



MONASH University

ISBaC: An automated pipeline for *In Silico* Bacterial Identification

By

Hira Javaid

Bioinformatics (Hons) (Computer Science)

The University of Agriculture Faisalabad, Pakistan

A thesis submitted for the degree of Master of Science (Research) at Monash University
Malaysia in 2021

School of Science

Copyright notice

©The author (2021). Except as provided in the Copyright act 1968, this thesis may not be reproduced in any form without the written permission of the author.

I certify that I have made all reasonable efforts to secure copyright permissions for the third-party content included in this thesis and have not knowingly added copyright content to my work without the owner's permission.

Abstract

The advancement in computer technologies has boosted computationally intensive data analysis in biology and medicine research. With various data processing algorithms, ready access to thousands of whole microbial genomes could facilitate the study of microbial diseases. The current high-throughput sequencing technologies and development of these sequence-based identification methods make the task of identification of microorganisms simpler and faster with better accuracy. Nevertheless, using different bioinformatics tools available for the identification of microbes is sometimes challenging, especially for the beginners. This is particularly notable in identifying mycobacterial species, common agents of many opportunistic infections in humans. Diseases caused by mycobacterial species are especially challenging because of trouble in getting suitable clinical samples, particularly from non-accessible sites and due to poor sensitivity of identifying methods. Most of the current bioinformatic workflows and pipelines available are difficult to implement because users may face some problems installing the required software's in the Linux system if users are not familiar with Bioinformatics skills. Users may also have problems performing sequencing alignment, and downloading proper reference genomes. Besides, user also required some programming language if the user wants to visualize the result properly through graphs and charts. Thus, it is not easy to implement the analysis separately to identify the identity of mycobacterial species if the users do not have any Bioinformatics knowledge.

So the designed ISBaC will be the first pipeline to automated the identification of Mycobacterium species. This pipeline can accurately identify mycobacterial species in less time is of paramount importance. The pipeline start from raw data quality check, trimming of poor-quality sequences, *de-novo* genome assembly, genome annotation, virulence gene prediction, and lastly, the identification of the species identity using *in silico* approaches: the *16S rRNA* gene analysis, multilocus sequence analysis (MLSA) and Average Nucleotide Identity (ANI) analysis. The complete analysis takes around one hour. The ISBaC pipeline script, which is written in Perl language, is user friendly with just single one command to execute the pipeline. The pipeline can analyse single end, paired end and whole genome sequence from different sequencing platforms. ISBaC pipeline showed arguably the best overall performance, combining high sensitivity with excellent specificity and accuracy by identifying mycobacterium species in the repeated samples.

Declaration

Monash University

Declaration for thesis based or partially based on conjointly published or unpublished work

General Declaration

In accordance with Monash University Doctorate Regulation 17.2 Doctor of Philosophy and Research Master's regulations the following declarations are made:

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and beliefs, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

The core theme of the thesis is to design an automated pipeline that mainly focuses on the identification of *mycobacterial* strains by chaining various processing and identification tools in a script that is easily accessible, user friendly and can be run with small number of commands. The ideas, development and writing up of the thesis were the principal responsibility of myself, the student, working with the School of Science under the supervision of Dr. Wee Wei Yee.

Student Signature: Hira Javaid

Name: Hira Javaid

Date 11/01/2021

Supervisor Signature: WWY

Name: Wee Wei Yee

Date: 11/01/2021

Acknowledgements

Firstly, I would like to thank you my supervisor Dr. Wee Wei Yee, Dr. Song Beng Kah, Dr. Md Zobaer Hasan for their counsel, support, and patience throughout Research Master's. I would have been lost without their knowledge and their guidance. My sincere thanks to the academic and administrative staff members in School of Science for being helpful and pleasant to me, placing me at ease throughout my Research Master's. I would like to express my gratefulness to Monash University Malaysia, for the scholarship.

To finish, a huge thank you to my family, particularly my parents and my brother, without their guidance, moral support, and constant support motivation, I would not be where I am at the present. Your faith and trust in me, made me adept to accomplish the year and more in future and made more courageous in life. I have in general relished the Research Master's experience, despite the sporadic stints of disappointments and weariness and hopefully I gain more experiences and discover my future research career as fulfilling as my Research Master's.

Contents

| | |
|---|------|
| Abstract..... | iii |
| Declaration | iv |
| Acknowledgements | v |
| List of Figure | vii |
| List of Table..... | vii |
| List of Abbreviations | viii |
| Chapter1. Introduction | 1 |
| 1.1 <i>Non-tuberculous mycobacteria</i> (NTM) | 2 |
| 1.2 Phenotyping Methods for identification of mycobacteria | 3 |
| 1.3 Genotyping methods of identification of mycobacteria | 4 |
| 1.3.1 PCR Restriction Enzyme Analysis | 4 |
| 1.3.2 Gene Sequencing | 4 |
| 1.3.3. Multilocus Sequencing Approaches (MLSA) | 6 |
| 1.3.4 Whole-genome sequencing | 6 |
| 1.4 Limitations of existing traditional and bioinformatics approaches..... | 7 |
| 1.5 Aims and significance of study | 9 |
| 1.6 Objectives | 10 |
| Chapter2. Material and methods | 11 |
| 2.1 Data mining | 11 |
| 2.2 Raw reads pre-processing and genome assembly..... | 11 |
| 2.3 Genome annotation | 12 |
| 2.4 Mycobacteria identification | 12 |
| 2.5 Virulence gene predication | 13 |
| 2.6 Statistical analysis..... | 14 |
| 2.7 Flowchart of Experimental Method..... | 15 |
| Chapter3. Results..... | 16 |
| 3.1 Raw reads pre-processing..... | 18 |
| 3.2 Genome assembly and annotation | 22 |
| 3.3 Mycobacterial identification using the <i>16S rRNA</i> gene analysis | 22 |
| 3.4 Mycobacterial Identification using Multilocus sequence analysis (MLSA)..... | 23 |
| 3.5 Mycobacterial Identification using whole genome sequence..... | 25 |
| 3.5.1 Mash: fast genome and metagenome distance estimation using MinHash..... | 25 |
| 3.5.2 Fast whole genome similarity estimation tool (FastANI) | 26 |
| 3.6 Virulence gene predication | 27 |
| 3.7 Calculation of ISBaC's accuracy through statistical analysis..... | 33 |
| 3.8 Determining the sensitivity of ISBaC pipeline by identifying closely related <i>mycobacterium</i> species | 36 |
| 3.9 Evaluating the reproducibility of ISBaC pipeline | 37 |
| Chapter4. Discussion | 38 |
| Conclusion and future work | 40 |
| Reference | 42 |
| Appendix..... | A |
| Appendix 1: List of <i>16S rRNA</i> genes..... | A |
| Appendix 2: List of whole-genome sequence | F |
| Appendix 3: Screening of virulence gene profile | L |

List of Figure

| | |
|--|----|
| Figure 3.1: Calling SoftwareInstallation.pl script that automatically install all required software's of ISBaC pipeline | 17 |
| Figure 3.2: ISBaC pipeline input options. | 18 |
| Figure 3.3: Final output directory folder containing all the ISBaC results..... | 19 |
| Figure 3.4: Per Base Sequence Quality in FastQC report..... | 21 |
| Figure 3.5: Bar chart of top 30 mycobacterial species..... | 23 |
| Figure 3.6: Bar chart of top 30 mycobacterial species..... | 24 |
| Figure 3.7: Similarity matrix of top 10 mycobacterial species | 24 |
| Figure 3.8: MinHash Results. | 25 |
| Figure 3.9: FastANI Heatmap. | 26 |
| Figure 3.10: Heatmap of comparative pathgenomic analysis..... | 33 |
| Figure 3.11: Error bar for <i>16S rRNA</i> , MLSA and ANI analysis calculated by the ISBaC result on 10 <i>M.chelonae</i> samples | 34 |
| Figure 3.12: Error bar for <i>16S rRNA</i> , MLSA and ANI analysis calculated by the ISBaC result on 10 different <i>mycobacterium</i> species samples | 36 |

List of Table

| | |
|---|----|
| Table 3.1: List of software's with link | 16 |
| Table 3.2: List of functional categories of virulence genes found in query genome..... | 27 |
| Table 3.3: Number of virulence gene in <i>mycobacterium</i> species | 30 |
| Table 3.4: Top hit and identity value for all the three analyses of 10 <i>M.chelonae</i> samples | 33 |
| Table 3.5: Top hit and identity value for all the three analyses of 10 different <i>mycobacterium</i> species samples | 35 |
| Table 3.6: ISBaC pipeline testing result for differentiating <i>M. malmoeense</i> form <i>M. szulgai</i> | 37 |
| Table 3.7: ISBaC pipeline testing result for differentiating <i>M. kansasii</i> from <i>M. gastri</i> | 37 |

List of Abbreviations

| | |
|---------------------|--|
| ISBaC | An automated pipeline for <i>In Silico</i> Bacterial Identification |
| MLSA | Multilocus sequencing approach |
| ANI | Average Nucleotide Identity |
| NTM | <i>Nontuberculous mycobacteria</i> |
| SGM | Slow-growing mycobacteria |
| RGM | Rapidly growing mycobacteria |
| <i>mce</i> | Mammalian cell entry |
| PCR | Polymerase chain reaction |
| MALDI-TOF MS | Matrix-assisted laser desorption/ionization-time of flight mass spectrometry |
| CLSI | Clinical and Laboratory Standards Institute |
| <i>ITS</i> | 23S rRNA internal transcribed spacer |
| <i>dnaJ</i> | 32-kDa protein |
| <i>sod</i> | Superoxide dismutase |
| <i>gyrB</i> | β subunit of DNA gyrase |
| <i>secA1</i> | Secretory pathway protein |
| <i>recA</i> | DNA recombination gene |
| SOS | Save our soul |
| hk | Housekeeping genes |
| WGS | Whole-genome sequencing |
| MinHash | Min-wise independent permutations locality sensitive hashing scheme |
| FastANI | Fast Whole-Genome Similarity Estimation |
| MICRA | Microbial identification and characterization through reads analysis |
| BIBI | Bioinformatics Bacterial Identification Tool |
| NCBI | National Centre for Biotechnology Information |
| ENA | European Nucleotide Archive |
| VFDB | Virulence Factor Database |
| SPSS | Statistical Package for the Social Sciences |
| CI | Confidence interval |

Chapter1. Introduction

Mycobacterium is a genus of phylum *Actinobacteria*, given its family name, *Mycobacteriaceae* (Hartmans Sybe and de Bont 2006). The genus *mycobacterium* is aerobic, gram-positive, non-sporulating and bacillary. *Mycobacteria* are found in water, soil, and bogs. *Mycobacteria* are hard to kill because their division pattern is asymmetric, leading to a population of cells that differ in growth rate and vulnerability to antibiotics, which resulted in their rise (Falkinham 2009). In *Mycobacteria* high content of lipid compound mycolic acid in the outer membrane is responsible for the poor absorption during staining procedure (Bergey's Manual® of Systematic Bacteriology 2012).

There are over 170 species in this genus, and most are associated with human diseases (Parte 2014). Previously genus *Mycobacteria* has been split into five main monophyletic clades: *Tuberculosis-Simiae* clade, *Terrae* clade, *Triviale* clade, *Fortuitum-Vaccae* clade, and *Abscessus-Cheloniae* clade. To better reflect the evolutionary relationship between the known species of mycobacteria, a proposal was made for phylogenetic classification. According to revised classification, *Mycobacterium* is split into five genera which are: *Mycobacterium* consisting of the members of *Tuberculosis-Simiae* clade, *Mycolicibacterium* gen. nov. consisting of *Fortuitum-Vaccae* clade, *Mycolicibacter* gen. nov. consisting of the members of *Terrae* clade, *Mycolicibacillus* gen. nov. consisting of the members of *Triviale* clade. *Mycobacteroides* gen. nov. consisting of the members of *Abscessus-Cheloniae* clade (Gupta, Lo, & Son 2018).

More than 20 mycobacterial species are identified as causative agent of human diseases, but *M. tuberculosis* is by far the most reported human pathogen (Bottai & Brosch 2009; Meehan et al. 2019). *Mycobacteria* can be categorized into various groups for diagnosis and treatment. But our main focus is to study *Nontuberculous mycobacteria* (NTM) species because the frequency of NTM disease is increasing worldwide and rapidly becoming a major public health problem (Ryu, Koh, & Daley 2016). There are two types of *non-tuberculous mycobacteria* (NTM), slow-growing mycobacteria (SGM), which takes more than seven days to form colonies on agar medium, and rapidly growing mycobacteria (RGM), which takes less than seven days to form colonies on agar medium (C. J. Kim et al. 2013).

1.1 *Non-tuberculous mycobacteria* (NTM)

Nontuberculous mycobacteria (NTM) are found in water and soil. Over 180 different species and subspecies are characterized under NTM; most species do not cause human disease, except vulnerable persons (Marras and Daley, 2002). When a person gets exposed to environmental sources of NTM, this microorganism enters the lungs and causes inflammation in the respiratory system. Although most individuals do not show symptoms, some vulnerable persons show a progressive lung infection that can be cured with continuous treatment of antibiotics for at least 12 months (Johnson & Odell 2014). The NTM are usually two types corresponding to their permit (Cook et al. 2009), which are slowly growing mycobacteria (7 days to form colonies) and rapidly growing mycobacteria (3 days to form colonies). Slowly growing mycobacteria have a prolonged growth rate and take almost seven days to form transparent colonies on agar. For Example, *M. avium*, *M. intracellulare*, *M. kansasii*, *M. xenopi* and *M. simiae*. The reason for slow growth of mycobacteria is outer membrane's impermeability to hydrophilic chemicals (Brennan 1995). Due to the hydrophobic nature of mycobacteria, they can form biofilms, and these biofilm-grown cells are also resistant to disinfectants. They can grow in low nutrient cultures, such as drinking water. They are even heat-resistant and can tolerate a temperature of 55°C (130°F) or higher (Steed & Falkinham 2006).

The rapidly growing mycobacteria usually take less than one week to form colonies in culture medium. Few species are also found to be associated with human disease (Schlossberg 2017). More than 80% of clinical isolates of RGM include *M. fortuitum*, *M. chelonae*, and *M. abscessus*. The most frequently isolated species, *M. abscessus* and *M. fortuitum* are isolated from clinical respiratory and non-respiratory specimens. Some RGM species, such as *M. smegmatis* and *M. chelonae* are disinfectant resistant. RGM cannot be treated with a standard therapeutic regimen. Therefore, specific antimicrobial susceptibility testing is required. Antimicrobial susceptibility can also be used to differentiate between RGM species. However, the outcomes of these therapies vary depending on the species and severity of the sickness.

RGM were classified into various groups based on nucleotide differences in the *16S rRNA* gene sequence: the *M. fortuitum* group included *M. fortuitum*, *M. peregrinum*, *M. houstonense*, and *M. neworleane*; the *M. peregrinum* group included *M. peregrinum*, *M. houstonense*, and *M. neworleane*; and the *M. septicum* group included *M. septicum*, *M. mageritense*, *M. mucogenicum*, and *M. senegalense*; the *M. chelonae abscessus* group included *M. abscessus*, *M. chelonae*, and *M. immunogenum*; and the *M. smegmatis* group included *M. smegmatis*, *M. wolinskyi*, and *M. goodii* (Brown-Elliott & Wallace 2002). However, there has been some debate about the taxonomic classification of *M. mucogenicum*,

because its biochemical profile and antibiotic-susceptibility pattern were found to be more closely related to those of members of the *M. chelonae-abscessus* group (*M. chelonae-abscessus* group) (*M. chelonae-abscessus* group) (Brown-Elliott & Wallace 2002). A phylogenetic tree constructed using the *16S rRNA* gene indicated that *M. mageritense* was more closely linked to the *M. smegmatis* group than the *M. fortuitum* group, despite the fact that it exhibited antibiotic sensitivity biochemical patterns similar to those of the *M. smegmatis* group (Brown-Elliott & Wallace 2002). *M. leprae* is more closely related to the NTM species *M. avium*, according to phylogenetic trees based on whole genome sequencing and *16S rRNA* gene sequences. Other approaches, on the other hand, confirm the existence of a close sister group to *M. tuberculosis* (Claeys & Robinson 2018).

The genetic reasons responsible for mycobacterial species growth rate are not well known. Recent findings imply that ancient mycobacteria exhibited a fast growth characteristic, which was followed by a single main evolutionary divergence into sub-genera that grew quickly and slowly. Among the genes they discovered were those encoding for amino acid transport/metabolism (e.g., the livFGMH operon) and transcription, as well as genes encoding for new ABC transporters. In slow-growing organisms, the loss of the livFGMH and ABC transporter operons implies that decreased cellular amino acid transport may be a limiting factor in growth. According to the results of a comparative genomic investigation, horizontal gene transfer from non-mycobacterial taxa may have contributed to the trait. It was discovered that the mammalian cell entry (*mce*) operon was present in all species, regardless of growth phenotype or pathogenicity, despite the fact that there was little protein sequence similarity between fast-growing and slow-growing species. As a result, this shows that the *mce* operon was present in ancient fast expanding species but was later repurposed by slower growing species to serve as a mechanism for establishing an intracellular lifestyle (Bachmann et al. 2020).

1.2 Phenotyping Methods for identification of mycobacteria

Phenotypes are the observable traits determine by genotypes, and they include morphological, physiological, and biochemical properties of the organism. Before the advancement in molecular techniques, bacterial taxonomy was solely based on comparative studies of the phenotypic features, which require pure laboratory cultures (Lagier et al. 2015). Traditionally, mycobacterium strains and species were also identified and classified using phenotyping methods. Some RGM and SGM were determined by phenotyping methods based on growth (Brown-Elliott & Wallace 2002). The other

parameters for identification were optimum growth temperature, acid fastness, gram staining, colony morphology and absence of pigmentation. But these traditional phenotypic identification methods are laborious, challenging, and time-consuming, require many weeks for proper growth. Sometimes, findings of these phenotyping methods may lead to inaccurate identification of species because different species share similar morphological and biochemical profiles. Different culture media are utilized for the isolation of mycobacteria. The most used method is “Löwenstein-Jensen medium”, which is based on eggs, and it also contains high concentrations of malachite green to overcome contamination with other bacteria. Another method is liquid culture media which is more sensitive than egg-based solid media for the isolation of mycobacteria from clinical specimens. One more method is Ziehl-Neelsen staining for the direct detection of mycobacteria. This method is also used to identify acid fast bacilli which appear bright red after staining (Palomino 2009).

1.3 Genotyping methods of identification of mycobacteria

The genotyping methods depends on what data is being looked for. Numerous strategies at first require amplification of the DNA sample, which is commonly done utilizing PCR. These methods are automated, and results are obtained very quickly, often with more precision than with phenotyping methods.

1.3.1 PCR Restriction Enzyme Analysis

The restriction fragment length polymorphisms is a powerful method for the identification of SGM and RGM (Steingrube et al. 1995). In today's world, new technology such as gene sequencing and other molecular procedures are gradually displacing these more traditional approaches. The Polymerase chain reaction-Restriction Enzyme analysis technique entails amplifying a gene encoding a protein using polymerase chain reaction (PCR), followed by subjecting the PCR products to particular restriction endonuclease digestion. Gel electrophoresis is used to acquire particular patterns of digested amplicons, which are then compared with patterns of digested amplicons obtained from known species to determine whether the patterns are related (Singh & Kashyap 2012).

1.3.2 Gene Sequencing

Most molecular methods use partial sequence targets, including the *16S rRNA* gene, *hsp65* gene, *rpoB*, and others to identify mycobacterium. The *16S rRNA* gene is the primary gene target for mycobacterial identification and differentiation. The main reason for using this gene in molecular taxonomic studies is that there is a well-organized and robust database of *16S rRNA* gene. The *16S rRNA* gene is a highly

conserved gene composed of nearly 1,500 nucleotides. The first 500 bps of 5' prime end of the *16S rRNA* gene comprises two major hypervariable domains, known as regions A and B. Region A contains species-specific sequence variations ("signature sequences"). Thus, the sequencing of this region is more informative for mycobacterial species identification, while region B might be confirmatory. *16S rRNA* gene sequences of members of the genus *Mycobacterium* are closely related, but they may show a difference of few base pairs. However, the *16S rRNA* gene cannot differentiate among some closely related NTM, including *M. abscessus*, *M. chelonae*, and some species within the *M. fortuitum* group (Turenne, Tschetter, Wolfe, & Kabani 2001). *M. abscessus* and *M. chelonae* vary by 4-bp within the *16S rRNA* gene but are indistinguishable within regions A and B. Thus, for those species identification which share highly similar *16S rRNA* gene sequence, sequencing outside of regions A and B or the addition of another gene target is compulsory (Brown-Elliott & Wallace 2002).

The Clinical and Laboratory Standards Institute (CLSI) has suggested the standards for identifying mycobacterial species by the *16S rRNA* gene sequencing (S. H. Kim & Shin 2018). According to CLSI, the reference and query sequences should be compared with at least a '300 bp quality sequence'. There must be a minimum of one region of the gene where variations are expected for reliable results. The 100% sequence identity is required with reference sequences for clear species identification (Bosshard & McDaniel 2010).

Next, a 441-bp hypervariable region in 65-kDa heat shock protein gene (*hsp65*) is another useful gene target for species-level identification. The *hsp65* gene is well conserved than the *16S rRNA* gene, thus allowing species-level identification of closely related species., such as *M. abscessus* and *M. chelonae*. These two species differ by 30bp in the 441-bp *hsp65* fragment. Additionally, many other species within the *M. fortuitum* group (*M. fortuitum*, *M. septicum*, *M. peregrinum*, *M. houstonense* and *M. senegalense*) are more rapidly identified by *hsp65* gene. However, the lack of a well-integrated and updated database of *hsp65* gene sequences is the major drawback of this identification methods.

The *rpoB* gene is a single copy gene that encodes beta subunit of the RNA polymerase. A 723-bp fragment in region V of the *rpoB* gene is most used in sequencing . Since it is less conserved, it can be useful for species discrimination that could not be differentiated by the *16S rRNA* gene or the *hsp65* gene sequence alone. *rpoB* gene being a single copy is more valuable because a single location without deletion or insertion is generally sufficient to identify many of the SGM and RGM to the species level. Thus, *ropB* can differentiate SGM and RGM at the intra- and interspecies levels (S. H. Kim & Shin 2018).

Several marker genes have been proposed for the identification of SGM and RGM including, 16S, 23S rRNA internal transcribed spacer (*ITS*) region, a 32-kDa protein gene (*dnaJ*) (Yamada-Noda et al. 2007), the superoxide dismutase (*sod*) gene (Zolg & Philippi-Schulz 1994), the *gyrB* gene encoding the β subunit of DNA gyrase (Dauendorffer et al. 2003), the *secA1* gene encoding a vital component of the major protein secretory pathway across the cytoplasmic membrane (Soini, Bottger, & Viljanen1 1994), and the *recA* gene which is crucial for homologous DNA recombination, DNA damage repair, and induction of the SOS (save our soul) response (Blackwood et al. 2000). However, the usefulness of these genetic targets is uncertain because of the insufficient literature about them and the lack of an appropriate database (Adékambi, Colson, & Drancourt 2003).

1.3.3. Multilocus Sequencing Approaches (MLSA)

Recently, a new sequencing approach has been proposed for mycobacterial identification: multilocus gene sequencing (i.e., sequencing portions of multiple genes) (Devulder, de Montclos, & Flandrois 2005). Usually, this usage of about 8 to 10 gene targets unlocks the opportunity to differentiate and identify mycobacterium at the species level (Macheras et al. 2009). However, this method does not apply to clinical diagnostic laboratories because it may end up with identical findings. But they may play a part in the evaluation and identification of new species. Case-by-case studies are required to confirm the efficiency of Multilocus sequencing approach (MLSA) for the identification of bacterial species. Multilocus sequencing analysis was used to classify genus *M. abscessus* and *Salinivibrio*. These studies suggested that MLSA can replace DNA-DNA hybridization because there is a sufficient degree of similarity between them (López-Hermoso et al. 2017). Multilocus sequencing approach (MLSA) contains higher discrimination between the mycobacteria genomes than *16S rRNA* gene analysis (Liu, Lai, & Shao 2017a). because it uses a set of housekeeping genes (hk) which helps to lower the chances of horizontal gene transfer.

1.3.4 Whole-genome sequencing

Recently, new methods of whole-genome sequencing (WGS) and phylogenomic analysis have emerged for analysing the genetic variations and population studies of bacteria and mycobacteria. The significant advantage of this study is that it allows analysis of multiple genetic regions associated with resistance to antibiotics and disinfectants and the general pathogenicity of strain or species. It can also identify the genetic factors responsible for species diversity and strain specificity (Chan et al. 2012). Although WGS is not widely available in clinical or reference laboratories because of lack of NGS data, but there is no doubt this technology is emerging rapidly (Brown-Elliott & Wallace 2015).

A molecular biology technique, DNA–DNA hybridization, is a gold standard for analysing the genomic similarity between pools of DNA sequences. Genetic distance between two organisms can be calculated using this technique (Woese 1987). Traditionally, a similarity index greater than 70% indicated that the strains being compared belong to distinct species. DNA-DNA hybridization provides more accurate results than 16S rDNA sequencing because this method uses 70% similarity criteria for distinguishing bacterial species (Wayne 1988). However, this method has not been practised much globally due to the laborious nature of pairwise cross-hybridizations, and it is not always easy to perform this technique in routine. Another disadvantage of this method is it requires a fully established central database (Cho & Tiedje 2001). DNA–DNA hybridization is currently performed *in silico* using entirely or partially sequenced genomes, allowing more in-depth classification and identification of bacteria (Castejon et al. 2018).

