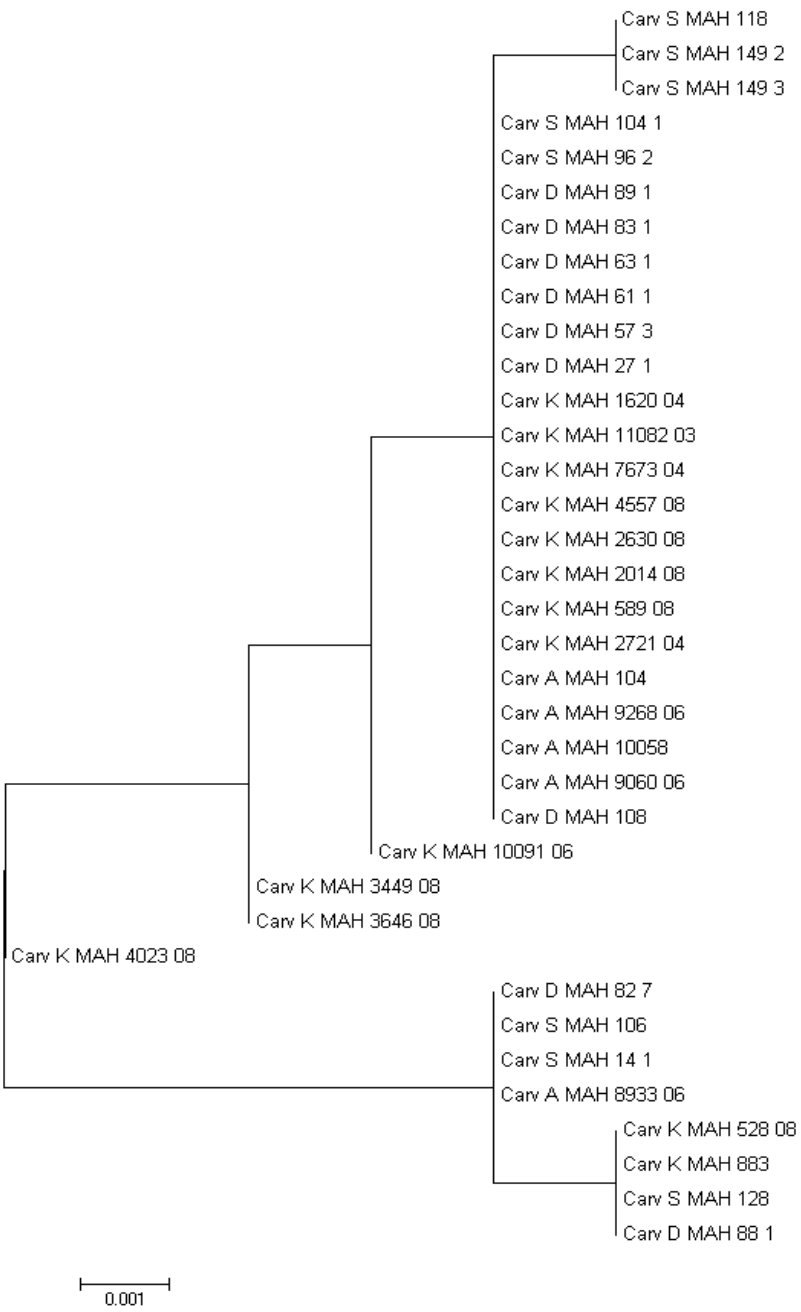


Figure 24: Phylogenetic analysis of MAV_0846 (Carveol dehydrogenase).

The illustration shows a phylogenetic analysis of gene MAV_0846 in MAH strains from dust (D), soil (S), adult with lung infection (A) and children suffering from lymphadenitis (K). The analysis shows that a majority of the MAH strains from dust and from children suffering from lymphadenitis share proximity to one another. However the cluster also comprises of other MAH strains hence showing that it is difficult to group the MAH strains according to their ecological origin.

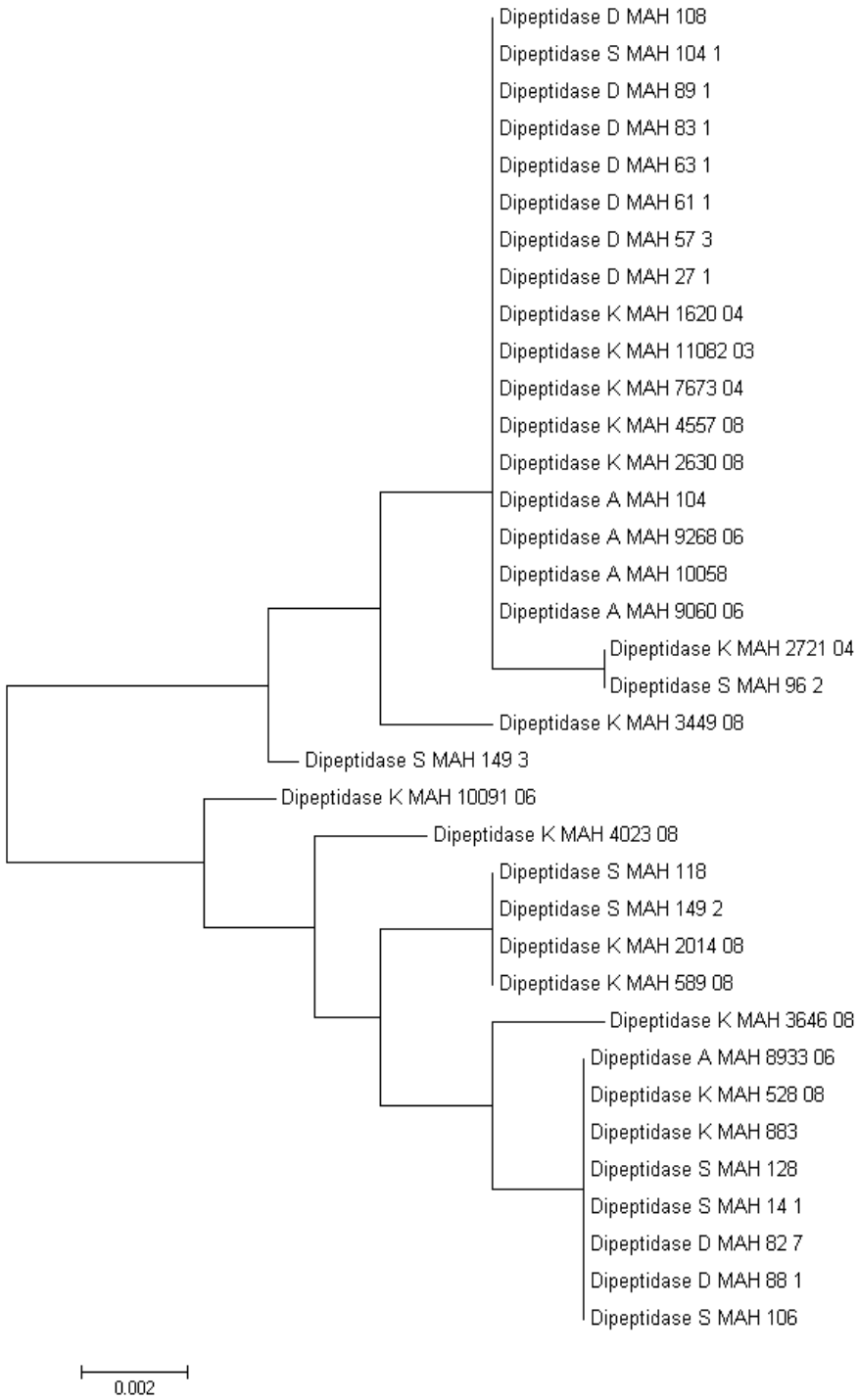


Figure 25: Phylogenetic analysis of MAV_0853 - ddpX - D-alanyl-D-alanine dipeptidase

The illustration shows a phylogenetic analysis of gene MAV_0853 in MAH strains from dust (D), soil (S), adult with lung infection (A) and children suffering from lymphadenitis (K). The tree shows that the MAH strains from environmental and clinical niches are unevenly distributed across different branches.

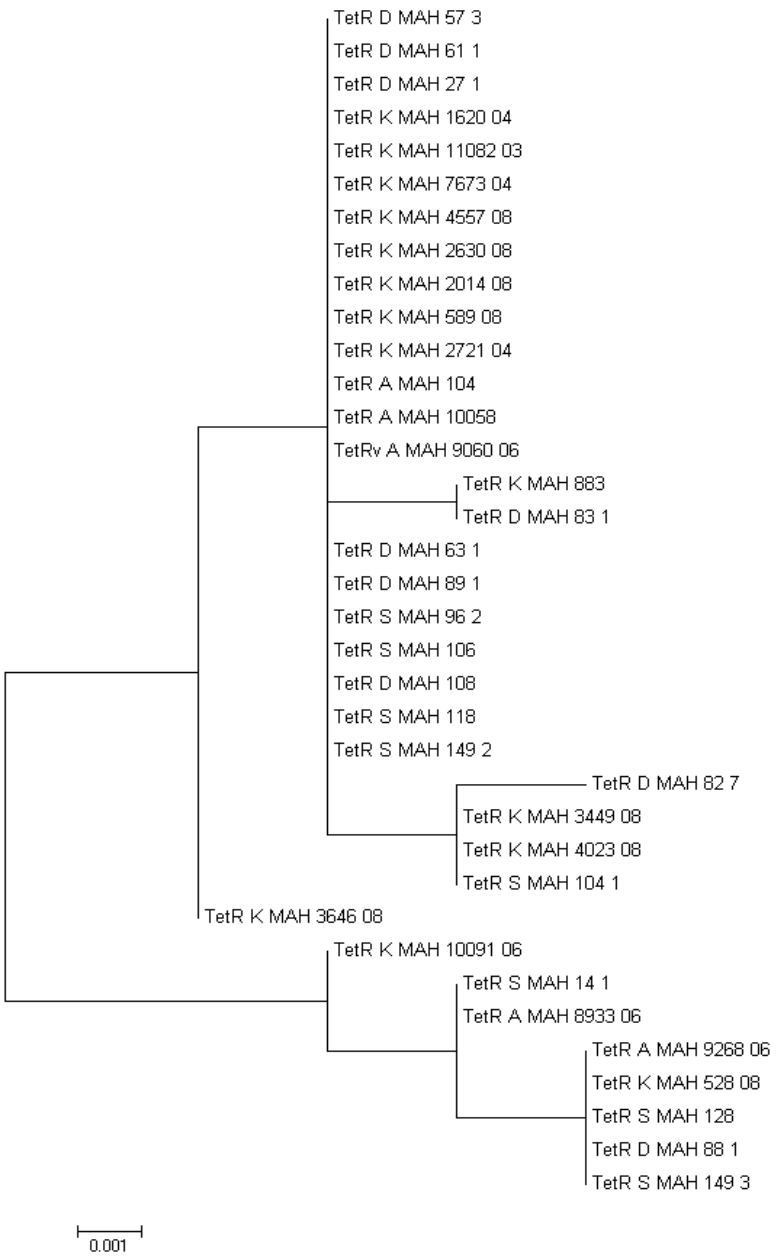


Figure 26: Phylogenetic analysis of MAV_0847 (TetR family transcriptional regulator)

The illustration shows a phylogenetic analysis of gene MAV_0847 in MAH strains from dust (D), soil (S), adult with lung infection (A) and children suffering from lymphadenitis (K). The phylogenetic analysis of MAV_0847 gene indicates that the MAH strains from children suffering from lymphadenitis could be closely related. The rest of the MAH strains from different niches are unevenly mixed across different branches.

