Comparative Genomics and Transcriptomic Analysis of

*Mycobacterium Kansasii*

Thesis by

Yara Alzahid

In Partial Fulfilment of the Requirements

For the Degree of

Master of Science (MSc)

King Abdullah University of Science and Technology, Thuwal,

Kingdom of Saudi Arabia

April 2014
EXAMINATION COMMITTEE APPROVALS

The dissertation/thesis of Yara Alzahid is approved by the examination committee.

**Committee Chairperson:** Dr. Arnab Pain  
**Committee Member:** Dr. Christoph Gehring  
**Committee Member:** Dr. Timothy Ravasi
ABSTRACT

Comparative Genomics and Transcriptomic Analysis of *Mycobacterium Kansasi*i

Yara Alzahid

The group of Mycobacteria is one of the most intensively studied bacterial taxa, as they cause the two historical and worldwide known diseases: leprosy and tuberculosis. Mycobacteria not identified as tuberculosis or leprosy complex, have been referred to by ‘environmental mycobacteria’ or ‘Nontuberculous mycobacteria (NTM). *Mycobacterium kansasii* (*M. kansasii*) is one of the most frequent NTM pathogens, as it causes pulmonary disease in immuno-competent patients and pulmonary, and disseminated disease in patients with various immuno-deficiencies. There have been five documented subtypes of this bacterium, by different molecular typing methods, showing that type I causes tuberculosis-like disease in healthy individuals, and type II in immune-compromised individuals. The remaining types are said to be environmental, thereby, not causing any diseases. The aim of this project was to conduct a comparative genomic study of *M. kansasii* types I-V and investigating the gene expression level of those types. From various comparative genomics analysis, provided genomics evidence on why *M. kansasii* type I is considered pathogenic, by focusing on three key elements that are involved in virulence of Mycobacteria: ESX secretion system, Phospholipase c (*plcb*) and Mammalian cell entry (*Mce*) operons. The results showed the lack of the *espA* operon in types II-V, which renders the ESX-1 operon dysfunctional, as *espA* is one of the key factors that control this secretion system. However, gene expression analysis showed this operon to be deleted in types II, III and IV. Furthermore, *plcB* was found to be truncated in types III and IV. Analysis of *Mce* operons (1-4) show that *mce*-1 operon is duplicated, *mce*-2 is absent and *mce*-3 and *mce*-4 is present in one copy in *M. kansasii* types I-V. Gene expression profiles of type I-IV, showed that the secreted proteins of ESX-1 were slightly upregulated in types II-IV when compared to type I and the secreted forms of ESX-5 were highly down regulated in the same types. Differentially expressed genes in types II-IV were also evaluated and validated by qPCR for selected genes. This study gave a general view of the genome of this bacterium and its types, highlighted some different aspects of its subtypes and supplemented by gene expression data.
I would like to express the deepest appreciation to my advisor Dr. Arnab Pain for giving me the opportunity to work on this project and his continuous support and advice throughout this thesis. I would also like to thank Dr. Abdallah Abdallah for his guidance and valuable ideas, without him, this thesis would have never been possible, as he taught me everything I need to know about Mycobacteria, and have been a constant source of inspiration during my time here for this project. I would also like to thank my committee members, Dr. Christoph Gehring and Dr. Timothy Ravasi for serving as my committee members. I would especially like to thank Shoaib Amini for generating the data for this project and Dr. Shwen Ho for all his guidance, support and ideas during the project, you always managed to make all the bioinformatics work simple for me to understand. Last but not least, I place a deep sense of gratitude to my family members and friends here at KAUST, who have truly become like family to me, without you guys I would have never made it this far with such high spirits.
TABLE OF CONTENT

List of Abbreviations ........................................................................................................... 9

List of figures ...................................................................................................................... 10

List of tables ....................................................................................................................... 12

CHAPTER 1. INTRODUCTION ................................................................................. 13

1.1 The mycobacterium genus ......................................................................................... 13

1.2 *M. tuberculosis* ........................................................................................................ 16

1.3 *M. kansasii* ............................................................................................................... 18

1.4 Types of *M. kansasii* ............................................................................................... 19

1.5 *M. kansasii* pathogenicity ....................................................................................... 21

1.6 *M. kansasii* compared to other mycobacteria species ........................................... 22

1.7 PE/PPE family ........................................................................................................... 23

1.8 Type VII (ESX) secretion system ............................................................................. 25

1.9 Mammalian cell entry (*Mce*) operons .................................................................... 28

1.10 Comparative genomics and Mycobacteria ......................................................... 29

1.11 Aim of thesis work ................................................................................................. 30

1.11.1 Genome Analysis of *M. kansasii* types I-V .................................................. 30

1.11.2 Comparative transcriptome analysis ................................................................. 30
CHAPTER 2. GENOME ANALYSIS OF *M. KANSASII* TYPES I-V ..............31

2.1 Material and methods .........................................................31

2.1.1 DNA Isolation .................................................................31

2.1.2 Construction, sequencing, and assembly of *M. kansasii* type I-V ......31

2.1.3 Genome analysis and annotation .........................................31

2.1.4 Comparative genomics of the 5 types of *M. kansasii* .................32

2.1.5 PCR Confirmation of the *espA* operon ................................33

2.2 Results and discussion ...........................................................34

2.2.1 General features of *M. kansasii* type I-V genome .........................34

2.2.2 Comparing genomes within types of *M. kansasii* ..................36

  Blastn similarities between *M. kansasii* I-V ................................36

  OrthoMCL analysis of the 5 types ............................................37

  Core *M. kansasii* genome ....................................................39

  Unique genes present within each type ..................................40

  PE/PPE .................................................................................40

  ESX system presence in *M. kansasii* .......................................41

  Phospholipase C (*Plc*) .........................................................46

  Mammalian cell entry (*Mce*) operons ......................................48

2.3 Future work .................................................................49
CHAPTER 3. mRNA PROFILE OF M. KANSASII

3.1 Material and methods

3.1.1 Bacteria cultures

3.1.2 RNA Extraction and Removal of rRNA

3.1.3 Strand Specific RNA Library

3.1.4 RNA-Seq

3.1.5 Validation by qPCR

3.2 Results and discussion

3.2.1 RNA sequencing profiles of M. kansasii types I-V

3.2.2 Comparative RNA seq of ESX-1 and 5

3.2.3 RNA seq validation (qRT-PCR) of selected differentially expressed genes

3.3 Future work

CONCLUSIONS

BIBLIOGRAPHY
List of Abbreviations

MTBC Mycobacterium tuberculosis complex
TB tuberculosis
Mtb *Mycobacterium tuberculosis*
MDR Multi-drug resistant
XDR Extremely drug-resistant
NTM nontuberculosis mycobacteria
MPT mitochondrial permeability transition
RFLP restriction fragment length polymorphism
PFGE pulsed-field gel electrophoresis
PRA PCR restriction analysis
HIV human immunodeficiency virus
ESAT-6 the 6 kDa early secretory antigenic target
CFP-10 10 kDa culture filtrate antigen
BRIG BLAST Ring Image Generator
MAUVE Multiple Alignment of Conserved Genomic Sequence With Rearrangements
NGS next generation sequencing technologies
Plc Phospholipase C
Mce Mammalian cell entry
TraDIS Transposon-directed insertion site sequencing
List of figures

Figure 1. Overall structure of the mycobacterial cell wall complex of *Mycobacterium tuberculosis* .................................................................14

Figure 2. Phylogenetic tree based on 16S rRNA gene sequences. This shows the relatedness of the species on neighboring branches .................................15

Figure 3. Detailed representation of the mode of transmission and pathogenesis of *M. tuberculosis* .................................................................17

Figure 4. PE and PPE family subgroups ..................................................23

Figure 5. The PE/PPE protein complex formed by known gene products of *M. tuberculosis* H37Rv strain. .................................................................24

Figure 6. General view of types I-IV secretion pathways .............................25

Figure 7. Predicted genetic organization of the 5 ESX loci in *M. tuberculosis* H37Rv strain ..................................................................................27

Figure 8. Organization of the four *mce* operons in H37Rv .........................28

Figure 9. The bioinformatics pipeline used to assemble *M. kansasii* type I-V ......32

Figure 10. General diagram explaining the OrthoMCL algorithm for clustering orthologous proteins .................................................................33

Figure 11. Blastn comparation of *M. kansasii* types II-V compared to *M. kansasii* type I..................................................................................36

Figure 12. Venn diagram showing orthologous clusters among five *M. kansasii* types (MK1-MK5) determined by OrthoMCL ........................................38

Figure 13. Aligned sequences of *M. kansasii* types I-V. ............................39

Figure 14. Number of PE/PPE genes in *M. kansasii* types I-V ..................41

Figure 15. Schematic representation of the ESX operons present in *M. kansasii* type I..................................................................................42

Figure 16. Artemis view of PCR-free libraries of *M. kansasii* types I-V mapped against type I .................................................................43

Figure 17. Confirmation of *espA* deletion in types II-V. ............................43
**Figure 18.** (A) ACT view of types I-V showing the insertion present in type III ESX-1 (B) schematic representation of the 20.5 kb region and the genes associated in this inserted region

**Figure 19.** ACT view of the (a) PlcABC operon, (b) reads of type III and IV mapped to type I and (c) plcD in types I-V.