There are some bioinformatic approaches available to check the genomic similarity of whole genomes. Average nucleotide identity (ANI) is a convenient and straightforward measure of genetic relatedness; it compares the nucleotide sequences of conserved shared genes between genomes. A standalone software fast genome and metagenome distance estimation using MinHash (min-wise independent permutations locality sensitive hashing scheme) is currently available that estimates the fast genome and metagenome distance (Ondov et al. 2016). Minash reduces large sequences and sequences sets to small, representative sketches, from which global mutation distances can be rapidly estimated. MinHash results strongly correlate with alignment-based measures such as the Average Nucleotide Identity (ANI). MinHash distance ≤ 0.05 equates to an ANI of $\geq 95\%$, and this threshold roughly corresponds to a 70 % DNA-DNA (Konstantinidis & Tiedje 2005). In 2021 MinHash was used to study the phylogeny of *Escherichia coli*. According to this study, MinHash reproduces known phylogroups and identified previously uncharacterized phylogroups in *E. coli* species (Abram et al. 2021). Another approach is the Fast Whole-Genome Similarity Estimation (FastANI) available. It estimates the average nucleotide identity between shared genomes (Jain et al. 2018). It calculates the ANI values identical to the alignment-based ANI values for complete and draft quality genomes related to 80 to 100% nucleotide identity range. The cut-off-value equal to or greater than 95% confirms that genomes descend from the same species (Goris et al. 2007).

1.4 Limitations of existing traditional and bioinformatics approaches

As mentioned earlier, traditional phenotyping identification methods are laborious, complex, and time-consuming as they may require many weeks for the proper growth of *mycobacterial* strains as it is

necessary to grow the bacterial culture sufficiently to obtain DNA for molecular analysis such as whole genome sequencing. Sometimes, the findings of these phenotyping methods may lead to inaccurate identification of species because many mycobacterial species share similar morphological and biochemical profiles. Similarly, genotyping methods also have few limitations, like these methods only identify closely related species and require specialized training and the apparatus. Moreover, these method results rely on up-to-date and high-quality databases. Based on nucleotide differences in the *16S rRNA* gene sequence, RGM were classified into three groups: the *M. fortuitum* group consisting of *M. fortuitum*, *M. peregrinum*, *M. houstonense*, *M. neworleansense*, *M. septicum*, *M. mageritense*, *M. mucogenicum* and *M. senegalense*; the *M. chelonae-abscessus* group consisting of *M. abscessus*, *M. chelonae* and *M. immunogenum*; and the *M. smegmatis* group consisting of *M. smegmatis*, *M. wolinskyi* and *M. goodii* (Brown-Elliott & Wallace 2002). However, there has been controversy regarding the taxonomic classification of *M. mucogenicum*, since its biochemical profile and antibiotic-susceptibility pattern were more closely related to those of members of the *M. chelonae-abscessus* group. A *16S rRNA* gene-based phylogenetic tree revealed that *M. mageritense* was more closely related to the *M. smegmatis* group than the *M. fortuitum* group, despite its antibiotic susceptibility biochemical patterns (Brown-Elliott & Wallace 2002). *16S rRNA* gene analysis also could not differentiate the species of *M. intercellulare* and *M. chimaera*; moreover, three strains of *M. peregrinum* were misidentified as *M. septicum* (Schweickert et al. 2008). WGS-based and *16S rRNA* gene-based phylogenetic trees support that *M. leprae* is more closely related to the NTM species *M. avium*. In contrast, other methods support a close sister clade to *M. tuberculosis* (Claeys & Robinson 2018).

Several open access web-based analysis is available such as Bacterial analysis pipeline (Thomsen et al. 2016), Microbial identification and characterization through reads analysis (MICRA) (Caboche et al. 2017), Bioinformatics Bacterial Identification Tool (BIBI) (Devulder et al., 2003), A comprehensive pipeline for whole genome sequence analysis of *Mycobacterium tuberculosis* complex isolates (MTBseq) (Kohl et al. 2018), A bioinformatics whole genome sequencing workflow for clinical *Mycobacterium tuberculosis* complex isolate analysis, validated using a reference collection extensively characterized with conventional methods and in silico approaches (Bogaerts et al. 2021), and a bioinformatics pipeline for *Mycobacterium tuberculosis* sequencing that cleans contaminant reads from sputum samples (Cuevas-Córdoba et al. 2021) are also available. The Bacterial analysis pipeline offers molecular typing tools and resistance and virulence gene predictions, and SNP-based phylogeny. While MICRA uses iterative mapping against reference genomes to identify genes and variations. Both the Bacterial analysis pipeline and MICRA are reference-based pipelines that cannot identify unknown bacterial species. BIBI

combines similarity search tools in the sequence databases and phylogeny display programs. But it does not offer pre-processing steps, and it also uses an online BLAST search tool which may give inaccurate results. MTBseq performs phylogenetic analysis of Illumina whole genome sequence data of *Mycobacterium tuberculosis* complex isolates. It does not cover all mycobacterium species available. A bioinformatics workflow available is also a web-based tool which perform species identification (16S, csb/RD, hsp65) and single nucleotide polymorphism (SNP)-based antimicrobial resistance prediction on *Mycobacterium tuberculosis* complex isolates. A one more bioinformatics pipeline identifies *Mycobacterium tuberculosis* reads from sputum by filtering another microorganism.

In addition, web-server-based analysis depends on the server load and needs a fast and steady internet connection to upload raw data files and requires an up-to-date database. There is also some open-source locally installable software available such as, NGSPanPipe which offers pan-genome identification using a reference sequence (Lambris and Paoletti 2018). Nullarbor produces complete public health microbiology reports from sequenced isolates and supports Illumina paired-end sequencing data from either Illumina or Ion Torrent (Seemann T 2018). However, none of the methods discussed earlier or web servers and software help to identify mycobacterial species.

1.5 Aims and significance of study

Using phenotyping and genotyping approaches to identify mycobacterial species is difficult since mycobacteria have a wide range of characteristics such as survival, growth, permeability, pathogenicity, and antibiotic resistance profiles. Furthermore, these approaches need specialised training as well as the necessary equipment. As a result of the advancement of high-throughput sequencing technology, microbial identification is no longer a difficult task. However, the individual use of these bioinformatics tools can be difficult, especially for those who are new to the field and have no prior understanding of bioinformatics. As a result, different tools must be used to identify mycobacteria using genomic information, when moving from one tool to another during a computational analysis of a mycobacterial genome: (1) pre-processing of the sequence data, (2) *de novo* genome assembly, (3) identification. This means that results will be less accurate when moving from one tool to another. Furthermore, if only one programme was used to do all of the necessary computing steps for this study, the approach might not be as efficient and rapid. Thus, having a pipeline that can run the whole mycobacterial characterisation process at once, without having to pause every time a new bioinformatics tool is required, would make this in-silico technique much easier to undertake in the first place.

As a result, the pipeline was designed to eliminate the inaccuracies that can occur while manually switching from one tool to another by including filtering scripts and automating the whole process. Pipeline script is written with the intention of being a user-friendly tool when used via command-line; that is, the user should be able to understand the manual provided by its developers without having difficulty understanding how to use it; this means that no additional research on how to use it should be required, and the code should not need to be modified. The designed pipeline (ISBaC) will be an open-source script capable of performing a wide range of analyses, beginning with raw data quality check, trimming of low-quality sequences, *de-novo* genome assembly, genome annotation, virulence gene prediction, and finally the identification of the species identity. These tools will provide % similarity data as well as identification of the mycobacterial species in the sample. As a result, with a single query, we may identify mycobacterial species of interest.

The promising way to test the accuracy of the pipeline is to run it with the sample where we already know the identity of the species, so if our pipeline gives the same identity at the end of the pipeline, that means our pipeline is working accurately. So we used a paired-end raw read sequence of *M.chelonae* (SSR10177528) downloaded from the public database European Nucleotide Archive (ENA) as an input. Signal end raw read sequence and assembled whole genome sequence can also be used as input. In the end, we compared ISBaC pipeline results with previously identified strain of *M.chenloae*.

1.6 Objectives

- To create an automated pipeline for the identification and annotation of mycobacterial species.
- To perform statistical analysis to determine the accuracy of the pipeline by comparing the result from the pipeline with prior published data.

Chapter2. Material and methods

2.1 Data mining

First, we have prepared filtered reference *16S rRNA* gene sequences database (that should be 1550bp long) (Clarridge 2004) of 170 mycobacterium species downloaded from National Centre for Biotechnology Information (NCBI) for the *16S rRNA* gene analysis. We have also downloaded five housekeeping gene sequences (*gyrA*, *gyrB*, *ropB*, *groEL*, *recA*) (which are highly conserved, not vulnerable to horizontal gene transfer, long enough to contain phylogenetically useful information and could predict whole genome relationship) (Rong & Huang 2014) from National Centre for Biotechnology Information (NCBI) of 170 mycobacterium species for multilocus sequencing analysis (MLSA) and 170 mycobacterium species reference whole genome sequences from National Centre for Biotechnology Information (NCBI) for the ANI analysis. The list of *Mycobacterium* species has been listed in Appendix 1 and 2.

In order to test the ISBaC pipeline, the paired-end raw sequence of *M.chelonae* ATCC_35752 (SRR4423483), *M.chimaera* CCUG_50989 (SRR10394508), *M.intercellulare* ATCC_13950 (SRR5052607), *M.kansasii* ATCC_12475 (SRR3319297), *M.tuberculosis* ATCC_27294 (SRR786668), *M.ulcerans* ATCC_19423 (SRR6346326), *M.franklinii* DSM_45524 (SRR3605312), *M.salmonihilum* ATCC_13758 (SRR3617054), *M.arupensis* DSM_44942 (SRR5052611) and *M.iranicum* DSM_45541 (SRR10143752) have been downloaded from European Nucleotide Archive (ENA), a public repository of the world's nucleotide sequencing information, covering raw sequencing data (Leinonen et al. 2011). The pipeline is suitable for all sequence data generated from different sequencing platforms for example Illumina, PacBio.

2.2 Raw reads pre-processing and genome assembly

The quality of the raw reads is assessed using FastQC (Andrews 2010). FastQC estimates multiple read quality statistics and calculates the Phred score of each position for the reads. If the observed mean quality is below 27, then a warning is issued and if it is below 20, then fail is given. Next, Trimmomatic (Bolger, Lohse, & Usadel 2014) is used to trim off the poor-quality sequences. Low-quality bases below the threshold quality of 20, lesser than 100 bp length, and adaptors sequences will be trimmed off. Raw reads pre-processing will ensure that the raw sequencing data is unbiased before drawing biological conclusions upon downstream analysis. SPAdes genome assembler (Bankevich et al. 2012) is a standard *de-novo* genome assembler for whole-genome sequencing data

of bacteria and other tiny microbes. The pipeline uses SPAdes to perform genome assembly. SPAdes use multi-sized de Bruijn graphs to construct the assembly graph also detects and removes chimeric reads. Next, for mapping the edges of the assembly graph, distances between the automatically selected *k*-mers (*K21*, *K33*, *K55*, *K77*, *K99*, *K127*) are calculated. In the end, a paired assembly graph is constructed, and the assembler gives contigs as output. SPAdes do not select a *k*-mer length but instead makes a combined assembly (by default) by using multiple pre-selected *k*-mer. Then the quality of assembled genome will be assessed using Quast (Gurevich, Saveliev, Vyahhi, & Tesler 2013). Quast evaluates genome assemblies both with a reference genome, as well as without a reference. QUAST produces numerous reports, outline tables and plots to assist researchers.

2.3 Genome annotation

Functional elements of the assembled genome are identified using Prokka (Seemann 2014). Prokka give many output files like, gene sequence, protein sequence and functional annotation files, which are later used in pipeline. Like .fna file was used to identify the 5-housekeeping gene (*gyrA*, *gyrB*, *rpoB*, *groEL*, and *recA*), and .faa file was used to align query genome against Virulence Factor Database (VFDB) in virulent gene predication (Chen et al. 2005).

2.4 Mycobacteria identification

The *16S rRNA* gene was extracted from the assembled query genome using barnap (Torsten 2014). *16S rRNA* gene sequences of 170 different mycobacterial species were first built as a local database using makeblastdb (Camacho 2008). The query *16S rRNA* gene sequence was aligned using the local *16S rRNA* reference gene database to check the regions of local similarity. The top 30 hits of blastn results were used to draw a bar chart by R using packages like ggplot2, data.table, ggpubr (R Core Team 2021). Furthermore, a similarity matrix based on the top 10 hits was drawn by R using packages like ggplot2, data.table, ggpubr, reshape2, dplyr, tidyr to check the similarity of inter and intraspecies (R Core Team 2021).

The query genome's five housekeeping genes (*gyrA*, *gyrB*, *rpoB*, *groEL*, and *recA*) were identified through gene alignment using the *M. tuberculosis* (H37Rv) housekeeping genes as reference because H37Rv has been utilized broadly in biomedical research and it still provide the backbone for most of TB related research. The reference MLSA database is created based on the presence of the housekeeping genes in the query genome. For example, if only three housekeeping genes are present

in the query genome, then the pipeline will concatenate only those three housekeeping gene files to create the MLSA reference database. Next, the extracted housekeeping gene set was compared with the MLSA database to check the regions of local similarity. The top 30 hits of blastn results were used to draw a bar chart using R (R Core Team 2021). Then again, a database was designed for top 10 hits to check the inter and intraspecies similarity; these similarity values were used to draw a similarity matrix using R (R Core Team 2021).

For whole-genome sequence analysis, a standalone software MinHash estimates the fast genome and metagenome distance (Ondov et al. 2016). First, MinHash sketch of all reference whole genome sequence of 170 mycobacterial species was constructed then query sequence was compared with MinHash sketched references to calculate MinHash distance. The resulting MinHash distances correlate well with ANI. Where $D \approx 1 - \text{ANI}$, another tool, Fast Whole-Genome Similarity Estimation tool (FastANI), was used, which calculates the average nucleotide identity between shared genomes (Jain et al. 2018). First, two files Query_path and Database_path, were created automatically, containing paths to query genome and reference database genome of 170 mycobacterial species, respectively, one per line. Then, ANI values were computed by comparing the query genome with the reference genome. ANI values were graphically represented through heatmap using R (R Core Team 2021).

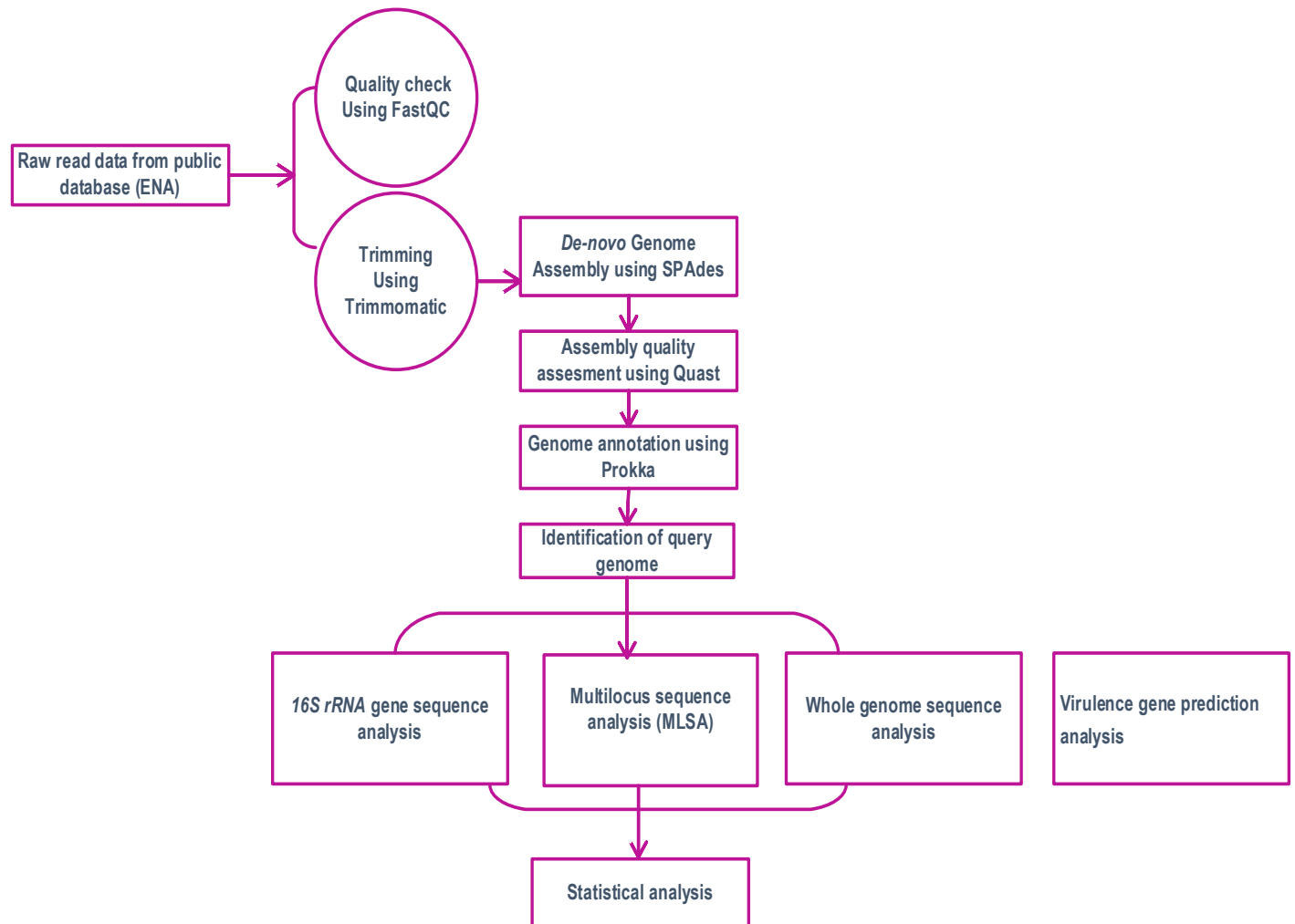
2.5 Virulence gene predication

Virulence genes in the query genome were searched using the Virulence Factor Database (VFDB) (Chen et al. 2005). The annotated protein sequence in the query genome had been blasted against the VFDB database. The aligned protein with coverage and identity of more than 50% was considered to be the homolog to virulence genes. The virulence genes in the query genome will be further compared to the virulence genes of 10 other mycobacterium species (*M.abscessus* ATCC 19977, *M.africanum* GM041182, *M.avium* 104, *M.bovis* AF2122/97, *M.gilvum* PYR-GCK, *M.indicus pranii* MTCC 9506, *M.intracellulare* ATCC 13950, *M.leprae* TN, *M.ulcerans* Agy99 and *M.tuberculosis* H37Rv).

2.6 Statistical analysis

The accuracy of the ISBaC pipeline was calculated using statistical analysis. First, we downloaded raw data of 10 different *M.chelonae* strains (SRR4423483, DRR015959, SRR3617056, SRR3617047, SRR3617048, SRR3617050, SRR3617057, SRR3617058, SRR3617059 and SRR3617060) and 10 different mycobacterium species (*M.chelonae* ATCC_35752 (SRR4423483), *M.chimaera* CCUG_50989 (SRR10394508), *M.intercellulare* ATCC_13950 (SRR5052607), *M.kansasii* ATCC_12475 (SRR3319297), *M.tuberculosis* ATCC_27294 (SRR786668), *M.ulcerans* ATCC_19423 (SRR6346326), *M.franklinii* DSM_45524 (SRR3605312), *M.salmoniohilum* ATCC_13758 (SRR3617054), *M.arupensis* DSM_44942 (SRR5052611) and *M.iranicum* DSM_45541 (SRR10143752). All the downloaded raw data was input into the ISBaC pipeline to get the similarity values for each characterization analysis. The similarity value of the top hit for each run has been recorded. Next, the mean, standard deviation and 95% confidence interval value for the *16S rRNA*, MLSA and ANI analysis were calculated, and the error bars were drawn using IBM Statistical Package for the Social Sciences (SPSS) version 26.0. Error bars showing the range of mean percentage identity values obtained from ISBaC, helps to understand either the range captures the threshold values for *16S rRNA* gene analysis, MLSA and ANI analysis in repeated samples.

2.7 Flowchart of Experimental Method



Chapter3. Results

To use ISBaC, users need to install the required software. First, the user is required to run the SoftwareInstallation.pl Perl script. The Perl script is downloaded for the installation of all the required software's into the users' system. The list of software required by ISBaC are FastQC, Trimmomatic, SPAdes, Quast, barrmap, ncbi-blast++, Prokka, R, seqkit, faSomeRecords, MinHash and FastANI with their download links are shown in (Table 3.1).

Table 3.1: List of software's with link

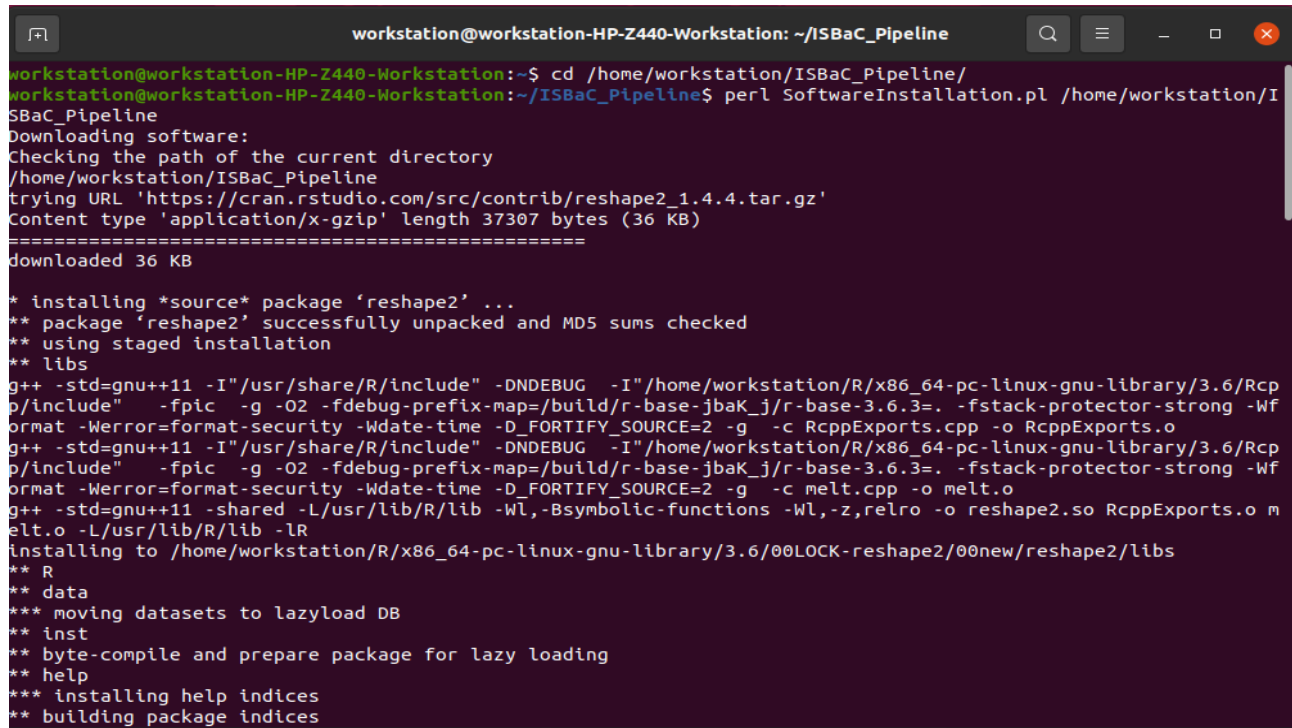
| List of software's | Link to the software's |
|------------------------|---|
| FastQC (v0.11.9) | http://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.11.9.zip |
| Trimmomatic (v0.33) | http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trimmomatic-0.33.zip |
| SPAdes (v3.12.0) | http://cab.spbu.ru/files/release3.12.0/SPAdes-3.12.0-Linux.tar.gz |
| Quast (v5.0.1) | https://sourceforge.net/projects/quast/files/quast-5.0.1.tar.gz |
| Barrmap (v0.9) | https://github.com/tseemann/barrnap/archive/0.9.tar.gz |
| ncbi blast++ (v2.13.0) | https://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/ncbi-blast-2.13.0+-x64-linux.tar.gz |
| Prokka (v1.14.5) | https://github.com/tseemann/prokka/archive/refs/tags/v1.14.5.tar.gz |
| R (v4.0.3) | https://cran.r-project.org/src/base/R-4/R-4.0.3.tar.gz |
| Seqkit (v0.14.0) | https://github.com/shenwei356/seqkit/releases/download/v0.14.0/seqkit_linux_386.tar.gz |
| faSomeRecords | https://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/faSomeRecords |
| MinHash (v2.3) | https://github.com/marbl/Mash/releases/download/v2.3/mash-Linux64-v2.3.tar |
| FastANI (v1.32) | https://github.com/ParBLiSS/FastANI/releases/download/v1.32/fastANI-Linux64-v1.32.zip |