3.4.2 Phylogenetic analysis of the concatenated genes across thirty six strains

The phylogenetic analysis of concatenated genes across 36 strains showed an irregular distribution. The irregular distribution made it difficult to decipher the route of infections. However there was one big cluster containing 6 MAH isolates from dust and 5 MAH isolates from children suffering from lymphadenitis and 3 MAH isolates from adults with lung infection. The cluster indicates that these MAH isolates are close to one another. However conclusions in this regard would require the analysis of much larger MAH populations.

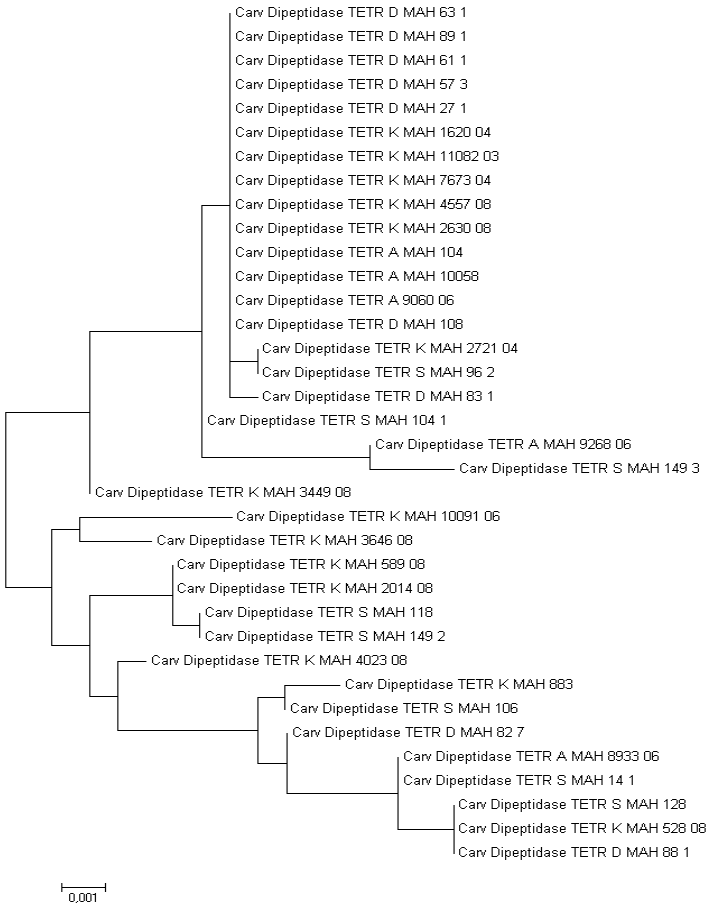


Figure 27: Phylogenetic analysis of concatenated genes in thirty six MAH strains

The illustration shows a phylogenetic analysis of the concatenation of three genes in MAH strains from dust (D), soil (S), adult with lung infection (A) and children suffering from lymphadenitis (K). The concatenated phylogenetic analysis of three genes in thirty six MAH strains shows an uneven distribution.

4. Discussion

Prokaryotic classification in practice has relied on a polyphasic approach to incorporate phenotypic, genotypic, and ecological information about a microbial species or subspecies. Any microbial species or subspecies though, defined phylogenetically may comprise of a variety of ecotypes. Since bacterial strains are commonly isolated from different environments, incorporation of fine-scale ecological information into taxonomic classifications is generally difficult [106]. Recent data using approaches like genome sequencing, SNP analysis, multilocus sequence type analysis and microarray analyses support the interpretation that speciation processes may occur at a subspecies level, both within ecological niches (ecovars) and different biogeography (geovars) [107].

4.1 Environmental MAH in Germany

Organismic and ecological properties of a bacterium can be judged through its habitat and its physical environment. *M. avium* is of relevant clinical importance in Germany as these members are the most common causative agent of NTM infections. Apart from infections in livestock and birds through MAP and MAA, the *M. avium* also causes pulmonic infections in elderly people, cervical lymphadenitis in children, lung infections in HIV infected or otherwise immune-compromised hosts. These infections are attributed to MAH. The NTM hospital study as mentioned in the personal communication in section 3.1.1 revealed that *M. avium* is the main cause of NTM disease (78%) in children in Germany [108]. Whether decreasing BCG vaccination rates and decreasing exposure to TB cases have an impact on the number of pediatric NTM infection could not be elucidated in this communication. This result was in accordance with the outcome of a similar study from the Netherlands on NTM disease, which also identified that *M. avium* was the main cause of NTM disease [109].

The sources of MAH infections for man has remained ambiguous though high resolution genotyping methods like Random fragment length polymorphism (RFLP) and Pulse field gel electrophoresis (PFGE) have been applied to compare isolates originating from different host origins [110]. Contaminated food originating from livestock was also discussed as potential source of infection for man. However no suitable evidences have been found. Studies of MAH strains from pigs, human and living environment by Iwamoto and colleagues revealed that regional factors or local specific source of infection and route of transmission could be

responsible for MAH infections in humans [87]. The environment rather than direct transmission between individuals has been suggested as the primary source of MAH infection. The public health authorities have little interest in national or regional surveillance programs and transmission based studies from environment to patients have henceforth been few.

The data concerning environmental reservoirs of MAH in Germany was inadequate. Hence reservoirs of MAH were examined in order to approach the question of possible sources of human infections by these bacteria in Germany.

Soil, dust, water and biofilms are the common environmental reservoirs of MAH. Differentiating between soil and dust is difficult and can be based on particulate size and density. Indoor dust constitutes of finer soil particles blown indoors through open windows and doors or particles drawn in by the indoor heating and air conditioning system. The source of indoor dust can be quite variable and can comprise of food residues, skin particles and hair. Hence a careful assessment of soil and dust was performed. Contents of vacuum cleaner bags were primarily considered as dust and those from garden soil, potted soil and other outdoor regions were categorized as soil.

Studies by Falkinham and colleagues have shown evidence of water being the source of *M. avium* pulmonary disease in patients in the United States [76]. A second example by Arbeit and colleagues has also shown that potable water systems colonized by *M. avium* had similar DNA fingerprints as that of *M. avium* isolated from AIDS patients who were exposed to these waters [111]. These above mentioned studies may imply that *M. avium* reservoirs play an important role in defining the profiles of patients that have *M. avium* infections. Studies performed in Australia show comparable results where NTM isolated from showerheads in patient households matched the profiles of the NTM affecting the patients. These evidences were found the strongest for *M. avium* and *M. kansasii* [112]. These studies point towards water being the most apparent source of infection as *M. avium* has been most frequently isolated from water in these countries.

Contrasting to the above mentioned findings, personal communication with Dr. Roland Schulze–Röbbecke had shown a low prevalence of *M. avium* (3%) in surface water and drinking water in Germany. Also, study of water samples in Berlin, Germany by Peters and

colleagues [113] revealed MAC in only 1.7 % of tap water samples. Such low *M. avium* isolation rates in household water samples in Berlin, Germany was the main reason behind the search for MAH in other alternate environmental habitats in Germany and we mainly focused on samples from Berlin. Our sample collection covered the substrates such as water, biofilms, soil, dust.

Additional confirmatory results from our study as presented in section 3.1.2 replicate similar results as Peters and colleagues and assert that MAH concentration in biofilms, surface water and drinking water is low in Germany. The sensitivity of our isolation method required 1000 MAH per ml of water implying either absence or low density of MAH in these samples. It therefore seems unlikely that water and biofilms are the main infection source for *M. avium*-induced infections in Germany. Alternative environmental niches of *M. avium* such as soil as outdoor environment and household dust taken from vacuum cleaners as indoor environment were considered for further analysis. Torvinen and colleagues have used dust from vacuum cleaners to assess the human exposure to NTM in the house [114]. Simultaneous to the same, we were able to frequently isolate *M. avium* from a relatively high percentage of soil (19%) and dust samples (33%). No recovery occurred in soils from forests or urban spaces, while potting soils, garden soils and soils from playgrounds allowed MAH recovery. Interestingly, all the positive samples came from environments close to human contact. MAH were isolated from 37.5 % of potting soil samples, 12.5 % of garden soil samples and 17 % of soil samples from play grounds (Unpublished data from this study). The high recovery rates of MAH from potting soils support the hypothesis of other authors who proposed work with potting soil as transmission route [80, 115]. All *M. avium* isolates belonged to the subspecies *M. avium hominissuis* (MAH) confirming MAH as environmental *M. avium* subspecies. The more frequent abundance of *M. avium* in soil and dust compared to water suggests that soil or dust could be the substrates transmitting *M. avium* infections in Germany. A very recent article by Falkinham supports the idea that the route of MAH infections can be numerous and can vary. Ingestion of soil by children leading to cervical lymphadenitis, inhalation of aerosols and dust from potting soil leading to pulmonary infections and oral ingestion of MAH in water by immune-compromised patients can be possible means of MAH infection [115]. This observation by Falkinham supports the reasoning behind MAH in German soil and dust being a potential infection source.

On the other hand, contradictory results of the MAH ecology, from our studies in Germany and those of USA and Australia provides an impression that geographical compartmentalization of the MAH may have an important role to play in determining their occurrence. *M. avium* numbers in the United States, for example have also been correlated with zinc levels in natural water and high levels of humic and fulvic acids in soil. Other favorable conditions for growth of MAC are low pH, and low dissolved oxygen content [20]. It is also known that NTM are capable of surviving in dormant conditions triggered by stress, low oxygen and nutrient deprivation such as low organic carbon levels (e.g., drinking water) and exposure to acidic substrates (e.g., peat bogs and swamps. Although most studies of mycobacterial dormancy have focused on MTb and tuberculosis disease, it is time that dormancy is considered as a factor influencing the ecology of MAH [116].

The MAH transmission dynamics may also be different. In this situation, it is quite imperative to study and explore the MAH microbiome in a whole and comparative fashion. Apart from prevalence in soil, dust water and biofilms, the infectivity and transmission potential is another important area to explore. Genetic apparatus of MAH populations from different habitats could certainly be different. More epidemiological research is obligatory to ascertain the specific environmental reservoirs for these infections.

4.2 Environmental NTM in India

Infections due to NTM are more often diagnosed in developed countries like Europe and America. Though the post-AIDS era brought about tremendous upsurge in NTM research, most of these were reported from TB non endemic countries. Developed countries have included the identification of NTM to species level as a part of routine laboratory practice. However accurately identifying NTM and determining their clinical significance in TB endemic countries needs improvement. The diagnosis of NTM infections gets missed in these countries because of the high burden of TB infections. The difference in the environmental and climatic conditions could very well be contributing factors to the low detection of NTM from these countries.