**Figure 20.** Correlation of the three M. kansasii types I-IV biological replicates

**Figure 21.** Heat map of ESX-1 and ESX-5 expression patterns.

**Figure 22.** Representation of qPCR results of selected down regulated (A) and up regulated (B) genes
List of tables

Table 1. Identifying *M. kansasii* subspecies using different genotyping approaches.........................................................20

Table 2. *espA* forward and reverse primers..................................................33

Table 3. General genome characteristics of *Mtbc* H37Rv and *M. kansasii* types I-V...................................................................................................................................35

Table 4. Summary of number of mce copies in *M. kansasii* types I-V..............48

Table 5. Primers used in qPCR validation.........................................................53

Table 6. RNA sequencing profiles for cultures of *M. kansasii*. ....................54

Table 7. Selected down and up regulated genes (in log2fold change) for qPCR validation......................................................................................................................59
CHAPTER 1. INTRODUCTION

1.1 The mycobacterium genus

The mycobacteria group are considered to be one of the most intensively studied bacterial taxa, as they cause the two historical and worldwide known diseases: leprosy and tuberculosis. Initially, interest in this taxa started with the work of the German physician Robert Koch, where he detected a tubercle bacillus in stained infected tissue and cultivated the organism on inspissated serum medium (Goodfellow and Magee, 1998). It was agreed that mycobacteria members are Gram-positive aerobic, asporogenous rods, which were usually acid-alcohol-fast at some stage in the growth cycle. They belong to the family Mycobacteriaceae within the order Actinomycetales. Organisms within this genus share a high genomic DNA G+C content (varying from 62–70%) and produce mycolic acids with closely related genera, Nocardia and Corynebacterium within Actinomycetales, and therefore possess a complex cell envelope that is rich in lipids and glycolipids. This complex cell envelope consists of a cytoplasmic membrane and a cell wall, and together, they make an efficient permeability barrier, which provides a crucial role in intrinsic drug resistance and in survival under harsh conditions. As mentioned before, mycobacteria produce mycolic acids, which is one of the diverse lipids that it makes, as well as very long fatty acids, accounting for 30% to 40% of the cell envelope mass (Niederweis et al, 2010). In the envelope, mycolic acids are connected, covalently, to peptidoglycan via an arabinogalactan polymer (a type of polysaccharide) and the mycolic acid-arabinogalactan-peptidoglycan polymer is finally arranged to form a hydrophobic layer along with other lipids. It has been observed that pore proteins are found in the mycobacterial cell wall. From structural analysis of these pores, the cell wall lipids are proven to be organized in an outer membrane. To date, there is ongoing research on the organization of lipids in the mycobacterial cell wall, which would aid in understanding the physiology and virulence of some of important mycobacteria species, such as M. tuberculosis (Niederweis et al, 2010). An illustration of the most recent views on mycobacteria cell wall can be seen in figure 1.
Figure 1. Overall structure of the mycobacterial cell wall complex of *Mycobacterium tuberculosis*. This complex contains the three different covalently linked structures: peptidoglycan (grey), arabinogalactan (blue) and mycolic acids (green). The mycolic acid layer, or Mycomambrane, is a hydrophobic layer, which was a result of the covalent linkage of mycolic acids, and it has many free lipids, some are considered to be specific for Mycobacteria. The capsule, which is the top layer, is a polysaccharide layer that mainly has two types of sugars: glucan and arabinomannan (Abdallah et al, 2007).

Phylogenetic analysis of genus mycobacteria is mainly based on homology of the 16S ribosomal gene sequence, as it is the case in most bacterial groups. The 16S rRNA homology shows how related the organisms are to each other, as seen in figure 2 (Saviola and Bishai, 2006; Tortoli, 2011). From these phylogenetic studies, the genus is broadly assigned based on their growth rates, i.e. slow and fast growing bacteria. As their name suggests, slow growing bacteria take weeks to form colonies in optimal conditions, whereas rapid or fast growing mycobacteria include species that produce colonies in solid media within 7 days, under optimal nutrient and temperature conditions (Goodfellow and Magee, 1998; Saviola and Bishai, 2006).
Figure 2. Phylogenetic tree based on 16S rRNA gene sequences. This shows the relatedness of the species on neighboring branches (Pittius et al., 2006). Tree was constructed using Paup 4.0b10 (heuristic search, gaps = fifth state), by implementing 1286 aligned nucleotides of the 16S rRNA DNA sequence of 80 Mycobacterium species with Gordonia aichiensis sequences as the outgroup. Rapid and slow growers are separated by the dotted lines.
Amongst the slow growers group of mycobacteria, the tubercle and leprosy bacilli are found, both of which account for principal human pathogens. Within this group of mycobacteria, a number of species are grouped into complexes, such as the *M. avium* and *M. tuberculosis* complexes, and these would include bacterial species with a high degree of genetic similarity, hence, causing similar disease. The reservoir, or habitats, of *Mycobacterium* species vary significantly. *Mycobacterium tuberculosis*, for example, has no significant environmental reservoirs, whereas *Mycobacterium bovis* (the tuberculosis agent for humans and cattle) has a natural reservoir in ruminants. Water can harbor many medically important mycobacteria, and that includes: *M. marinum*, *M. cheloneae*, *M. fortuitum*, *M. kansasii* and *M. avium*. In addition, soil may also serve as a reservoir for mycobacteria, such as: *M. cheloneae*, *M. fortuitum*, and *M. avium* (Saviola and Bishai, 2006).

### 1.2 *Mycobacterium tuberculosis*

One of the well-studied pathogens of the *Mycobacterium* genus is *Mycobacterium tuberculosis*. It is an obligate pathogen of human that causes the well-known disease tuberculosis. The route of transmission for this particular bacterium is via person-to-person contact, and mainly through aerosolization of infectious particles (Goodfellow and Magee, 1998). When the aerosolized droplets are inhaled, the tubercle bacilli travel to the terminal bronchioles and alveoli, and are subsequently phagocyted by alveolar macrophages, where most of the bacilli are killed in the phagosomes (via acidification), however, some of the invading *M. tuberculosis*, can survive these initial host defenses, because *M. tuberculosis* is an intracellular pathogen and can efficiently inhibit phagosome-lysosome fusion, hence, making it survive and grow within the macrophages and are released when the macrophages die. When the new and un-activated macrophages ingest the bacterium that has been liberated from an old macrophage, it grows within the un-activated macrophages for approximately 3 weeks, and finally, the bacilli lyse the macrophages and spill out into the host tissue (Gagneux, 2013; Gordon, 2009). As *M. tuberculosis* is the most intensively studied species of *Mycobacterium*, bellow is a figure explaining the mode of dispersal and pathogenesis of this organism (figure 3).
Early sequencing efforts were made on the strains that cause TB, and it was said that the variation was negligible, as it was limited compared to other bacteria, and hence, not clinically relevant. In the 1960s, differences in virulence was detected in clinical strains that have been infected into guinea pigs, however, no molecular tools were available to detect these differences. With the rise of genotyping methods in the 1990s, scientists have shown the *M. tuberculosis* differed in propensity to spread between individuals, and their abilities in causing prolonged outbreaks. Tuberculosis (TB) is known to be biologically complex, and strains that are know to cause TB are variable, and the variability is not a result of a few genomic differences, but there would be a combined effect of various genomic features of a given strain, which would eventually causes TB. These complex features and interactions can only be detected by systems biology approaches, which would offer a way forward (Saviola and Bishai, 2006; Gagneux, 2013).
1.2 M. tuberculosis

*M. tuberculosis* is a part of the *M. tuberculosis* complex (MTBC), which is comprised of *M. bovis*, *M. bovis*, *M. microti*, *M. canetti*, *M. caprae*, *M. pinnipedii* and *M. africanum*, however, only *M. tuberculosis* and *M. bovis* are major threats to humans. *M. africanum* causes up to 50% of human TB, although, it is limited to West Africa for unknown reasons. Members of the MTBC are known to be a clonal population structure, exhibiting 0.01–0.03% synonymous nucleotide variations in their genomes, and do not show any evidence of horizontal gene transfer (Gordon et al, 2009). MTBC is hypothesized to be derived from a pool of ancestral tubercle bacilli, named *Mycobacterium prototuberculosis*. As they are very similar, the way to distinguish them is done via a limited number of phenotypic or genotypic characteristics, however, they differ in two main aspects: host range and pathogenicity. For example, *M. microti* is almost exclusively a rodent pathogen, whereas *M. bovis*, as mentioned before, infects a wide variety of mammalian species, including humans (Wirth et al, 2008).

