Preprints of the 12th IFAC Symposium on Computer Applications in Biotechnology
The International Federation of Automatic Control
16-18, 2013, December. Mumbai, India

Comprehensive Phylogenetic Analysis of Mycobacteria

Arun N. Prasanna. Sarika Mehra

Genomics and Systems Biology Laboratory, Department of Chemical Engineering,
IIT Bombay 400 076, India (Tel: +9122-2576-7221; e-mail: sarika@che.iitb.ac.in)

Abstract: The genus mycobacterium encompasses both pathogens and non-pathogens, alternatively both slow and rapid growers. They are the source of a variety of infectious diseases in a range of hosts. Comparative genome analyses provide useful information to understand the genome feature of each pathogenic species to its unique niche. In this work, we report the phylogenetic analysis of 47 mycobacterium species, whose genome sequences are complete and available. Trees were constructed using two approaches namely single sequence and genome feature based methods. While single sequence based tree cannot distinguish between MTB complex genomes, trees based on genome features were able to resolve them better. Gene order based phylogeny highlights distinct evolutionary characteristics as illustrated by the shift in the relative position of drug susceptible and resistant *M. tuberculosis* complex species. Thus, phylogenetic relationship between closely related organisms can be resolved by genome feature based tree methods.

Keywords: Mycobacteria, comparative genomics, orthologs, phylogeny

1. INTRODUCTION

Pathogenic species of mycobacterium genus cause a variety of infectious diseases such as tuberculosis, leprosy and skin ulcers. Availability of whole genome sequences has opened the possibility of using various techniques to identify vaccine and therapeutic targets. A wide variety of techniques are currently available including pan-genomics, transcriptomics, proteomics, functional genomics and comparative genomics. Pan-genomics analyzes the genome of several organisms of same species to detect an antigenic target that represents the diversity of an organism. Pan-genome analysis of eight group B Streptococcus isolates revealed the presence of four proteins and their combination as potential vaccine targets (Maione *et al*, 2005). Transcriptomics is the study of gene expression profiles as a function of RNA transcript expressed by an organism under specific conditions. Reason for enhancement of transmission of pathogens during epidemic spread was found with analysis of transcriptome of *Vibrio cholera.* Isolates from human stool revealed the high induction of genes belonging to nutrient acquisition and motility and expression of chemotaxis genes to low levels (Merrell *et al*, 2002). Similar to transcriptomics, proteomics directly analyzes the expression of protein sets under specific conditions. For example, group A streptococcus was screened for surface exposed proteins for their use as vaccine target (Rodriguez *et al*, 2006). On the other hand, Functional genomics identifies candidate genes required for survival of an organism. 47 genes of *Helicobacter pyroli* that are essential for gastric colonization were identified and verified *via* mutant studies (Kavermann et al, 2003). Finally, (but not limited to) comparative genomics is powerful tool that can identify virulence genes that are present in pathogens but absent in non-pathogens (Rasko *et al*, 2008). Comparisons that can be done are infinite and flexible. Besides, such studies also shed light on their evolutionary relationship of closely related organisms.

In this study, comparative genomics of non-pathogenic (Appendix A) and pathogenic (Appendix B & C) species is performed. Most of the studies construct evolutionary relationships based on protein or nucleotide sequences of house-keeping genes such as 16S and dnaN. However, recently it has been shown that, trees based on gene content and gene order provide good resolution against conventional sequence based methods (Boore *et al*, 2008; Luo *et al*, 2009). In this work, we extend our previous study on 10 mycobacterium species (Prasanna AN, Mehra S, 2013) to include more closely related species. We employ different methods and show that, the combination approach works better in resolving relationship among mycobacteria.