MAH is a well-documented important opportunistic pathogen in Germany as NTM studies have identified *M. avium* as the most common causal agent of lymphadenitis in children in

Germany [108]. In India, less is known about the clinical NTM. Hence survey about the most prominent NTM in Indian environment was necessary.

Also multi-country NTM retrospective studies have been carried out in the past. One of the most interesting of them was the study by Martín-Casabona and colleagues in 2004 where patterns of NTM isolations were studied across 14 countries like those of Germany, Czech Republic, Switzerland, Spain, Portugal, Brazil, Belgium, Iran, Italy, Turkey, United Kingdom and many more [117]. Some of the countries were however not well represented in these studies because of lack of records or very meagre amounts of records which would neither reveal the true incidence of NTM disease in the total population nor the relationship between the isolation of NTM and disease. Since India remains under-represented in such studies, it was quite imperative that the NTM and the MAH scenario in India be explored and revealed.

India represents an endemic zone of mycobacterial infections. Although a lot of research and surveillance effort is directed at Tb, not much information on the existence, surveillance and transmission potential of NTM exists from this region. Region-specific NTM epidemiology studies have been performed [118-123] but the nationwide data is missing. Contrasting soil and climate types makes it difficult to perceive the precise picture of the geographical distribution of NTM within the Indian subcontinent. Four major capitals cities namely Hyderabad, Bhubaneswar, Kolkata and New Delhi and their surrounding areas were selected for sample collection points. While New Delhi is the capital of India, Hyderabad is the state capital of Andhra Pradesh, Kolkata of West Bengal and Bhubaneswar of Odisha. Hyderabad, Kolkata and New Delhi are amongst the largest cities in India. Bhubaneswar as a capital city was selected as a sampling point as it lies intermediary between the states of Andhra Pradesh and West Bengal. Having Hyderabad, Bhubaneswar and Kolkata as sampling points covers three-fourth of the capital cities on the Indian eastern coastal front by the Bay of Bengal.

Studies of MAH in these three capital cities provide a perspective of the NTM distributions in along the eastern coastal front of India. Also previous studies with regard to NTM prevalence have been performed in the neighboring states of Amritsar (Punjab), Karnataka (Bangalore district) [124] and that of Tamil Nadu (Pondicherry) [120], Maharashtra (Nagpur) [123] and Gujarat [122].

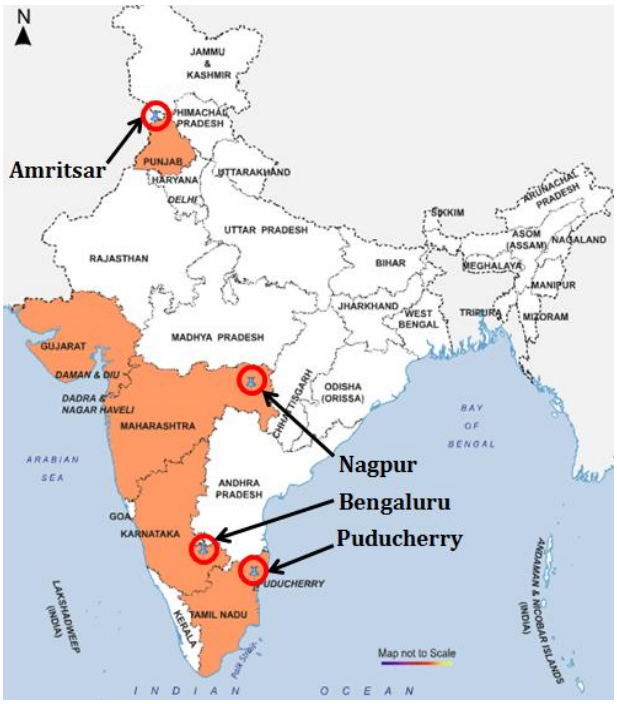


Figure 28: Map of India with regions having performed NTM epidemiology studies

The map of India indicates the regions where epidemiology studies with respect to the occurrence of NTM have been already performed. These regions are marked in orange color and involve studies across 5 different states or counties in India.

The samples collected in India were transported to Germany for further processing of NTM. Sample collections in India involved only dust and soil samples. Water and biofilms were not considered because of difficulties in transportation to Germany. Experiences with MAH isolations studies from German water reservoirs have proven that effective MAH isolation from water requires a minimum of 90 ml – 500 ml of water. Transportation of bulk volumes of water was difficult. Use of water filters was initially considered as an alternative but the time lapse between filtration of the water samples collected and processing of the samples, once transported to Germany was perceived to give ambiguous results. It was hence decided that that only soil and dust would be analyzed for NTM identification in India.

Challenges were faced during the collection of dust samples in India as carpeted floors are not a common occurrence in India. Hence vacuum cleaners are not used adequately. This is the underlying reason behind the low percentage of dust sample collection (16%) from India

in comparison to soil. Substrate accumulation on air filters of air conditioning and contents of vacuum cleaner bags were considered as dust.

Our studies showed that no MAH was identified in soil and dust from India. Due to an agrarian economy, direct contact with plant and animal materials and excreta could have been a potential factor in providing a saprophytic mycobacterial exposure to the Indian population for centuries. Less rigorous standards of food, water and community hygiene could be added to the reasoning along with BCG vaccination background that could make the Indian population somewhat 'immune' to the NTM infections. Cross immunity induced by TB should also be taken under consideration [125-127]. Immunological correlates of such adaptive advantage, if any, need to be worked out.

No MAH was identified in India. Either the ecology of MAH in India is different from the MAH ecology in Germany or the MAH strains from India were genetically distinct from those from Germany resulting in ineptness of our typing methods if applied to Indian MAH isolates. Since no MAH were identified in India, more weightage was given to identifying the commonly occurring NTM in India. High proportion of isolates from *M. fortuitum, M. terrae, M. asiaticum and M. intracellulare* were identified in our study. 17% of the isolates were *M. fortuitum, M. terrae complex, and M. asiaticum*. 12% of the isolates also contained *M. intracellulare* and strains belonging *to M. scrofulaceum and M. parascrofulaceum*. *M. terrae complex, M. asiaticum* and strains belonging *to M. scrofulaceum* and *M. parascrofulaceum* are the less commonly known infectious NTM. *M. fortuitum and M. intracellulare* are the more commonly occurring infectious NTMs and were identified from our studies from India.

Another interesting observation was the high proportion of *M. asiaticum* in the soil samples collected from New Delhi and around. These bacteria were reported for the very first time in monkeys from India that were exported to a research institute in Hungary [128, 129]. Several reports by other researchers like Blacklock, Taylor and Dawson have suggested these bacteria to cause pulmonary diseases, bursitis and flexor tenosynovitis [129-133]. Studies have also suggested tropical and sub-tropical climate to play an important role in the distribution of *M. asiaticum* [129]. There has however been no reporting about *M. asiaticum* in the environment or as a cause of pulmonary infections in India. The nature, extent and spectrum of diseases

caused by *M. asiaticum* in India remain unknown though they have been abundantly identified in our studies from New Delhi.

Is the high abundance of *M. fortuitum and M. intracellulare* in Indian soil and dust reflected in high rates of infections caused by *M. intracellulare* and *M. fortuitum* in India? This question cannot be answered at the moment since precise epidemiological data regarding the incidence of NTM in the whole of India remains largely unavailable due to lack of facilities and expertise. Recent cross sectional epidemiology studies from Pakistan which share similar biogeographical and climatic conditions with India reveal that *M. fortuitum, M. mucogenicum* and *M. smegmatis* are the most commonly isolated RGM and MAC are the most commonly occurring SGM in Pakistan [134]. However the epidemiology and in-depth analysis of subspecies of MAC members that cause disease in Pakistan remain ambiguous. Under such circumstances it is not exactly clear whether *M. avium* or *M. intracellulare* is actively involved as a causative agent of NTM disease. Studies from species level identification in South-Indian BCG trial areas have suggested *M. intracellulare* (22.6% of all NTM), *M. terrae* (12.5%) and *M. scrofulaceum* (10.5%) as the most frequently isolated NTM from sputum specimens [135]. These results from clinical sputum samples match the environmental profiles of NTM identified from our studies in India. However more recent epidemiological data is missing. The study from Shenai and colleagues on clinical NTM in India has identified *M. fortuitum* (41%) and *M. abscessus* (59%) as the predominant RGM. SGM in clinical specimens predominantly covered 40% *M. intracellulare*, followed by *M. simiae* (35%), *M. kansasii* (6%), *M. gordonae* (4%) *M. szulgai* (2%) and *M. avium* (1%) [136]. Studies from Khatter and colleagues [119] have also shown that 57% of the samples isolated from HIV seropositive individuals were NTM. These studies point towards the impression that regardless of the high endemicity of tuberculosis in India, the presence of NTM cannot be ruled out, especially in HIV seropositive individuals. NTM studies by Chauhan in Bangalore in Karnataka have identified *M. phlei* (17%) *M.gordonae* (13%), *M. scrofuleceum* (3%), MAC (3%), *M. fortuitum* (3%) and *M. xenopi* (3%) as the commonly found NTM in sputum specimens [124]. Recent reports in 2012 have described pulmonary infections due to *M. massilienese* for the very first time in India. [137]

TB infections have an overwhelming clinical importance in India and NTM are not in the focus of diagnosis and treatment. Clinical isolations of NTM are reported rare in India and

their role as etiological agents has hardly been studied [120]. It is not clear whether there is a genuine increase in NTM prevalence or the availability of more sensitive laboratory isolations and identification techniques have resulted in their upsurge.

BCG vaccination may play an important role in the epidemiology of NTM disease in India as BCG had an estimated coverage of 99% in 2011 according to WHO statistics [138]. It was demonstrated that BCG vaccination confers immunity to leprosy because of the cross-immunity of mycobacterial antigens [125]. Several studies have proposed a relationship between declining rates of BCG vaccination, decreased exposition to TB cases and increasing numbers of NTM cases [139-141]. This was supported by a rise in incidence in NTM disease observed in countries where BCG vaccination ceased (e.g. Sweden, Netherlands). However, an increasing trend was also observed in countries where BCG vaccination has never been implemented wide-scale (e.g. Canada) [142]. Several hypothesis have also suggested that the lack of protective efficacy of BCG vaccination in certain rural parts of South India has been attributed to the role of NTM [143] but no definitive proofs exist.

These organisms are treated as contaminants, colonization or indolent infections most of the time and hence discarded [136]. Very often, the typing of clinical isolates is not extended to the species level, meaning that *M. avium* and *M. intracellulare* are collectively referred to as MAC as is also shown by the NTM studies in Pakistan [27, 119, 134, 142]. It is plausible that the mycobacterial repertoire of the soil and dust could well be distinct from the profiles observed in the temperate west due to its typical geo-climatic conditions. The geographical differences and the different environmental stress conditions may be guiding factors for the prevalence of different NTM species in India.

4.3 Sequencing of two MAH genomes

Evolving sequencing technologies generate a large number of short reads of length 30 – 200 bp without much costs involved. However the error rates are relatively high (0.5%–1.0% error per raw base) [144] and low coverage assemblies have lower contiguity. These cause problems with identifying rearrangements, duplications, and repetitive elements, the resultant of which are miscalled bases and erroneous insertions and deletions [145, 146] . Standard genotype-calling algorithms depend on redundant sequencing of each base to discriminate between sequencing errors and true polymorphisms [91, 93]. Hence deep sequencing is

necessary to ensure polymorphisms and variations and facilitate better comparative genomics.