The closest relatives to this complex are two environmental bacteria, known as *Mycobacterium marinum* and *Mycobacterium ulcerans*. *M. marinum* shares a >85% nucleotide identity with *M. tuberculosis*, and *Mycobacterium ulcerans* and *M. marinum* have a nucleotide identity of 99.6%, however, they are phenotypically distinct and cause different diseases, as *M. marinum* is an ectothermic pathogen rarely causes infection, and *M. ulcerans* causes a cutaneous disease that was a part of an epidemic in some areas of West Africa (Gordon et al, 2009) (Stinear et al, 2008)(Wirth et al, 2008).

**1.3 M. kansasii**

*Mycobacterium kansasii* is a slow-growing acid-fast bacillus that belongs to the group of environmental mycobacteria, or nontuberculosis mycobacteria (NTM). Hauduroy, the French scientist, first recognized this bacterium as a new species in 1953. The overall morphology of *M. kansasii* shows that it is a large bacillus with rough colonies, which in most cases develops an intense yellow pigmentation following the light exposure. Their growth rate is considered to be slow and requires 2-3 weeks at temperatures ranging from 30 to 40°C. After *M. avium*, *Mycobacterium kansasii* is
most frequently responsible for disease due to nontuberculous mycobacteria, although, very little is known about its pathogenicity, route of transmission, and natural reservoir. It is thought that major reservoir of this bacterium is local water supplies and is often recovered from municipal tap water, and occasionally from river or lake water, and very rarely from other environmental sources (e.g. soil and dust) (Alcaide et al, 1997). In the state of Texas at the United States, for example, a \textit{M. kansasii} infection is considered as an urban disease, whereby the organism has been recovered in areas with high prevalence of \textit{M. kansasii} from piped water systems. In terms of \textit{M. kansasii} infections, it is speculated that human infections can be acquired from the environment; however, a definite epidemiological link between those natural reservoirs and human disease is yet to be established (Zhang et al, 2004).

Previously, the standard laboratory tests used to identify \textit{M. kansasii} were mainly biochemical and growth tests, which includes photochromogenicity, catalase production, urease activity, tween hydrolysis, and nitrate reduction. In the early 1990s, the genetic identification of \textit{M. kansasii} was possible by the release of a commercial DNA test and the isolation of a specific 2.2-kbDNA probe (pMK1-9) (Huang et al, 1991), and following that probe, more DNA probe technologies helped develop more probes for the identification of \textit{M. kansasii}. As much as we know now about this bacterium, clear answers on fundamental questions, such as: reservoir, transmission routes, and pathogenicity of this bacterium is complicated by the evidence of heterogeneity within the \textit{M. kansasii} species (Tortoli, 2003).

1.4 Types of \textit{M. kansasii}

The notion that \textit{M. kansasii} is a heterogeneous species with several distinct subtypes is further supported by phylogenetic and molecular analysis studies, namely molecular typing methods. In 1992, Yang and colleagues published a paper that showed one of the earliest studies of the \textit{M. kansasii} types, where they investigated isolates from Australia, Belgium, Japan, South Africa, and Switzerland. The results
have shown that 20 of 105 isolates were pMK1-9 probe negative, and interestingly, there were some differences in the 16S rRNA sequence and restriction fragment length polymorphism (RFLP) profile in 19 of the 20 pMK1-9-negative isolates from that of *M. kansasii*. These findings suggest that *M. kansasii* has subtypes, or the possibility of the existence of a new species. Another study by Picardeau et al. (1997) analyzed 38 clinical *M. kansasii* strains and 24 *M. kansasii* strains from water samples. They used molecular-based techniques, such as RFLP analysis with the major polymorphic tandem repeat probe and IS1652, pulsed-field gel electrophoresis (PFGE), and PCR restriction analysis (PRA) of the *hsp*-65 gene, which showed the clinical and environmental isolates of *M. kansasii* are five different *M. kansasii* subspecies. The results obtained in the study by Picardeau and colleagues was further validated by Alcaide et al, (1997). Table 1 shows the different genotyping approaches that are performed in order to distinguish within *M. kansasii* species (Tortoli, 2003).

**Table 1.** Identifying *M. kansasii* subspecies using different genotyping approaches (Tortoli, 2003)

<table>
<thead>
<tr>
<th>Type</th>
<th>DNA probe</th>
<th>INNO LiPA</th>
<th>Sequencing</th>
<th>Molecular typing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AccuProbe</td>
<td></td>
<td></td>
<td>RFLP</td>
</tr>
<tr>
<td></td>
<td>1st</td>
<td>2nd</td>
<td></td>
<td>ITS</td>
</tr>
<tr>
<td>i</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ii</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>iii</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>iv</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>v</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
</tbody>
</table>

^a amplified fragment length polymorphism; S, single pattern; M, multiple patterns

In more detail, *M. kansasii* subtype 1 is the most frequent subtype to be isolated from humans, and is rarely isolated from the environment. *M. kansasii* subtype 2 has shown to cause infection only in immunocompromised patients, where they show susceptibility to this subtype, and hence, acts as an opportunistic pathogen. For the other defined subtypes, type III, IV and V, very little is known about their pathogenic roles; however, they are generally isolated from environmental sources (Han et al, 2010).
1.5 *M. kansasii* pathogenicity

Ever since the identification of *M. kansasii*, its pathogenic role has been overlooked due to the large scales tuberculosis epidemics. With better understanding of mycobacteria, knowledge about NTM, such as *M. kansasii*, has been growing. It is estimated that the annual infection rate of *M. kansasii* ranges from 0.5 to 1 per 100,000 people, supporting its importance and its significance as a human pathogen (Tortoli, 2003). In terms of epidemiology, or prevalence, this pathogen shows geographic variability, ranging from a very low frequency in Australia and Japan to a very high one in several states of USA and in central Europe (Tortoli, 2003). Similar to *M. tuberculosis*, *M. kansasii* causes pulmonary disease, or fibrocavitary lung disease in immuno-competent patients. In terms of immunodeficient patients, *M. kansasii* can cause pulmonary and extra-pulmonary disease in patients with various immunodeficiencies, particularly in human immunodeficiency virus (HIV) infected individuals. In more detail, common symptoms of a *M. kansasii* infection are: chronic bronchopulmonary disease (in adults with chronic obstructive pulmonary disease or cystic fibrosis), skeletal infections, skin and soft tissue infection, cervical or other lymphadenitis, and disseminated infection (which occurs exclusively in immunocompromised patients). Disseminated *M. kansasii* infection has been reported to occur in the lungs, liver, spleen, bone marrow, lymph node, bowels, central nervous system, pericardium, and pleura or kidneys. Of note, the disseminated infection associated with skin has not been reported (Han et al, 2010).