Users need to provide the path to the desired working directory. The example command to run the Perl script is:

```
perl SoftwareInstallation.pl /home/workstation/ISBaC_Pipeline
```

“/home/workstation/ISBaC_Pipeline “is the directory where the user would like the software to be installed. The whole installation process takes around 9 minutes to complete with 4 core CUPs and

128gb RAM, it can be faster with higher operating system specifications. The user is only required to run the SoftwareInstallation.pl script once (Figure 3.1).



```

workstation@workstation-HP-Z440-Workstation: ~/ISBaC_Pipeline
workstation@workstation-HP-Z440-Workstation:~$ cd /home/workstation/ISBaC_Pipeline/
workstation@workstation-HP-Z440-Workstation:~/ISBaC_Pipeline$ perl SoftwareInstallation.pl /home/workstation/I
SBaC_Pipeline
Downloading software:
Checking the path of the current directory
/home/workstation/ISBaC_Pipeline
trying URL 'https://cran.rstudio.com/src/contrib/reshape2_1.4.4.tar.gz'
Content type 'application/x-gzip' length 37307 bytes (36 KB)
=====
downloaded 36 KB

* installing *source* package 'reshape2' ...
** package 'reshape2' successfully unpacked and MD5 sums checked
** using staged installation
** libs
g++ -std=gnu++11 -I"/usr/share/R/include" -DNDEBUG -I"/home/workstation/R/x86_64-pc-linux-gnu-library/3.6/Rcp
p/include" -fpic -g -O2 -fdebug-prefix-map=/build/r-base-jbaK_j/r-base-3.6.3=. -fstack-protector-strong -Wf
ormat -Werror=format-security -Wdate-time -D_FORTIFY_SOURCE=2 -g -c RcppExports.cpp -o RcppExports.o
g++ -std=gnu++11 -I"/usr/share/R/include" -DNDEBUG -I"/home/workstation/R/x86_64-pc-linux-gnu-library/3.6/Rcp
p/include" -fpic -g -O2 -fdebug-prefix-map=/build/r-base-jbaK_j/r-base-3.6.3=. -fstack-protector-strong -Wf
ormat -Werror=format-security -Wdate-time -D_FORTIFY_SOURCE=2 -g -c melt.cpp -o melt.o
g++ -std=gnu++11 -shared -L/usr/lib/R/lib -Wl,-Bsymbolic-functions -Wl,-z,relro -o reshape2.so RcppExports.o m
elt.o -L/usr/lib/R/lib -lR
installing to /home/workstation/R/x86_64-pc-linux-gnu-library/3.6/00LOCK-reshape2/00new/reshape2/libs
** R
** data
*** moving datasets to lazyload DB
** inst
** byte-compile and prepare package for lazy loading
** help
*** installing help indices
** building package indices

```

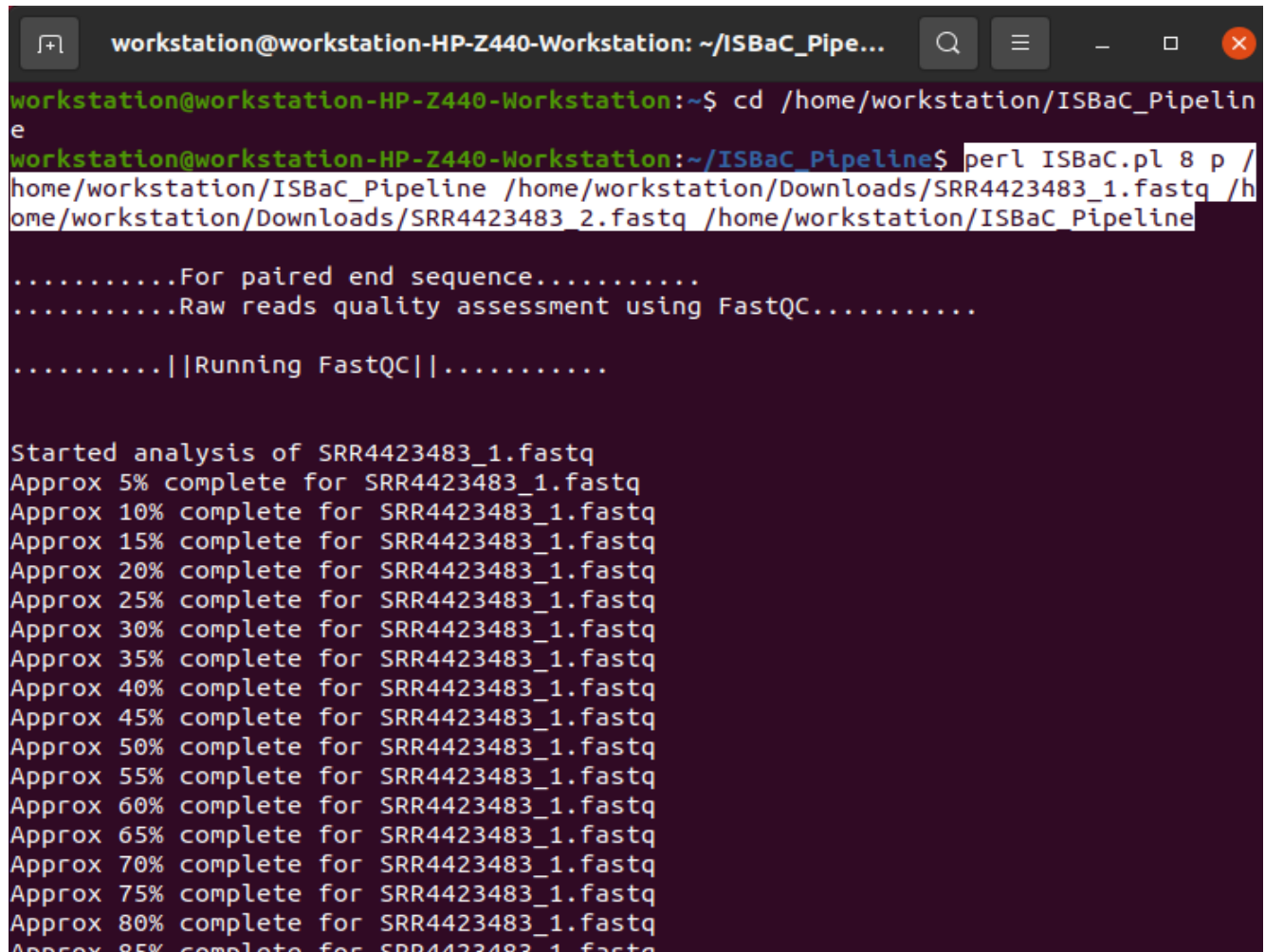
Figure 3.1: Calling SoftwareInstallation.pl script that automatically install all required software's of ISBaC pipeline

After completed the software installation, we can begin ISBaC by running the ISBaC.pl Perl script. ISBaC required only one command to execute the whole pipeline analysis. The example of the command to run ISBaC is shown below:

“perl ISBaC.pl 8 p /home/workstation/ISBaC_Pipeline /home/workstation/Downloads/SRR4423483_1.fastq /home/workstation/Downloads/SRR4423483_2.fastq /home/workstation/ISBaC_Pipeline”

To start with pipeline user, need to call ISBaC.pl script which is followed by few arguments. The first argument indicated the number of threads (“8”), Next is the type of raw reads sequencing (“s” indicated single end and “p” indicated paired-end sequences), directory to the installed software, path to the input raw data (/home/workstation/Downloads/SRR4423483_1.fastq /home/workstation/Downloads/SRR4423483_2.fastq) and lastly is the output directory (/home/workstation/ISBaC_Pipeline) (Figure 3.2). ISBaC takes around 48 minutes to 60 minutes to complete the whole process with

4 core CUPs and 128gb RAM, it can be faster with higher operating system specifications. The example of the output results can be seen in Figure 3.3.



```
workstation@workstation-HP-Z440-Workstation: ~/ISBaC_Pipe...
workstation@workstation-HP-Z440-Workstation:~/ISBaC_Pipeline$ perl ISBaC.pl 8 p /
/home/workstation/ISBaC_Pipeline /home/workstation/Downloads/SRR4423483_1.fastq /h
ome/workstation/Downloads/SRR4423483_2.fastq /home/workstation/ISBaC_Pipeline

.....For paired end sequence.....
.....Raw reads quality assessment using FastQC.....

.....||Running FastQC||.....

Started analysis of SRR4423483_1.fastq
Approx 5% complete for SRR4423483_1.fastq
Approx 10% complete for SRR4423483_1.fastq
Approx 15% complete for SRR4423483_1.fastq
Approx 20% complete for SRR4423483_1.fastq
Approx 25% complete for SRR4423483_1.fastq
Approx 30% complete for SRR4423483_1.fastq
Approx 35% complete for SRR4423483_1.fastq
Approx 40% complete for SRR4423483_1.fastq
Approx 45% complete for SRR4423483_1.fastq
Approx 50% complete for SRR4423483_1.fastq
Approx 55% complete for SRR4423483_1.fastq
Approx 60% complete for SRR4423483_1.fastq
Approx 65% complete for SRR4423483_1.fastq
Approx 70% complete for SRR4423483_1.fastq
Approx 75% complete for SRR4423483_1.fastq
Approx 80% complete for SRR4423483_1.fastq
Approx 85% complete for SRR4423483_1.fastq
```

Figure 3.2: ISBaC pipeline input options.

3.1 Raw reads pre-processing

Inside the “1.FastQC_Results” folder is the result of the processed raw read. We checked the quality of the paired-end raw data of *M. chelonae* ATCC_35752 (SRR4423483) using FastQC. The FastQC results show that a total of 922,725 sequences have a length that lies between 36-301 bp. The GC percentage is 63, and no sequences was flagged as poor quality. The per base sequence quality boxplot of *M. chelonae* shows the average range of the sequence’s quality values across the read. The higher the quality score, the better the base call. The graph divides the Y-axis into excellent quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). In *M. chelonae* the quality of calls falls into red towards the end of a read.

Once the quality of the raw reads has been checked, the pipeline will proceed with trimming poor-quality bases using Trimmomatic. The trimmonatic results will be stored in the folder

“2.Trimmomatic_Results” (Figure 3.3). Although the input sequence reads of *M. chelonae* are of average quality, the subsequent steps in the analysis will benefit from pre-processing, which includes trimming of adaptors, low-quality bases. After completing the adapter trimming and quality filtering steps, the sequencing reads of *M. chelonae* were assessed for the quality a second time. The per base sequence quality plot of trimmed reads shows excellent improvement in quality as the all the low-quality bases below the threshold quality of 20 and lesser than 100 bp length were trimmed off from both reads (R1 & R2). It took around two to three minutes for a single end sequence and five to six minutes for the paired-end sequence with 4 core CPUs and 128Gb RAM to perform these steps. The per base sequence quality plots before and after trimming are shown in (Figure 3.4A, 4B).

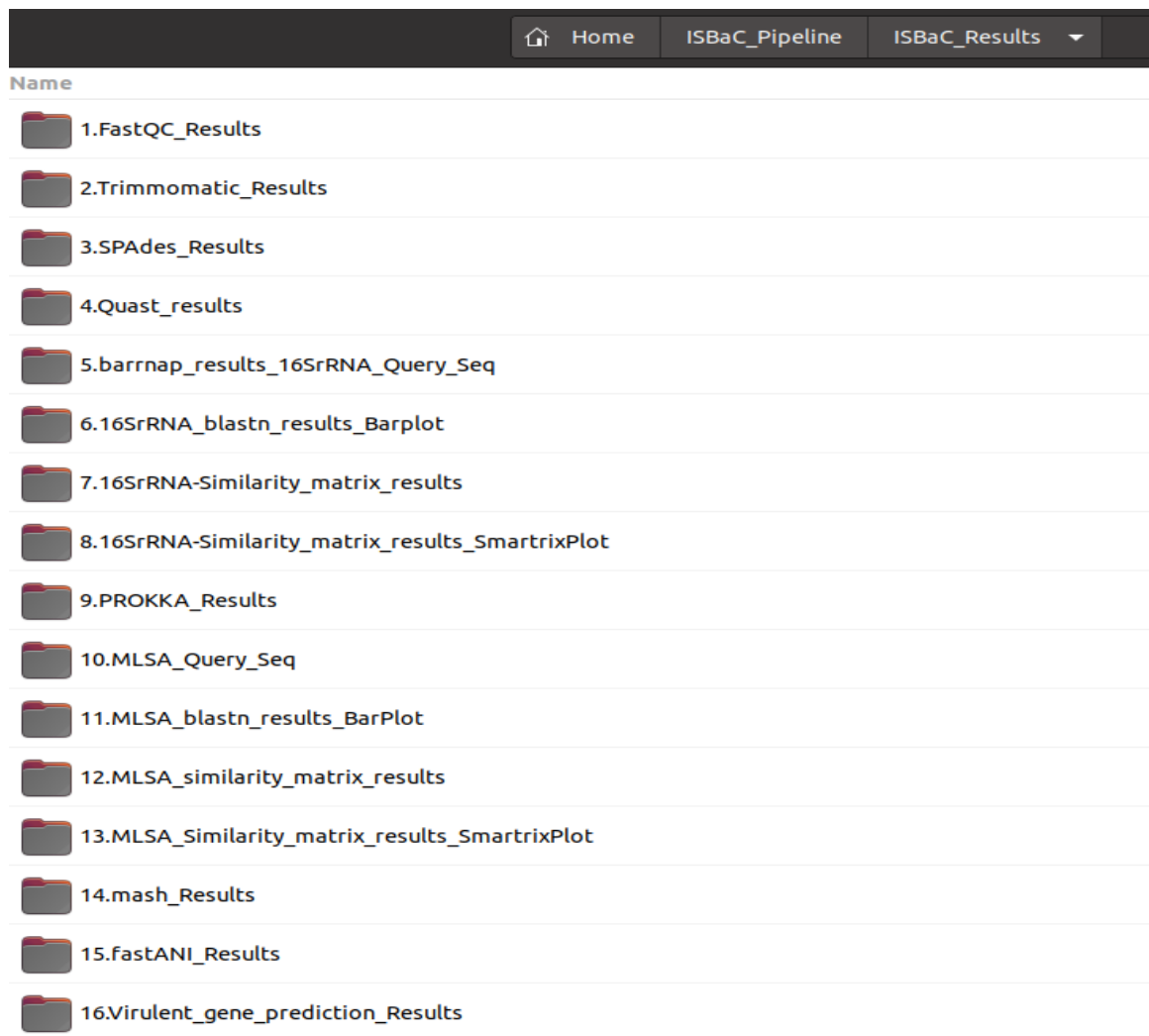
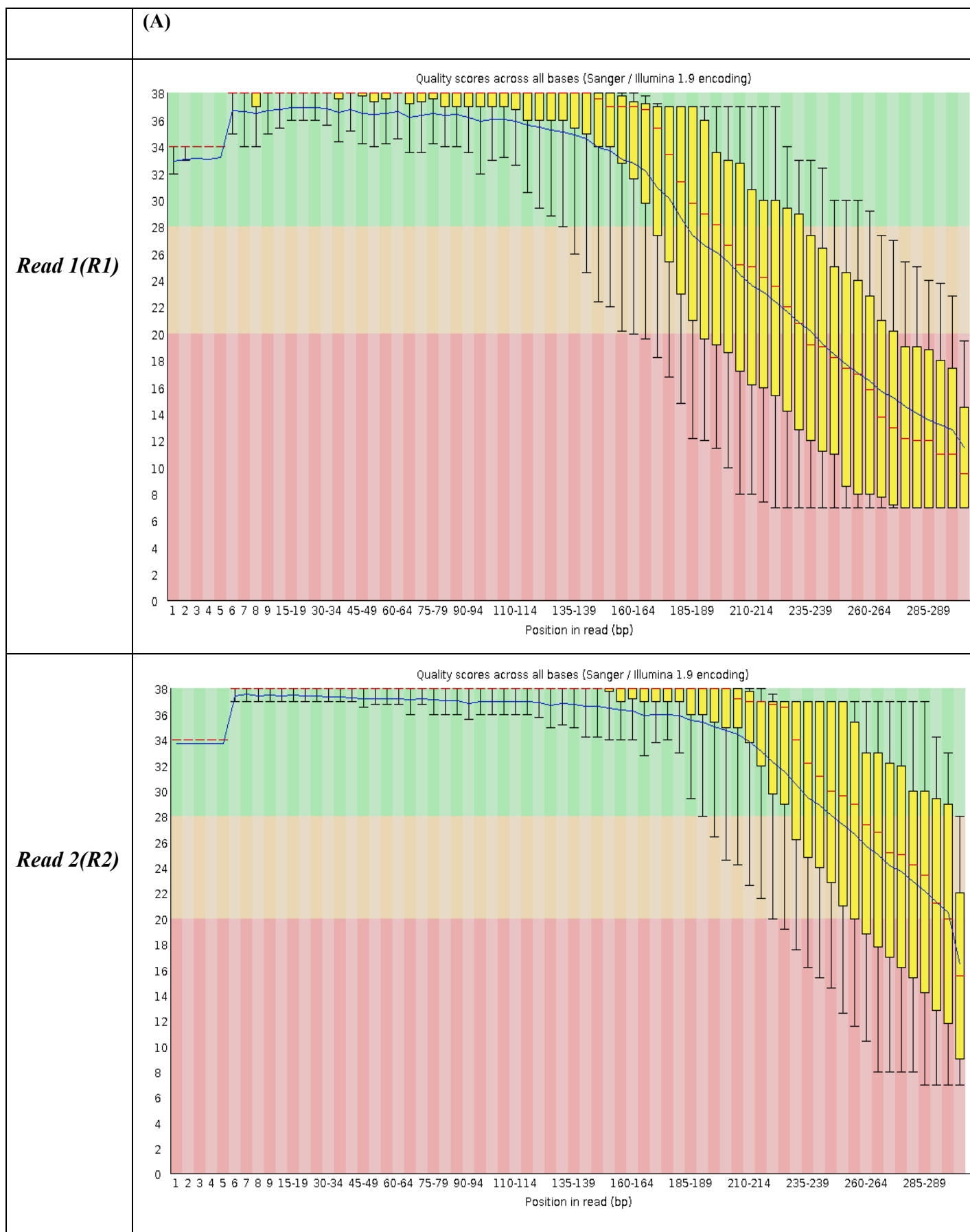


Figure 3.3: Final output directory folder containing all the ISBaC results.



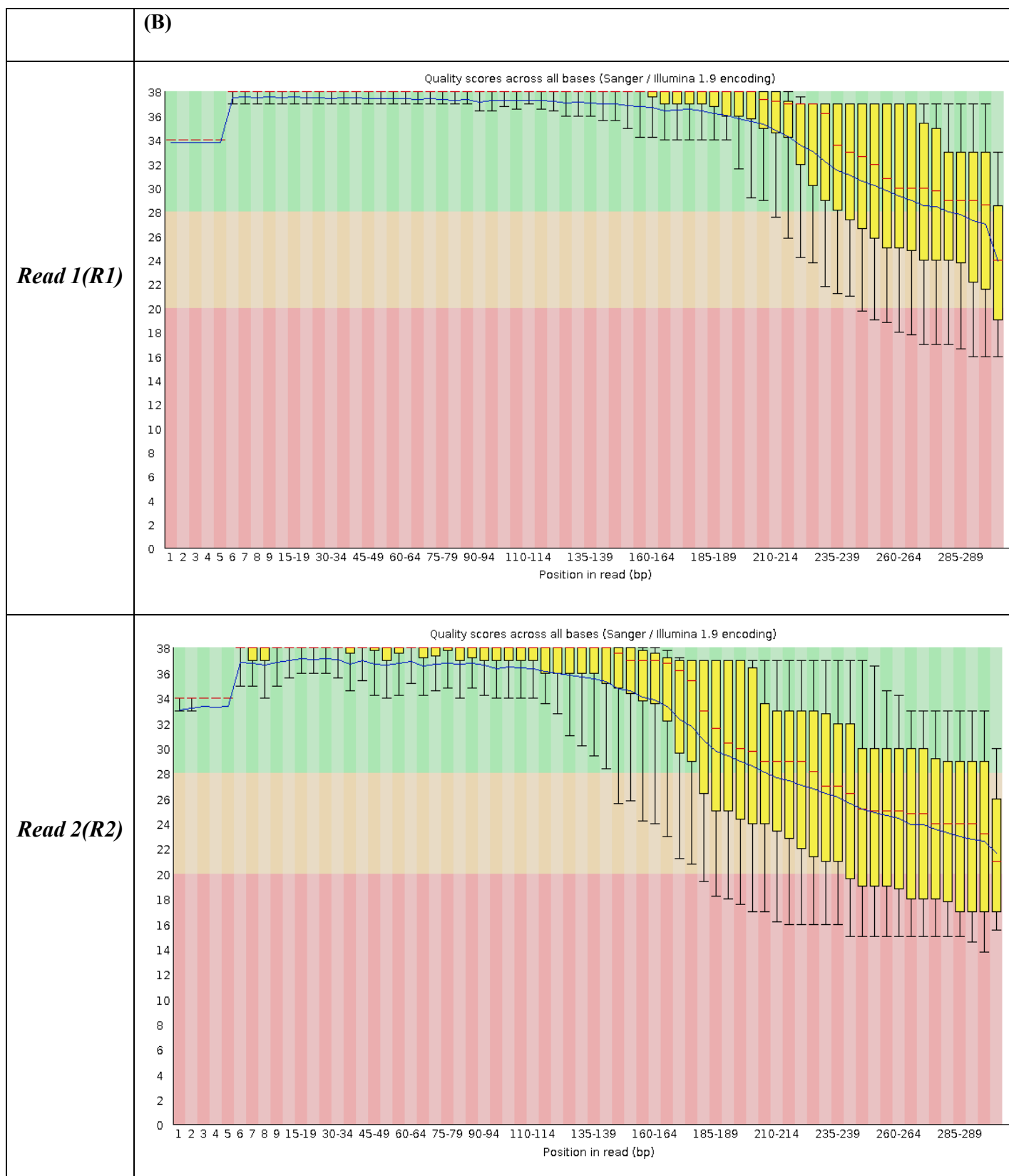


Figure 3.4: Per Base Sequence Quality in FastQC report.

(A) Per base quality plot for paired end raw data of *M. chelonae* (R1-R2) before trimming and filtering. (B) Per base quality plot for paired end raw data of *M. chelonae* (R1-R2) after trimming and filtering

3.2 Genome assembly and annotation

Next, inside the “3.SPAdes_Results” folder (Figure 3.3) is the genome assembly result using SPAdes. After the pre-processing step, the paired-end sequence of *M. chelonae* was assembled using the SPAdes assembler. The genome assembly analysis took 8 to 10 minutes with 4 core CPUs and 128GB RAM. The quality of the assembled genome was assessed using Quast. The Quast results were stored in “4.Quast_Results” (Figure 3.3). The Quast report shows that number of contig is 15 for the assembled genome. The assembled genome size is 5,035,181 bp. The largest contig size is 1,148,810 bp. The value of N50 is 981,998. The GC value is 63.9%, and there is zero number of mismatches and misassemblies.

The assembled draft genome from ISBaC has been compared with *M. chelonae* ATCC_35752 in the NCBI, and our assembled genome has better quality in terms of the N50 value. The assembled genome contains an N50 value of 981,998 compared with the N50 value of 936,739 for *M. chelonae* ATCC_35752 in the NCBI (Hasan et al. 2015). Next, the assembled genome was annotated using Prokka. The annotation results were stored in the “9.PROKKA_Results” folder (Figure 3.3). Prokka result contains predicted gene sequence in FASTA format with .fna extension, predicted protein sequences in FASTA format with .faa extension, master functional annotation file with .gff extension and a text file with statistics of annotated features.

3.3 Mycobacterial identification using the *16S rRNA* gene analysis

The *16S rRNA* gene sequence from the query genome is aligned to the locally created reference database of mycobacterium species to identify the regions of local similarity. The threshold level of 98% (Beye, Fahsi, Raoult, & Fournier 2018) indicates the identity of the same species. The BLAST result is stored in the “6.16SrRNA_blastn_results_Barplot” folder (Figure 3.3). The BLAST result shows that both *M. chelonae* and *M. franklini* have 100% percentage identity with the query sequence (Nogueira et al. 2015). Following that, *M. saopaulense* has 99.78%, *M. abscessus* has 99.73%, *M. abscessus subsp bolletii* has 99.73%, and *M. salmoniphilum* has 99.66%. These 6-identified species are all above the threshold level of 98% because many mycobacterial species share highly similar *16S rRNA* gene sequences (Figure 3.5), so *16S rRNA* gene analysis cannot differentiate among some closely related mycobacterium species (Turenne, Tschetter, Wolfe, & Kabani 2001). Hence, from the *16S rRNA* gene analysis, the query genome can be any of the 6-identified species. Again, to check the inter and intraspecies similarity, we selected the top 10 hits from previous results and drew a similarity matrix of those values using R.