Eight MAH genomes were initially selected for sequencing by the Roche 454 sequencing platform. Though the length of the reads generated by Roche platform was large, the coverage of the reads was poor as is elucidated in the Table 14. Since mycobacterial genomes have a large number of repetitive units, deciphering the MAH genome would have been reasonably difficult. Two MAH strains from dust and child suffering from lymphadenitis were thereafter selected for further sequencing with Illumina and Ion torrent platform. The Illumina sequencing platform generated small high quality reads and the Ion torrent platform generated comparatively larger reads. Since Ion torrent reads have higher error rates in comparison to Illumina reads, the ion torrent reads were mapped to Illumina reads while building the hybrid assemblies. These helped with diminishing assembly errors. The low quality bases were trimmed at the start of the assembly to ensure that the coverage of the contigs obtained was high. The final results were screened to ensure that no gaps existed and the sizes of contigs were greater than 200 bp. These contigs were submitted to the NCBI as draft genomes.

MAH 104 had been available in the genome databases since 1980s. It was only recently in 2013 that five new MAH genomes (one complete and four draft genomes) were introduced into the databases. These sudden incorporations can be attributed to the increased MAH infections in various parts of the world. The complete genome from MAH TH135 isolated from a HIV negative patient form Japan was introduced in May 2013 to aid better comparative analysis of clinical HIV and non HIV isolates of MAH. Four MAH draft genomes have been available in the WGS format since December 2013. These are the MAH 10-5606 and MAH 10-4249. MAH 10-5606 is a MAH isolate from pig whereas MAH 10-4249 is an isolate from a deer in Colorado, Unites States. The other two draft genomes available in the database belong to this study and are MAH isolates from dust and child suffering from lymphadenitis namely the MAH 27-1 and MAH 2721. While the genome from dust is the first of its kind in the genome banks for being an environmental MAH strain, the MAH genome from a child suffering from lymphadenitis adds additional data towards clinical MAH genomes. It is already clear from the studies of MAH in Germany and India that environment has a strong role to play in shaping MAH disease dynamics. Hence it

becomes increasingly necessary to study the environmental MAH strains as understanding the differences between the clinical MAH strains and related less pathogenic environmental MAH strains is expected to reveal key bacterial virulence mechanisms and provide opportunities to understand host resistance to MAH infections better.

However more refined genomic studies with MAH isolates from soil, water, biofilms is needed to provide a new outlook on the macro and micro evolution of the MAH. This would also help us understand the evolutionary forces that shape the differences between these bacteria [11].

4.4 Identification of a genome island in MAH 104 which represents a zone of diversity

Comparative analysis of genomes from different bacteria has helped with facilitating the understanding of bacterial evolution. Most bacterial pathogens have a stable core genome that encodes factors essential for growth, survival and adaptation in specific hosts and different environments and a flexible gene pool, that encodes virulence traits, resistance determining factors and genes that confer mobility such as transposons, and insertion sequences [147]. The flexible genome often contains bacteriophages, IS-elements, plasmids, and transposons that contributes to the virulence of a particular organism [148]. The success of a bacterium is critically dependent on its variability which is determined by the flexible gene pool. It has hence become increasingly important to study the impact of genome variability on the evolution of a bacteria and its virulence. Several inter-species LSP studies have been performed by Behr and colleagues to identify host specific variants of MAC [89] but the complete plethora of LSPs in MAC have not been explored completely.

Seven regions of diversity were identified in this study by whole genome comparative analysis of MAH 104, MAP-*K10* and MAA ATCC 25291. These results were analogous to a study performed by Uchiya and colleagues in Japan who performed a similar study with MAH 104 and their newly sequenced genome from HIV negative MAH strain, the MAH *TH135* [90]. Eleven regions specific to the MAH 104 were identified in their comparative analysis with MAH TH135 and seven of the eleven regions were identical to the regions of specificity identified by this study. The regions SR11, SR12, SR14, SR15, SR16, SR19 and SR20 identified in the study by Uchiya and colleagues are similar to the regions 1, 2, 3, 4, 5,

6 and 7 as shown in this study in Table 21. Region 3 from this study was given specific importance as this region was identified as an island of diversity in this study.

The 47 Kb island identified from region 3 in the reference strain MAH 104 encompasses 63 genes starting from MAV_0779 to MAV_0841 and shares 53% homology with the genome of *M. xenopi* RIVM700367. Other MAC strains have shown higher homology to this region but however are not typed to the species level. Comparative analysis with the MAH TH135 revealed that the 47 Kb insert was absent in the TH135 strain. Instead the region has been replaced by a smaller island of size 1.71 Kb comprising of two hypothetical genes namely the MAH_0684 and MAH_0685. Parallel to these results, the region of diversity was analyzed in the two genomes sequenced in this study the MAH 27-1 from dust and the MAH 2721, isolated from a child suffering from lymphadenitis. The results revealed a completely different 14.2 Kb island in MAH 27-1 which is represented in the form of contigs 32 and 38 in the MAH 27-1 WGS draft genome. The 14.2 Kb island shares 99% homology with a *M. intracellulare* isolate, the MOTT-36Y. The MOTT 36Y is a clinical *M. intracellulare* strain identified from a clinical subject in Korea [149, 150]. This strain shares a unique hsp65 genotype specific to Korean patients but *M. intracellulare* identical ITS1 and 16S r-RNA genotype. The island identified in the MAH 27-1 shares homology with 14 genes in *M. intracellulare* MOTT 36Y ranging from *W7S_03350* to *W7S_3410* and a t-RNA.

The MAH 2721 revealed an island of atleast 31.5 Kb. The island was found in two contigs namely the contig 338 and the contig 341 in the MAH 2721 WGS draft genome. However the connecting link between the two contigs was not identified by the reassembly of raw reads obtained from Illumina and Ion torrent data. Hence the actual size of the island cannot be predicted. Long range PCR kits are currently being optimized to cover the gap existing between the two above mentioned contigs. The size of the island in actuality remains unknown. Simultaneous to the same the identification of 2 more islands in the deer isolate and the pig isolate depict an example of a flexible gene pool.

The regions between the O' methoxytransferase and carveol dehydrogenase clearly show a region of diversity in different MAH strains. Only six MAH strains have been looked into and identification of additional island in various MAH strains could better define the implications of this diversity. Functional studies to identify the relevance of such regions

within the MAH genomes are missing. Insertional inactivation, mutagenesis and knock out may be considered as strategies for functional studies related to the island.

Genomic studies and lessons from comparative genomics of *M. tuberculosis* present a portrait of reductive genomics but the pattern of reductive genomics may not always be applicable to other mycobacteria. It is thought that majority of mycobacteria demonstrate a general capacity to acquire genetic material due to the presence of selective pressures [151]. Despite the large amount of genomic distinctions revealed for most gene deletions and acquisitions, the phenotypic consequence are yet to be established [152]. It is clear from the above study that more MAH genome sequences would be required to acquire an understanding of these bacteria.

4.5 SNP analysis of three MAH genes and their implications towards identifying plausible routes of infections.

The advent of next-generation DNA sequencing has facilitated the ease of numerous genome sequencing projects of multiple isolates from a same species or subspecies. SNPs have proven to be powerful tools for inferring phylogenies, strain classification and measuring evolutionary distances between strains. An example of such a study was the one performed by Dominguez and colleagues where SNPs in insertion element IS901 was used for diagnostic and epidemiological purposes in MAP strains [153]. Several SNP typing schemes have been developed in the recent years due to growing number of genomes.

The global phylogenetic diversity of MAH remains unknown as new MAH genomes are only recently being sequenced and introduced into databases. There is increasing evidence that both environmental factors and host genetics play an important part in strain variation of both clinical and environmental MAH strains and these variations play an important role in the outcome of MAH infections and disease. Hence, there is a need to better understand the global diversity of MAH, and determine if and how this diversity has relevance to disease [154].

Since eight genomes from different niches (both environmental and patient specific) were sequenced during this study, whole genome SNP analysis was performed with eight MAH genomes. The SNP studies were performed to investigate the presence of SNPs that could

define environmental strains from patient strains and also define the route of infections in clinical subjects. Gene wise SNP analysis was performed and several genes with interesting SNPs were screened. The genes were selected based on SNPs showing relevant similarities between MAH isolates from dust and MAH isolates from child suffering from lymphadenitis. These genes were further checked for virulence associations in MAH. Three genes namely the MAV_0846 (Carveol dehydrogenase), MAV_0847 (TetR family transcriptional regulator), and MAV_0853 (ddpX; D-alanyl-D-alanine dipeptidase) were selected for further analysis. However since the coverage of the Roche 454 sequencing data was low, SNPs identified in the three genes were reconfirmed by additional sequencing of 36 MAH strains isolated or collected during the course of this study.

Nor region specific similarities neither possible route of infections could be inferred from the SNP analysis results of the three genes. Since no environmental or clinical subject specific SNPs were identified from this study, it can be hypothesized that the selection of genes could not show the specific route of MAH infection. The occurrence of MAH is highly diverse as is evident from our studies of MAH environmental reservoirs in section 4.1 and the concentration of MAH and the amount of MAH exposure has an important role to play in the progression of MAH infection and disease. The host specific factors like immunity, underlying preexisting lung conditions can also not be overruled. Additionally, different countries exhibit different vaccination strategies: in Germany, BCG vaccination was no longer recommended since 1998, while in India BCG had an estimated coverage of 99% in 2011 according to WHO statistics. Different BCG vaccination strategies also confer varied levels of cross-immunity of mycobacterial antigens and hence may also offer protective advantages to many hosts. It is unclear at this stage whether MAH infections have more specific environmental relevance but better comparative genomics aided by more MAH genome sequencing would help us unravel this opportunistic mycobacterial pathogen.

5. Summary

Non tuberculosis mycobacteria (NTM) are ubiquitous opportunistic environmental bacteria. Commonality of infections due to NTM in the developed countries like Germany has attained significant importance. Little is known about NTM prevalence in developing countries like India.

Since *Mycobacterium avium* subsp. *hominissuis* (MAH) was identified as the most predominant NTM in clinical samples in Germany, the ecological niche of this bacterium in Germany was explored in this study and it was concluded that soil and dust may be relevant sources for MAH infections in Germany. No MAH was identified in soil and dust samples from India. Instead a plethora of different NTM residing in Indian soils (e.g. *M. terrae, M. fortuitum* and *M. asiaticum*) was identified. The study revealed that the biogeography of a place has an important role in defining the habitat and occurrence of NTM.

Two draft MAH genomes from isolates from dust and from a child suffering from lymphadenitis were introduced in the public databases to facilitate better comparative genomics and understand the underlying molecular mechanisms responsible for this environmental bacterium to evolve as an opportunistic pathogen.

The study also proposed a genome island in MAH which represents a highly dynamic zone. This identified dynamic region points to the opinion that the genomes of MAH are open to DNA rearrangements and the flexible gene pools have an imperative role to play in offering adaptive advantages to these microbes.