As most *mycobacterium* infections, the first line of defense in the host is the phagocyte of the innate immune system, also known as macrophages. Virulent mycobacteria, such as *M. tuberculosis*, reprogram the macrophage antimicrobicidal activity, which helps the bacterium to survive and cause the infections within the host. Several studies suggest that cell death (either by apoptosis or necrosis) occurs when the mycobacterium interacts with the macrophages. In a study by Schaible and colleagues (2003), they demonstrated that the apoptotic bodies formed during the end stages of apoptosis of infected macrophages with *mycobacterium* kills the intracellular mycobacteria and aids in transporting the mycobacterial antigen to dendritic cells to cause a more broader effect. However, if necrosis occurs, it directly causes extensive growth of the *mycobacterium*, therefore, these pathogens tend to suppress apoptosis.
and induce more necrosis in macrophages. One of the key effects of macrophage and mycobacterial infections is the damage of the mitochondria, as it irreversibly opens the mitochondrial permeability transition (MPT) pore, resulting in significant mitochondrial transmembrane potential loss, which then leads to necrosis. *M. kansasii* is said to cause apoptosis in alveolar macrophages, however, there is not a lot known about apoptosis or necrosis during *M. kansasii*-macrophage interaction (Schaible et al, 2003).

1.6 *M. kansasii* compared to other mycobacteria species

A recent study by Matveychuk and colleagues in 2012 has analyzed features of *M. kansasii* infections compared to other NTM infections in Israeli citizens. Results have shown that *M. kansasii* patients tended to be of young age, native-born Israelis, and had symptoms of chest pain, cough, and hemoptysis. In addition, *M. kansasii* infections in patients have presented more cavitations, unilateral disease, and a higher likelihood of right upper lobe disease in the chest radiographs. Other NTM infections have shown to persist in old patients that were mostly immigrants, and interestingly, were said to have a history of tuberculosis, and presented with weight loss, fever, and sweating. Due to the differences in natives and non-natives presenting different degrees of NTM infections, as mentioned before, this could show that the geographic heterogeneity of the NTM species, ethnic differences, interaction with relative frequencies of TB in the *M. kansasii* versus non-*M. kansasii* groups, or some combination of these factors can have a big role in the infection process.

As mentioned earlier, non-*M. kansasii* infections tend to have a history of TB, which means that they had higher rate of pre-existing lung disease. In the underlined study, this was explained, partially, by demographic changes. Other notes that were taken from the study was that NTM species groups are less pathogenic than *M. kansasii*, hence, the other NTM agents need a pre-existing lung infection, such as TB, in order to breach in the normal local host defenses, thereby causing its effects. *M. kansasii*, on the other hand, can establish an infection in relatively normal lung tissue (Matveychuk et al, 2012).
1.7 The PE and PPE families

From genomic studies on Mycobacteria, the group of PE/PPE genes were noticed to be exclusively present in the mycobacteria genus (Kohli et al, 2012), and expanded in the slow-growing mycobacteria group. The PE/PPE genes correspond to genes producing proteins that carry an N-terminal of ProGlu and ProProGlu motifs. The protein family with proline and glutamine in their N-termini, or PE family, has three main subgroups. The first group has the PE domain on its own, and it comprises 29 members in H37Rv. The second group has contains the PE domain followed by a unique sequence, with 8 existing members. Lastly, the third group has a PE domain followed by polymorphic GC-rich repetitive sequences (PGRS) domain, which have GlyGlyAla or GlyGlyAsn tandem repeats. The PPE family comprises of 69 members with a conserved region of approximately 180 amino acids in their N-termini sequence, but the C-terminal sequence is variable (Stinear et al, 2008; Sampson, 2010; Voskuil et al, 2004). Figure 4 shows a simplified diagram of the different structures known of PE/PPE proteins identified in H37Rv.

![Diagram of PE and PPE family subgroups](image)

**Figure 4.** PE and PPE family subgroups (Sampson, 2010).

As there is variability within the PE/PPE families, it may promote the antigenic diversity in many of the pathogenic *Mycobacteria* family members, such as the well-known *M. tuberculosis*. In addition to its possible role in antigenic variability, functions in immune evasion, cell interactions and virulence have been suggested, and more importantly, their functional role in the ESX secretion systems, which will be discussed further in the following chapter. In terms of structural studies of these proteins, it was shown to be very difficult to study. However, it was solved due to the
observation that PE and PPE proteins often pair together and form complexes, hence, co-expression of known PE and PPE proteins together helped in solving the crystal structure. Strong et al, (2006) showed that the crystal structure observed was an overall extended α-helical structure. The PE protein formed an antiparallel pair of helices with two of the five helices of the PPE protein, as seen in figure 5.

**Figure 5.** The PE/PPE protein complex formed by known gene products of H37Rv strain. (PPE, in green) and (PE, in orange) (Strong et al, 2006).

The power of comparative genomics has uncovered the close association of PE/PPE genes with esx regions, which is a crucial region in *Mtb* that encodes type VII or ESX secretion systems 5 (Abdallah et al, 2009). Comparative genomics has also showed that PE_PGRS and PPE_MPTR subgroups are the most phylogenetically recent, as they were most repetitive and variable family members (McEvoy et al, 2012). PE_PGRS and PPE_MPTR have also been restricted to pathogenic mycobacterial species, and the expansion of these genes have been linked with the expansion and duplication of esx-5 (Abdallah et al, 2009). All of the aforementioned lines of evidence suggest an association of PE/PPE families and the ESX secretion systems (Van Pittius et al, 2006). However, the extent of this association, or functional link, between the PE and PPE genes and ESX clusters is still unclear (Abdallah et al, 2007). Further details of the ESX secretion systems will be in the subsequent chapter.
1.8 Type VII (ESX) secretion system

As mentioned earlier, the key to the pathogenic mycobacteria survival within the host is its unique cell wall. The transport of proteins across the complex cell wall is done through several pathways, such as the SecA1-mediated general secretory pathway, and alternatively, the SecA2-operated pathway, the twin-arginine translocation (TAT) system, and more importantly, the specialized secretion pathways type I–VI. For these secretion systems, it can either be a one-step process, whereby the cell envelope is crossed in one step, or by two steps, which needs a specific machinery to cross the outer membrane of the bacteria (figure 6). In more detail, the one-step process involves three elements: the inner membrane (IM) ATPase-binding cassette (ABC) transporter, a membrane-fusion protein and an outer membrane (OM) pore, which ultimately helps in the crossing of the cell envelope in one step. The two-step system, on the other hand, is more complex, as it contain an amino-terminal signal sequence to mediate translocation across the IM by the general Sec- or Tat-translocons. In this system, a structure termed the secreton is needed to ensure the proteins fold in the periplasm before the translocation across the OM. This structure has a conserved OM pore (the secretin) and a pilus-like structure in the IM, which is hypothesized to act as a piston to push substrates through the secretin.

![Figure 6. General view of types I-VI secretion pathways (Abdallah et al, 2007). Type I is considered to be a one-step secretion system. Type two uses the two-step secretion system. Types III and IV can secrete proteins by one-step or a two-step mechanisms. The type V system uses a simple two-step mechanism, and type IV uses a mechanism that is still unknown.](image)

In this thesis we will be focusing on type VII secretion system. The Type VII, or
ESX, secretion system has been intensively studied in recent years. It was identified when researchers have noticed that the ESX-1 gene cluster secreted the 6 kDa ESAT-6 protein (Abdallah et al, 2007). Overall, ESX genes encode many proteins, but mostly, immunodominant T cell antigens, and these genes are arranged in tandem pairs at 10 genomic loci (ESX-1–ESX-5) (figure 7). It can be noticed that the esx genes are flanked by genes coding for components of secretion machineries involved in the export of the corresponding ESX proteins, and as mentioned before, PE/PPE genes are also present with the esx clusters. Other genes are noted to be localized in these core ESX regions, which could be required for the type VII secretion system, such as rv3616c-rv3614c, ESAT-6 and CFP-10 in ESX-1, however, rv3616c-rv3614c are needed for ESX functionality. Approximately seven ESX-1 secreted proteins have been identified (Bordin et al, 2006; Serfani et al, 2009), and this particular ESX operon is considered to be the best characterized ESX operons. Generally, the proteins are either secreted in the extracellular medium or attached to the cell-surface, which indicates that a specific channel must be present in order to pump, or transport, these proteins across the outer membrane. ESAT-6 forms a complex with a 10kD protein termed culture filtrate protein (CFP-10), which involves hydrophobic interaction, and they both have a leading role in M. tuberculosis pathogenicity. Several studies have stated that ESAT-6 and CFP-10 secretion is essential for Mtb pathogenicity (Abdallah et al, 2007). Interestingly, the attenuation of BCG and M. microti vaccines was partly due to the absence of ESAT-6, which is encoded in the region of deletion 1 (RD1) (Abdallah et al, 2007; Simeone et al, 2009). These two secreted proteins are a part of the WXG100 family of small-secreted proteins that have a helical structure, and form a hairpin by the conserved Trp-Xaa-Gly (W-X-G) motif. ESX-1, which encompasses these two proteins, is present in several mycobacteria species, including M. kansasii, although the literature focuses mainly on its presence in M. marinum and M. smegmatis, due to the homology of the ESX present in these organisms to ESX in M. tuberculosis. For ESX-2 and 4, very limited information is known about them, however, studies on the ESX-3 have showed that ESX-3 might be involved in iron/zinc homeostasis (Bitter et al, 2009; Serfani et al, 2009), and ESX-5 has also shown its involvement in PE/PPE secretion (Bitter et al, 2009). There is growing interest in the ESX field of research in mycobacteria, as it is considered to be relatively new.
Figure 7. Predicted genetic organization of the 5 ESX loci in *M. tuberculosis* H37Rv strain (A) and the ESX gene products’ proposed cellular localization and their interactions (B). Abbreviations: ecc stands for esx conserved component and esp stands for ESX-1 secretion-associated proteins. To note: the channel drawn in (B) refers to a hypothetical pore that has not been experimentally demonstrated (Bitter et al, 2009).
1.9 Mammalian cell entry (Mce) operons