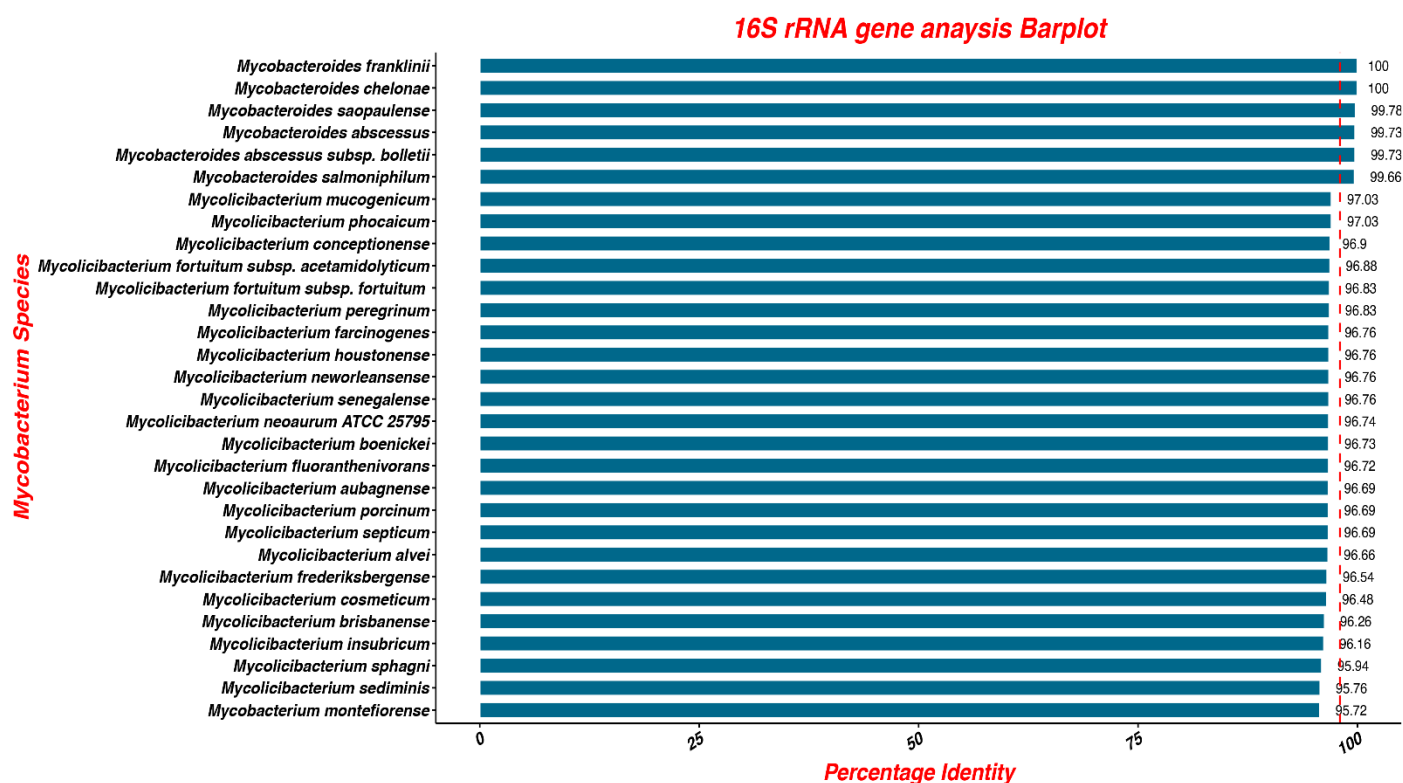


Figure 3.5: Bar chart of top 30 mycobacterial species

Bar chart of blastn results showing the top 30 hits with their percentage identity values after comparing the query sequence of *M. chelonae* with the local database. The red line indicates the threshold value of 98%, which measures species-level mycobacterial identity using *16S rRNA* gene analysis.

3.4 Mycobacterial Identification using Multilocus sequence analysis (MLSA)

The 5-housekeeping genes (*gyrA*, *gyrB*, *rpoB*, *groEL*, and *recA*) in the query genome were identified. The extracted housekeeping genes were further compared with the MLSA reference genes database to check the regions of local similarity with 170 other mycobacterial species. The output file was stored in folder “11.MLSA_blastn_results_BarPlot” (Figure 3.3). The threshold level of 97% indicates the identity of species (Liu, Lai, & Shao 2017a). Our MLSA result shows that our query genome has the highest similarity to *M. chelonae* (100%), followed by *M. saopaulense* (92.21%), *M. franklinii* (91.38%) and *M. salmoniphilum* (91.34%). The result shows that only *M. chelonae* is above the threshold level of 97% (Figure 3.6). Thus, the MLSA result had identified the identity of the query genome as *M. chelonae*. The result has shown that ISBaC can identify the specie correctly with the MLSA sequence analysis. The similarity matrix file is stored in the folder “13.MLSA_Similarity_matrix_results_SmatrixPlot” (Figure 3.3). The result shows the percentage identity values of every species against all ten hits (Figure 3.7). It takes 6 to 9 minutes to perform these steps.

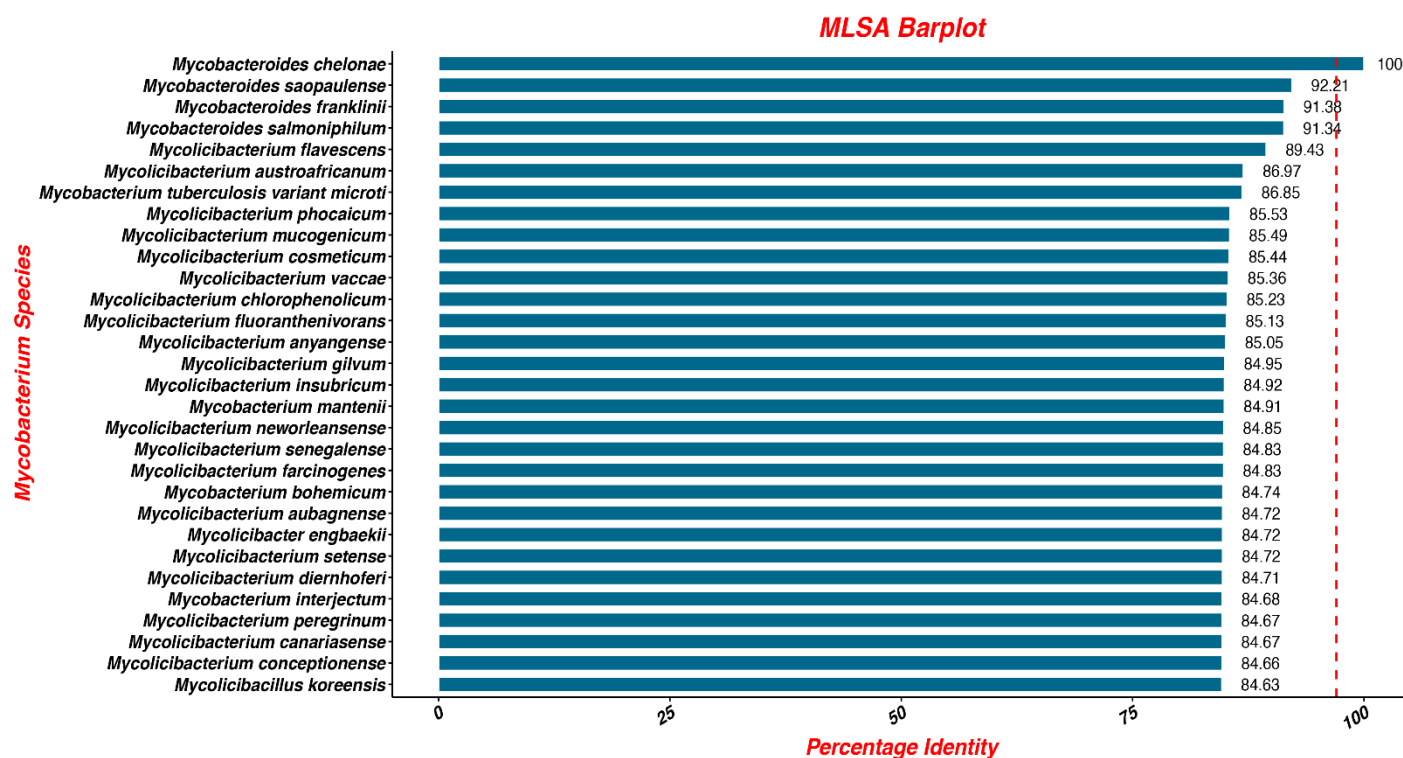


Figure 3.6: Bar chart of top 30 mycobacterial species

Bar chart of blastn results showing the top 30 hits with their percentage identity values after comparing the query sequence of *M. chelonae* with the local database. The red line indicates the threshold value of 97%, which measures species-level mycobacterial identity using multilocus sequence analysis.

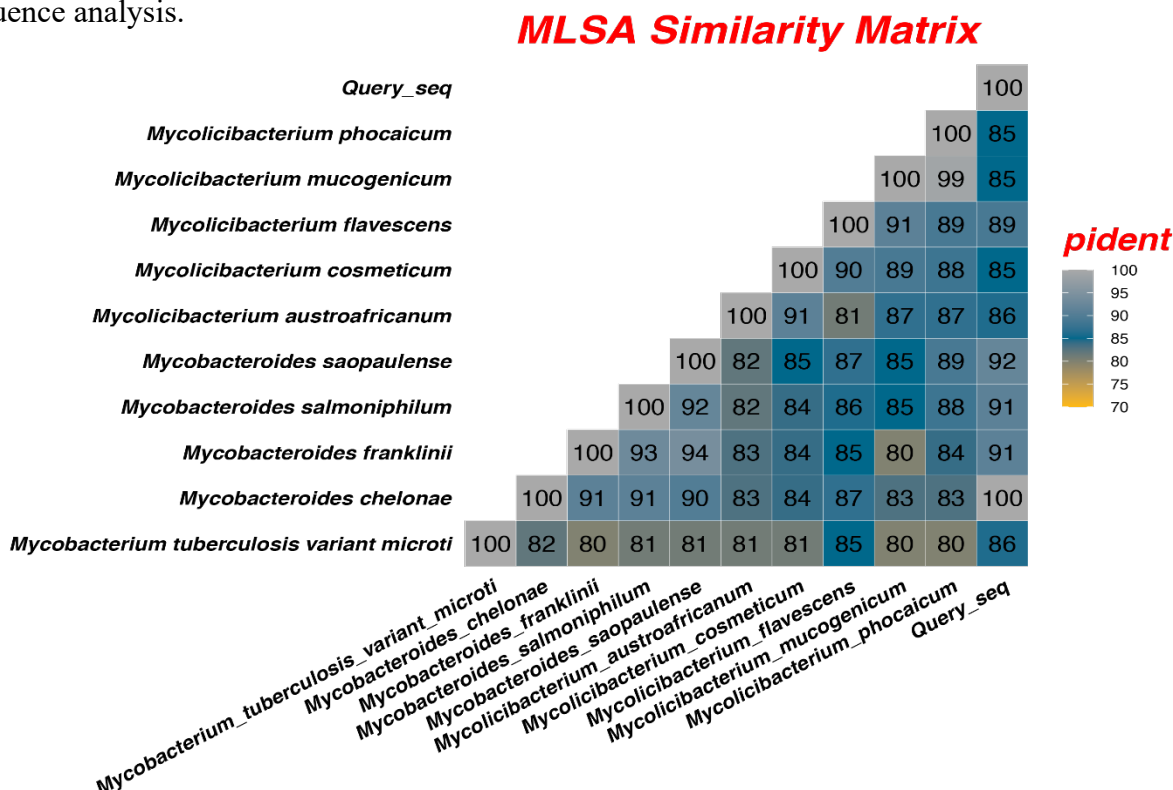


Figure 3.7: Similarity matrix of top 10 mycobacterial species

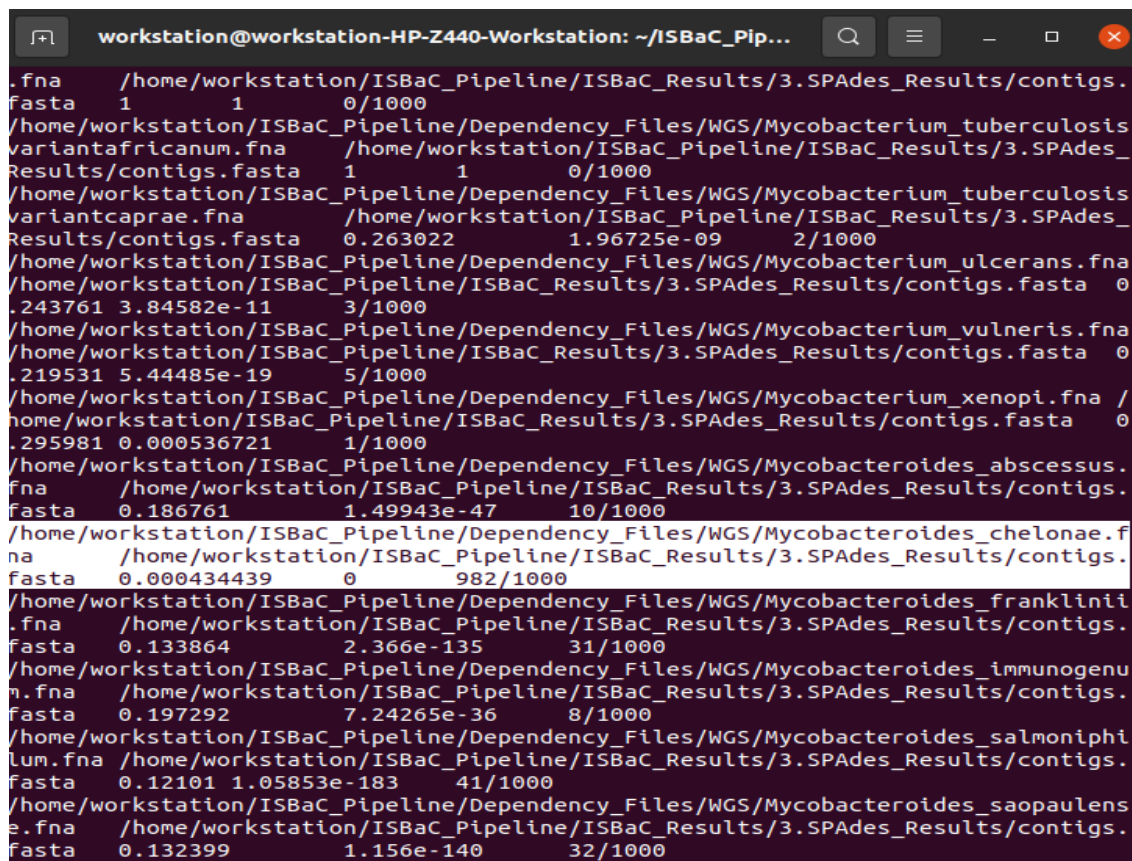
The similarity matrix of the top 10 hits of blastn results indicates the inter and intraspecies level percentage identity values, ranging from 70 to 100.

3.5 Mycobacterial Identification using whole genome sequence

ISBaC included 2 softwares to identify the identity of the mycobacterial species using whole-genome sequence. The 2 softwares are Mash and FastANI.

3.5.1 Mash: fast genome and metagenome distance estimation using MinHash

Mash estimates the distance from the query sequence to each database reference sequence (Ondov et al. 2016). First MinHash sketch (clustering of reference genome) of all the genome sequence of 170 reference mycobacterial species was constructed. Next, the query genome sequence was compared with the reference MinHash sketched to calculate the Mash distance. The results of MinHash were in tab-delimited lists of Reference-ID, Query-ID, Mash-distance, P-value, and Matching-Hashes stored in folder “14.mash_Results” (Figure 3.3). These MinHash and representative sketches are ANI and approximate one minus MinHash distance...eg $1 - 0.000434439 = 0.999565561 * 100 = 99.95\%$ (Figure 3.8).



```

workstation@workstation-HP-Z440-Workstation: ~/ISBaC_Pip...
.fna /home/workstation/ISBaC_Pipeline/ISBaC_Results/3.SPAdes_Results/contigs.
fasta 1 1 0/1000
/home/workstation/ISBaC_Pipeline/Dependency_Files/WGS/Mycobacterium_tuberculosis
variantafricanum.fna /home/workstation/ISBaC_Pipeline/ISBaC_Results/3.SPAdes_
Results/contigs.fasta 1 1 0/1000
/home/workstation/ISBaC_Pipeline/Dependency_Files/WGS/Mycobacterium_tuberculosis
variantcaprae.fna /home/workstation/ISBaC_Pipeline/ISBaC_Results/3.SPAdes_
Results/contigs.fasta 0.263022 1.96725e-09 2/1000
/home/workstation/ISBaC_Pipeline/Dependency_Files/WGS/Mycobacterium_ulcerans.fna
/home/workstation/ISBaC_Pipeline/ISBaC_Results/3.SPAdes_Results/contigs.fasta 0
.243761 3.84582e-11 3/1000
/home/workstation/ISBaC_Pipeline/Dependency_Files/WGS/Mycobacterium_vulneris.fna
/home/workstation/ISBaC_Pipeline/ISBaC_Results/3.SPAdes_Results/contigs.fasta 0
.219531 5.44485e-19 5/1000
/home/workstation/ISBaC_Pipeline/Dependency_Files/WGS/Mycobacterium_xenopi.fna /
home/workstation/ISBaC_Pipeline/ISBaC_Results/3.SPAdes_Results/contigs.fasta 0
.295981 0.000536721 1/1000
/home/workstation/ISBaC_Pipeline/Dependency_Files/WGS/Mycobacteroides_abscessus.
fna /home/workstation/ISBaC_Pipeline/ISBaC_Results/3.SPAdes_Results/contigs.
fasta 0.186761 1.49943e-47 10/1000
/home/workstation/ISBaC_Pipeline/Dependency_Files/WGS/Mycobacteroides_chelonae.f
na /home/workstation/ISBaC_Pipeline/ISBaC_Results/3.SPAdes_Results/contigs.
fasta 0.000434439 0 982/1000
/home/workstation/ISBaC_Pipeline/Dependency_Files/WGS/Mycobacteroides_franklinii
.fna /home/workstation/ISBaC_Pipeline/ISBaC_Results/3.SPAdes_Results/contigs.
fasta 0.133864 2.366e-135 31/1000
/home/workstation/ISBaC_Pipeline/Dependency_Files/WGS/Mycobacteroides_immunogenu
n.fna /home/workstation/ISBaC_Pipeline/ISBaC_Results/3.SPAdes_Results/contigs.
fasta 0.197292 7.24265e-36 8/1000
/home/workstation/ISBaC_Pipeline/Dependency_Files/WGS/Mycobacteroides_salmoniphi
lum.fna /home/workstation/ISBaC_Pipeline/ISBaC_Results/3.SPAdes_Results/contigs.
fasta 0.12101 1.05853e-183 41/1000
/home/workstation/ISBaC_Pipeline/Dependency_Files/WGS/Mycobacteroides_saopaulens
e.fna /home/workstation/ISBaC_Pipeline/ISBaC_Results/3.SPAdes_Results/contigs.
fasta 0.132399 1.156e-140 32/1000

```

Figure 3.8: MinHash Results.

MinHash result of *M. chelonae* which shows Mash distance of 0.000434439.

3.5.2 Fast whole genome similarity estimation tool (FastANI)

FastANI calculates the average nucleotide identity between genomes (Jain et al. 2018). ANI values are computed by comparing the whole genome sequence with the reference genome sequences. The output of FastANI is stored in the folder “15.fastANI_Results” (Figure 3.3). The result shows that the estimated ANI value between the query genome sequence to *M. chelonae* reference genome is 99.9882%. The second highest match is *M. salmoniphilum* which is 87.49%. Strains with ANI value of $\geq 95\%$ will be considered as the same species (Konstantinidis & Tiedje 2005). Thus, our result show that ISBaC identify the identity of the query genome correctly as *M. chelonae*.

ISBaC has shown the top 30 hits with the highest ANI values through a heatmap. The colour in the heatmap varies according to the percentage ANI values among the mycobacterium species, and *M.chelonae* have the highest percentage identity value of 99%, and it shows green colour (Figure 3.9).

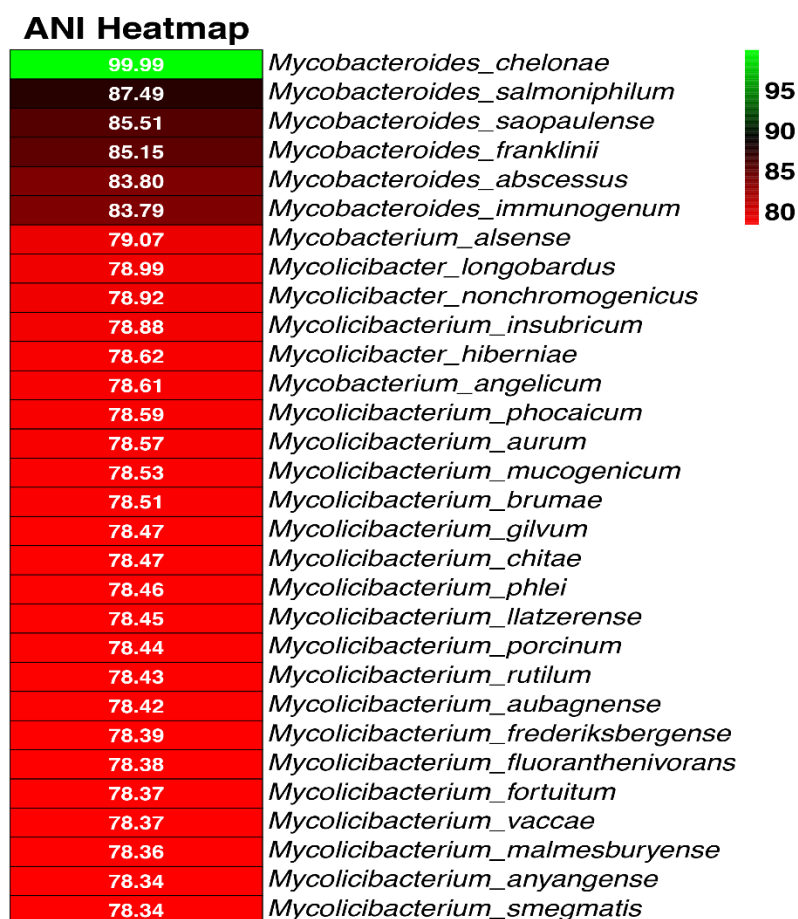


Figure 3.9: FastANI Heatmap.

The heatmap is showing that *M. chelonae* have the highest percentage identity value. Genome with ANI values range from 75 to 85 show red colour, genome with ANI values ranging from 85 to 95 show black colour and lastly genome with ANI values ranging from 95 to 100 show green colour in the heatmap.

3.6 Virulence gene predication

ISBaC also predicts the virulence profile of the query genome. The annotated protein sequences of the query genome from Prokka (SRR786668) were queried using the VFDB database. Result showed that a total of 267 non-redundant virulence genes were predicted across the genome. These virulence genes can be categorized into functional categories including amino acid and purine metabolism, anaerobic respiration and stress adaptation, lipid and fatty acid metabolism, phagosome arresting, culture filtrate proteins, transcriptional regulators, mammalian cell entry operons, cell envelope proteins and metal transporter proteins as shown in Table 3.2 (Ripoll et al. 2009).