2. METHODS

2.1 Identification of Homologs

Complete genome sequences for 47 mycobacteria was available and downloaded from NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). A bidirectional blast hit approach was carried out for every pair of mycobacteria. Standalone version of BLAST program (Altschul *et al*, 1990) was used to align gene/protein sequences. To perform the bidirectional BLAST, a subject database was created from one genome and the second genome was queried against this database. The BLAST was repeated by interchanging the subject and query genomes. Overall, 2162 comparison files were generated [n*(n-1); where n=47].

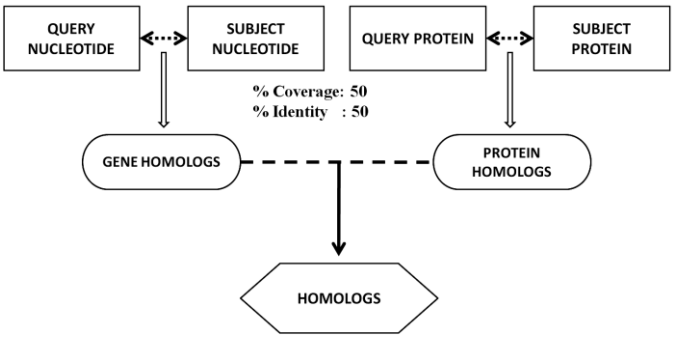

Figure. 1. Methodology for identifying homologs.

The nucleotide and protein blast parameters are given in table 1 below. For every gene, the threshold of significance for a BLAST hit was that, overall percentage identity was greater than or equal to 50% over the alignment length greater than or equal to 50% of the reference gene in the database. A gene pair is called homolog, only if both the nucleotide and protein sequences satisfy the above criteria.

Table 1. BLAST Parameters

Parameters	Nucleotide	Protein
E-value	10^{-6}	10^{-6}
Word length	11	3
Gap penalty	5	11
Extension cost	2	1

2.2 Identification of Core Orthologs

An ortholog is a pair of homologs, which are the best hit of each other. To identify orthologs from the list of homologs, the best hit in terms of average % identity is identified among the similar pair of genes. From the list of orthologs, a set of genes, in which each member has an ortholog from all the other genomes, were defined as the Core orthologs (C.O). The workflow involved in determination of core ortholog is given in figure 2 below.

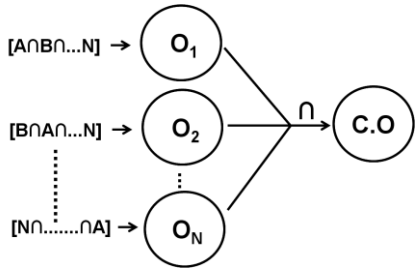

Figure. 2. Workflow of identification of Core Orthologs. A, B and N represent organisms and O_i represents the Ortholog set.

2.3 Phylogenetic Analysis

2.3.1 Sequences trees

Phylogenetic analysis for trees based on single and concatenated sequences was performed using MEGA 5.1 software (Tamura *et al*, 2011). For sequence based tree construction such as 16S and *dnaN*, alignment was done with clustalW program. Distance was computed using the jukes-cantor (Jukes, 1969) model. Trees were built using neighbour-joining, minimum evolution and UPGMA methods. To test the reliability of the tree branches, a bootstrap analysis with 1000 replicates was performed.

For trees based on core orthologs, protein sequences of the genes conserved across all mycobacteria was used. Protein sequences of the core orthologs were concatenated, and the resulting 47 sequences each representing one organism, were subjected to multiple sequence alignment on the MAFFT web server (Katoh *et al*, 2002). BLOSUM62 was used as the scoring matrix. Poorly aligned regions were removed using the software trimalv1.2 (Capella *et al*, 2009). Default settings were used to run the programs. Resulting alignment file was used as the input file for tree construction. The distance between two protein sequences was computed based on Jones-Taylor-Thornton (JTT) model (Jones *et al*, 1992). Finally, the tree was constructed as mentioned above.