The habitat of the bacteria is one important factor adding up to the list of several others like infectivity, transmission potential of the bacteria and intensity of MAH exposures that could be responsible for MAH infections. SNP analysis in three MAH genes revealed no correlation between the SNPs from MAH in different environmental and clinical habitats.

6. Zusammenfassung

Nicht-tuberkulöse Mykobakterien (NTM) sind ubiquitäre, opportunistische Umweltbakterien. In den Industriestaaten wie beispielsweise in Deutschland haben die Infektionsraten von NTM eine signifikante Bedeutung erreicht. Über die NTM-Prävalenzen in Entwicklungs- und Schwellenländern wie beispielsweise Indien ist wenig bekannt.

Da *Mycobacterium avium* subsp. *hominissuis* (MAH) als das häufigste aus klinischen Proben isolierte NTM in Deutschland identifiziert wurde, wurde in der vorliegenden Arbeit die ökologische Nische dieses Bakteriums in Deutschland ermittelt und Erde und Staub als mögliche relevante Quellen für MAH Infektionen identifiziert. MAH konnte nicht aus Erd- und Staubproben aus Indien isoliert werden. Stattdessen wurde eine Fülle verschiedener NTM (z.B. *M. terrae, M. fortuitum* and *M. asiaticum*) in Erdproben aus Indien gefunden. Diese Studie konnte aufzeigen, dass die Biogeographie eines Ortes sowohl das Habitat als auch das Vorkommen von NTM entscheidend definiert.

Zwei MAH draft Genome, von einem Umweltisolat aus Staub und einem Patientenisolat eines Kindes mit Lymphadenitis, wurden den öffentlichen Datenbanken hinzugefügt, um eine bessere vergleichende Genomanalyse zu gewährleisten. Dies ermöglicht ein besseres Verständnis der molekularen Mechanismen, die für die Evolution eines Umweltkeines zu einem opportunistischen Krankheitserreger verantwortlich sind.

Des Weiteren wurde eine Genominsel in MAH identifiziert, die sich als stark dynamische Sequenzregion darstellt. Diese dynamische Region deutet darauf hin, dass das MAH Genom Reorganisationen unterliegt und dass diese flexiblen Genpools eine wichtige Rolle bei der Anpassung der Bakterien an Umweltveränderungen spielen.

Das Habitat der Bakterien ist als wichtiger Faktor neben anderen Faktoren wie Infektiosität, Übertragungspotential und Expositionsstärke mitverantwortlich für eine MAH Infektion. Untersuchungen von drei MAH Genen mittels SNP-Analyse zeigten keine Korrelation zwischen den SNPs von MAH und dem ökologischen oder klinischen Habitat.

7. References

1. Veyrier FJ, Dufort A, Behr MA: **The rise and fall of the Mycobacterium tuberculosis genome**. *Trends Microbiol* 2011, **19**(4):156-161.

2. Lalvani A, Behr MA, Sridhar S: **Innate immunity to TB: A druggable balancing act**. *Cell* 2012, **148**(3):389-391.

3. Ahmed N, Dobrindt U, Hacker J, Hasnain SE: **Genomic fluidity and pathogenic bacteria: Applications in diagnostics, epidemiology and intervention**. *Nature Reviews Microbiology* 2008, **6**(5):387-394.

4. Veyrier F, Pletzer D, Turenne C, Behr MA: **Phylogenetic detection of horizontal gene transfer during the step-wise genesis of Mycobacterium tuberculosis**. *BMC Evol Biol* 2009, **9**(1).

5. Levy-Frebault VV, Portaels F: **Proposed minimal standards for the genus Mycobacterium and for description of new slowly growing Mycobacterium species**. *Int J Syst Bacteriol* 1992, **42**(2):315-323.

6. Euzéby JP: **List of bacterial names with standing in nomenclature: A folder available on the internet**. *Int J Syst Bacteriol* 1997, **47**(2):590-592.

7. Jang J, Becq J, Gicquel B, Deschavanne P, Neyrolles O: **Horizontally acquired genomic islands in the tubercle bacilli**. *Trends Microbiol* 2008, **16**(7):303-308.

8. WHO: **Global tuberculosis report**. In: *WHO Tech Rep Ser.* 2013.

9. WHO: **Global Tuberculosis Report**. *WHO Tech Rep Ser* 2012.

10. Lienhardt C, Ogden JA: **Tuberculosis control in resource-poor countries: Have we reached the limits of the universal paradigm?** *Trop Med Int Health* 2004, **9**(7):833-841.

11. Behr MA: **Evolution of Mycobacterium tuberculosis**. *Adv Exp Med Biol* 2013, **783**:81-91.

12. Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, Musser JM: **Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination**. *Proceedings of the National Academy of Sciences of the United States of America* 1997, **94**(18):9869-9874.

13. Smith NH, Hewinson RG, Kremer K, Brosch R, Gordon SV: **Myths and misconceptions: The origin and evolution of Mycobacterium tuberculosis**. *Nature Reviews Microbiology* 2009, **7**(7):537-544.

14. Stinear TP, Seemann T, Harrison PF, Jenkin GA, Davies JK, Johnson PDR, Abdellah Z, Arrowsmith C, Chillingworth T, Churcher C *et al*: **Insights from the complete genome**

sequence of Mycobacterium marinum on the evolution of Mycobacterium tuberculosis. *Genome Res* 2008, **18**(5):729-741.

15. Wolinsky E: **Nontuberculous mycobacteria and associated diseases**. *American Review of Respiratory Disease* 1979, **119**(1):107-159.

16. Primm TP, Lucero CA, Falkinham Iii JO: **Health Impacts of Environmental Mycobacteria**. *Clin Microbiol Rev* 2004, **17**(1):98-106.

17. M Cristina G, Brisse S, Brosch R, Fabre M, Omaïs B, Marmiesse M, Supply P, Vincent V: **Ancient origin and gene mosaicism of the progenitor of Mycobacterium tuberculosis**. *PLoS Path* 2005, **1**(1):0055-0061.

18. Gutierrez MC, Supply P, Brosch R: **Pathogenomics of mycobacteria**. In., vol. 6; 2009: 198-210.

19. Arnold C: **Molecular evolution of Mycobacterium tuberculosis**. *Clin Microbiol Infect* 2007, **13**(2):120-128.

20. Falkinham III JO: **Epidemiology of infection by nontuberculous mycobacteria**. *Clin Microbiol Rev* 1996, **9**(2):177-215.

21. Khan K, Wang J, Marras TK: **Nontuberculous mycobacterial sensitization in the United States: National trends over three decades**. *Am J Respir Crit Care Med* 2007, **176**(3):306-313.

22. Horsburgh Jr CR: **Epidemiology of mycobacterial diseases in AIDS**. *Res Microbiol* 1992, **143**(4):372-377.

23. Horsburgh Jr CR, Selik RM: **The epidemiology of disseminated nontuberculous mycobaterial infection in the acquired immunodeficiency syndrome (AIDS)**. *American Review of Respiratory Disease* 1989, **139**(1):4-7.

24. Marras TK, Daley CL: **Epidemiology of human pulmonary infection with nontuberculous mycobacteria**. *Clinics in Chest Medicine* 2002, **23**(3):553-567.

25. Prevots DR, Shaw PA, Strickland D, Jackson LA, Raebel MA, Blosky MA, De Oca RM, Shea YR, Seitz AE, Holland SM *et al*: **Nontuberculous mycobacterial lung disease prevalence at four integrated health care delivery systems**. *Am J Respir Crit Care Med* 2010, **182**(7):970-976.

26. Grange JM, Yates MD, Pozniak A: **Bacteriologically confirmed non-tuberculous mycobacterial lymphadenitis in south east England: A recent increase in the number of cases**. *Arch Dis Child* 1995, **72**(6):516-517.

27. Gopinath K, Singh S: **Non-Tuberculous mycobacteria in TB-endemic countries: Are we neglecting the danger?** *PLoS Neglected Tropical Diseases* 2010, **4**(4).

28. Merle CS, Cunha SS, Rodrigues LC: **BCG vaccination and leprosy protection: Review of current evidence and status of BCG in leprosy control**. *Expert Review of Vaccines* 2010, **9**(2):209-222.

29. Whittington RJ, Marshall DJ, Nicholls PJ, Marsh IB, Reddacliff LA: **Survival and dormancy of Mycobacterium avium subsp. paratuberculosis in the environment**. *Appl Environ Microbiol* 2004, **70**(5):2989-3004.

30. Falkinham Iii JO, Norton CD, Lechevallier MW: **Factors Influencing Numbers of Mycobacterium avium, Mycobacterium intracellulare, and Other Mycobacteria in Drinking Water Distribution Systems**. *Appl Environ Microbiol* 2001, **67**(3):1225-1231.

31. Le Dantec C, Duguet JP, Montiel A, Dumoutier N, Dubrou S, Vincent V: **Occurrence of mycobacteria in water treatment lines and in water distribution systems**. *Appl Environ Microbiol* 2002, **68**(11):5318-5325.

32. Dailloux M, Albert M, Laurain C, Andolfatto S, Lozniewski A, Hartemann P, Mathieu L: **Mycobacterium xenopi and Drinking Water Biofilms**. *Appl Environ Microbiol* 2003, **69**(11):6946-6948.

33. Ojha A, Anand M, Bhatt A, Kremer L, Jacobs Jr WR, Hatfull GF: **GroEL1: A dedicated chaperone involved in mycolic acid biosynthesis during biofilm formation in mycobacteria**. *Cell* 2005, **123**(5):861-873.

34. Stewart PS, Costerton JW: **Antibiotic resistance of bacteria in biofilms**. *Lancet* 2001, **358**(9276):135-138.

35. Taylor RH, Falkinham Iii JO, Norton CD, LeChevallier MW: **Chlorine, chloramine, chlorine dioxide, and ozone susceptibility of Mycobacterium avium**. *Appl Environ Microbiol* 2000, **66**(4):1702-1705.

36. Rosenblueth M, Martinez-Romero JC, Reyes-Prieto M, Rogel MA, Martinez-Romero E: **Environmental mycobacteria: A threat to human health?** *DNA Cell Biol* 2011, **30**(9):633-640.

37. Runyon EH: **Typical Mycobacteria: Their Classification**. *The American review of respiratory disease* 1965, **91**:288-289.

38. Kim CJ, Kim NH, Song KH, Choe PG, Kim ES, Park SW, Kim HB, Kim NJ, Kim EC, Park WB *et al*: **Differentiating rapid- and slow-growing mycobacteria by difference in time to growth detection in liquid media**. *Diagn Microbiol Infect Dis* 2013, **75**(1):73-76.

39. Hoefsloot W, Van Ingen J, Andrejak C, Ängeby K, Bauriaud R, Bemer P, Beylis N, Boeree MJ, Cacho J, Chihota V *et al*: **The geographic diversity of nontuberculous mycobacteria isolated from pulmonary samples: An NTM-NET collaborative study**. *Eur Respir J* 2013, **42**(6):1604-1613.