Table 3.2: List of functional categories of virulence genes found in query genome

| Functional categories of virulence genes | Virulence genes found in query genome | Function of virulence genes found in query genome |
|---|--|--|
| Amino acid and purine metabolism | <i>glnA1, leuD, lysA, proC, purC, trpD</i> | These genes encode for enzymes that helps in biosynthesis of some amino acids and purines. |
| Anaerobic respiration and stress adaptation | <i>narX, narG, narH, narI, narK2, katG, sodC, soda, ompA</i> | These genes encode for enzymes superoxide dismutases and catalases and these enzymes are crucial for the body's response to various external oxidative stresses. |
| Lipid and fatty acid metabolism | <i>icl, lipE, panC, panD, sapM, plcA, plcB, plcC, plcD, mqtC</i> | These genes help in modulation of lipid biosynthesis in mycobacteria. |
| Phagosome arresting | <i>PE PGRS30, ptpA, mpa</i> | These genes are involved with the arrest of phagosome trafficking. |

| | | |
|------------------------------------|--|--|
| Culture filtrate proteins | <i>hspX, fbpA, fbpB, fbpC, eis, pknG, secA2, PE35, PPE68, eccA1, eccB1, eccCa1, eccCb1, eccD1, eccE1, espA, espB, espC, espD, espE, espF, espG1, espH, espI, espJ, espK, espL, espR, esxA, esxB, mycP1, PE36, PPE69, eccA2, eccB2, eccC2, eccD2, eccE2, espG2, esxC, esxD, mycP2, PE5, PPE4, eccA3, eccB3, eccC3, eccD3, eccE3, espG3, esxG, esxH, eccB4, eccC4, eccD4, esxT, esxU, mycP4, PE18, PE19, PPE25, PPE26, PPE27, PPE41, cyp143, eccA5, eccB5, eccCa5, eccCb5, eccD5, eccE5, esxM, esxN, mycP5, ahpC</i> | <p>These proteins are expected to be exposed to the environment in which bacteria grows. They encode for the enzymes that degrades ROIs and important for survival of mycobacteria during infection.</p> |
| Transcriptional Regulators | <i>devR/dosR, devS, mosR, mprA, mprB, phoP, phoR, prrA, regX3, senX3, sigA/rpoV, sigD, sigE, sigF, sigH, sigL, sigM, whiB3, lpqH</i> | <p>These genes encode for transcriptional regulators that control the transcription of many genes. By inactivating these regulatory genes through a targeted mutational approach, we can uncover some virulence factors.</p> |
| Mammalian cell entry (mce) operons | <i>Mce1 (mce1B, (mce1C, (mce1D, (mce1E, (mce1F) Mce2 (mce2A, mce2B, mce2C, mce2D, mce2E, mce2F) Mce3 (mce3A, mce3B, mce3C, mce3D, mce3E, mce3F) Mce4 (mce4A, mce4B, mce4C, mce4D, mce4E, mce4F) Mce5</i> | <p>These proteins enable mycobacteria to enter mammalian cells and survive in the macrophages.</p> |

| | | |
|----------------------------|--|---|
| | <p>(<i>mce5A, mce5B, mce5C, mce5D, mce5E, mce5F</i>) <i>Mce6</i> (<i>mce6A, mce6B, mce6C, mce6D, mce6E, mce6F</i>) <i>Mce7</i> (<i>mce7A, mce7B, mce7C, mce7D, mce7E, mce7F</i>) <i>Mce8</i> (<i>mce8A, mce8B, mce8C, mce8D, mce8E, mce8F</i>) <i>Mce9</i> (<i>mce9A, mce9B, mce9C, mce9D, mce9E, mce9F</i>)</p> | |
| Cell envelope proteins | <p><i>caeA, erp, fad23, fadE5, fnt, gap, gtfI, gtf2, mbtH, mmpL4a, mmpL4b, mmpS4, mps1, papA3, pe, pks, rmlA, rmlB, hbhA, lprG, mmaA4, cmaA2, adhD, chp, lipR, mymA, sadH, tgs4, drrC, fadD22, fadD26, fadD28, fadD29, lppx, lppx, mas, mmpL7, papA5, pks15/1, ppsA, ppsB, ppsC, ppsD, ppsE, tesA, kefB, pcaA, mmpL8, papA1, papA2, pks2, stf0, chp1, fad23, icl2, lpqY, sap, sugA, sugB, sugC, ctpV</i></p> | <p>These genes encode for surface proteins, and they assist in the adhesion of mycobacteria to the surface and promotes their entry into the host cell.</p> |
| Metal transporter proteins | <p><i>mceA, irtA, irtB, exiT, fxbA, fxbBC, fxuA, fxuB, fxuC, fxuD, mmpL11, mmpl3, ideR, fadD33, fadE14, mbtA, mbtB, mbtC, mbtD, mbtE, mbtF, mbtG, mbtH, mbtI, mbtJ, mbtK, kasB</i></p> | <p>These genes encode for proteins that are involved in iron and magnesium acquisition and causes the attenuation of virulence</p> |

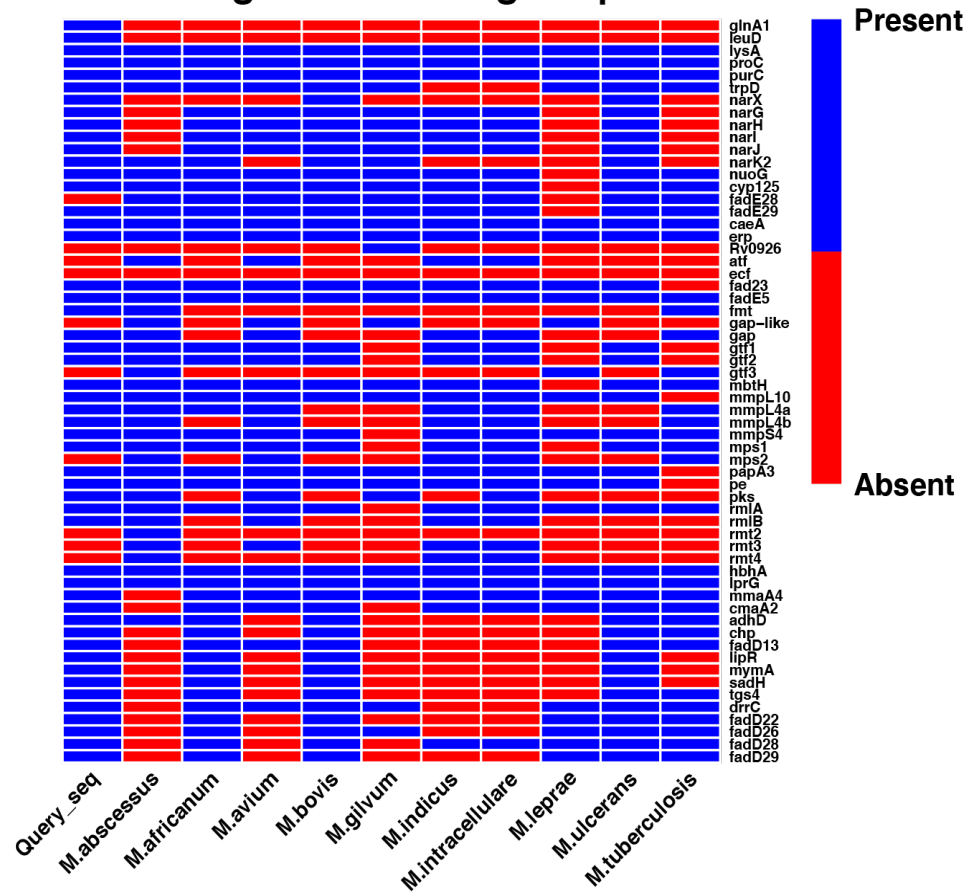
Comparative pathogenomics

ISBaC further perform the comparative pathogenomic analysis by comparing the virulence genes in the query genome (*M.tuberculosis*) with 10 other mycobacterium species. ISBaC create a gene presence and absence matrix, and graphically presented it through a heatmap using R. The results were stored in the “16. Virulence_gene_prediction_Results” folder (Figure 3.3). Each row in the heatmap represent a virulence gene. Blue colour in the heatmap indicates the presence of virulence genes while red colour indicates the absence of virulence gene (Figure 3.10). Different number of virulence genes have been observed in the query genome (*M.tuberculosis*) compare to the 10 mycobacterium species as shown in Table 3.3 and Appendix 3. 234 virulent genes have been found in query genome (*M.tuberculosis*) which is almost similar to the number of virulent genes in *M.tuberculosis* reference genome. 36 virulence genes (*lysA*, *proC*, *purC*, *caeA*, *erp*, *fadE5*, *hbhA*, *lprG*, *lpqY*, *sugA*, *sugB*, *sugC*, *mmpL11*, *mmpL3*, *ideR*, *mpa*, *relA*, *mprA*, *mprB*, *prpA*, *senX*, *sigA/rpoV*, *sigE*, *lpqH*, *pknG*, *secA2*, *espR*, *PPE4*, *eccA3*, *eccB3*, *eccD3*, *eccE3*, *espG3*, *esxH*, *mycP3*, *sodC*) were found to be shared among all the *mycobacterium* species and 2 virulence genes (*glnA1*, *leuD*) were found only in the query sequence.

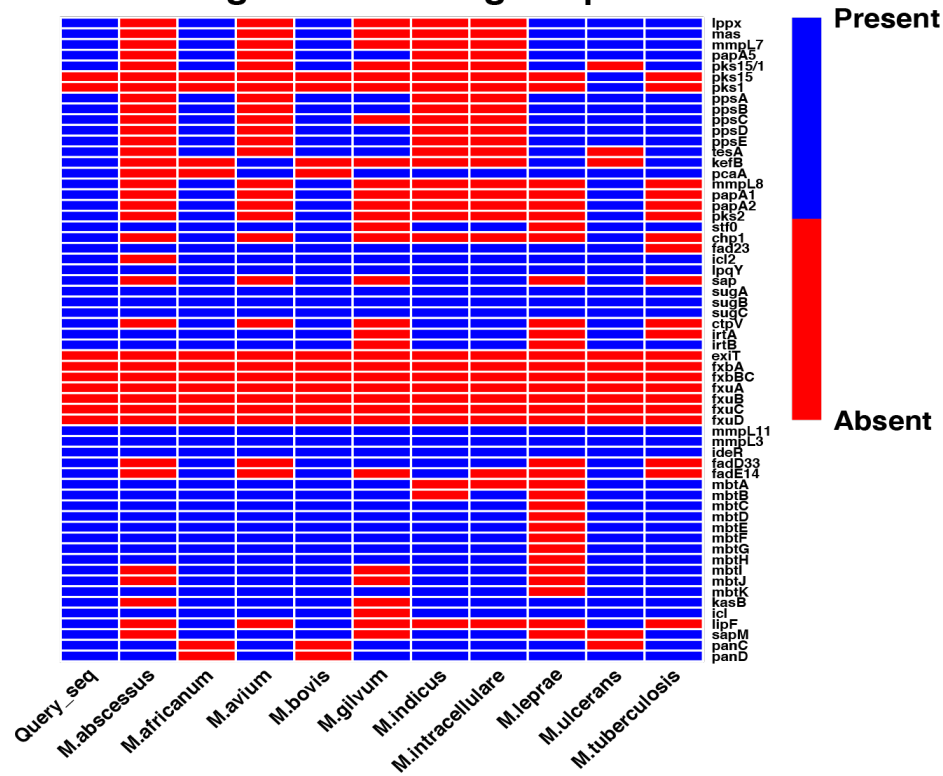
Table 3.3: Number of virulence gene in *mycobacterium* species

| <i>Mycobacterium</i> species | Number of virulence genes |
|------------------------------|---------------------------|
| Query genome (SRR786668) | 234 |
| <i>M.abscessus</i> | 125 |
| <i>M.africanum</i> | 216 |
| <i>M.avium</i> | 181 |
| <i>M.bovis</i> | 214 |
| <i>M.gilvum</i> | 151 |
| <i>M.indicus pranii</i> | 172 |
| <i>M.intracellulare</i> | 173 |
| <i>M.leprae</i> | 108 |
| <i>M.ulcerans</i> | 201 |
| <i>M.tuberculosis</i> | 226 |

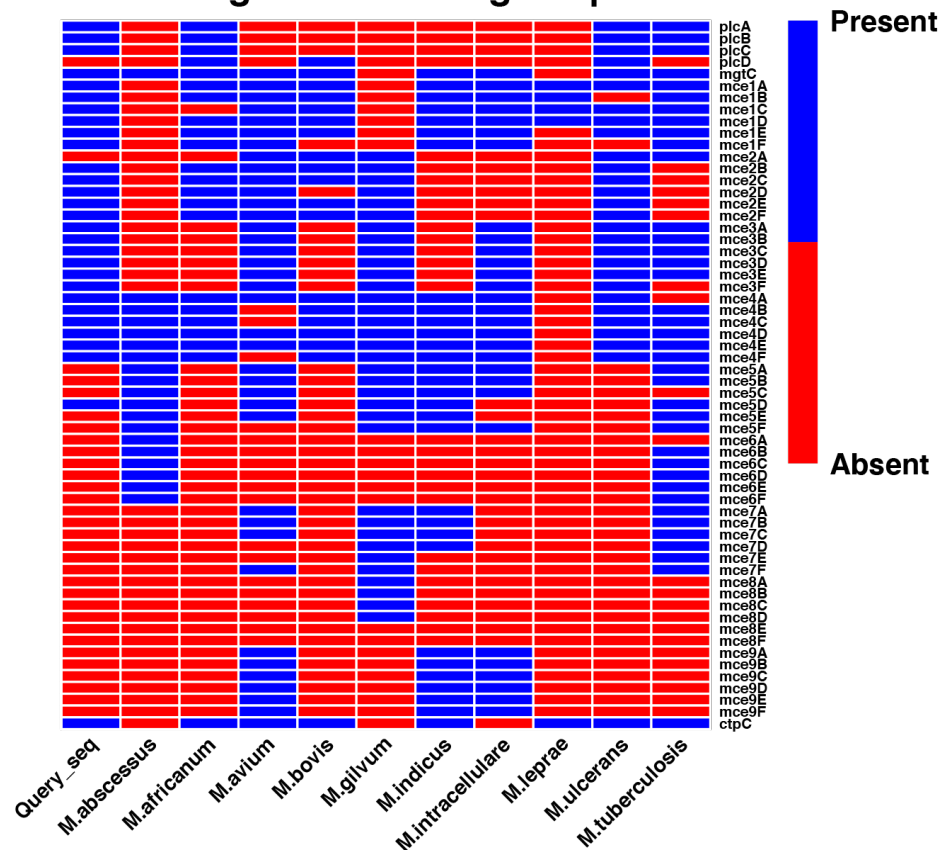
Screening of virulence gene profile



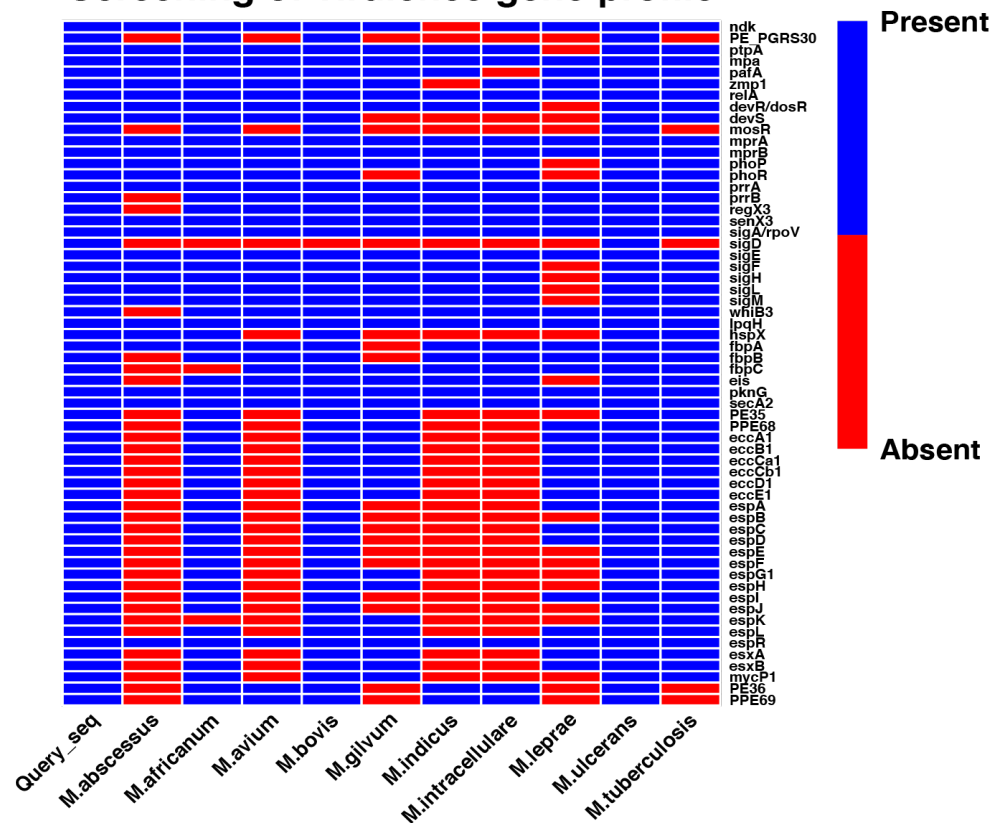
Screening of virulence gene profile



Screening of virulence gene profile



Screening of virulence gene profile



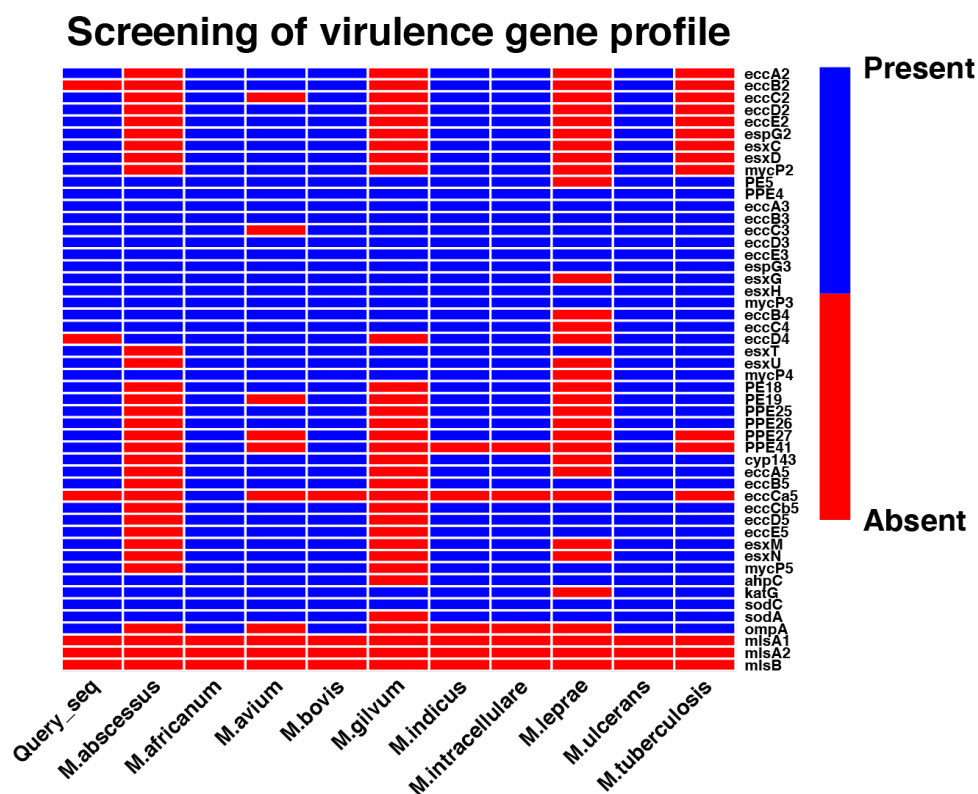


Figure 3.10: Heatmap of comparative pathgenomic analysis.

Blue color indicates presence of particular virulence gene and red color indicates absence of the virulence gene.

3.7 Calculation of ISBaC's accuracy through statistical analysis

The accuracy of ISBaC was first tested on the ten different *M. chelonae* samples as input. ISBaC can identify *M. chelonae* as the top hit in all the three analyses for each run. Table 3.4 shows the top hit and identity value for the three analyses in each run.

Table 3.4: Top hit and identity value for all the three analyses of 10 *M.chelonae* samples

| Input sequence | Top hit | 16S rRNA | MLSA | ANI |
|---|---------------------------------|----------|---------|---------|
| <i>Mycobacteroides chelonae</i> (ATCC35752) | <i>Mycobacteroides chelonae</i> | 100.00% | 100.00% | 100.00% |
| <i>Mycobacteroides chelonae</i> (DRS014067) | <i>Mycobacteroides chelonae</i> | 99.79% | 99.50% | 99.34% |
| <i>Mycobacteroides chelonae</i> (D16R3) | <i>Mycobacteroides chelonae</i> | 99.79% | 99.50% | 99.34% |
| <i>Mycobacteroides chelonae</i> (96-1705) | <i>Mycobacteroides chelonae</i> | 99.79% | 99.50% | 99.34% |

| | | | | |
|---|---------------------------------|---------|---------|--------|
| <i>Mycobacteroides chelonae</i> (96-1717) | <i>Mycobacteroides chelonae</i> | 99.79% | 100.00% | 99.97% |
| <i>Mycobacteroides chelonae</i> (D16Q24) | <i>Mycobacteroides chelonae</i> | 100.00% | 99.60% | 99.59% |
| <i>Mycobacteroides chelonae</i> (D16R7) | <i>Mycobacteroides chelonae</i> | 100.00% | 99.60% | 99.59% |
| <i>Mycobacteroides chelonae</i> (D16R9) | <i>Mycobacteroides chelonae</i> | 99.79% | 99.50% | 99.34% |
| <i>Mycobacteroides chelonae</i> (96-1720) | <i>Mycobacteroides chelonae</i> | 99.79% | 99.50% | 99.34% |
| <i>Mycobacteroides chelonae</i> (D16R10) | <i>Mycobacteroides chelonae</i> | 100.00% | 99.97% | 99.99% |

Based on the identity value from each run, we calculated the confidence interval (CI) value with 0.05 significant level for all the three analyses. Results show that the 95% CI for *16S rRNA* gene analysis was 99.80 to 99.95, 95% CI for MLSA was 99.51 to 99.83% and 95% CI for ANI was 99.37 to 99.79%. So, statistically, we were able to say that: say that: in repeated bacterial samples, 95% of the C.I.s for the mean share identity of ten totally different samples of *M. chelonae* determined by ISBaC pipeline is higher than the cutoff value of *16S rRNA* = 98, MLSA = 97, ANI = ≥ 95 (Figure 3.11). Thus, ISBaC was able to capture the true identity value for all *M.chelonae* samples for all the three analyses.

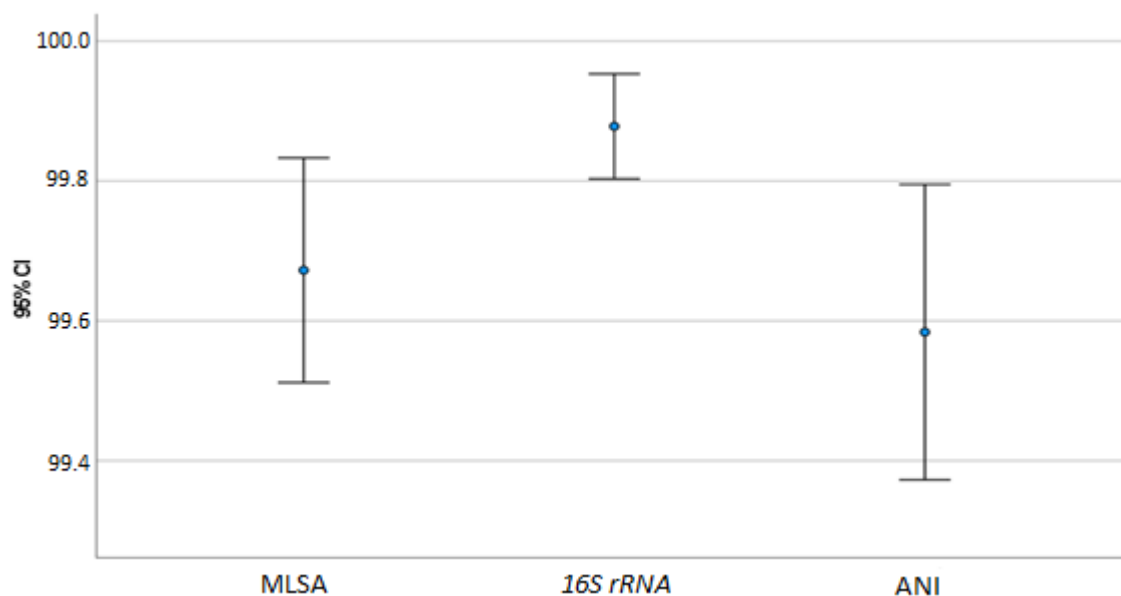


Figure 3.11: Error bar for *16S rRNA*, MLSA and ANI analysis calculated by the ISBaC result on 10 *M.chelonae* samples

Next, the same statistical analysis was performed on ten different mycobacterium species samples as input. The Table 3.5 below show the top hit and identity value for all the three analyses in each run.

Table 3.5: Top hit and identity value for all the three analyses of 10 different *mycobacterium* species samples

| Input sequence | Top hit | <i>16S rRNA</i> | MLSA | ANI |
|--------------------------------------|--------------------------------------|-----------------|---------|---------|
| <i>Mycobacterium tuberculosis</i> | <i>Mycobacterium tuberculosis</i> | 100.00% | 100.00% | 100.00% |
| <i>Mycobacterium kansasii</i> | <i>Mycobacterium kansasii</i> | 99.79% | 99.50% | 99.34% |
| <i>Mycobacteroides franklinii</i> | <i>Mycobacteroides franklinii</i> | 99.79% | 99.50% | 99.34% |
| <i>Mycobacteroides salmoniphilum</i> | <i>Mycobacteroides salmoniphilum</i> | 99.79% | 99.50% | 99.34% |
| <i>Mycobacterium intracellulare</i> | <i>Mycobacterium intracellulare</i> | 99.79% | 100.00% | 99.97% |
| <i>Mycolicibacter arupensis</i> | <i>Mycolicibacter arupensis</i> | 100.00% | 99.60% | 99.59% |
| <i>Mycolicibacterium iranikum</i> | <i>Mycolicibacterium iranikum</i> | 100.00% | 99.60% | 99.59% |
| <i>Mycobacterium chimaera</i> | <i>Mycobacterium chimaera</i> | 99.79% | 99.50% | 99.34% |
| <i>Mycobacterium ulcerans</i> | <i>Mycobacterium ulcerans</i> | 99.79% | 99.50% | 99.34% |
| <i>Mycobacteroides chelonae</i> | <i>Mycobacteroides chelonae</i> | 100.00% | 99.97% | 99.99% |

Again, ISBaC was able to identity the correct mycobacterium species samples as top hit in all the three analyses for each run. Based on the identity value, we calculated the 95% CI for *16S rRNA*: (99.55%, 100.14%), MLSA: (97.58%, 100.30%) and ANI: (97.45%,100.23%). Similarly, we can say that: in repeated samples, 95% of the C.I.s for the mean percentage identity of different mycobacterium species samples determined by ISBaC pipeline were above the threshold value of *16S rRNA* = 98, MLSA = 97 and ANI = ≥ 95 (Figure 3.12). Thus, ISBaC was able to capture the true identity value for different mycobacterium species samples in all the three analyses.