2.3.2 Distance matrix trees

Phylogenetic analysis for trees based on distance matrices such as gene content and gene order were performed using PHYLIP program (Felsenstein, 1989). Gene content is defined as the ratio of the number of orthologs shared between two genomes to the total number of genes in the smallest of the genome. Gene content is used as a measure of distance between the two organisms. The smaller genome defines the maximum possible shared orthologs. Consensus trees were obtained based on majority rule (extended) using the program CONSENSE in PHYLIP (Margush, 1981).

To map the genome rearrangements, start position was used as the marker to determine the position of core orthologs on the chromosome. Genomes are represented as a signed permutation, 1,…,n, where a positive sign indicates coding strand and vice-versa. The order of genes on any one organism is considered as a reference and this process was repeated till all 47 organisms were considered. The number of reversal steps in one genome that would result in the same order of genes in the second genome is defined as the gene order distance. To compute the reversal distance in GRIMM (Tesler *et al*, 2008), all genomes were assumed to be unichromosomal and circular. The distance matrix, thus obtained was used to construct a tree. Jacknife resampling approach (Shi *et al*, 2010) was performed to obtain statistical support for the tree branches. 40% of the genes were removed randomly from the initial core orthologs to obtain 50 jackknife sets. Tree was constructed from each of these sets. Finally, consensus tree was obtained using CONSENSE program as described earlier.

3. RESULTS AND DISCUSSION

A phylogenetic tree was constructed based on the 16S and dnaN nucleotide sequence. Figure 3 shows the consensus tree based on the neighbor-joining method. Overall, the pathogenic and non-pathogenic mycobacteria form two distinct groups. The non-pathogenic mycobacterium species lie close to each other. *M. abscessus* and *M. massiliense* lies at the boundary of the two groups. It is closer to non-pathogens than to the pathogens. However, *M. tuberculosis*

KZN4207 lies farthest from the rest of *M. tuberculosis* strains.

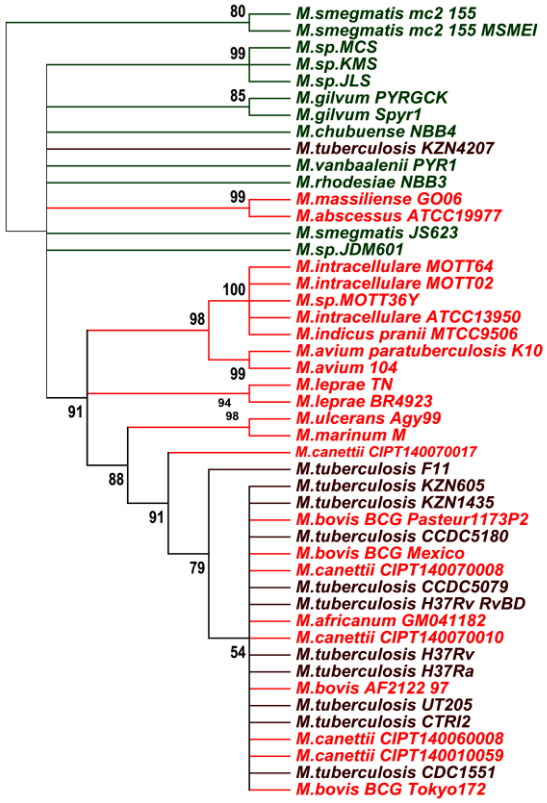

Figure.3. Phylogenetic tree based on 16S ribosomal RNA. Branches with less than 50% confidence values were merged together. Clades colored in red indicates pathogenic species, while maroon represents MTB strains and, green indicates non-pathogens.

A tree was also constructed based on the nucleotide sequence of *dnaN*, a DNA polymerase gene that is conserved across all the genomes. The tree is shown in Figure 4. Similar to the 16S based tree, the pathogenic and non-pathogenic mycobacteria form two distinct groups. In the tree based on *dnaN*, *M. tuberculosis* KZN4207 lies within MTB complex clade unlike the 16S tree. Also, the bootstrap values at many branch points indicate higher statistical significance compared to 16S rRNA tree.