40. Thorel MF, Krichevsky M, Levy-Frebault VV: **Numerical taxonomy of mycobactin-dependent mycobacteria, emended description of Mycobacterium avium, and description of Mycobacterium avium subsp. avium subsp. nov., Mycobacterium avium subsp. paratuberculosis subsp. nov., and Mycobacterium avium subsp. silvaticum subsp. nov**. *Int J Syst Bacteriol* 1990, **40**(3):254-260.

41. Turenne CY, Collins DM, Alexander DC, Behr MA: **Mycobacterium avium subsp. paratuberculosis and M. avium subsp. avium are independently evolved pathogenic clones of a much broader group of M. avium organisms**. *J Bacteriol* 2008, **190**(7):2479-2487.

42. Dvorska L, Bull TJ, Bartos M, Matlova L, Svastova P, Weston RT, Kintr J, Parmova I, Van Soolingen D, Pavlik I: **A standardised restriction fragment length polymorphism (RFLP) method for typing Mycobacterium avium isolates links IS901 with virulence for birds**. *J Microbiol Methods* 2003, **55**(1):11-27.

43. Feller M, Huwiler K, Stephan R, Altpeter E, Shang A, Furrer H, Pfyffer GE, Jemmi T, Baumgartner A, Egger M: **Mycobacterium avium subspecies paratuberculosis and Crohn's disease: a systematic review and meta-analysis**. *Lancet Infect Dis* 2007, **7**(9):607-613.

44. Ignatov D, Kondratieva E, Azhikina T, Apt A: **Mycobacterium avium-triggered diseases: Pathogenomics**. *Cell Microbiol* 2012, **14**(6):808-818.

45. Matveychuk A, Fuks L, Priess R, Hahim I, Shitrit D: **Clinical and radiological features of Mycobacterium kansasii and other NTM infections**. *Respir Med* 2012, **106**(10):1472-1477.

46. Kaustová J, Chemlík M, Ettlová D, Hudec V, Lazarcvá H, Richtrová S: **Disease due to Mycobacterium kansasii in the Czech Republic: 1984-89**. *Tubercle Lung Dis* 1995, **76**(3):205-209.

47. Lortholary O, Deniel F, Boudon P, Le Pennec MP, Mathieu M, Soilleux M, Le Pendeven C, Loiseau P, Vincent V, Valeyre D *et al*: **Mycobacterium kansasii infection in a Paris suburb: Comparison of disease presentation and outcome according to human immunodeficiency virus status**. *Int J Tuberc Lung Dis* 1999, **3**(1):68-73.

48. Asija A, Prasad A, Eskridge E: **Disseminated mycobacterium gordonae infection in an immunocompetent host**. *American Journal of Therapeutics* 2011, **18**(3):e75-e77.

49. Weinberger M, Berg SL, Feuerstein IM, Pizzo PA, Witebsky FG: **Disseminated infection with Mycobacterium gordonae: Report of a case and critical review of the literature**. *Clin Infect Dis* 1992, **14**(6):1229-1239.

50. Marx CE, Fan K, Morris AJ, Wilson ML, Damiani A, Weinstein MP: **Laboratory and clinical evaluation of Mycobacterium xenopi isolates**. *Diagn Microbiol Infect Dis* 1995, **21**(4):195-202.

51. Abdallah AM, Rashid M, Adroub SA, Elabdalaoui H, Ali S, van Soolingen D, Bitter W, Pain A: **Complete genome sequence of Mycobacterium xenopi type strain RIVM700367**. *J Bacteriol* 2012, **194**(12):3282-3283.

52. Slosarek M, Kubin M, Jaresova M: **Water-borne household infections due to Mycobacterium xenopi**. *Central European Journal of Public Health* 1993, **1**(2):78-80.

53. Contreras MA, Cheung OT, Sanders DE, Goldstein RS: **Pulmonary infection with nontuberculous mycobacteria**. *American Review of Respiratory Disease* 1988, **137**(1):149-152.

54. Wallace Jr RJ: **Recent changes in taxonomy and disease manifestations of the rapidly growing mycobacteria**. *European Journal of Clinical Microbiology and Infectious Diseases* 1994, **13**(11):953-960.

55. Griffith DE, Girard WM, Wallace Jr RJ: **Clinical features of pulmonary disease caused by rapidly growing mycobacteria: An analysis of 154 patients**. *American Review of Respiratory Disease* 1993, **147**(5):1271-1278.

56. Fitzgerald DA, Smith AG, Lees A, Yee L, Cooper N, Harris SC, Gibson JA: **Cutaneous infection with Mycobacterium abscessus**. *Br J Dermatol* 1995, **132**(5):800-804.

57. Repath F, Seabury JH, Sanders CV, Domer J: **Prosthetic valve endocarditis due to Mycobacterium chelonei**. *Southern Medical Journal* 1976, **69**(9):1244-1246.

58. Ingram CW, Tanner DC, Durack DT, Kernodle Jr GW, Corey GR: **Disseminated infection with rapidly growing mycobacteria**. *Clin Infect Dis* 1993, **16**(4):463-471.

59. Kirschner P, Springer B, Vogel U, Meier A, Wrede A, Kiekenbeck M, Bange FC, Bottger EC: **Genotypic identification of mycobacteria by nucleic acid sequence determination: Report of a 2-year experience in a clinical laboratory**. *J Clin Microbiol* 1993, **31**(11):2882-2889.

60. Patel JB, Leonard DGB, Pan X, Musser JM, Berman RE, Nachamkin I: **Sequence-based identification of Mycobacterium species using the MicroSeq 500 16S r-RNA bacterial identification system**. *J Clin Microbiol* 2000, **38**(1):246-251.

61. Pontiroli A, Khera TT, Oakley BB, Mason S, Dowd SE, Travis ER, Erenso G, Aseffa A, Courtenay O, Wellington EMH: **Prospecting Environmental Mycobacteria: Combined Molecular Approaches Reveal Unprecedented Diversity**. *PLoS ONE* 2013, **8**(7).

62. Springer B, Stockman L, Teschner K, Roberts GD, Bottger EC: **Two-laboratory collaborative study on identification of mycobacteria: Molecular versus phenotypic methods**. *J Clin Microbiol* 1996, **34**(2):296-303.

63. Hughes MS, Skuce RA, Beck LA, Neill SD: **Identification of mycobacteria from animals by restriction enzyme analysis and direct DNA cycle sequencing of polymerase chain reaction-amplified 16S r-RNA gene sequences**. *J Clin Microbiol* 1993, **31**(12):3216-3222.

64. Devallois A, Khye Seng G, Rastogi N: **Rapid identification of mycobacteria to species level by PCR-restriction fragment length polymorphism analysis of the hsp65 gene and proposition of an algorithm to differentiate 34 mycobacterial species**. *J Clin Microbiol* 1997, **35**(11):2969-2973.

65. Fiss EH, Chehab FF, Brooks GF: **DNA amplification and reverse dot blot hybridization for detection and identification of mycobacteria to the species level in the clinical laboratory**. *J Clin Microbiol* 1992, **30**(5):1220-1224.

66. Telenti A, Marchesi F, Balz M, Bally F, Bottger EC, Bodmer T: **Rapid identification of mycobacteria to the species level by polymerase chain reaction and restriction enzyme analysis**. *J Clin Microbiol* 1993, **31**(2):175-178.

67. Katoch VM, Parashar D, Chauhan DS, Singh D, Sharma VD, Ghosh S: **Rapid identification of mycobacteria by gene amplification restriction analysis technique targeting 16S-23S ribosomal RNA internal transcribed spacer & flanking region**. *Indian Journal of Medical Research* 2007, **125**(2):155-162.

68. Roth A, Fischer M, Hamid ME, Michalke S, Ludwig W, Mauch H: **Differentiation of phylogenetically related slowly growing mycobacteria based on 16S-23S rRNA gene internal transcribed spacer sequences**. *J Clin Microbiol* 1998, **36**(1):139-147.

69. Turenne CY, Tschetter L, Wolfe J, Kabani A: **Necessity of quality-controlled 16S r-RNA gene sequence databases: Identifying nontuberculous Mycobacterium species**. *J Clin Microbiol* 2001, **39**(10):3637-3648.

70. Kim H, Kim SH, Shim TS, Kim MN, Bai GH, Park YG, Lee SH, Chae GT, Cha CY, Kook YH *et al*: **Differentiation of Myocobacterium species by analysis of the heat-shock protein 65 gene (hsp65)**. *Int J Syst Evol Microbiol* 2005, **55**(4):1649-1656.

71. Kwok AYC: **Species identification and phylogenetic relationships based on partial HSP60 gene sequences within the genus Staphylococcus**. *Int J Syst Bacteriol* 1999, **49**(3):1181-1192.

72. Gürtler V, Stanisich VA: **New approaches to typing and identification of bacteria using the 16S-23S rDNA spacer region**. *Microbiology* 1996, **142**(1):3-16.

73. De Smet KAL, Brown IN, Yates M, Ivanyi J: **Ribosomal internal transcribed spacer sequences are identical among Mycobacterium avium-intracellulare complex isolates from AIDS patients, but vary among isolates from elderly pulmonary disease patients**. *Microbiology* 1995, **141**(10):2739-2747.

74. Turenne CY, Wallace Jr R, Behr MA: **Mycobacterium avium in the Postgenomic Era**. *Clin Microbiol Rev* 2007, **20**(2):205-229.

75. Bellamy R, Sangeetha S, Paton NI: **Causes of death among patients with HIV in Singapore from 1985 to 2001: Results from the Singapore HIV Observational Cohort Study (SHOCS)**. *HIV Med* 2004, **5**(4):289-295.

76. Falkinham III JO, Iseman MD, de Haas P, van Soolingen D: **Mycobacterium avium in a shower linked to pulmonary disease**. *J Water Health* 2008, **6**(2):209-213.

77. Glover N, Holtzman A, Aronson T, Froman S, Berlin OGW, Dominguez P, Kunkel KA, Overturf G, Stelma Jr G, Smith C *et al*: **The isolation and identification of Mycobacterium avium complex (MAC) recovered from Los Angeles poTable water, a possible source of infection in AIDS patients**. *Int J Environ Health Res* 1994, **4**(2):63-72.

78. Kim SY, Lee ST, Jeong BH, Jeon K, Kim JW, Shin SJ, Koh WJ: **Clinical significance of mycobacterial genotyping in Mycobacterium avium lung disease in Korea**. *Int J Tuberc Lung Dis* 2012, **16**(10):1393-1399.

79. Álvarez J, García IG, Aranaz A, Bezos J, Romero B, De Juan L, Mateos A, Gómez-Mampaso E, Domínguez L: **Genetic diversity of Mycobacterium avium isolates recovered from clinical samples and from the environment: Molecular characterization for diagnostic purposes**. *J Clin Microbiol* 2008, **46**(4):1246-1251.

80. De Groote MA, Pace NR, Fulton K, Falkinham Iii JO: **Relationships between Mycobacterium isolates from patients with pulmonary mycobacterial infection and potting soils**. *Appl Environ Microbiol* 2006, **72**(12):7602-7606.