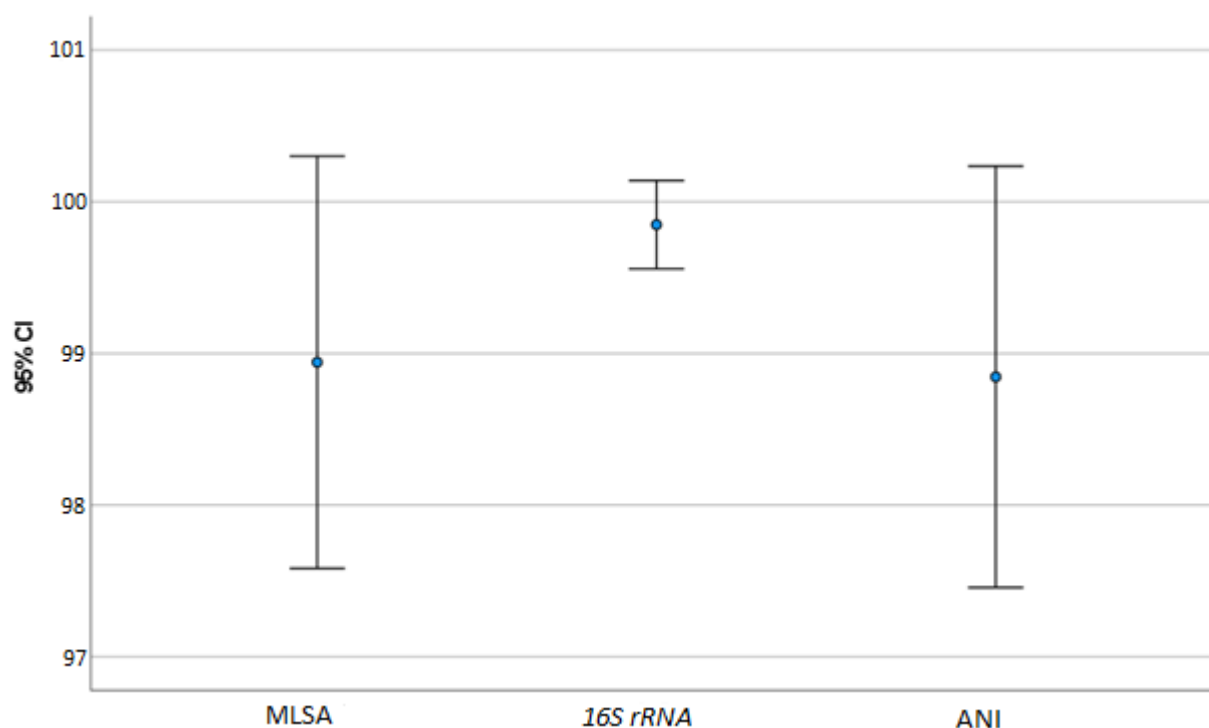


Figure 3.12: Error bar for *16S rRNA*, MLSA and ANI analysis calculated by the ISBaC result on 10 different mycobacterium species samples

3.8 Determining the sensitivity of ISBaC pipeline by identifying closely related *mycobacterium* species

Some traditional methods like Polymerase chain reaction (PCR) and Matrix-assisted laser desorption/ionization-time of flight mass spectrometry (MALDI-TOF MS) cannot properly identify between closely related *mycobacterium* species (Prammananan 2005; Rychert 2019). Such a *M. kansasii* and *M. gastri*, *M. malmoense* and *M. szulgai* who share a highly similar sequence so often traditional methods led to their misidentification (Beye, Fahsi, Raoult, & Fournier 2018; Jagielski et al. 2020).

Thus, to test the efficiency of ISBaC, we performed a comparison between *M. gastri* with *M. kansasii* and *M. malmoense* with *M. szulgai*. First, we used *M. malmoense* (SRR6046816) as the input sequence and ran it through ISBaC. Table 3.6 show the ISBaC result. ISBaC was able to identify SRR6046816 as *M. malmoense* correctly after the identity values in the three analyses (*16S rRNA*, MLSA and FastANI) are all above the threshold value. ISBaC was able to distinguish between *M. malmoense* and *M. szulgai* through the MLSA and FastANI result.

Table 3.6: ISBaC pipeline testing result for differentiating *M. malmoeense* from *M. szulgai*.

| <i>Mycobacterium</i> species | <i>16S rRNA</i> gene analysis | MLSA | FastANI |
|---|-------------------------------|--------|---------|
| <i>M. malmoeense</i> / <i>M. malmoeense</i> | 99.59% | 98.75% | 98.56% |
| <i>M. malmoeense</i> / <i>M. szulgai</i> | 99.04% | 97.27% | 96.54% |

Next, we tested ISBaC on the *M. kansasii* (SRR3319297) raw sequence. Table 3.7 show the ISBaC result using *M. kansasii* (SRR3319297) raw data as the input sequences and it was noticed that ISBaC clearly identifies (SRR3319297) as *M. kansasii*, as the percentage identify values obtained from ISBaC were above the threshold values in three analyses (*16S rRNA*, MLSA, FastANI.). So, ISBaC was able to differentiate between *M. kansasii* and *M. gastri* through MLSA and FastANI.

Table 3.7: ISBaC pipeline testing result for differentiating *M. kansasii* from *M. gastri*

| <i>Mycobacterium</i> species | <i>16S rRNA</i> gene analysis | MLSA | FastANI |
|---|-------------------------------|--------|---------|
| <i>M. kansasii</i> / <i>M. kansasii</i> | 99.86% | 98.92% | 98.96% |
| <i>M. kansasii</i> / <i>M. gastri</i> | 99.84% | 97.41% | 96.67 % |

3.9 Evaluating the reproducibility of ISBaC pipeline

We evaluated the reproducibility of ISBaC pipeline by using *M. chelonae* ATCC_35752 (SRR4423483) as query sequence. We ran the same query sequence twice on the same and separate computational setting, respectively and each time it gave the same results by taking same amount of time as shown in Table 3.8.

Table 3.8: Evaluating the reproducibility of ISBaC pipeline by using *M. chelonae*

| Query sequence | Computational setting | Time taken |
|--|-------------------------|------------|
| <i>M. chelonae</i> ATCC_35752 (SRR4423483) | 4 core CPUs & 128Gb RAM | 1 hour |
| <i>M. chelonae</i> ATCC_35752 (SRR4423483) | 2 core CPUs & 64 Gb RAM | 1.5 hour |

Chapter4. Discussion

Precise identification of microorganisms is required in almost every aspect of the research. Phenotypic methods present some limitations in terms of sensitivity, specificity, and time. These limitations are more evident for some types of slow-growing bacteria. Moreover, the time needed to identify a pathogen based on its phenotypic characteristics is the first challenge, as the sample must be seeded and incubated for at least 24 hours. Then, conventional biochemical tests must be performed at least another 24 hours, conditions that delay results and compromise the patient's health. Phenotypic methods cannot always identify the microorganism at species level and much less at strain level (Georghiou et al. 1995). This technique as a rule needs more than 48 hours after a colony isolation and more fourteen days are expected for the identification of many slow growing. In certain conditions, no identification can be made following quite a while of investigation, even by an accomplished technologist. Still, several aspects convolute their application in the microbial science research facility: the trouble in the segregation, the colony development, the expenses of the tests, and poor identification of bacterial species coming from complex samples, among others. They are likewise normally not accessible at research facilities of public emergency clinics and don't have an all-inclusive execution (Castro-Escarpulli et al. 2015).

The total time required for *16S rRNA* gene sequencing, Multi Locus Sequence Typing (MLST) and whole genome sequencing are comparable to conventional phenotyping methods because both methods require some amount of time to grow bacterial culture for experiment. The 16S rRNA gene sequencing analysis takes around 9 hours per sample without bioinformatics work, and DNA-DNA hybridization takes 18 hours to give results (Ahmad et al. 2013; Gee et al. 2004). Besides, Multi Locus Sequence Typing (MLST) begin with PCR amplification step utilizing primers particular to the loci of the MLST scheme, followed by DNA sequencing. The strategy is both expensive and time consuming. In this modern time of high throughput sequencing, it may be more sound to utilize whole genome sequencing (WGS) information for genotyping. The cost of DNA sequencing has consistently gone down generally 10-fold each 5 a long time (Land et al. 2015), and the advancement of next- and third-generation sequencing strategies has provided equally great reductions in equipment investments, thus making the technology accessible to individual investigators and routine clinical and microbial laboratories. The challenge, however, is to extract the relevant information from the massive amount of knowledge generated by these techniques (Larsen et al. 2012). Several web-based analysis are available such as bacterial analysis pipeline (Thomsen et al. 2016), microbial identification and characterization through reads analysis (MICRA) (Caboche et al. 2017) and a Bioinformatics Bacterial Identification Tool (BIBI) (Devulder, Perrière, Baty, & Flandrois 2003).

However, these tools lack of user customization and typically require the user to enter their data into their servers, and the user can only upload raw data with limited size. The available genome analysis of a mycobacterial genome requires the usage of many bioinformatics tools. However, these tools might be complex to handle and give less accurate results during data transition from one software to another software because user need to separately set the parameters for every software.

Taken all together, the project's main objective is to develop an accurate and user-friendly pipeline for the identification of mycobacterial species. ISBaC is precise and user-friendly as the pipeline automates the whole identification process through only single command. ISBaC will take around an hour with 4 core CPUs and 128Gb RAM (can be faster depending on the computer core and system) to identify the mycobacterium species once users get the raw sequencing data. ISBaC is also designed to run on the user's computer without an internet connection, access accounts, and additional requirements. The user can upload any size of data depending on the specification of the users' system. ISBaC is designed to simplify and accelerate the bioinformatics analysis, especially for microbiology researchers and clinicians who lack of bioinformatics knowledge.

ISBaC can accept raw data from both single and paired-end sequencing generated from different sequencing platforms for example Illumina, PacBio. ISBaC includes all the required analyses to process the whole genome raw data and identify the identity of the Mycobacteria species. The pipeline includes raw reads pre-processing, genome assembly, genome annotation and mycobacterial identification using *16S rRNA* gene analysis, Multilocus sequence analysis (MLSA) and ANI analysis. All the analysis in ISBaC is automated and designed to be user-friendly. Users are required to key in only a single command to run the whole pipeline. Other than that, ISBaC is also easy to install as all the software's in ISBaC will be automatically downloaded in the user-specified directory by running a Perl script.

ISBaC integrates many mycobacterial identification steps like *16S rRNA* gene analysis, which identifies mycobacterium species based on the percentage identity values obtained from blastn. The threshold value for species identification through *16S rRNA* gene analysis is 98%. However, the *16S rRNA* gene sequences is not enough to discriminate between mycobacteria species because many mycobacterial species share highly similar *16S rRNA* gene sequences (Clarridge 2004). Thus, to further support the *16S rRNA* gene analysis result, the subsequent multilocus sequencing analysis (MLSA) has been performed. ISBaC will concatenate the housekeeping genes (*gyrA*, *gyrB*, *rpoB*, *groEL*, and *recA*) which are highly conserved, not vulnerable to horizontal gene transfer, long enough to contain phylogenetically useful information and could predict whole genome relationship (Rong & Huang 2014). ISBaC can handle the different combinations of the housekeeping genes according

to the availability of the housekeeping genes in the query genome. MLSA contains higher discrimination between the mycobacteria genomes than *16S rRNA* gene analysis (Liu, Lai, & Shao 2017b).

Other than gene level, ISBaC also identify *mycobacterium* species on the genome level. Single gene or housekeeping genes (hk) are still subjected to deletions, duplications, and mutations. The genome complexity of an organism is the complete history of genetic recombination's and other nucleotide sequence drifts that occurred during evolution (Sentausa & Fournier 2013). Thus, MinHash and FastANI have been integrated into ISBaC. ANI value of $\geq 95\%$ corresponds to a 70 % DNA-DNA hybridization (Konstantinidis & Tiedje 2005). Moreover, ISBaC provides results in the form of bar chart and heatmap for easier interpretation of results.

Other than the Mycobacterium identification analysis, ISBaC also contains genome annotation analysis. The assembled genome will be annotated using Prokka. Gene in the genome will be predicted, and the function of each gene will also be annotated. The predicted gene and protein sequence will be provided in FASTA format. Taken all together, ISBaC is designed to be executed using a simple command and is suitable for researchers lacking of bioinformatic background. In addition to that, ISBaC is intrinsically flexible to allow different customization.

In the last, we checked the accuracy and specificity of pipeline, and we found that ISBaC can identify the identity of mycobacterium species with 95% Confidence interval (CI) and it also captures the threshold values for *16S rRNA* gene analysis, MLSA and ANI analysis in repeated samples. In addition to that ISBaC also showed higher sensitivity by differentiating among closely related mycobacterium species.

Conclusion and future work

An automated pipeline for *In silico* Bacterial Identification (ISBaC) has been developed to identify *mycobacterial* species. is a handy approach that can be used to identify bacterial species. ISBaC pipeline requires just one hour to identify Mycobacterium species with a 4 core CPUs and 128Gb RAM. The whole process can become even faster with higher core CPUs and RAM. ISBaC simplifies the user's task of analysing large amounts of raw read and whole-genome sequence data and provides an automated method for identification of mycobacterial species. Other than that, ISBaC also provides extra genome analysis such as genome annotation and virulence gene prediction. ISBaC pipeline is user-friendly by chains up all the required analyses to identify mycobacteria's identity with just a

single command. ISBaC has been tested with several validation cases and ISBaC manage to identify correctly the real identity of *Mycobacterium* species. ISBaC is designed in a way suitable for all students or microbiology researchers that lack of bioinformatics knowledge. ISBaC allows them to run Bioinformatics analysis more easily. All the embedded tools in ISBaC helps to better identify poorly described, rarely isolated *mycobacterial* species. ISBaC offers robust resolution between closely related mycobacterial species and can be routinely used for identification of mycobacteria.

Some aspects of this study still need some improvements as the pipeline can be tested with more different. Multilocus sequence analysis uses a combination of five housekeeping genes which can be increased to seven to eight genes for better accuracy. Thus multilocus sequence analysis can be refined further by incorporating more housekeeping gene data from five genes to seven to eight genes in future. Next, the developed database for the study just contains sequence data of *Mycobacterium* species which limits ISBaC to just the identification of *Mycobacterium* species. Therefore, ISBaC can be integrated into different genera by developing a database for that specific genus so the same analysis could be used for the identification of that genus. Lastly, ISBaC pipeline should also be transformed from a command-line tool to a pipeline with a graphical user interface (GUI).

Reference

- Abram, K, Udaondo, Z, Bleker, C, Wanchai, V, Wassenaar, TM, Robeson, MS, & Ussery, DW, 2021, "Mash-based analyses of *Escherichia coli* genomes reveal 14 distinct phylogroups," *Communications Biology*, vol. 4, no. 1, doi: 10.1038/s42003-020-01626-5.
- Adékambi, T, Colson, P, & Drancourt, M, 2003, "rpoB-Based Identification of Nonpigmented and Late-Pigmenting Rapidly Growing Mycobacteria," *Journal of Clinical Microbiology*, vol. 41, no. 12, pp. 5699–5708, doi: 10.1128/JCM.41.12.5699-5708.2003.
- Ahmad, R, Hansen, GÅ, Hansen, H, Hjerde, E, Pedersen, HL, Paulsen, SM, Nyrud, MLJ, Strauss, A, Willassen, NP, & Haugen, P, 2013, "Prediction, microarray and northern blot analyses identify new intergenic small RNAs in *aliivibrio salmonicida*," *Journal of Molecular Microbiology and Biotechnology*, vol. 22, no. 6, pp. 352–360, doi: 10.1159/000345769.
- Andrews, S, 2010, "FastQC: a quality control tool for high throughput sequence data, Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>," .
- Bachmann, NL, Salamzade, R, Manson, AL, Whittington, R, Sintchenko, V, Earl, AM, & Marais, BJ, 2020, "Key Transitions in the Evolution of Rapid and Slow Growing Mycobacteria Identified by Comparative Genomics," *Frontiers in Microbiology*, vol. 10, doi: 10.3389/fmicb.2019.03019.
- Bankevich, A, Nurk, S, Antipov, D, Gurevich, AA, Dvorkin, M, Kulikov, AS, Lesin, VM, Nikolenko, SI, Pham, S, Pribelski, AD, Pyshkin, A v., Sirotkin, A v., Vyahhi, N, Tesler, G, Alekseyev, MA, & Pevzner, PA, 2012, "SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing," *Journal of Computational Biology*, vol. 19, no. 5, pp. 455–477, doi: 10.1089/cmb.2012.0021.
- Bergey's Manual® of Systematic Bacteriology, 2012, Bergey's Manual® of Systematic Bacteriology, Springer New York, doi: 10.1007/978-0-387-68233-4.
- Beye, M, Fahsi, N, Raoult, D, & Fournier, PE, 2018, "Careful use of 16S rRNA gene sequence similarity values for the identification of Mycobacterium species," *New Microbes and New Infections*, vol. 22, pp. 24–29, doi: 10.1016/j.nmni.2017.12.009.
- Blackwood, KS, He, C, Gunton, J, Turenne, CY, Wolfe, J, Kabani, AM, & Canada, H, 2000, "Evaluation of recA Sequences for Identification of Mycobacterium Species", *Journal of Clinical Microbiology*, vol. 38.
- Bogaerts, B, Delcourt, T, Soetaert, K, Boarbi, S, Ceyssens, PJ, Winand, R, van Braekel, J, de Keersmaecker, SCJ, Roosens, NHC, Marchal, K, Mathys, V, & Vanneste, K, 2021, "A bioinformatics whole-genome sequencing workflow for clinical Mycobacterium tuberculosis complex isolate analysis, validated using a reference collection extensively characterized with conventional methods and in silico approaches," *Journal of Clinical Microbiology*, vol. 59, no. 6, doi: 10.1128/JCM.00202-21.
- Bolger, AM, Lohse, M, & Usadel, B, 2014, "Trimmomatic: A flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, doi: 10.1093/bioinformatics/btu170.
- Bosshard, P & McDaniel, G, 2010, "CLSI guideline addresses identification of bacteria and fungi using DNA target sequencing," *Laboratory Medicine*, doi: 10.1309/LMP7HXYMIEYYOMG8.

- Bottai, D & Brosch, R, 2009, "Mycobacterial PE, PPE and ESX clusters: Novel insights into the secretion of these most unusual protein families," *Molecular Microbiology*, doi: 10.1111/j.1365-2958.2009.06784.x.
- Brennan, P J, 1995, "The envelope of mycobacteria", *Annu. Rev. Biochem* , www.annualreviews.org.
- Brown-Elliott, BA & Wallace, RJ, 2002, "Clinical and taxonomic status of pathogenic nonpigmented or late-pigmenting rapidly growing mycobacteria," *Clinical Microbiology Reviews*, doi: 10.1128/CMR.15.4.716-746.2002.
- Brown-Elliott, BA & Wallace, RJ, 2015, "Mycobacterium : Clinical and Laboratory Characteristics of Rapidly Growing Mycobacteria ," in *Manual of Clinical Microbiology* , pp. 595–612, ASM Press, doi: 10.1128/9781555817381.ch32.
- Caboche, S, Even, G, Loywick, A, Audebert, C, & Hot, D, 2017, "MICRA: An automatic pipeline for fast characterization of microbial genomes from high-throughput sequencing data," *Genome Biology*, vol. 18, no. 1, doi: 10.1186/s13059-017-1367-z.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, & Madden T.L, 2008, "BLAST+: architecture and applications". *BMC Bioinformatics* 10:421.
- Castejon, M, Menéndez, MC, Comas, I, Vicente, A, & Garcia, MJ, 2018, "Whole-genome sequence analysis of the Mycobacterium avium complex and proposal of the transfer of Mycobacterium yongonense to mycobacterium intracellulare subsp. Yongonense subsp. nov.," *International Journal of Systematic and Evolutionary Microbiology*, vol. 68, no. 6, pp. 1998–2005, doi: 10.1099/ijsem.0.002767.
- Castro-Escarpulli, G, Maribel Alonso-Aguilar, N, Rivera Sánchez, G, Bocanegra-Garcia, V, Guo, X, Juárez-Enríquez, SR, Luna-Herrera, J, Majalca Martínez, C, Ma Guadalupe, A-A, & Guadalupe Aguilera Arreola, M, 2015, "Archives of Clinical Microbiology Citation: Aguilera-Arreola MG. Identification and Typing Methods for the Study of Bacterial Infections: a Brief Review and Mycobacterial as Case of Study," *Arch Clin Microbiol*, vol. 7, p. 1, Retrieved from <http://www.imedpub.com/>.
- Chan, J, Halachev, M, Yates, E, Smith, G, & Pallen, M, 2012, "Whole-genome sequence of the emerging pathogen mycobacterium abscessus strain 47J26," *Journal of Bacteriology*, vol. 194, no. 2, pp. 549–549, doi: 10.1128/JB.06440-11.
- Chen, L, Yang, J, Yu, J, Yao, Z, Sun, L, Shen, Y, & Jin, Q, 2005, "VFDB: A reference database for bacterial virulence factors," *Nucleic Acids Research*, vol. 33, no. DATABASE ISS., doi: 10.1093/nar/gki008.
- Cho, JC & Tiedje, JM, 2001, "Bacterial Species Determination from DNA-DNA Hybridization by Using Genome Fragments and DNA Microarrays," *Applied and Environmental Microbiology*, vol. 67, no. 8, pp. 3677–3682, doi: 10.1128/AEM.67.8.3677-3682.2001.
- Claeys, TA & Robinson, RT, 2018, "The Many Lives of Nontuberculous Mycobacteria" , <https://doi.org/10.1128/CMR.17.4.840-862.2004>.
- Clarridge, JE, 2004, "Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases," *Clinical Microbiology Reviews*, doi: 10.1128/CMR.17.4.840-862.2004.