Next, phylogenetic trees were constructed based on genome features such as protein sequences of core orthologs, gene order and gene content. The number of core orthologs shared by 47 mycobacteria was found to be 598. Tree constructed by concatenation of all the 598 proteins (Figure 5) is remarkably similar to the *dnaN* tree discussed above. Compared to the sequence based tree, this tree is better resolved in terms of confidence values. Two sub-groups are identified in the tree; one comprising of the pathogenic species and the other comprising of the non-pathogenic species. *M. abscessus* and *M. massiliense* form a separate clade that is closer to non-pathogens than to pathogens. On the other hand, in tree constructed based on gene order (Figure 6), the MTB complex do not group together as observed in previous trees. This suggests that, the various strains from different origin have undergone gene rearrangements. For example, the gene order of *M. tuberculosis* KZN species from South Africa is distinct from that of *M. tuberculosis* CTRI2 from Russia. It is interesting to note that while the former are drug resistant strains, the CTR12 strain is drug susceptible. Coincidentally, strains also showed groups based on geographical location. Thus, gene order method captures distinct phenotypic characteristics of the organisms compared. Finally, the tree based on gene content (Figure 7) also showed two groups with all strains in *M. tuberculosis* complex together except two strains (*M. tuberculosis* CCDC 5079 and CCDC 5180).

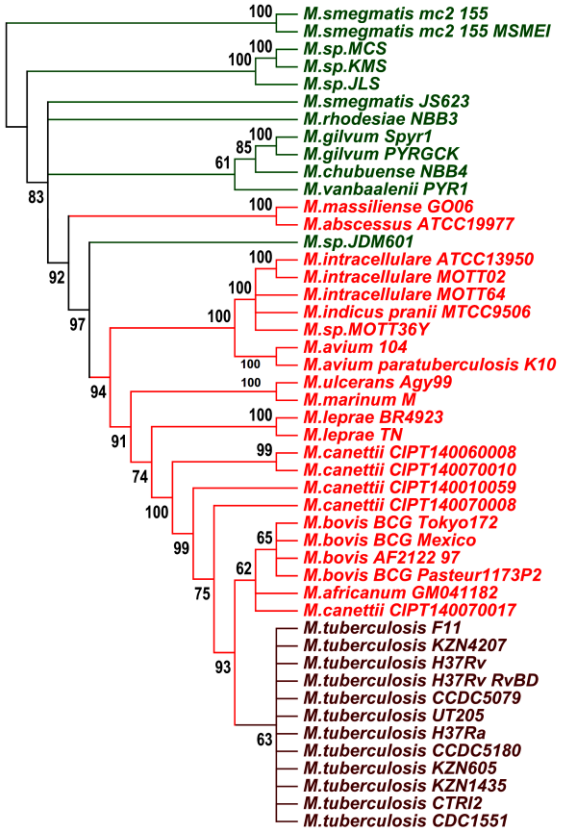

Figure.4. Phylogenetic tree based on *dnaN* nucleotide sequences. Branches with less than 50% confidence values were merged together.

Phylogenetic analyses of mycobacterial genomes based on genome features show that, single gene sequence based methods are sometimes unable to resolve the differences between closely related organisms, whereas, the trees based on gene order and concatenation show clear distinction. For instance, the tree based on 16S and *dnaN* could not differentiate strains within MTB complex. Similarly the 16S tree does not distinguish between MTB complex and *M. canetti*. However, the tree based on gene concatenation is better in resolving them into distinct groups. Also, the tree based on gene order suggests that the drug resistant and susceptible strains within MTB complex possess different gene rearrangements.