This family of proteins is for the virulence of *Mtb*, as it is closely involved in the invasion and prolonged existence in host macrophages. To date, there are four mce operons, namely mce1 to mce4, which have similar arrangement and a 450 bp core sequence and 8–13 genes in each operon. These operons were essentially identified from the complete genome sequence of *Mtb* H37Rv. In *Mtb*, mce loci comprises of two *yrbE* (*yrbEA* and *yrbEB*) and six *mce* genes (*mceA*, *mceB*, *mceC*, *mceD*, *mceE*, and *mceF*) (figure 8). In terms of predicted functions of these operons, they seem to be involved in lipid metabolism or redox reactions; however, their functions still remain largely unknown (Zhang and Xie, 2011). The first evidence that showed the involvement of *mce* genes in virulence of *Mtb* is when the recombinant *E. coli* harboring *mce1A* gene enabled this non-pathogenic bacterium to invade and survive in the macrophages (Casali and Riley, 2007).

![Figure 8. Organization of the four mce operons in H37Rv (Zhang and Xie, 2011).](image)

The black arrows represent proximal transcription regulators, while the stripe arrow represents *fadD5*. Dots represent non-coding sequences. White arrows represent *yrbE* genes, while grey arrows *mce* genes, and the griding represents Rv0590A.

To note, the function of *yrbE* contained in all four *mce* locus is still unknown, however, it may play a role in host cell invasion, as it is thought to be an integral membrane protein. Other than *yrbE*, *fadD5* is present in mce1, and is predicted to catalyze the first step in the fatty acid degradation. To fully understand these operons, more research needs to be made on the four *mce* locus (Zhang and Xie, 2011).
1.10 Comparative genomics and Mycobacteria

Genome sequencing techniques have had a key role in providing detailed information of many bacterial agents. The *M. tuberculosis* sequence of the H37Rv strain was completely sequenced and publicly available for almost a decade, however, its biology still remains a mystery. Several other *M. tuberculosis* strains and other closely related Mycobacteria have also been sequenced and published, which gives an opportunity of comparative genomics between these different strains and other mycobacteria species, in order to understand the evolution, pathogenesis and host-pathogen interactions of the members of this family. Kato-Maeda and colleagues in 2001 have published a paper that studied the mycobacterial evolution and pathogenesis by the use of t array-based comparative genomics for 19 isolates of *M. tuberculosis* that investigated the small-scale genomic deletions. Results have shown that deletions occurred within mycobacterial clones were identical, but varied between different clones. These deletions were suggested to have unnecessary ancestral genes that have lost the need to be present in the organism. They have also noted that with the increase of genomic deletion in a given isolate, the less ability it would have to cause the hallmark of TB, i.e., pulmonary cavitation.

Later in 2002, Fleischmann et al have published a paper showing the complete sequence of the *M. tuberculosis* clinical strain CDC1551 compared to the previously sequenced H37Rv strain. They focused on the polymorphic sequences between the compared genomes that were large-sequence and single-nucleotide polymorphisms in numerous genes, it mainly included phospholipase C, a membrane lipoprotein, members of an adenylate cyclase gene family, and members of the PE/PPE gene family. Generally, the paper has concluded the extensiveness of polymorphism events amongst *M. tuberculosis* strains that has the potential to be relevant to disease pathogenesis, immunity, and evolution (Fleischmann et al, 2002).

McGuire et al, (2012) published a more recent study that utilizes comparative genomics for the study of mycobacterial bacteria, by using 31 genomes, which were 8 strains of Mtb and *M. bovis*, 11 additional Mycobacteria (ranging from obligate parasites to free-living soil bacteria), 4 *Corynebacteria*, 2 *Streptomyces*, *Rhodococcus jostii* RHA1, *Nocardia farcinia*, *Acidothermus cellulolyticus*, *Rhodobacter sphaeroides*, *Propionibacterium acnes*, and *Bifidobacterium longum*. Reasons behind
choosing other non-mycobacteria members were to provide further insight into the 
*Mycobacterium* cluster and evolutionary trends. This study has showed the 
importance of lipid metabolism and its regulation, by using a compendium of 
microarray gene expression experiments, and a list of genes were found to be 
upregulated in the presence of different fatty acid sources. Interestingly, under 
unsaturated fatty acids conditions, the upregulated genes had uniform phylogenetic 
profiles, whilst the upregulated genes in the saturated fatty acid conditions showed 
expansion through duplications in pathogenic Mycobacteria. Results have also 
highlighted that DNA repair and molybdopterin cofactors (which are required in 
enzymes that could have physiological functions in the metabolism of reactive oxygen 
species during stress response) are important in pathogenic Mycobacteria, by using 
dN/dS analysis. Additionally, the study has used sequence conservation (or 
phylogenetic footprinting) and gene expression data that have identified 
approximately 400 conserved noncoding regions. Overall, this study has confirmed 
some gene families that are associated with the adaptation of environmental 
Mycobacteria to obligate pathogenesis, which were mainly the gene families 
mentioned above.

1.11 Aim of thesis work

1.11.1 Genome Analysis of *M. kansasii* types I-V

The aim of this part of the thesis is to do a comparative genomics study of *M. kansasii* 
sub-types I-V, in order to identify and highlight possible similarities and differences 
between the different subtypes, as very little is known about this heterogeneous 
species.

1.11.2 Comparative transcriptome analysis

To complement the genome studies, RNA sequencing technology will be used to 
study the gene expression level of *M. kansasii* sub-types, as RNA-seq allows the rapid 
profiling and deep investigation of the transcriptome in a very high-throughput and 
quantitative manner.
CHAPTER 2. GENOME ANALYSIS OF M. KANSASII TYPES I-V

2.1 Material and methods

2.1.1 DNA Isolation

The *M. kansasii* bacterial strains were obtained from Jakko van Ingen (Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands). DNA isolation of *kansasii* types I-V was done using bead beater phenol/chloroform protocol. Briefly, bacteria were resuspended in a 1:1 mixture of 10 mM Tris-1mM EDTA-100 mM NaCl (pH8.0) and phenol-chloroform-isoamylalcohol (50:48:2), and disrupted by bead-beating with 0.1 mm Zirconia/Silica beads (Biospec Products). After centrifugation, supernatants were extracted with chloroform and centrifuged. DNA was precipitated from the supernatant with isopropanol in the presence of 300 mM NaAc. Precipitated DNA was washed with 70% ethanol, air-dried and dissolved in H2O.

2.1.2 Construction, sequencing, and assembly of *M. kansasii* type I-V

Genomic DNA from *M. kansasii* types (I-V) was sheared into approximately 500 bps fragments using Covaris™. Paired-end, Nextera Matepair and PCR-Free libraries were performed following the instructions of their respective manufacturers and sequenced using HiSeq 2000™. De-novo assembly using IDBA (Chen et al, 2009) was performed, followed by scaffolding with SSPACE (Boetzer et al, 2010), gap filling using IMAGE and Gap filler, and the genome was evaluated using REAPER. The schematic in figure 9 shows the pipeline used to generate the genome. The mapped reads of all the aforementioned libraries were mapped into the final assembly to manually curate the genome by looking into SNPs, insert size and scaffold pairing.