- Cook, GM, Berney, M, Gebhard, S, Heinemann, M, Cox, RA, Danilchanka, O, & Niederweis, M, 2009, "Physiology of Mycobacteria," *Advances in Microbial Physiology*, Academic Press, doi: 10.1016/S0065-2911(09)05502-7.
- Cuevas-Córdoba, B, Fresno, C, Haase-Hernández, JI, Barbosa-Amezcu, M, Mata-Rocha, M, Muñoz-Torrico, M, Salazar-Lezama, MA, Martínez-Orozco, JA, Narváez-Díaz, LA, Salas-Hernández, J, González-Covarrubias, V, & Soberón, X, 2021, "A bioinformatics pipeline for Mycobacterium tuberculosis sequencing that cleans contaminant reads from sputum samples," *PLoS ONE*, vol. 16, no. 10 October, doi: 10.1371/journal.pone.0258774.
- Dauendorffer, JN, Guillemin, I, Aubry, A, Truffot-Pernot, C, Sougakoff, W, Jarlier, V, & Cambau, E, 2003, "Identification of mycobacterial species by PCR sequencing of quinolone resistance-determining regions of DNA gyrase genes," *Journal of Clinical Microbiology*, vol. 41, no. 3, pp. 1311–1315, doi: 10.1128/JCM.41.3.1311-1315.2003.
- Devulder, G, de Montclos, MP, & Flandrois, JP, 2005, "A multigene approach to phylogenetic analysis using the genus Mycobacterium as a model," *International Journal of Systematic and Evolutionary Microbiology*, vol. 55, no. 1, pp. 293–302, doi: 10.1099/ijls.0.63222-0.
- Devulder, G, Perrière, G, Baty, F, & Flandrois, JP, 2003, "BIBI, a bioinformatics bacterial identification tool," *Journal of Clinical Microbiology*, vol. 41, no. 4, pp. 1785–1787, doi: 10.1128/JCM.41.4.1785-1787.2003.
- Prammananan, 2005, "Evaluation of polymerase chain reaction and restriction enzyme", *Journal of Tropical Medicine and Public Health*, Vol. 36, no. 5, pp. 1252.
- Falkinham, JO, 2009, "Surrounded by mycobacteria: Nontuberculous mycobacteria in the human environment," *Journal of Applied Microbiology*, doi: 10.1111/j.1365-2672.2009.04161.x.
- Gee, JE, De, BK, Levett, PN, Whitney, AM, Novak, RT, & Popovic, T, 2004, "Use of 16S rRNA gene sequencing for rapid confirmatory identification of Brucella isolates," *Journal of Clinical Microbiology*, vol. 42, no. 8, pp. 3649–3654, doi: 10.1128/JCM.42.8.3649-3654.2004.
- Georghiou, PR, Hamill, RJ, Wright, CE, Versalovic, J, Koeuth, T, Watson, DA, & Lupski, JR, 1995, "Molecular Epidemiology of infections due to Enterobacter aerogenes: Identification of Hospital Outbreak-Associated Strains by Molecular Techniques", vol. 20, , <https://about.jstor.org/terms>.
- Goris, J, Konstantinidis, KT, Klappenbach, JA, Coenye, T, Vandamme, P, & Tiedje, JM, 2007, "DNA-DNA hybridization values and their relationship to whole-genome sequence similarities," *International Journal of Systematic and Evolutionary Microbiology*, vol. 57, no. 1, pp. 81–91, doi: 10.1099/ijls.0.64483-0.
- Gupta, RS, Lo, B, & Son, J, 2018, "Phylogenomics and comparative genomic studies robustly support division of the genus Mycobacterium into an emended genus Mycobacterium and four novel genera," *Frontiers in Microbiology*, vol. 9, no. FEB, doi: 10.3389/fmicb.2018.00067.
- Gurevich, A, Saveliev, V, Vyahhi, N, & Tesler, G, 2013, "QUAST: Quality assessment tool for genome assemblies," *Bioinformatics*, vol. 29, no. 8, pp. 1072–1075, doi: 10.1093/bioinformatics/btt086.
- Hartmans Sybe and de Bont, JAM and SE, 2006, "The Genus Mycobacterium–Nonmedical," in S. and R. E. and S. K.-H. and S. E. Dworkin Martin and Falkow (ed.), *The Prokaryotes: Volume*

- 3: *Archaea. Bacteria: Firmicutes, Actinomycetes*, pp. 889–918, Springer New York, New York, NY, doi: 10.1007/0-387-30743-5_33.
- Hasan, NA, Davidson, RM, de Moura, VCN, Garcia, BJ, Reynolds, PR, Elaine Epperson, L, Farias-Hesson, E, DeGroote, MA, Jackson, M, & Strong, M, 2015, “Draft genome sequence of *Mycobacterium chelonae* type strain ATCC 35752,” *Genome Announcements*, vol. 3, no. 3, doi: 10.1128/genomeA.00536-15.
- Jagielski, T, Borówka, P, Bakula, Z, Lach, J, Marciniak, B, Brzostek, A, Dziadek, J, Dziurzyński, M, Pennings, L, van Ingen, J, Žolnir-Dovč, M, & Strapagiel, D, 2020, “Genomic Insights Into the *Mycobacterium kansasii* Complex: An Update,” *Frontiers in Microbiology*, vol. 10, doi: 10.3389/fmicb.2019.02918.
- Jain, C, Rodriguez-R, LM, Phillippy, AM, Konstantinidis, KT, & Aluru, S, 2018, “High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries,” *Nature Communications*, vol. 9, no. 1, doi: 10.1038/s41467-018-07641-9.
- Johnson, MM & Odell, JA, 2014, “Nontuberculous mycobacterial pulmonary infections,” *Journal of Thoracic Disease*, Pioneer Bioscience Publishing, doi: 10.3978/j.issn.2072-1439.2013.12.24.
- Kim, CJ, Kim, NH, Song, KH, Choe, PG, Kim, ES, Park, SW, Kim, H bin, Kim, NJ, Kim, EC, Park, WB, & Oh, M don, 2013, “Differentiating rapid- and slow-growing mycobacteria by difference in time to growth detection in liquid media,” *Diagnostic Microbiology and Infectious Disease*, vol. 75, no. 1, pp. 73–76, doi: 10.1016/j.diagmicrobio.2012.09.019.
- Kim, SH & Shin, JH, 2018, “Identification of nontuberculous mycobacteria using multilocus sequence analysis of 16S rRNA, hsp65, and rpoB,” *Journal of Clinical Laboratory Analysis*, vol. 32, no. 1, doi: 10.1002/jcla.22184.
- Kohl, TA, Utpatel, C, Schleusener, V, de Filippo, MR, Beckert, P, Cirillo, DM, & Niemann, S, 2018, “MTBseq: A comprehensive pipeline for whole genome sequence analysis of *Mycobacterium tuberculosis* complex isolates,” *PeerJ*, vol. 2018, no. 11, doi: 10.7717/peerj.5895.
- Konstantinidis, KT & Tiedje, JM, 2005, “Genomic insights that advance the species definition for prokaryotes,” *PNAS February*, vol. 15, , www.tigr.org.
- Lagier, JC, Edouard, S, Pagnier, I, Mediannikov, O, Drancourt, M, & Raoult, D, 2015, “Current and past strategies for bacterial culture in clinical microbiology,” *Clinical Microbiology Reviews*, vol. 28, no. 1, pp. 208–236, doi: 10.1128/CMR.00110-14.
- Lambris, JD & Paoletti, R, 2018, “Advances in Experimental Medicine and Biology” , <http://www.springer.com/series/5584>.
- Land, M, Hauser, L, Jun, SR, Nookaew, I, Leuze, MR, Ahn, TH, Karpinets, T, Lund, O, Kora, G, Wassenaar, T, Poudel, S, & Ussery, DW, 2015, “Insights from 20 years of bacterial genome sequencing,” *Functional and Integrative Genomics*, Springer Verlag, doi: 10.1007/s10142-015-0433-4.
- Larsen, M v., Cosentino, S, Rasmussen, S, Friis, C, Hasman, H, Marvig, RL, Jelsbak, L, Sicheritz-Pontén, T, Ussery, DW, Aarestrup, FM, & Lund, O, 2012, “Multilocus sequence typing of total-genome-sequenced bacteria,” *Journal of Clinical Microbiology*, vol. 50, no. 4, pp. 1355–1361, doi: 10.1128/JCM.06094-11.

- Leinonen, R, Akhtar, R, Birney, E, Bower, L, Cerdeno-Tárraga, A, Cheng, Y, Cleland, I, Faruque, N, Goodgame, N, Gibson, R, Hoad, G, Jang, M, Pakseresht, N, Plaister, S, Radhakrishnan, R, Reddy, K, Sobhany, S, Hoopen, P ten, Vaughan, R, Zalunin, V, & Cochrane, G, 2011, "The European nucleotide archive," *Nucleic Acids Research*, vol. 39, no. SUPPL. 1, doi: 10.1093/nar/gkq967.
- Liu, Y, Lai, Q, & Shao, Z, 2017a, "A multilocus sequence analysis scheme for phylogeny of *Thioclava* bacteria and proposal of two novel species," *Frontiers in Microbiology*, vol. 8, no. JUL, doi: 10.3389/fmicb.2017.01321.
- Liu, Y, Lai, Q, & Shao, Z, 2017b, "A multilocus sequence analysis scheme for phylogeny of *Thioclava* bacteria and proposal of two novel species," *Frontiers in Microbiology*, vol. 8, no. JUL, doi: 10.3389/fmicb.2017.01321.
- López-Hermoso, C, de la Haba, RR, Sánchez-Porro, C, Papke, RT, & Ventosa, A, 2017, "Assessment of multilocus sequence analysis as a valuable tool for the classification of the genus *Salinivibrio*," *Frontiers in Microbiology*, vol. 8, no. JUN, doi: 10.3389/fmicb.2017.01107.
- Macheras, E, Roux, AL, Ripoll, F, Sivadon-Tardy, V, Gutierrez, C, Gaillard, JL, & Heym, B, 2009, "Inaccuracy of single-target sequencing for discriminating species of the *Mycobacterium abscessus* group," *Journal of Clinical Microbiology*, vol. 47, no. 8, pp. 2596–2600, doi: 10.1128/JCM.00037-09.
- Meehan, CJ, Goig, GA, Kohl, TA, Verboven, L, Dippenaar, A, Ezewudo, M, Farhat, MR, Guthrie, JL, Laukens, K, Miotto, P, Ofori-Anyinam, B, Dreyer, V, Supply, P, Suresh, A, Utpatel, C, van Soolingen, D, Zhou, Y, Ashton, PM, Brites, D, Cabibbe, AM, de Jong, BC, de Vos, M, Menardo, F, Gagneux, S, Gao, Q, Heupink, TH, Liu, Q, Loiseau, C, Rigouts, L, Rodwell, TC, Tagliani, E, Walker, TM, Warren, RM, Zhao, Y, Zignol, M, Schito, M, Gardy, J, Cirillo, DM, Niemann, S, Comas, I, & van Rie, A, 2019, "Whole genome sequencing of *Mycobacterium tuberculosis*: current standards and open issues," *Nature Reviews Microbiology*, Nature Publishing Group, doi: 10.1038/s41579-019-0214-5.
- Nogueira, CL, Simmon, KE, Chimara, E, Cnockaert, M, Palomino, JC, Martin, A, Vandamme, P, Brown-Elliott, BA, Wallace, RJ, & Leão, SC, 2015, "*Mycobacterium franklinii* sp. nov., a species closely related to members of the *Mycobacterium chelonae*–*Mycobacterium abscessus* group," *International Journal of Systematic and Evolutionary Microbiology*, vol. 65, no. 7, pp. 2148–2153, doi: 10.1099/ij.s.0.000234.
- Ondov, BD, Treangen, TJ, Melsted, P, Mallonee, AB, Bergman, NH, Koren, S, & Phillippy, AM, 2016, "Mash: Fast genome and metagenome distance estimation using MinHash," *Genome Biology*, vol. 17, no. 1, doi: 10.1186/s13059-016-0997-x.
- Palomino, JC, 2009, "Molecular detection, identification and drug resistance detection in *Mycobacterium tuberculosis*," *FEMS Immunology and Medical Microbiology*, doi: 10.1111/j.1574-695X.2009.00555.x.
- Parte, AC, 2014, "LPSN - List of prokaryotic names with standing in nomenclature," *Nucleic Acids Research*, vol. 42, no. D1, doi: 10.1093/nar/gkt1111.
- R Core Team, 2021, "R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria."

- Ripoll, F, Pasek, S, Schenowitz, C, Dossat, C, Barbe, V, Rottman, M, Macheras, E, Heym, B, Hermann, JL, Daffé, M, Brosch, R, Risler, JL, & Gaillard, JL, 2009, “Non mycobacterial virulence genes in the genome of the emerging pathogen *Mycobacterium abscessus*,” *PLoS ONE*, vol. 4, no. 6, doi: 10.1371/journal.pone.0005660.
- Rong, X & Huang, Y, 2014, “Multi-locus sequence analysis. Taking prokaryotic systematics to the next level,” in *Methods in Microbiology*, vol. 41, pp. 221–251, Academic Press Inc., doi: 10.1016/bs.mim.2014.10.001.
- Rychert, J, 2019, “Commentary: Benefits and Limitations of MALDI-TOF Mass Spectrometry for the Identification of Microorganisms”, *Journal of Infectiology Mini Review Journal of Infectiology*, vol. 2.
- Ryu, YJ, Koh, WJ, & Daley, CL, 2016, “Diagnosis and treatment of nontuberculous mycobacterial lung disease: Clinicians’ perspectives,” *Tuberculosis and Respiratory Diseases*, Korean National Tuberculosis Association, doi: 10.4046/trd.2016.79.2.74.
- Schweickert, B, Goldenberg, O, Richter, E, Göbel, UB, Petrich, A, Buchholz, P, & Moter, A, 2008, “Occurrence and clinical relevance of *Mycobacterium chimaera* sp. nov., Germany,” *Emerging Infectious Diseases*, vol. 14, no. 9, pp. 1443–1446, doi: 10.3201/eid1409.071032.
- Seemann, T, 2014, “Prokka: Rapid prokaryotic genome annotation,” *Bioinformatics*, vol. 30, no. 14, pp. 2068–2069, doi: 10.1093/bioinformatics/btu153.
- Seemann T, Goncalves da Silva A, Bulach DM, Schultz MB, Kwong JC, Howden BP, 2018, “Nullarbor” *Github*.
- Sentausa, E & Fournier, PE, 2013, “Advantages and limitations of genomics in prokaryotic taxonomy,” *Clinical Microbiology and Infection*, Blackwell Publishing Ltd, doi: 10.1111/1469-0691.12181.
- Singh, A & Kashyap, VK, 2012, “Specific and rapid detection of mycobacterium tuberculosis complex in clinical samples by polymerase chain reaction,” *Interdisciplinary Perspectives on Infectious Diseases*, vol. 2012, doi: 10.1155/2012/654694.
- Soini, H, Bottger, EC, & Viljanen, MK, 1994, “Identification of Mycobacteria by PCR-Based Sequence Determination of the 32-Kilodalton Protein Gene”, *Journal of Clinical Microbiology*, vol. 32.
- Steed, KA & Falkinham, JO, 2006, “Effect of growth in biofilms on chlorine susceptibility of *Mycobacterium avium* and *Mycobacterium intracellulare*,” *Applied and Environmental Microbiology*, vol. 72, no. 6, pp. 4007–4011, doi: 10.1128/AEM.02573-05.
- Steingrube, VA, Gibson, JL, Brown, BA, Zhang, Y, Wilson, RW, Rajagopalan, M, & Wallace, RJ, 1995, “PCR Amplification and Restriction Endonuclease Analysis of a 65-Kilodalton Heat Shock Protein Gene Sequence for Taxonomic Separation of Rapidly Growing Mycobacteria”, *Journal of Clinical Microbiology*, vol. 33.
- Thomsen, MCF, Ahrenfeldt, J, Cisneros, JLB, Jurtz, V, Larsen, MV, Hasman, H, Aarestrup, FM, & Lund, O, 2016, “A bacterial analysis platform: An integrated system for analysing bacterial whole genome sequencing data for clinical diagnostics and surveillance,” *PLoS ONE*, vol. 11, no. 6, doi: 10.1371/journal.pone.0157718.

- Torsten S, 2014, “Barrnap: Bacterial ribosomal RNA predictor”. Available online at: <https://github.com/tseemann/barrnap>”.
- Scholssberg, 2017, “Tuberculosis and nontuberculous mycobacterial infections seventh edition”, *Washington,DC:ASM Press*.
- Turenne, CY, Tschetter, L, Wolfe, J, & Kabani, A, 2001, “Necessity of quality-controlled 16S rRNA gene sequence databases: Identifying nontuberculous Mycobacterium species,” *Journal of Clinical Microbiology*, vol. 39, no. 10, pp. 3637–3648, doi: 10.1128/JCM.39.10.3638-3648.2001.
- Wayne, LG, 1988, “International Committee on Systematic Bacteriology: Announcement of the Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics, System”. *Appl. Microbiol*, vol. 10.
- Woese, CR, 1987, “Bacterial Evolution”, *Microbiological Reviews*, vol. 51.
- Yamada-Noda, M, Ohkusu, K, Hata, H, Shah, MM, Nhung, PH, Sun, XS, Hayashi, M, & Ezaki, T, 2007, “Mycobacterium species identification - A new approach via dnaJ gene sequencing,” *Systematic and Applied Microbiology*, vol. 30, no. 6, pp. 453–462, doi: 10.1016/j.syapm.2007.06.003.
- Zolg, JW & Philippi-Schulz, S, 1994, “The Superoxide Dismutase Gene, a Target for Detection and Identification of Mycobacteria by PCR”, *Journal of Clinical Microbiology*.

Appendix

Appendix 1: List of 16S rRNA genes

| Mycobacterium Species | 16S rRNA gene Size (bp) |
|---|--------------------------------|
| <i>Mycobacterium mantenii</i> strain NLA000401474 | 1571 |
| <i>Mycobacterium marinum</i> strain ATCC 927 | 1562 |
| <i>Mycobacterium marseillense</i> strain 5356591 | 1570 |
| <i>Mycobacterium microti</i> | 1584 |
| <i>Mycobacterium monacense</i> strain B9-21-178 | 1573 |
| <i>Mycobacterium montefiorensis</i> | 1731 |
| <i>Mycobacterium moriokaense</i> | 1593 |
| <i>Mycobacterium mucogenicum</i> strain ATCC 49650 | 1582 |
| <i>Mycobacterium murale</i> | 1559 |
| <i>Mycobacterium nebraskense</i> strain UNMC-MY1349 | 1506 |
| <i>Mycobacterium alvei</i> | 1565 |
| <i>Mycobacterium neoaurum</i> | 1570 |
| <i>Mycobacterium neworleansense</i> strain ATCC 49404 | 1583 |
| <i>Mycobacterium nonchromogenicum</i> | 1566 |
| <i>Mycobacterium noviomagense</i> strain NLA000500338 | 1578 |
| <i>Mycobacterium novocastrensis</i> | 1517 |
| <i>Mycobacterium obuense</i> | 1558 |
| <i>Mycobacterium pallens</i> strain czh-8 | 1535 |
| <i>Mycobacterium palustre</i> strain E84 | 1823 |
| <i>Mycobacterium paraense</i> strain IEC26 | 1580 |
| <i>Mycobacterium paraffinicum</i> strain ATCC 12670 | 1592 |
| <i>Mycobacterium angelicum</i> strain DSM 45057T | 1510 |
| <i>Mycobacterium parafortuitum</i> | 1560 |
| <i>Mycobacterium paragordoniae</i> strain 49061 | 1593 |
| <i>Mycobacterium parakoreense</i> strain 299 | 1565 |
| <i>Mycobacterium parascrofulaceum</i> strain BAA-614 | 1568 |
| <i>Mycobacterium paraseoulensis</i> strain 31118 | 1522 |
| <i>Mycobacterium parmense</i> | 1533 |
| <i>Mycobacterium peregrinum</i> strain CIP 105382 | 1583 |
| <i>Mycobacterium phlei</i> | 1561 |
| <i>Mycobacterium phocaicum</i> strain CIP 108542 | 1582 |

| | |
|--|------|
| <i>Mycobacterium pinnipedii</i> | 1581 |
| <i>Mycobacterium anyangense</i> strain QIA-38 | 1560 |
| <i>Mycolicibacterium porcinum</i> strain CIP 105392 | 1583 |
| <i>Mycobacterium poriferae</i> | 1550 |
| <i>Mycobacterium pseudoshottsii</i> | 1553 |
| <i>Mycobacterium psychrotolerans</i> strain WA101T | 1517 |
| <i>Mycobacterium pulveris</i> strain DSM 44222T | 1592 |
| <i>Mycobacterium</i> sp. DSM 44605 | 1581 |
| <i>Mycobacterium rhodesiae</i> strain DSM 44223T | 1585 |
| <i>Mycobacterium riyadhense</i> strain NLA000201958 | 1575 |
| <i>Mycobacterium rufum</i> strain JS14 | 1559 |
| <i>Mycobacterium rutilum</i> strain czh-117 | 1578 |
| <i>Mycobacterium arabiense</i> strain YIM 121001 | 1535 |
| <i>Mycobacterium salmoniphilum</i> strain ATCC 13758 | 1956 |
| <i>Mycobacterium saopaulense</i> strain EPM10906 | 1584 |
| <i>Mycobacterium saskatchewanense</i> strain NRCM 00-250 | 1824 |
| <i>Mycobacterium scrofulaceum</i> | 1566 |
| <i>Mycobacterium sediminis</i> strain YIM M13028 | 1515 |
| <i>Mycolicibacterium senegalense</i> strain CIP 104941 | 1583 |
| <i>Mycobacterium senuense</i> strain 05-832 | 1527 |
| <i>Mycobacterium seoulense</i> strain 03-19 | 1522 |
| <i>Mycolicibacterium septicum</i> strain DSM 44393 | 1583 |
| <i>Mycobacterium setense</i> strain ABO-M06 | 1566 |
| <i>Mycobacterium aromaticivorans</i> strain JS19b1 | 1571 |
| <i>Mycobacterium sherrisi</i> | 1510 |
| <i>Mycobacterium shimoidei</i> | 1805 |
| <i>Mycobacterium shinjukuense</i> strain: GTC 2738 | 1505 |
| <i>Mycobacterium</i> sp. M175 | 1591 |
| <i>Mycobacterium simiae</i> | 1584 |
| <i>Mycobacterium smegmatis</i> strain ATCC 19420 | 1587 |
| <i>Mycobacterium sphagni</i> strain DSM44076T | 1505 |
| <i>Mycobacterium stomatepiae</i> strain DSM 45059T | 1571 |
| <i>Mycobacterium szulgai</i> | 1562 |
| <i>Mycobacterium terrae</i> | 1563 |
| <i>Mycobacterium arosiense</i> strain T1921 | 1593 |
| <i>Mycobacterium thermoresistibile</i> | 1564 |

| | |
|---|------|
| <i>Mycobacterium tokaiense</i> | 1551 |
| <i>Mycobacterium triplex</i> strain 90-1019 | 1574 |
| <i>Mycobacterium triviale</i> strain TMC 1453 | 1562 |
| <i>Mycobacterium tuberculosis</i> | 2538 |
| <i>Mycobacterium</i> sp. FI-25796 | 1589 |
| <i>Mycobacterium ulcerans</i> strain: ATCC 19423 | 1575 |
| <i>Mycobacterium vaccae</i> | 1539 |
| <i>Mycobacterium vanbaalenii</i> | 1908 |
| <i>Mycobacterium vulneris</i> strain NLA000700772 | 1571 |
| <i>Mycobacterium arupense</i> strain AR30097 | 1587 |
| <i>Mycolicibacterium wolinskyi</i> strain ATCC 700010 | 1585 |
| <i>Mycobacterium xenopi</i> | 1580 |
| <i>Mycobacterium yongonense</i> 05-1390 | 1595 |
| <i>Mycobacterium asiaticum</i> | 1566 |
| <i>Mycolicibacterium aubagnense</i> strain CIP 108543 | 1582 |
| <i>Mycobacterium aurum</i> | 1558 |
| <i>Mycobacterium austroafricanum</i> | 1562 |
| <i>Mycobacterium avium</i> subsp. <i>avium</i> | 1572 |
| <i>Mycobacterium paratuberculosis</i> | 1563 |
| <i>Mycobacterium avium</i> subsp. <i>silvaticum</i> strain ATCC 49884 | 1542 |
| <i>Mycobacterium bonickei</i> strain W5998 | 1598 |
| <i>Mycobacterium bohemicum</i> | 1516 |
| <i>Mycobacterium botniense</i> | 1787 |
| <i>Mycobacterium bouchedurhonense</i> strain 4355387 | 1598 |
| <i>Mycobacterium bourgelatii</i> | 1584 |
| <i>Mycobacterium branderi</i> | 1569 |
| <i>Mycobacterium brisbanense</i> strain W6743 | 1599 |
| <i>Mycobacterium brumae</i> | 1549 |
| <i>Mycobacterium canariasense</i> | 1533 |
| <i>Mycobacterium caprae</i> | 1524 |
| <i>Mycobacterium celatum</i> | 1460 |
| <i>Mycobacterium celeriflavum</i> strain AFPC-000207 | 1562 |
| <i>Mycobacteroides chelonae</i> strain CIP 104535 | 1581 |
| <i>Mycobacterium chimaera</i> strain FI-0169T | 1766 |
| <i>Mycobacterium chitae</i> | 1557 |
| <i>Mycobacterium chlorophenolicum</i> DSM 43826 | 1566 |

| | |
|--|------|
| <i>Mycobacteroides abscessus</i> strain CIP 104536 | 1581 |
| <i>Mycobacterium chubuense</i> | 1572 |
| <i>Mycobacterium colombiense</i> train:10B | 1807 |
| <i>Mycolicibacterium conceptionense</i> strain CIP 108544 | 1583 |
| <i>Mycobacterium confluentis</i> strain DSM 44017T | 1504 |
| <i>Mycobacterium conspicuum</i> | 1593 |
| <i>Mycobacterium cookii</i> | 1559 |
| <i>Mycobacterium cosmeticum</i> strain LTA-388 | 1507 |
| <i>Mycobacterium crocinum</i> strain czh-42 | 1598 |
| <i>Mycobacterium diernhoferi</i> | 1558 |
| <i>Mycobacterium doricum</i> | 1550 |
| <i>Mycobacteroides abscessus subsp. bolletii</i> strain CIP 108541 | 1581 |
| <i>Mycobacterium duvalii</i> | 1502 |
| <i>Mycobacterium elephantis</i> | 1517 |
| <i>Mycobacterium engbaekii</i> | 1568 |
| <i>Mycobacterium europaeum</i> strain DSM 45397 | 1796 |
| <i>Mycobacterium fallax</i> | 1570 |
| <i>Mycolicibacterium farcinogenes</i> strain NCTC 10955 | 1583 |
| <i>Mycobacterium flavescens</i> | 1554 |
| <i>Mycobacterium florentinum</i> strain FI-93171T | 1788 |
| <i>Mycobacterium fluoranthenivorans</i> strain FA-4 | 1594 |
| <i>Mycobacterium fortuitum subsp. acetamidolyticum</i> strain DSM442 | 1505 |
| <i>Mycobacterium africanum</i> | 1533 |
| <i>Mycolicibacterium fortuitum subsp. fortuitum</i> DSM 46621 | 1583 |
| <i>Mycobacterium fragae</i> strain HF8705 | 1552 |
| <i>Mycobacterium franklinii</i> strain CV002 | 1506 |
| <i>Mycobacterium frederiksbergense</i> strain DSM 44346 | 1574 |
| <i>Mycobacterium gadium</i> | 1556 |
| <i>Mycobacterium gastris</i> | 1569 |
| <i>Mycobacterium genavense</i> | 1549 |
| <i>Mycobacterium gilvum</i> | 1528 |
| <i>Mycobacterium goodii</i> | 1517 |
| <i>Mycobacterium gordonae</i> | 1561 |
| <i>Mycobacterium agri</i> strain DSM 44515T | 1556 |
| <i>Mycobacterium haemophilum</i> strain DSM 44634 | 1526 |
| <i>Mycobacterium hassiacum</i> | 1591 |

| | |
|--|------|
| <i>Mycobacterium heckeshornense</i> | 1527 |
| <i>Mycobacterium heidelbergense</i> | 1545 |
| <i>Mycobacterium heraklionense</i> strain NCTC 13432 | 1527 |
| <i>Mycobacterium hiberniae</i> | 1519 |
| <i>Mycobacterium hippocampi</i> strain BFLP-6T | 1573 |
| <i>Mycobacterium hodleri</i> | 1559 |
| <i>Mycobacterium holsaticum</i> strain 1406 | 1526 |
| <i>Mycolicibacterium houstonense</i> strain ATCC 49403 | 1583 |
| <i>Mycobacterium aichiense</i> | 1556 |
| <i>Mycobacterium insubricum</i> strain FI-06250 | 1995 |
| <i>Mycobacterium interjectum</i> ATCC:51457 | 1531 |
| <i>Mycobacterium intermedium</i> | 1541 |
| <i>Mycobacterium intracellulare</i> | 1540 |
| <i>Mycobacterium iranicum</i> strain M05 | 1550 |
| <i>Mycobacterium kansasii</i> | 1570 |
| <i>Mycobacterium komossense</i> | 1562 |
| <i>Mycobacterium koreense</i> strain 01-305 | 1574 |
| <i>Mycobacterium kubicae</i> | 1561 |
| <i>Mycobacterium kumamotonense</i> strain: CST7274 | 1564 |
| <i>Mycobacterium algericum</i> DSM 45454 | 1521 |
| <i>Mycobacterium kyorinense</i> strain: KUM 060204 | 1570 |
| <i>Mycobacterium lacus</i> | 1570 |
| <i>Mycobacterium lentiflavum</i> | 1552 |
| <i>Mycobacterium leprae</i> | 1585 |
| <i>Mycobacterium litorale</i> strain F4 | 1580 |
| <i>Mycobacterium llatzerense</i> | 1597 |
| <i>Mycobacterium longobardum</i> strain DSM 45394 | 1560 |
| <i>Mycobacterium madagascariense</i> | 1570 |
| <i>Mycobacterium mageritense</i> strain DSM 44476 | 1597 |
| <i>Mycobacterium malmoense</i> | 1563 |
| <i>Mycobacterium alsense</i> strain TB 1906T | 1565 |