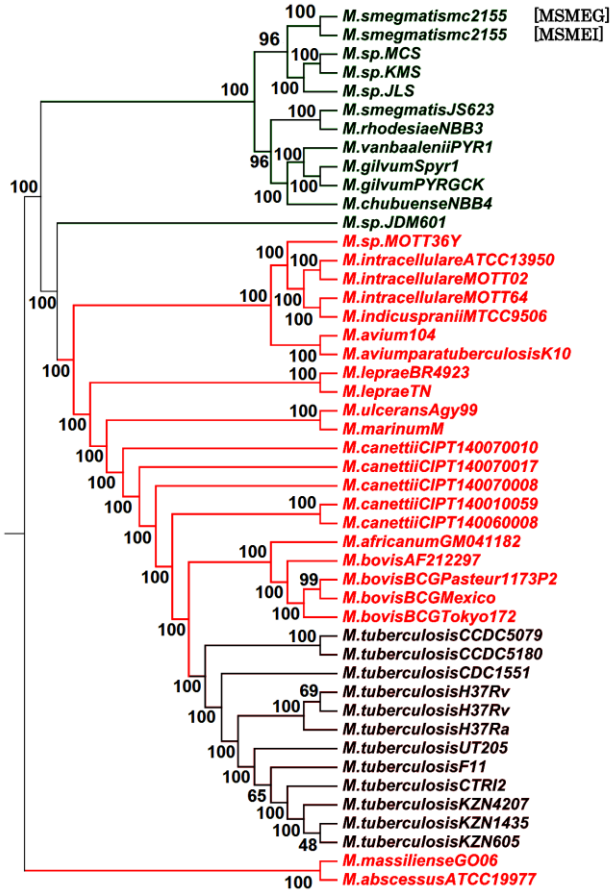

Figure.5. Phylogenetic tree based on gene concatenation

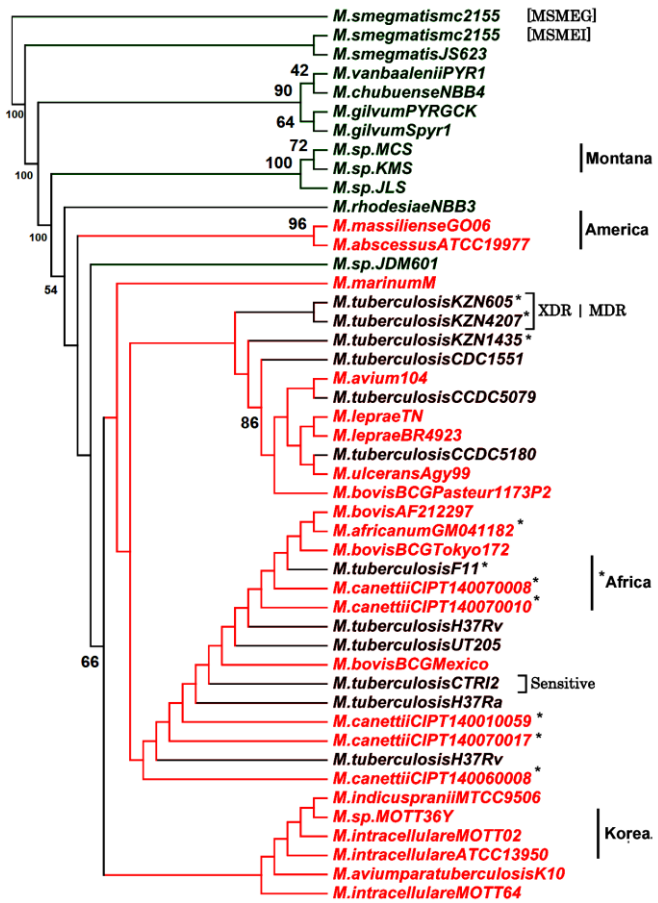

Figure.6. Phylogenetic tree based on gene order of core orthologs.