2.1.3 Genome analysis and annotation

Automated gene prediction for *M. kansasii* types I-V was completed by combining results from the Prokka (Seemann, 2014), RAST (Aziz et al, 2008) packages. Prediction of RNA genes was generated by using Prokka for non-
coding RNA, tRNAscan-SE (Lowe and Eddy, 1997) for tRNAs, and RNAmmer (Lagesen et al, 2007) for rRNA genes.

**Figure 9.** The bioinformatics pipeline used to assemble *M. kansasii* type I-V

### 2.1.4 Comparative genomics of the 5 types of *M. kansasii*

To visualize a reference-based BLAST comparisons of types I-V, type I being the reference, BLAST Ring Image Generator (BRIG) was used (Alikhan et al, 2011). Paralogous groups were determined using OrthoMCL (Li et al, 2003), which did an all against all blast of the 5 *M. kansasii* types and *Mtb* strain (H37Rv), with a threshold identity of 30% and the inflation parameter (I) of 1.2. Figure 10 shows the workflow of OrthoMCL. Multiple Alignment of Conserved Genomic Sequence With Rearrangements (MAUVE) (Darling et al, 2003) was used to align the genome of types I-V
Figure 10. General diagram explaining the OrthoMCL algorithm for clustering orthologous proteins (Li et al, 2003).

2.1.5 PCR Confirmation of the espA operon

The PCR reaction conditions included: 1µl PfuTurbo Cx hotstart DNA polymerase (2.5 U/µl), 1X Pfu reaction buffer with MgSO₄, DNA template, dNTPs, and primers in a final volume of 100µl. espA primer sequence is provided in table 2.

Table 2. espA forward and reverse primers.

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence (5'-3')</th>
</tr>
</thead>
<tbody>
<tr>
<td>espA forward primer</td>
<td>GTTCGTCTCGATTTCGCAGC</td>
</tr>
<tr>
<td>espA freverse primer</td>
<td>GAATCACGCGCCTTGATGAC</td>
</tr>
</tbody>
</table>
2.2 Results and discussion

2.2.1 General features of *M. kansasii* type I-V genome

The advances in high throughput sequencing, or next generation sequencing technologies (NGS), have revolutionised our understanding of the genome of many different bacterial species. The genome of *M. kansasii* subtypes I-V was sequenced using the illumina platform (HiSeq™ 2000), and have subsequently undergone de novo assembly, followed by automated annotations by Prokka and RAST. Table 3 represents the general genomic features of *M. kansasii* subtypes compared to *Mtb H37Rv*. *M. kansasii* type I comprises a 6,495,905-bp genome with G+C content of 65.45%. The chromosome is predicted to have 6165 protein-coding DNA sequences; three rRNA and 46 tRNAs were identified. As for types II-V, they comprise of 6,394,377-bp, 6,079,840-bp, 6,451,556-bp, 7,111,868-bp, respectively. G+C content slightly varied amongst types II-V, which were 65.13, 64.85, 62.24 and 62.3 respectively, with 6280, 5959, 6223 and 6767 CDS respectively. Ribosomal RNA numbers were consistent in all types, as they all had 3 rRNAs, but tRNAs were predicted to be 47, 46, 47 and 46 for types II-V respectively. OrthoMCL compared the protein sequences of H37Rv to all *M. kansasii*; unique genes of each type were identified, as shown in table 3 bellow.
Table 3. General genome characteristics of *Mtb* H37Rv and *M. kansasii* types I-V

<table>
<thead>
<tr>
<th>Feature</th>
<th>H37Rv</th>
<th>Type I</th>
<th>Type II</th>
<th>Type III</th>
<th>Type IV</th>
<th>Type V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome size (base pairs)</td>
<td>4,411,532</td>
<td>6,495,905</td>
<td>6,394,377</td>
<td>6,079,840</td>
<td>6,451,556</td>
<td>6,500,905</td>
</tr>
<tr>
<td>G+C (%)</td>
<td>65.61</td>
<td>65.45</td>
<td>65.13</td>
<td>64.85</td>
<td>62.24</td>
<td>62.3</td>
</tr>
<tr>
<td>Gene density (base pairs per gene)</td>
<td>1110</td>
<td>1066</td>
<td>1032</td>
<td>1035</td>
<td>1050</td>
<td>1061</td>
</tr>
<tr>
<td>Average CDS length</td>
<td>1009</td>
<td>953</td>
<td>915</td>
<td>908</td>
<td>887</td>
<td>897</td>
</tr>
<tr>
<td>Protein-coding sequences (CDS)</td>
<td>3974</td>
<td>6165</td>
<td>6280</td>
<td>5959</td>
<td>6223</td>
<td>6767</td>
</tr>
<tr>
<td>Conserved CDS with assigned function</td>
<td>3049 (77%)</td>
<td>3535 (58.01%)</td>
<td>3538 (57.13%)</td>
<td>3962 (56.49%)</td>
<td>3469 (56.49%)</td>
<td>3702 (55.27%)</td>
</tr>
<tr>
<td>Unique genes**</td>
<td>22 (0.74%)</td>
<td>2 (0.032%)</td>
<td>8 (0.129%)</td>
<td>4 (0.068%)</td>
<td>6 (0.097%)</td>
<td>70 (1.045%)</td>
</tr>
<tr>
<td>rRNAs</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>tRNAs</td>
<td>45</td>
<td>46</td>
<td>47</td>
<td>46</td>
<td>47</td>
<td>46</td>
</tr>
</tbody>
</table>

The table shows general features about the genome of *M. kansasii* I-V, compared to the well-studied *Mtb* strain (H37Rv). **Unique genes were identified using OthoMCL comparison of *Mtb* and *M. kansasii* types I-V, hence, the numbers indicated here only show what is unique to that particular organism when *Mtb* and *M. kansasii* types are compared against each other.
2.2.2 Comparing genomes within types of *M. kansasii* types

Blastn similarities between *M. kansasii* I-V

BLAST Ring Image Generator (BRIG) was used to visualize and compare types II-V against the reference (type I), as shown in figure 11. As this is a reference-based approach, this part of the analysis was only able to show regions of the reference sequence that are present or absent in query sequences (types II-V) mapped onto type I.

![Figure 11. Blastn comparison of *M. kansasii* types II-V compared to type I.](image)

From the BRIG figure above, 70% identity was chosen to be the lower percentage limit, and 100% was the highest, both of which are colour-coded. Mostly, the types are very similar to type I, as the predominant colour observed is black (100% similarity) and dark grey (90% similarity), however, some areas were not present in types II-V, for example, the area of the genome at 4 Mbp, which could potentially show areas of interest, or unique genes.
OrthoMCL analysis of the 5 types

Extracting orthologous proteins can help identify novel and/or shared genes between organisms, which is the main aim for any comparative genomics studies. Orthologs are defined as homologous proteins related via speciation, owing to them having similar structure and perform similar biological function(s), however, paralogs is another term describing genes that were created via gene duplication and are prone to diversification, which means that they are prone to acquire distinct biological functions. Computational methods have been developed in order to identify orthology between several organisms, and these methods can be categorized in two categories: tree-based and pair-based. As the name suggests, tree-based techniques build a phylogenetic tree in a set of similar genes of different genomes, and these trees are then used to distinguish between orthologs and paralogs. Pair-based methods use an empirical approach that identifies pairs of similar genes in different genome sets, which are then organized to orthologous groups by performing a subsequent clustering step to filter out some co-orthologous pairs. OrthoMCL, which is the technique of choice to identify orthologous groups in this project, is an example of a pair-based approach (Li et al, 2003). Proteins sequences of the five types of *M. kansasii* were first used for all-against-all BLASTP comparisons. Orthology is then identified between pairs of genomes by reciprocal best similarity matches, and paralogs are identified as sequences within the same genome that are (reciprocally) more similar to each other than either is to any sequence from another genome. This is then converted into a graph with nodes showing protein sequences and the weighted edges represent their relationships. The MCL algorithm is then applied to represent a symmetric similarity matrix of the aforementioned graph.
The below Venn diagram shows the comparison of the *M. kansasii* types I-V OrthoMCL clusters.