81. Iwamoto T, Nakajima C, Nishiuchi Y, Kato T, Yoshida S, Nakanishi N, Tamaru A, Tamura Y, Suzuki Y, Nasu M: **Genetic diversity of Mycobacterium avium subsp. hominissuis strains isolated from humans, pigs, and human living environment**. *Infect, Genet Evol* (0).

82. Falkinham JO: **Impact of human activities on the ecology of nontuberculous mycobacteria**. *Future Microbiology* 2010, **5**(6):951-960.

83. Shin SJ, Lee BS, Koh WJ, Manning EJB, Anklam K, Sreevatsan S, Lambrecht RS, Collins MT: **Efficient differentiation of Mycobacterium avium complex species and subspecies by use of five-target multiplex PCR**. *J Clin Microbiol* 2010, **48**(11):4057-4062.

84. Moravkova M, Hlozek P, Beran V, Pavlik I, Preziuso S, Cuteri V, Bartos M: **Strategy for the detection and differentiation of Mycobacterium avium species in isolates and heavily infected tissues**. *Res Vet Sci* 2008, **85**(2):257-264.

85. Inderlied CB, Kemper CA, Bermudez LEM: **The Mycobacterium avium complex**. *Clin Microbiol Rev* 1993, **6**(3):266-310.

86. Despierres L, Cohen-Bacrie S, Richet H, Drancourt M: **Diversity of Mycobacterium avium subsp. hominissuis mycobacteria causing lymphadenitis, France**. *European Journal of Clinical Microbiology and Infectious Diseases* 2012, **31**(7):1373-1379.

87. Iwamoto T, Nakajima C, Nishiuchi Y, Kato T, Yoshida S, Nakanishi N, Tamaru A, Tamura Y, Suzuki Y, Nasu M: **Genetic diversity of Mycobacterium avium subsp.**

hominissuis strains isolated from humans, pigs, and human living environment. *Infect, Genet Evol* 2012, **12**(4):846-852.

88. Horan KL, Freeman R, Weigel K, Semret M, Pfaller S, Covert TC, Van Soolingen D, Leão SC, Behr MA, Cangelosi GA: **Isolation of the genome sequence strain Mycobacterium avium 104 from multiple patients over a 17-year period**. *J Clin Microbiol* 2006, **44**(3):783-789.

89. Semret M, Turenne CY, De Haas P, Collins DM, Behr MA: **Differentiating host-associated variants of Mycobacterium avium by PCR for detection of large sequence polymorphisms**. *J Clin Microbiol* 2006, **44**(3):881-887.

90. Uchiya KI, Takahashi H, Yagi T, Moriyama M, Inagaki T, Ichikawa K, Nakagawa T, Nikai T, Ogawa K: **Comparative Genome Analysis of Mycobacterium avium Revealed Genetic Diversity in Strains that Cause Pulmonary and Disseminated Disease**. *PLoS ONE* 2013, **8**(8).

91. Shendure J, Ji H: **Next-generation DNA sequencing**. *Nat Biotechnol* 2008, **26**(10):1135-1145.

92. Kim J, Lee S, Shin H, Kim SC, Cho BK: **Elucidation of bacterial genome complexity using next-generation sequencing**. *Biotechnology and Bioprocess Engineering* 2012, **17**(5):887-899.

93. Mardis ER: **Next-generation DNA sequencing methods**. In., vol. 9; 2008: 387-402.

94. Treangen TJ, Salzberg SL: **Repetitive DNA and next-generation sequencing: Computational challenges and solutions**. *Nature Reviews Genetics* 2012, **13**(1):36-46.

95. Croucher NJ, Harris SR, Grad YH, Hanage WP: **Bacterial genomes in epidemiology-present and future**. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2013, **368**(1614).

96. Dobrindt U, Hochhut B, Hentschel U, Hacker J: **Genomic islands in pathogenic and environmental microorganisms**. *Nature Reviews Microbiology* 2004, **2**(5):414-424.

97. Fernández-Gómez B, Fernàndez-Guerra A, Casamayor EO, González JM, Pedrós-Alió C, Acinas SG: **Patterns and architecture of genomic islands in marine bacteria**. *BMC Genomics* 2012, **13**(1).

98. Schmidt H, Hensel M: **Pathogenicity Islands in Bacterial Pathogenesis**. *Clin Microbiol Rev* 2004, **17**(1):14-56.

99. Danelishvili L, Wu M, Stang B, Harriff M, Cirillo S, Cirillo J, Bildfell R, Arbogast B, Bermudez LE: **Identification of Mycobacterium avium pathogenicity island important for macrophage and amoeba infection**. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(26):11038-11043.

100. Stratmann J, Strommenger B, Goethe R, Dohmann K, Gerlach GF, Stevenson K, Li LL, Zhang Q, Kapur V, Bull TJ: **A 38-Kilobase Pathogenicity Island Specific for Mycobacterium avium subsp. paratuberculosis Encodes Cell Surface Proteins Expressed in the Host**. *Infect Immun* 2004, **72**(3):1265-1274.

101. Varma-Basil M, Garima K, Pathak R, Dwivedi SKD, Narang A, Bhatnagar A, Bose M: **Development of a novel pcr restriction analysis of the hsp65 gene as a rapid method to screen for the mycobacterium tuberculosis complex and nontuberculous mycobacteria in high-burden countries**. *J Clin Microbiol* 2013, **51**(4):1165-1170.

102. Hall BG: **Building phylogenetic trees from molecular data with MEGA**. *Mol Biol Evol* 2013, **30**(5):1229-1235.

103. Edgar RC: **MUSCLE: A multiple sequence alignment method with reduced time and space complexity**. *BMC Bioinformatics* 2004, **5**.

104. Edgar RC: **MUSCLE: Multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**(5):1792-1797.

105. Falkinham III JO: **Nontuberculous mycobacteria from household plumbing of patients with nontuberculous mycobacteria disease**. *Emerging Infect Dis* 2011, **17**(3):419-424.

106. Preheim SP, Timberlake S, Polz MF: **Merging taxonomy with ecological population prediction in a case study of Vibrionaceae**. *Appl Environ Microbiol* 2011, **77**(20):7195-7206.

107. Staley JT: **The bacterial species dilemma and the genomic-phylogenetic species concept**. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2006, **361**(1475):1899-1909.

108. Reuss AM, Wiese-Posselt M, Weißmann B, Siedler A, Zuschneid I, An Der Heiden M, Claus H, Von Kries R, Haas WH: **Incidence rate of nontuberculous mycobacterial disease in immunocompetent children a prospective nationwide surveillance study in Germany**. *Pediatr Infect Dis J* 2009, **28**(7):642-644.

109. Van Ingen J, Hoefsloot W, Dekhuijzen PNR, Boeree MJ, Van Soolingen D: **The changing pattern of clinical Mycobacterium avium isolation in the Netherlands**. *Int J Tuberc Lung Dis* 2010, **14**(9):1176-1180.

110. Möbius P, Lentzsch P, Moser I, Naumann L, Martin G, Köhler H: **Comparative macrorestriction and RFLP analysis of Mycobacterium avium subsp. avium and Mycobacterium avium subsp. hominissuis isolates from man, pig, and cattle**. *Vet Microbiol* 2006, **117**(2-4):284-291.

111. Von Reyn CF, Maslow JN, Barber TW, Falkinham Iii JO, Arbeit RD: **Persistent colonisation of poTable water as a source of Mycobacterium avlum infection in AIDS**. *Lancet* 1994, **343**(8906):1137-1141.

112. Thomson R, Tolson C, Carter R, Coulter C, Huygens F, Hargreaves M: **Isolation of nontuberculous mycobacteria (NTM) from household water and shower aerosols in patients with pulmonary disease caused by NTM**. *J Clin Microbiol* 2013, **51**(9):3006-3011.

113. Peters M, Müller C, Rüsch-Gerdes S, Seidel C, Göbel U, Pohle HD, Ruf B: **Isolation of atypical mycobacteria from tap water in hospitals and homes: Is this a possible source of disseminated MAC infection in AIDS patients?** *J Infect* 1995, **31**(1):39-44.

114. Torvinen E, Torkko P, Nevalainen A, Rintala H: **Real-time PCR detection of environmental mycobacteria in house dust**. *J Microbiol Methods* 2010, **82**(1):78-84.

115. Falkinham JO: **Ecology of nontuberculous mycobacteria-where do human infections come from?** *Seminars in Respiratory and Critical Care Medicine* 2013, **34**(1):95-102.

116. Archuleta RJ, Hoppes PY, Primm TP: **Mycobacterium avium enters a state of metabolic dormancy in response to starvation**. *Tuberculosis* 2005, **85**(3):147-158.

117. Martín-Casabona N, Bahrmand AR, Bennedsen J, Østergaard Thomsen V, Curcio M, Fauville-Dufaux M, Feldman K, Havelkova M, Katila ML, Köksalan K *et al*: **Non-tuberculous mycobacteria: Patterns of isolation. A multi-country retrospective survey**. *Int J Tuberc Lung Dis* 2004, **8**(10):1186-1193.

118. Jesudason MV, Gladstone P: **Non tuberculous mycobacteria isolated from clinical specimens at a tertiary care hospital in South India**. *Indian Journal of Medical Microbiology* 2005, **23**(3):172-175.

119. Khatter S, Singh UB, Arora J, Rana T, Seth P: **Mycobacterial infections in human immuno-deficiency virus seropositive patients: role of non-tuberculous mycobacteria**. *The Indian journal of tuberculosis* 2008, **55**(1):28-33.

120. Sivasankari P, Khyriem AB, Venkatesh K, Parija SC: **Atypical mycobacterial infection among HIV seronegative patients in Pondicherry**. *The Indian journal of chest diseases & allied sciences* 2006, **48**(2):107-109.

121. Aggarwal M, Jindal N, Arora R, Aggarwal NP, Arora S: **Non-tuberculous mycobacteria: The changing scenario at Amritsar**. *Indian Journal of Tuberculosis* 1993, **40**(1):25-27.

122. Trivedi SS, Desai SG, Trivedi SB: **Non-tuberculous lung mycobacteriosis in Gujarat**. *Indian Journal of Tuberculosis* 1986, **33**(4):175-178.

123. Hardas UD, Jayaraman VS: **Differential identification of mycobacteria**. *Indian Journal of Tuberculosis* 1984, **31**(1):11-13.

124. Chauhan MM: **Non-tuberculous mycobacteria isolated from an epidemiological survey in rural population of Bangalore district**. *Indian Journal of Tuberculosis* 1993, **40**(4):195-197.

125. Lietman T, Porco T, Blower S: **Leprosy and tuberculosis: The epidemiological consequences of cross- immunity**. *Am J Public Health* 1997, **87**(12):1923-1927.

126. Pulickal AS, Fernandez GVJ: **Comparison of the prevalence of tuberculosis infection in BCG vaccinated versus non-vaccinated school age children**. *Indian Pediatrics* 2007, **44**(5):344-347.

127. Stanford JL, Sheikh N, Bogle G, Baker C, Series H, Mayo P: **Protective effect of BCG in Ahmednagar, India**. *Tubercle* 1987, **68**(3):169-176.

128. Karassova V, Weissfeiler J, Krasznay E: **Occurrence of atypical mycobacteria in Macacus rhesus**. *Acta microbiologica Academiae Scientiarum Hungaricae* 1965, **12**(3):275-282.