4. CONCLUSIONS

In this work, the genomes of 47 mycobacterium species are compared. The chosen set consists of both pathogenic and non-pathogenic species. We have focussed on the overall difference between pathogens and non-pathogens, based on single sequence, gene concatenation, gene content and gene order phylogeny. We found that, there are 598 core orthologs shared between all these mycobacteria. All the phylogenetic methods were able to distinguish pathogens and non-pathogens into two groups. However, trees based on single sequence are unable to differentiate between strains of *M. tuberculosis* complex. In contrast, the gene order and gene concatenation based trees were able to clearly resolve the difference between them. Gene order trees capture distinct phenotypic characteristics such as drug sensitivity or resistance. Finally, we conclude that, the comparative analysis using a combination of genomic features provides improved resolution among the mycobacteria.

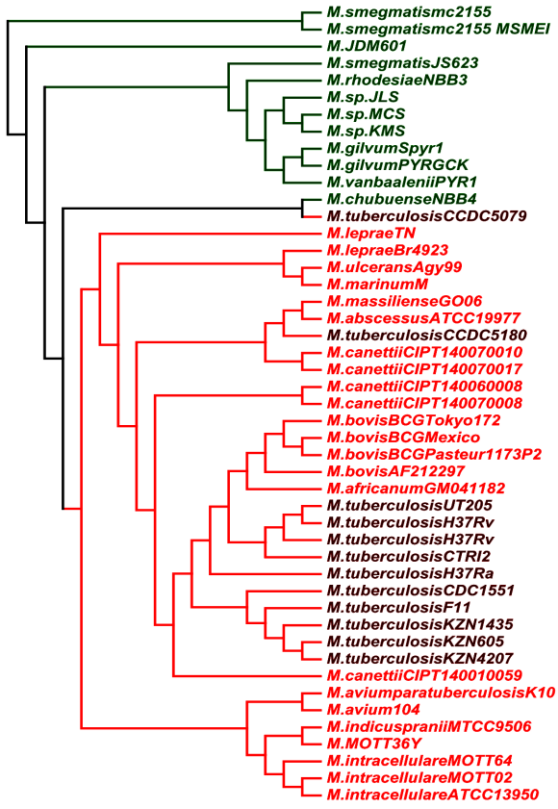

Figure.7. Phylogenetic tree based on gene content

REFERENCES

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

Boore JL, Fuerstenberg SI (2008) Beyond linear sequence comparisons: the use of genome-level characters for phylogenetic reconstruction. Philos Trans R Soc Lond B Biol Sci 363: 1445-1451.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972-1973.

Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166.

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8: 275-282.

Jukes TH, Cantor CR (1969) Evolution of Protein Molecules. Mammalian Protein Metabolism: 21-132.

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30: 3059-3066.

Kavermann, H., B. P. Burns, et al. (2003) Identification and characterization of Helicobacter pylori genes essential for gastric colonization. J Exp Med 197(7): 813-22.

Luo H, Sun Z, Arndt W, Shi J, Friedman R, et al. (2009) Gene order phylogeny and the evolution of methanogens. PLoS One 4: e6069.

Maione, D., I. Margarit, et al. (2005) Identification of a universal Group B streptococcus vaccine by multiple genome screen. Science 309(5731): 148-50.

Margush T, McMorris FR (1981) Consensus n-trees. Bltn Mathcal Biology 43: 239-244.

Merrell,D.S.,S.M. Butler, et al. (2002) Host-induced epidemic spread of the cholera bacterium. Nature 417(6889): 642-5.

Prasanna AN, Mehra S (2013) Comparative Phylogenomics of Pathogenic and Non-Pathogenic Mycobacterium. PLoS ONE 8(8): e71248.

Rasko, D. A., M. J. Rosovitz, et al. (2008) The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. J Bacteriol 190(20): 6881-93.

Rodriguez-Ortega,M.J., N. Norais, et al. (2006) Characterization and identification of vaccine candidate proteins through analysis of the group A Streptococcus surface proteome. Nat Biotechnol 24(2): 191-7.

Shi J, Zhang Y, Luo H, Tang J (2010) Using jackknife to assess the quality of gene order phylogenies. BMC Bioinformatics 11: 168.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731-2739.