![Venn diagram showing orthologous clusters among five *M. kansasii* types (MK1-MK5) determined by OrthoMCL.](image)

Figure 12. Venn diagram showing orthologous clusters among five *M. kansasii* types (MK1-MK5) determined by OrthoMCL.

From figure 12, orthologous clusters can be seen between *M. kansasii* types I-V, which shows the clusters both conserved and unique between the five types. The total number of clusters that is present in all types (2446) has been further exploited, and number of CDSs that is present within this cluster for types I-V is 5075, 5092, 5000, 5043 and 5313, respectively. To focus our study, further analysis in the RNA profile chapter was done on the 5075 genes of type I. Unique genes present within the unique cluster shown in figure 12 in each type have also been investigated and will be further discussed later.
Core *M. kansasii* genome

Multiple whole genome alignment of types I-V was done using MAUVE. In figure 13, the generated whole genome alignment identified blocks of sequence homology, and they are assigned to unique colour, this facilitated the comparison of the 5 types, as the sequences can be visualized according to coloured sequence blocks, and therefore, the blocks present in figure 13, show the most similar sequences in the genomes analyzed and, more importantly, deduce the most conserved sequences amongst the whole genomes of types I-V.

**Figure 13.** Aligned sequences of *M. kansasii* types I-V.
Unique genes present within each type

From comparing the 5 *M. kansasii* types using OrthoMCL (figure 12), unique orthologous groups were identified. In type I, there are 9 unique orthologous groups, which contained 11 CDS. Similarly, type II has 11 orthologous groups with 17 CDS amongst those groups. Type III, however, has 9 groups with 19 CDS within these groups. Type IV has 21 unique orthologous groups containing 28 CDS. Type V had the most unique orthologous groups, as it had 38 groups with 81 CDS within the groups. Most of the unique genes in types I-V were hypothetical proteins and all of them were manually checked for their uniqueness with BLAST, which gave very few or no hits at all, therefore, supporting the fact that they are unique to each type. These unique genes do not correspond to the unique genes identified in table 3, as the unique genes identified in the Venn diagram show what is unique when comparing the 5 types against each other, which the unique genes in table 3 show when the 5 types are also compared with H37Rv.

PE/PPE

As mentioned before, PE/PPE proteins are acronyms for proline-glutamate (PE), or proline-proline-glutamate (PPE) N-terminal motif proteins. As shown in figure 14, there are 92, 93, 88, 83 and 90 PE genes for types I-V, respectively. For PPE genes, there are 135, 140, 114, 98 and 126 in types I-V, respectively. The PE/PPE genes were automatically annotated by a combination of Prokka and RAST. On average, there are 212 predicted PE/PPE genes amongst all *M. kansasii* types, which is considered to be slightly higher than other mycobacterial species, such as *M. tuberculosis* (169 genes) and *M. ulcerans* (115 genes). Amongst the 5 types, type II followed by type I have the highest PE/PPE genes, which could be consistent with the hypothesis supported by (Gey van Pittius et al. 2006) that PE/PPE genes coevolved with the ESX loci, particularly ESX-5. This could be an interesting finding, as types I and II are pathogenic and the remaining types are non-pathogenic, and the high number of PE/PPE genes would mean that there is an
effect on the ESX-5 operon in these types when compared to others in terms on expression and its functionality.

**Figure 14.** Number of PE/PPE genes in *M. kansasii* types I-V

ESX system in *M. kansasii*

As mentioned in previous chapters, the ESX system consists of 5 major gene clusters (ESX1-5) that are dependent on ATP to export specific members of the 6-kDa early-secreted antigenic target (ESAT-6) protein family, thereby having a distinct role in mycobacterial pathogenicity. In *M. kansasii* types I-V, *esx* genes were predicted to be 33, 19, 18, 14 and 23, respectively. Types I and V had the most *esx* genes, and when compared to *M. marinum* and H37Rv, they show very similar numbers, as *M. marinum* has 29 and H37Rv has 23 genes. These *esx* genes are either within five ESX loci, as represented in figure 15, which shows these 5 loci in *M. kansasii* type I, and some are not within these loci. ESX-1 is hypothesized to have a crucial function in mediating virulence, as several data support its role in: macrophage production of IFN-β, activate the inflammasome, modulate macrophage cytokine
production and signaling, and escape from the phagolysosome (Garces et al., 2010) (McLaughlin et al., 2007). Interestingly, deletion of this locus attenuates the virulent Mtb, thereby, inhibiting their growth in macrophages and animals; however, its clear role in pathogenicity remains unknown (Garces et al., 2010).

As ESX-1 of Mtb is encoded by the extended RD1 region (extRD1), which secretes ESAT-6 and CFP-10 proteins, this region was shown to be deleted in the vaccine strains M. bovis BCG and M. microti (Brodin et al., 2006).

![Figure 15. Schematic representation of the ESX operons present in M. kansasii type I.](image)

Amongst the ESX_1 secreted proteins is the espA (MKANI_1032) protein, which is of unknown function. This espA gene is a part of an operon that also has 2 extra genes EspC (MKANI_1033) and EspD (MKANI_1034). Several studies, notably Fortune et al, have indicated that secretion of ESAT6 by ESX-1 is dependent on the espA operon, as the deletion of espA resulting in loss of ESAT-6 secretion and attenuation. One of the striking findings from sequencing M. kansasii types I-V was that the espA operon was deleted in types II-V, as seen in figure 16 bellow.
Figure 16. Artemis view of Illumina reads from PCR-free libraries of *M. kansasii* types I-V mapped against type I. From this view, it can be clearly noticed that the *espA* operon is deleted in types II-V. The colors of reads indicate type II in red, type III in green, type IV in blue and type V in magenta. The other cloned window (reads in blue) show type I alone mapped against type I.

This absence was also confirmed by PCR as shown in Figure 17. Furthermore, the mapped reads indicated in figure 16, the *espA* gene is only present in type I and absent in the remaining types.

Figure 17. Confirmation of *espA* deletion in types II-V. (M) indicates the ladder, (+) is the positive control for each type, (-) is the negative control and 1, 2, 3, 4 and 5 shows DNA from types I-V amplified with the *espA* primer. The expected product of the designed primer is 199 bp. In the lane marked 1,
there is a clear product present, which the other lanes (2, 3, 4 and 5) do not show the same product.

As mentioned earlier, depletion of espA retains ESX-1 function and attenuates the pathogenic *mycobacterium*, making espA one of the virulence determinants. As Pang and colleagues stated: “Secretion of espA and ESAT-6 is mutually dependent, with deletion of espA resulting in loss of ESAT-6 secretion and attenuation” (Pang et al, 2013). As for the remaining 2 genes in the operon, EspC is said to have a signal sequence required for secretion of espA and ESAT-6, and EspD has a role in stabilizing the levels of espA and C (Graces et al, 2010; Pang et al, 2013). Other implications of losing this operon is cell wall defects, as seen in the study by Garces et al (2010), but the attenuated *M. tuberculosis* strain in the study did not affect its interactions with host cells *in vitro*.

From ACT analysis of the 5 types, it was noticed that type III has an insertion of a group of genes that accounts for 20.25 kb (figure 18), which is highlighted in the black square in figure 18A. This insertion in ESX-1 is positioned between EccD (MKANIII_5413) and espK (MKANIII_5432) (figure 18B), which could be another reason as to why ESX-1 operon is inactive in this type, which was predicted earlier due to the loss of the espA operon.
Figure 18. (A) ACT view of types I-V showing the insertion present in type III within the ESX-1 operon, and is indicated by the black square. (B) Schematic representation of the 20.5 kb region and the genes associated in this inserted region.
Phospholipase C (Plc)

Phospholipases have been implicated in host-pathogen interactions in tuberculosis, as phospholipids in general have shown to mimic host lipids and halt the endocytic maturation process that is crucial for the pathogen-containing vacuole (Gomez, 2001). These phospholipases are divided into four groups (A1, A2, C and D) based on the position of the bond they hydrolyse on the phospholipid substrate. Notably, phospholipase C has been the most prominent phospholipase involved in mycobacterial pathogenesis. When the H37Rv strain was sequenced, it was found that plcA, plcB and plcC genes are clustered on the chromosome in one operon, and plcD was located in a different region. Early work by Raynaud et al. (2002), has shown that Mtb strains were attenuated when plc genes were knocked out in mouse models at the late phases of infection. Interestingly, some mycobacteria have naturally lost these phospholipases (PlcABC) such as the causative agent of bovine TB, M. bovis