Appendix 2: List of whole-genome sequence

| Mycobacterium species | Genome Size (bp) |
|--|-------------------------|
| <i>Mycobacterium alsense</i> strain E2978 | 5656398 |
| <i>Mycobacterium angelicum</i> strain DSM 45057 | 6662911 |
| <i>Mycobacterium aquaticum</i> strain RW6 | 7927592 |
| <i>Mycobacterium arosiense</i> ATCC BAA-1401 | 5980206 |
| <i>Mycobacterium asiaticum</i> strain 1081914.2 | 6035124 |
| <i>Mycobacterium avium</i> subsp. <i>paratuberculosis</i> K-10 | 4829781 |
| <i>Mycobacterium bohemicum</i> strain DSM 44277 | 5420516 |
| <i>Mycobacterium botniense</i> strain JCM 17322 | 4335050 |
| <i>Mycobacterium bouchedurhonense</i> strain DSM 45439 | 5897311 |
| <i>Mycobacterium bourgelatii</i> strain JCM 30725 | 6912804 |
| <i>Mycobacterium branderi</i> strain DSM 44624 | 5903509 |
| <i>Mycobacterium canettii</i> CIPT 140010059 | 4482059 |
| <i>Mycobacterium celatum</i> strain ATCC 51131 | 4662053 |
| <i>Mycobacterium colombiense</i> CECT 3035 | 5579559 |
| <i>Mycobacterium conspicuum</i> strain DSM 44136 | 6200614 |
| <i>Mycobacterium cookii</i> strain JCM 12404 | 5318517 |
| <i>Mycobacterium europaeum</i> strain DSM 45397 | 5630674 |
| <i>Mycobacterium florentinum</i> strain DSM 44852 | 6178353 |
| <i>Mycobacterium fragae</i> strain DSM 45731 | 4731047 |
| <i>Mycobacterium gastri</i> strain DSM 43505 | 5816659 |
| <i>Mycobacterium genavense</i> ATCC 51234 | 4936071 |
| <i>Mycobacterium gordonae</i> strain CTRI 14-8773 | 7552315 |
| <i>Mycobacterium haemophilum</i> strain UC3 | 4363991 |
| <i>Mycobacterium heckeshornense</i> strain RLE | 5010173 |
| <i>Mycobacterium heidelbergense</i> strain DSM 44471 | 4999105 |
| <i>Mycobacterium intermedium</i> strain HMC2_M5 | 6854888 |
| <i>Mycobacterium intracellulare</i> ATCC 13950 | 5328562 |

| | |
|---|---------|
| <i>Mycobacterium intracellulare subsp. chimaera</i> strain FLAC0070 | 5408332 |
| <i>Mycobacterium kansasii</i> ATCC 12478 | 6432277 |
| <i>Mycobacterium kubicae</i> strain ACS1160 | 5618076 |
| <i>Mycobacterium lacus</i> strain DSM 44577 | 4905288 |
| <i>Mycobacterium leprae</i> TN | 3268203 |
| <i>Mycobacterium liflandii</i> 128FXT | 6208955 |
| <i>Mycobacterium mantenii</i> strain E152 | 5836558 |
| <i>Mycobacterium marinum</i> M | 6636827 |
| <i>Mycobacterium marseillense</i> strain 1165549.7 | 5255222 |
| <i>Mycobacterium montefiorensense</i> strain BS | 5742797 |
| <i>Mycobacterium noviomagense</i> strain DSM 45145 | 4739740 |
| <i>Mycobacterium palustre</i> strain DSM 44572 | 6037522 |
| <i>Mycobacterium paraense</i> strain IEC26 | 5619528 |
| <i>Mycobacterium paraffinicum</i> strain M11 | 6474701 |
| <i>Mycobacterium paragordoniae</i> strain 49061 | 6730319 |
| <i>Mycobacterium paraintracellulare</i> | 5501090 |
| <i>Mycobacterium paraseoulense</i> strain DSM 45000 | 6078492 |
| <i>Mycobacterium parmense</i> strain DSM 44553 | 5891740 |
| <i>Mycobacterium pseudoshottsii</i> JCM 15466 | 6061597 |
| <i>Mycobacterium riyadhense</i> strain DSM 45176 | 6269850 |
| <i>Mycobacterium saskatchewanense</i> strain DSM 44616 | 5930935 |
| <i>Mycobacterium scrofulaceum</i> strain E3039 | 5536898 |
| <i>Mycobacterium seoulense</i> strain JCM 16018 | 5531300 |
| <i>Mycobacterium sherrisii</i> strain BC1_M4 | 5685834 |
| <i>Mycobacterium shimoidei</i> strain HMC_M2 | 4720739 |
| <i>Mycobacterium shinjukuense</i> strain CCUG 53584 | 4409896 |
| <i>Mycobacterium shottsii</i> strain JCM 12657 | 5973149 |
| <i>Mycobacterium simiae</i> strain MsiGto | 6686819 |
| <i>Mycobacterium simulans</i> strain FB-527 | 6234132 |
| <i>Mycobacterium stomatepieae</i> strain JCM 17783 | 6210822 |

| | |
|---|---------|
| <i>Mycobacterium szulgai</i> strain DSM 44166 | 6672659 |
| <i>Mycobacterium timonense</i> strain CCUG 56329 | 6009989 |
| <i>Mycobacterium triplex</i> strain DSM 44626 | 6366285 |
| <i>Mycobacterium tuberculosis</i> H37Rv | 4411532 |
| <i>Mycobacterium tuberculosis variant africanum</i> K85 supercont1.1 | 4432952 |
| <i>Mycobacterium tuberculosis variant caprae</i> strain MB2 | 4278564 |
| <i>Mycobacterium ulcerans</i> Agy99 | 5631606 |
| <i>Mycobacterium vulneris</i> strain DSM 45247 | 6266718 |
| <i>Mycobacterium xenopi</i> RIVM700367 | 4434836 |
| <i>Mycobacteroides abscessus</i> strain FLAC013 | 5074222 |
| <i>Mycobacterium chelonae</i> | 4898027 |
| <i>Mycobacteroides franklinii</i> strain 1559 | 5023635 |
| <i>Mycobacteroides immunogenum</i> strain SMUC14 | 5566491 |
| <i>Mycobacteroides salmoniphilum</i> strain D16Q15 | 4934017 |
| <i>Mycobacteroides saopaulense</i> strain EPM10906 | 4649175 |
| <i>Mycolicibacillus koreensis</i> strain HMC_M3 | 3873226 |
| <i>Mycolicibacillus trivialis</i> strain DSM 44153 | 3591083 |
| <i>Mycolicibacter algericus</i> DSM 45454 | 4619277 |
| <i>Mycolicibacter arupensis</i> strain GUC1 | 4441410 |
| <i>Mycolicibacter engbaekii</i> strain ATCC 27353 | 4521435 |
| <i>Mycolicibacter heraklionensis</i> strain Davo | 5109649 |
| <i>Mycolicibacter hiberniae</i> strain ATCC 49874 | 4342192 |
| <i>Mycolicibacterium aichiense</i> strain JCM 6376 | 5925482 |
| <i>Mycolicibacterium alvei</i> strain JCM 12272 | 5712683 |
| <i>Mycolicibacterium anyangense</i> strain JCM 30275 | 5696751 |
| <i>Mycolicibacterium arabiense</i> strain JCM 18538 | 6017160 |
| <i>Mycolicibacterium aromaticivorans</i> JS19b1 = JCM 16368 strain JS19b1 | 6297623 |
| <i>Mycolicibacterium aubagnense</i> strain DSM 45150 6191255 | 6191255 |
| <i>Mycolicibacterium aurum</i> isolate liquid 6038730 | 6038730 |

| | |
|---|---------|
| <i>Mycolicibacterium boenickei</i> strain CCUG47580 | 6544811 |
| <i>Mycolicibacterium brisbanense</i> strain JCM15654 | 7387429 |
| <i>Mycolicibacterium brumae</i> strain CIP1034565 | 3878367 |
| <i>Mycolicibacterium canariasense</i> strain JCM15298 | 6734449 |
| <i>Mycolicibacterium celeriflavum</i> strain 852002-51296_SCH5728562-a | 4927128 |
| <i>Mycolicibacterium chitae</i> strain JCM 12403 | 5482061 |
| <i>Mycolicibacterium chlorophenolicum</i> strain DSM 43826 | 7379285 |
| <i>Mycolicibacterium chubuense</i> NBB4 | 5583723 |
| <i>Mycolicibacterium conceptionense</i> strain MLE | 7098887 |
| <i>Mycolicibacterium confluentis</i> strain DSM 44017 | 5841691 |
| <i>Mycolicibacterium cosmeticum</i> strain DSM 44829 | 6446106 |
| <i>Mycolicibacterium diernhoferi</i> strain Bard | 5981922 |
| <i>Mycolicibacterium doricum</i> strain DSM 44339 | 3952103 |
| <i>Mycolicibacterium duvalii</i> strain IP141180004 | 5603691 |
| <i>Mycolicibacterium elephantis</i> strain Lipa | 5187616 |
| <i>Mycolicibacterium fallax</i> strain DSM 44179 | 4232450 |
| <i>Mycolicibacterium farcinogenes</i> strain DSM 43637 | 6062162 |
| <i>Mycolicibacterium flavescens</i> strain M6 | 5972870 |
| <i>Mycolicibacterium fluoranthenvivorans</i> strain DSM 44556 | 6408998 |
| <i>Mycolicibacterium fortuitum</i> subsp. <i>fortuitum</i> DSM 46621 = ATCC 6841 strain DSM 46621 | 6300050 |
| <i>Mycolicibacterium frederiksbergense</i> strain LB 501T | 6086872 |
| <i>Mycolicibacterium gadium</i> strain JCM 12688 | 5964999 |
| <i>Mycolicibacterium gilvum</i> PYR-GCK | 5619607 |
| <i>Mycolicibacterium goodii</i> strain X7B | 7105933 |
| <i>Mycolicibacterium hassiacum</i> DSM 44199 | 5000164 |
| <i>Mycolicibacterium helvum</i> strain JCM 30396 | 6400811 |
| <i>Mycolicibacterium hippocampi</i> strain JCM 30996 | 6160733 |
| <i>Mycolicibacterium hodleri</i> strain S5.20 | 6384412 |
| <i>Mycolicibacterium holsaticum</i> strain M7 | 5748336 |

| | |
|---|---------|
| <i>Mycolicibacterium houstonense</i> strain ATCC 49403T | 5525743 |
| <i>Mycolicibacterium insubricum</i> strain DSM 45130 | 4553680 |
| <i>Mycolicibacterium iranicum</i> strain H39 | 6484789 |
| <i>Mycolicibacterium komanii</i> strain GPK 1020 | 5378970 |
| <i>Mycolicibacterium litorale</i> strain CGMCC 4.5724 Ga0104440_101 | 2126661 |
| <i>Mycolicibacterium llatzerense</i> strain CLUC14 | 6091591 |
| <i>Mycolicibacterium madagascariense</i> strain JCM 13574 | 5712088 |
| <i>Mycolicibacterium mageritense</i> strain JCM 12375 | 8006721 |
| <i>Mycolicibacterium malmesburyense</i> strain WCM 7299 | 5470555 |
| <i>Mycolicibacterium monacense</i> strain 852013-50142_SCH4511227 | 5627631 |
| <i>Mycolicibacterium moriokaense</i> strain CIP105393 | 6217364 |
| <i>Mycolicibacterium mucogenicum</i> strain CSUR P2099 | 6210127 |
| <i>Mycolicibacterium murale</i> strain JCM 13392 | 6838280 |
| <i>Mycolicibacterium neoaurum</i> VKM Ac-1815D | 5421267 |
| <i>Mycolicibacterium neworleansense</i> strain ATCC 49404T | 3252791 |
| <i>Mycolicibacterium novocastrense</i> strain GA-2617 | 5639806 |
| <i>Mycolicibacterium obuense</i> strain UC1 | 6381451 |
| <i>Mycolicibacterium parafortuitum</i> strain CCUG 20999 | 6136108 |
| <i>Mycolicibacterium peregrinum</i> strain CSUR P2098 | 7109636 |
| <i>Mycolicibacterium phlei</i> RIVM601174 | 5681954 |
| <i>Mycolicibacterium phocaicum</i> strain DSM 45104 | 5771543 |
| <i>Mycolicibacterium porcinum</i> strain ACS3670 | 6778270 |
| <i>Mycolicibacterium poriferae</i> strain JCM 12603 | 5712830 |
| <i>Mycolicibacterium psychrotolerans</i> strain JCM 13323 | 5732362 |
| <i>Mycolicibacterium pulveris</i> strain JCM 6370 | 5484749 |
| <i>Mycolicibacterium rhodesiae</i> NBB3 | 6415739 |
| <i>Mycobacterium rufum</i> strain JS14 | 6176413 |
| <i>Mycolicibacterium rutilum</i> strain DSM 45405 | 5987931 |
| <i>Mycolicibacterium sarraceniae</i> strain JCM 30395 | 4828255 |
| <i>Mycolicibacterium sediminis</i> strain JCM 17899 | 6250641 |

| | |
|---|---------|
| <i>Mycolicibacterium senegalense</i> strain CK1 | 6738555 |
| <i>Mycolicibacterium septicum</i> DSM 44393 strain type strain: DSM 44393 | 6872299 |
| <i>Mycolicibacterium setense</i> strain Manresensis | 6065527 |
| <i>Mycolicibacterium smegmatis</i> MC2 155 | 6988209 |
| <i>Mycolicibacterium sphagni</i> strain ATCC 33027 | 6066978 |
| <i>Mycolicibacterium thermoresistibile</i> ATCC 19527 | 4870742 |
| <i>Mycolicibacterium tokaiense</i> strain JCM 6373 | 6328149 |
| <i>Mycolicibacterium tusciae</i> JS617 | 7306213 |
| <i>Mycolicibacterium vaccae</i> ATCC 25954 | 6223660 |
| <i>Mycolicibacterium vanbaalenii</i> PYR-1 | 6491865 |
| <i>Mycolicibacterium wolinskyi</i> strain CDC_01 | 7449739 |
| <i>Mycolicibacter kumamotonensis</i> strain Roo | 5329013 |
| <i>Mycolicibacter longobardus</i> strain DSM 45394 | 4812247 |
| <i>Mycolicibacter minnesotensis</i> strain DSM 45633 | 4187822 |
| <i>Mycolicibacter nonchromogenicus</i> strain DSM 44164 | 4465329 |
| <i>Mycolicibacter senuensis</i> strain DSM 44999 | 4534628 |
| <i>Mycolicibacter sinensis</i> | 4643668 |
| <i>Mycolicibacter terrae</i> strain CIP 104321 | 4524815 |

Appendix 3: Screening of virulence gene profile

| Relate genes | Query _seq | <i>M.abscessus</i> | <i>M.africanum</i> | <i>M.avium</i> | <i>M.bovis</i> | <i>M.gilvum</i> | <i>M.indicus</i> | <i>M.intracellulare</i> | <i>M.leprae</i> | <i>M.ulcerans</i> | <i>M.tuberculosis</i> |
|-----------------|------------|--------------------|--------------------|----------------|----------------|-----------------|------------------|-------------------------|-----------------|-------------------|-----------------------|
| <i>glnA1</i> | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>leuD</i> | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>lysA</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>proC</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>purC</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>trpD</i> | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| <i>narX</i> | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| <i>narG</i> | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| <i>narH</i> | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| <i>narI</i> | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| <i>narJ</i> | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| <i>narK2</i> | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| <i>nuoG</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>cyp125</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>fadE28</i> | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>fadE29</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>caeA</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>erp</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>Rv0926</i> | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| <i>atf</i> | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| <i>ecf</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>fad23</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| <i>fadE5</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>fmt</i> | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| <i>gap-like</i> | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| <i>gap</i> | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |

| | | | | | | | | | | | |
|---------------|---|---|---|---|---|---|---|---|---|---|---|
| <i>gtf1</i> | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| <i>gtf2</i> | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| <i>gtf3</i> | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| <i>mbtH</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>mmpL10</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| <i>mmpL4a</i> | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| <i>mmpL4b</i> | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| <i>mmpS4</i> | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| <i>mps1</i> | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| <i>mps2</i> | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| <i>papA3</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| <i>pe</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| <i>pks</i> | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| <i>rmlA</i> | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| <i>rmlB</i> | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| <i>rmt2</i> | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>rmt3</i> | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| <i>rmt4</i> | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| <i>hbhA</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>lprG</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>mmaA4</i> | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>cmaA2</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| <i>adhD</i> | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| <i>chp</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| <i>fadD13</i> | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| <i>lipR</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| <i>mymA</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| <i>sadH</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| <i>tgs4</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| <i>drnC</i> | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |

| | | | | | | | | | | | |
|----------------|---|---|---|---|---|---|---|---|---|---|---|
| <i>fadD22</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| <i>fadD26</i> | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| <i>fadD28</i> | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| <i>fadD29</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| <i>lppx</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| <i>mas</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| <i>mmpL7</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| <i>papA5</i> | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| <i>pks15/1</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| <i>pks15</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| <i>pks1</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| <i>ppsA</i> | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| <i>ppsB</i> | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| <i>ppsC</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| <i>ppsD</i> | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| <i>ppsE</i> | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| <i>tesA</i> | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |
| <i>kefB</i> | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| <i>pcaA</i> | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>mmpL8</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| <i>papA1</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| <i>papA2</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| <i>pks2</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| <i>stf0</i> | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| <i>chp1</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| <i>fad23</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| <i>icl2</i> | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>lpqY</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>sap</i> | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| <i>sugA</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

| | | | | | | | | | | | |
|---------------|---|---|---|---|---|---|---|---|---|---|---|
| <i>sugB</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>sugC</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>ctpV</i> | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| <i>irtA</i> | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| <i>irtB</i> | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| <i>exiT</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>fxbA</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>fxbBC</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>fxuA</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>fxuB</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>fxuC</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>fxuD</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>mmpL11</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>mmpL3</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>ideR</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>fadD33</i> | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| <i>fadE14</i> | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| <i>mbtA</i> | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| <i>mbtB</i> | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| <i>mbtC</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>mbtD</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>mbtE</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>mbtF</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>mbtG</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>mbtH</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>mbtI</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| <i>mbtJ</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| <i>mbtK</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>kasB</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| <i>icl</i> | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |

| | | | | | | | | | | | |
|--------------|---|---|---|---|---|---|---|---|---|---|---|
| <i>lipF</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| <i>sapM</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 |
| <i>panC</i> | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 |
| <i>panD</i> | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>plcA</i> | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| <i>plcB</i> | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| <i>plcC</i> | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| <i>plcD</i> | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| <i>mgtC</i> | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| <i>mce1A</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| <i>mce1B</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 |
| <i>mce1C</i> | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| <i>mce1D</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| <i>mce1E</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| <i>mce1F</i> | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| <i>mce2A</i> | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| <i>mce2B</i> | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| <i>mce2C</i> | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| <i>mce2D</i> | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| <i>mce2E</i> | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| <i>mce2F</i> | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| <i>mce3A</i> | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| <i>mce3B</i> | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| <i>mce3C</i> | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| <i>mce3D</i> | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| <i>mce3E</i> | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| <i>mce3F</i> | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| <i>mce4A</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| <i>mce4B</i> | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>mce4C</i> | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |

| | | | | | | | | | | | |
|--------------|---|---|---|---|---|---|---|---|---|---|---|
| <i>mce4D</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>mce4E</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>mce4F</i> | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>mce5A</i> | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 |
| <i>mce5B</i> | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 |
| <i>mce5C</i> | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| <i>mce5D</i> | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| <i>mce5E</i> | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| <i>mce5F</i> | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 |
| <i>mce6A</i> | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>mce6B</i> | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| <i>mce6C</i> | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| <i>mce6D</i> | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| <i>mce6E</i> | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| <i>mce6F</i> | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| <i>mce7A</i> | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| <i>mce7B</i> | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| <i>mce7C</i> | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| <i>mce7D</i> | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| <i>mce7E</i> | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| <i>mce7F</i> | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| <i>mce8A</i> | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| <i>mce8B</i> | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| <i>mce8C</i> | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| <i>mce8D</i> | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| <i>mce8E</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>mce8F</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>mce9A</i> | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| <i>mce9B</i> | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| <i>mce9C</i> | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |

| | | | | | | | | | | | |
|-----------------------|---|---|---|---|---|---|---|---|---|---|---|
| <i>mce9D</i> | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| <i>mce9E</i> | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| <i>mce9F</i> | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| <i>ctpC</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| <i>ndk</i> | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| <i>PE_PGR S30</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| <i>ptpA</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>mpa</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>pafA</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| <i>zmp1</i> | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| <i>relA</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>devR/dos R</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>devS</i> | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| <i>mosR</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| <i>mprA</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>mprB</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>phoP</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>phoR</i> | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| <i>prpA</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>prpB</i> | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>regX3</i> | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>senX3</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>sigA/rpo V</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>sigD</i> | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| <i>sigE</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>sigF</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>sigH</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>sigL</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |

| | | | | | | | | | | | |
|---------------|---|---|---|---|---|---|---|---|---|---|---|
| <i>sigM</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>whiB3</i> | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>lpqH</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>hspX</i> | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| <i>fbpA</i> | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| <i>fbpB</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| <i>fbpC</i> | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>eis</i> | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>pknG</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>secA2</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>PE35</i> | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| <i>PPE68</i> | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| <i>eccA1</i> | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| <i>eccB1</i> | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| <i>eccCa1</i> | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| <i>eccCb1</i> | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| <i>eccD1</i> | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| <i>eccE1</i> | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| <i>espA</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| <i>espB</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| <i>espC</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| <i>espD</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| <i>espE</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| <i>espF</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| <i>espG1</i> | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| <i>espH</i> | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| <i>espI</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| <i>espJ</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| <i>espK</i> | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| <i>espL</i> | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |

| | | | | | | | | | | | |
|--------------|---|---|---|---|---|---|---|---|---|---|---|
| <i>espR</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>esxA</i> | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| <i>esxB</i> | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| <i>mycP1</i> | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| <i>PE36</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| <i>PPE69</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| <i>eccA2</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| <i>eccB2</i> | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| <i>eccC2</i> | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| <i>eccD2</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| <i>eccE2</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| <i>espG2</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| <i>esxC</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| <i>esxD</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| <i>mycP2</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| <i>PE5</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>PPE4</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>eccA3</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>eccB3</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>eccC3</i> | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>eccD3</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>eccE3</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>espG3</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>esxG</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>esxH</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>mycP3</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>eccB4</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>eccC4</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>eccD4</i> | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| <i>esxT</i> | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

| | | | | | | | | | | | |
|---------------|---|---|---|---|---|---|---|---|---|---|---|
| <i>esxU</i> | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>mycP4</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>PE18</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| <i>PE19</i> | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| <i>PPE25</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| <i>PPE26</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| <i>PPE27</i> | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| <i>PPE41</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| <i>cyp143</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| <i>eccA5</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| <i>eccB5</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| <i>eccCa5</i> | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| <i>eccCb5</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| <i>eccD5</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| <i>eccE5</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| <i>esxM</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| <i>esxN</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| <i>mycP5</i> | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| <i>ahpC</i> | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| <i>katG</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| <i>sodC</i> | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| <i>sodA</i> | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| <i>ompA</i> | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| <i>mlsA1</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>mlsA2</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>mlsB</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