Tesler G, Bourque G (2008) Computational tools for the analysis of rearrangements in mammalian genomes. Methods Mol Biol 422: 145-170.

Appendix A. LIST OF NON-PATHOGENS

LOCUS-ID	ORGANISM NAME	# PROTEINS
Mycch	*Mycobacterium chubuense* NBB4	5181
Mflv	*Mycobacterium gilvum* PYR GCK	5241
Mspyr1	*Mycobacterium gilvum* Spyr1	5130
MIP	*Mycobacterium indicus* pranii MTCC 9506	5254
JDM601	*Mycobacterium* JDM601	4346
Mjls	*Mycobacterium* JLS	5739
Mkms	*Mycobacterium* KMS	5460
Mmcs	*Mycobacterium* MCS	5391
MycrhN	*Mycobacterium rhodesiae* NBB3	6147
Mvan	*Mycobacterium vanbaalenii* PYR 1	5979
MSMEG	*Mycobacterium smegmatis* MC2 155	6717
MSMEI	*Mycobacterium smegmatis* MC2 155	6690
Mycsm	*Mycobacterium smegmatis* JS623	6186

Appendix B. LIST OF PATHOGENS

LOCUS-ID	ORGANISM NAME	# PROTEINS
MAB	*Mycobacterium abscessus* ATCC 19977	4920
MAF	*Mycobacterium africanum* GM041182	3830
MAV	*Mycobacterium avium* 104	5120
MAP	*Mycobacterium avium* paratuberculosis K 10	4350
Mb	*Mycobacterium bovis* AF2122 97	3918
BCGMEX	*Mycobacterium bovis* BCG Mexico	3952

BCG	*Mycobacterium bovis* BCG Pasteur 1173P2	3949
BN44	*Mycobacterium canettii* CIPT 140060008	3981
BN43	*Mycobacterium canettii* CIPT 140070008	3986
BN42	*Mycobacterium canettii* CIPT 140070010	3941
BN45	*Mycobacterium canettii* CIPT 140070017	4009
JTY	*Mycobacterium bovis* BCG Tokyo 172	3944
MCAN	*Mycobacterium canettii* CIPT 140010059	3861
OCU	*Mycobacterium intracellulare* ATCC 13950	5144
OCO	*Mycobacterium intracellulare* MOTT 02	5149
OCQ	*Mycobacterium intracellulare* MOTT 64	5249
MLBr	*Mycobacterium leprae* Br4923	1604
ML	*Mycobacterium leprae* TN	1605
MMAR	*Mycobacterium marinum* M	5423
MYCMA	*Mycobacterium massiliense* GO 06	2626
W7S	*Mycobacterium* MOTT36Y	5128
MUL	*Mycobacterium ulcerans* Agy99	4160
TBFG	*Mycobacterium tuberculosis* F11	3941
MRA	*Mycobacterium tuberculosis* H37Ra	4034
Rv	*Mycobacterium tuberculosis* H37Rv	4003
RvBD	*Mycobacterium tuberculosis* H37Rv	4111
TBXG	*Mycobacterium tuberculosis* KZN 605	4001
TBMG	*Mycobacterium tuberculosis* KZN 1435	4059
TBSG	*Mycobacterium tuberculosis* KZN 4207	3996
MRGA327	*Mycobacterium tuberculosis* RGTB327	3691
MRGA423	*Mycobacterium tuberculosis* RGTB423	3622
UDA	*Mycobacterium tuberculosis* UT205	3796

Appendix C. LIST OF MTB STRAINS

LOCUS-ID	**ORGANISM NAME**	**# PROTEINS**
CCDC5079	*Mycobacterium tuberculosis* CCDC5079	3646
CCDC5180	*Mycobacterium tuberculosis* CCDC5180	3590
MT	*Mycobacterium tuberculosis* CDC1551	4189
MTCTRI2	*Mycobacterium tuberculosis* CTRI 2	3944