129. Weiszfeiler G, Karasseva V, Karczag E: **A new mycobacterium species: Mycobacterium asiaticum n. sp**. *Acta microbiologica Academiae Scientiarum Hungaricae* 1971, **18**(4):247-252.

130. Dawson DJ, Kane DW, McEvoy D: **Mycobacterium asiaticum as a potential pulmonary pathogen for humans. A clinical and bacteriologic review of five cases**. *American Review of Respiratory Disease* 1983, **127**(2):241-244.

131. Taylor LQ, Williams AJ, Santiago S: **Pulmonary disease caused by Mycobacterium asiaticum**. *Tubercle* 1990, **71**(4):303-305.

132. Dawson DJ, Blacklock ZM, Ashdown LR, Bottger EC: **Mycobacterium asiaticum as the probable causative agent in a case of olecranon bursitis**. *J Clin Microbiol* 1995, **33**(4):1042-1043.

133. Foulkes GD, Floyd JCP, Stephens JL: **Flexor tenosynovitis due to mycobacterium asiaticum**. *Journal of Hand Surgery* 1998, **23**(4):753-756.

134. Ahmed I, Jabeen K, Hasan R: **Identification of non-tuberculous mycobacteria isolated from clinical specimens at a tertiary care hospital: A cross-sectional study**. *BMC Infect Dis* 2013, **13**(1).

135. Paramasivan CN, Govindan D, Prabhakar R, Somasundaram PR, Subbammal S, Tripathy SP: **Species level identification of non-tuberculous mycobacteria from South Indian BCG trial area during 1981**. *Tubercle* 1985, **66**(1):9-15.

136. Shenai S, Rodrigues C, Mehta A: **Time to identify and define non-tuberculous mycobacteria in a tuberculosis-endemic region**. *International Journal of Tuberculosis and Lung Disease* 2010, **14**(8):1001-1008.

137. Mitra S, Tapadar SR, Banerjee D, Bhattacharjee S, Dey S, Kundu S: **Pulmonary disease due to Mycobacterium massiliense**. *The Indian journal of chest diseases & allied sciences* 2012, **54**(1):53-57.

138. Radhakrishna S, Murthy BN, Nair NGK, Ezhil R, Venkatasubramanian S, Ramalingam N, Periannan V, Ganesan R: **A concurrent comparison of a WHO-recommended 30-cluster survey and a modified version of it under Indian conditions in the estimation of immunization coverages**. *Indian Pediatrics* 1995, **32**(3):383-390.

139. Romanus V, Hallander HO, Wahlen P, Olinder-Nielsen AM, Magnusson PHW, Juhlin I: **Atypical mycobacteria in extrapulmonary disease among children. Incidence in Sweden from 1969 to 1990, related to changing BCG-vaccination coverage**. *Tubercle and Lung Disease* 1995, **76**(4):300-310.

140. Petrini B, Bennet R: **Cervical mycobacterial lymphadenitis in Swedish children during the post-BCG vaccination era [2]**. *Acta Paediatrica, International Journal of Paediatrics* 2007, **96**(1):146-147.

141. Katila ML, Brander E, Backman A: **Neonatal BCG vaccination and mycobacterial cervical adenitis in childhood**. *Tubercle* 1987, **68**(4):291-296.

142. Jindal N, Devi B, Aggarwal A: **Mycobacterial cervical lymphadenitis in childhood**. *Indian Journal of Medical Sciences* 2003, **57**(1):12-15.

143. Challu VK, Chandrasekaran S, Sreenivas TR, Chauhan MM, Jones B, Rajalakshmi R, Mahadev B, Balasangameshwara VH, Chaudhuri K: **Role of nontuberculous infection in immunization against tuberculosis**. *Indian Journal of Tuberculosis* 1992, **39**(3):165-170.

144. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR: **Low-coverage sequencing: Implications for design of complex trait association studies**. *Genome Res* 2011, **21**(6):940-951.

145. Green P: **2x Genomes - Does depth matter?** *Genome Res* 2007, **17**(11):1547-1549.

146. Margulies EH, Vinson JP, Blakesley RW, Bouffard GG, Hansen NF, Maskeri B, Thomas PJ, McDowell JC, Miller W, Jaffe DB *et al*: **An initial strategy for the systemic identification of functional elements in the human genome by low-redundancy comparative sequencing**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(13):4795-4800.

147. Ohlsen K, Hacker J, Ziebuhr W: **Genome variability in gram-positive pathogenic bacteria - Impact on virulence and evolution**. *Curr Genomics* 2004, **5**(7):589-600.

148. Hacker J, Blum-Oehler G, Hochhut B, Dobrindt U: **The molecular basis of infectious diseases: Pathogenicity islands and other mobile genetic elements - A review**. *Acta Microbiologica et logica Hungarica* 2003, **50**(4):321-330.

149. Kim BJ, Choi BS, Choi IY, Lee JH, Chun J, Hong SH, Kook YH, Kim BJ: **Complete genome sequence of Mycobacterium intracellulare clinical strain MOTT-36Y, belonging to the INT5 genotype**. *J Bacteriol* 2012, **194**(15):4141-4142.

150. Park JH, Shim TS, Lee SA, Lee H, Lee IK, Kim K, Kook YH, Kim BJ: **Molecular characterization of Mycobacterium intracellulare-related strains based on the sequence analysis of hsp65, internal transcribed spacer and 16S r-RNA genes**. *J Med Microbiol* 2010, **59**(9):1037-1043.

151. Kinsella RJ, Fitzpatrick DA, Creevey CJ, McInerney JO: **Fatty acid biosynthesis in Mycobacterium tuberculosis: Lateral gene transfer, adaptive evolution, and gene duplication**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(18):10320-10325.

152. Behr MA: **Mycobacterium du jour: what's on tomorrow's menu?** *Microb Infect* 2008, **10**(9):968-972.

153. Castellanos E, Aranaz A, De Juan L, Álvarez J, Rodríguez S, Romero B, Bezos J, Stevenson K, Mateos A, Domínguez L: **Single nucleotide polymorphisms in the IS900 sequence of Mycobacterium avium subsp. paratuberculosis are strain type specific**. *J Clin Microbiol* 2009, **47**(7):2260-2264.

154. Stucki D, Gagneux S: **Single nucleotide polymorphisms in Mycobacterium tuberculosis and the need for a curated database**. *Tuberculosis* 2013, **93**(1):30-39.

8. Appendix

8.1 List of abbreviations

°C	Degree centigrade
µl	Microliter
AIDS	Acquired Immuno Deficiency Syndrome
bp	base pair
BCG	Bacillus Calmette–Guérin
BLAST	Basic Local Alignment Search Tool
CTAB	Cetyl-trimethyl-ammonium-bromide
dH_2O	Distilled water
DDBJ	DNA Database of Japan
DNA	Deoxyribonucleic acid
DOTS	Directly Observed Treatment Strategy
EDTA	Ethelene diaminetetraacetic acid
FAFLP	Fluorescent Amplified Fragment Length Polymorphism
FI	Fitness Island
GI	Genome Island
GP	Glycoprotein
HCl	Hydrochloric acid
HGT	Horizontal gene transfer
HSP	Heat Shock Protein
HPLC	High Pressure Liquid Chromatography
IS	Insertion Sequence
ITS	Internal Transcribed Spacer
Kb	Kilobase
LSP	Long sequence polymorphisms
M	Molar
Mb	Mega base
MAA	*Mycobacterium avium avium*
MAC	*Mycobacterium avium* complex
MAIC	*Mycobacterium avium intracellulare* complex
MAH	*Mycobacterium avium hominissuis*
MAP	*Mycobacterium avium paratuberculosis*
MAS	*Mycobacterium avium silvaticum*
Min	Minutes
ml	Milliliter
ML	Maximum Likelihood
MLST	Multi Locus Sequence Typing
MOTT	Mycobacteria Other Than Tuberculosis

MTBC	*Mycobacterium tuberculosis* complex
MTb	*Mycobacterium tuberculosis*
NaCl	Sodium chloride
NaOH	Sodium Hydroxide
NET	NTM-Network European Trials
NGS	Next Generation Sequencing
nm	Nanometer
NRCM	National Reference Centre for Mycobacteria
NTM	Non Tuberculosis Mycobacteria
NTPs	National TB control programs
OADC	Oleic acid albumin dextrose
OD	Optimum Density
PCR	Polymerase chain reaction
PE	Paired End
PFGE	Pulse Field Gel Electrophoresis
Q score	Phred quality score
RAST	Rapid Annotation using Subsystem Technology
RFLP	Restriction Fragment Length Polymorphism
RGM	Rapid Growing Mycobacteria
RKI	Robert Koch Institute
r-RNA	Ribosomal ribonucleic acid
SDS	Sodium dodecyl sulfate
Sff	Standard flowgram format
SGM	Slow Growing Mycobacteria
SNP	Single Nucleotide Polymorphism
TE Buffer	Tris EDTA Buffer
TB	Tuberculosis
TBE buffer	Tris Borate EDTA buffer
TIGR	The Institute for Genomic Research
TSB	Tryptic Soya Broth
t-RNA	Transfer ribonucleic acid
WGS	Whole Genome Shotgun
WHO	World Health Organisation

8.2 Marker used during gel electrophoresis

The marker or ladder that is shown in the gel images (results) is presented here in details in the Figure 29. 100bp GeneRuler Plus DNA Ladder (Thermo Scientific, Schwerte, Germany) was used during gel electrophoresis.

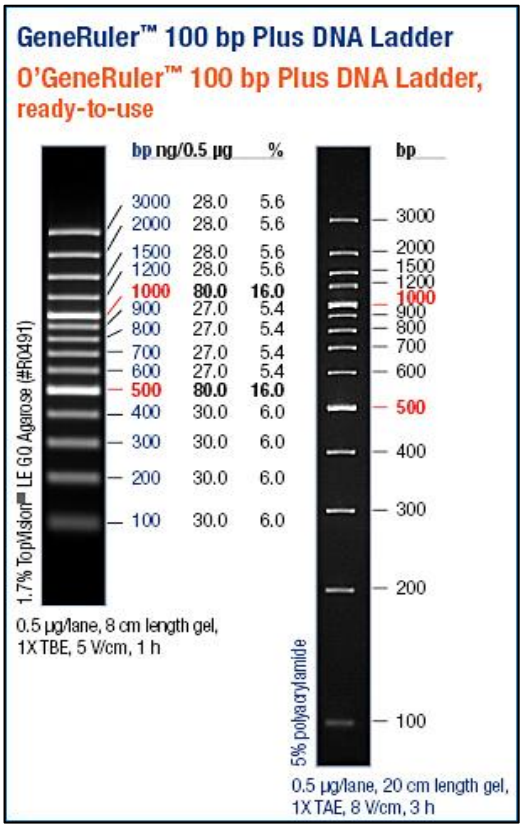


Figure 29: 100 bp plus DNA ladder used in gel electrophoresis

Above Figure represents the marker that has been shown in the gel images in the results section (Section-3). This marker provides a reference to estimate the size of the bands attained by gel electrophoresis.