Here, it was found that in type III and IV in M. kansasii, plcB gene, which is approximately 1.6kb in type I, was truncated, rendering the whole operon inactive. This was confirmed by mapping reads of types III and IV to type I, and viewing this alignment file, which showed that there were no reads of type IV to the region of plcB and type III reads only mapped to approximately 0.8kb of that gene (figure 19a). This was then visualized using ACT, which proved the alignment file findings (figure 19b). As for plcD, it is truncated in types IV and V (figure 19c). These observations may contribute to the degree of virulence in each type, as previously mentioned, types I and II being the pathogenic types and the remaining types being environmental. Research in the role of Plc enzymes in TB infection claims that Plcs may have a role in lysing the phagolysosomal membrane, however, their exact role is still not confirmed in mycobacterial pathogenesis (Raynaud et al, 2002).
Figure 19. (a) Artemis view of Illumina reads from PCR-free libraries of *M. kansasii* types III-IV mapped against type I, showing the absence of type IV reads of that particular gene and reads of type III mapping to a partial region of the *plcB* gene. (b) ACT view of types I-V, showing the truncated *plcB* gene in types III and IV by the indicated black square, (c) ACT view of types I-V, showing the presence of *plcD* gene in all the types by the indicated black square.
Mammalian cell entry (Mce) operons

In 1998, when the first sequence of H37Rv was published, four operons of mce genes (1-4) were noticed, arranged in operons containing approximately eight genes (Cole et al, 1998). From previous work, these operons were predicted to be involved in mycobacterium virulence (Klepp et al, 2012). Several studies have showed that knocking out mce1-3 operons in H37Rv altered their ability to persist within the host and mce4 operon is involved in the transport of cholesterol in M. tuberculosis (Klepp et al, 2012). From analyzing the sequences and annotations of M. kansasii types I-V, twenty five CDSs were accounted for mce genes in type I, which was the same number identified in H37Rv. Type II had 24 genes, and 18 mce genes in type III, which is considered less than type I and II, as type III lost the mce3 operon completely. Type IV had 25 mce genes, and type V had the largest number of mce genes, as it had 41 genes, which was difficult to explain from genomics data. On average, mce-1 operon is duplicated, mce-2 is absent and mce-3 and mce-4 is present in one copy in types I-V. Interestingly, type III, lacks the mce-3 operon. Table 4 bellow summarizes the number of copies of mce operons present in M. kansasii types I-V.

Table 4. Summary of number of mce copies in M. kansasii types I-V

<table>
<thead>
<tr>
<th></th>
<th>Type I</th>
<th>Type II</th>
<th>Type III</th>
<th>Type IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mce-1</td>
<td>2 copies</td>
<td>2 copies</td>
<td>2 copies</td>
<td>2 copies</td>
</tr>
<tr>
<td>Mce-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mce-3</td>
<td>1 copy</td>
<td>1 copy</td>
<td>-</td>
<td>1 copy</td>
</tr>
<tr>
<td>Mce-4</td>
<td>1 copy</td>
<td>1 copy</td>
<td>1 copy</td>
<td>1 copy</td>
</tr>
</tbody>
</table>
2.3 Future work

To provide further details and for this project to be taken further, future work would include the following:

Firstly, confirming the unique genes present within each *M. kansasii* type to develop a possible molecular typing method based on those unique genes, which would give us more details about how to identify which *M. kansasii* type is present in any given sample. By establishing this, distinguishing between each type can be done, which would ultimately help in diagnostic purposes. This was one of the initial aims of this study, however, due to lack of available time for this thesis, it was difficult to test and further investigate these unique genes and develop a molecular approach.

Secondly, biochemical testing, such as western blots, could help validate the finding of the absence of *espA* in *M. kansasii* types II-V, as previously described in Garces et al, 2010 and Fortune et al, 2005. Biochemical testing could also confirm the absence of *PlcB* in type III and IV, such as previously described in Raynaud et al, 2002. The absence of *mce-2* in all *M. kansasii* types would also be confirmed by biochemical means, and the loss of *mce-3* in type III.

Thirdly, in order to confirm the role of *espA* in *kansasii* pathogenicity and its effects of ESX-1, a knock-in approach would help validate this theory. Generally, animal studies would help tremendously to validate a specific point, such as the role of *espA* in virulence.

Lastly, another genetic approach that could add more valuable information about the underlined bacterium that is to identify the essential genes in *M. kansasii* type I, as it was used as a reference in most experiments done in this project. Such approach could be the Transposon-directed insertion site sequencing (TraDIS) technique. Ideally, TraDIS uses the presence/absence of transposons to assess the essentiality of genes within given genome. It investigates the sequence flanking each transposon insertion, which were generated by Illumina sequencing, thereby comparing the number of specific reads derived from inocula and output pools recovered from animals, or a certain cell type, which would be provided with a numerical measure of the extent to which mutants were selected *in vivo*, and that would reveal much information, such as the essentiality of the gene(s) present. This approach could also be of use in terms of checking for the essentiality of the functional groups investigated in this thesis.
(PE/PPE gene, ESX, Plc operon and Mce operons), and therefore, would enable us to draw more conclusions about their presence within the *M. kansasii* genome. To note, this method was started at the beginning of the project, as the libraries of *M. kansasii* type I was made, however, there were some difficulties generating the data in terms of library preparation and sequencing techniques used.
CHAPTER 3. mRNA PROFILE OF M. KANSASII

3.1 Material and methods

3.1.1 Bacterial cultures

A swab of DNA from each type was grown in Middlebrook 7H10 agar supplemented with 0.05% of Tween 80, Glycerol and enriched with 10% OADC. Liquid cultures were then done on all kansasii strains by taking colonies from the agar culture and added in Middlebrook 7H9 broth supplemented with 0.05% Tween 80, Glycerol and enriched with ADC. The liquid cultures were left for 3-4 weeks to grow to reach an OD$_{600}$ of 0.6-1.0 and each type was then divided into three tubes to form three biological replicates for further analysis.

3.1.2 RNA Extraction and removal of ribosomal RNA

Liquid cultures were centrifuged at 3500 RPM at RTP for 10 minutes, resuspended with 1 ml of Trizol (Invitrogen) and left for 5 minutes, and then transferred into ~500µl zirconia beads. The beads were then subjected to the bead beat 6x at maximum speed for 30 seconds and then placed on ice for 1 minute each time. Samples were then centrifuged for 3 minutes at 4°C, and the supernatant was then incubated at RTP for 5 minutes and 200µl of chloroform (Sigma) was added, shaken and left for RTP for 3 minutes, followed by centrifugation for 15 minutes at 4°C. The aqueous phase of all the samples was removed and an equal amount of isopropanol was added to each sample and were shaken and left at RTP for 10 minutes, followed by centrifugation for 20 minutes at 4°C, discarding the supernatant and 1 ml of 80% ethanol was added. This was then centrifuged for 10 minutes at 4°C and the supernatant was discarded and pellet was air-dried, followed by resuspension with 50µl nuclease free water and incubated at 55-60°C for 10 minutes. Concentration of each sample was then determined by Qubit, and stored in -80°C. Genomic DNA was removed by Turbo Dnase (2U/µl) (Ambion) based on the manufacturers guidelines. Ribosomal RNA was removed by using RiboMinus for bacteria kits (Invitrogen) as stated by the manufacturers guide, followed by
ethanol precipitation for all samples that were selected for strand specific library preparation.

3.1.3 Strand Specific RNA Library
Strand specific libraries were done on all RNA samples (three biological replicates for each types, with exception of type V that had two replicates) (20µl of RNA each) as stated by the TrueSeq manual (Part # 15031048 Rev. E).

3.1.4 RNA-Seq
Strand specific RNA libraries were sequenced using Hiseq 2000, then reads of types II-V were mapped against type I using Bowtie (Langmead et al, 2009). Cluster generation of RNA-seq data was done based in the core genes extracted from OrthoMCL analysis (5075 genes) using the model-based clustering (Si et al, 2014). Differentially expressed genes were identified using Cufdiff (Tranpell et al, 2012), assuming that any gene with a p value bellow 0.05 is considered differentially expressed.

3.1.5 Validation by qPCR
Up and down regulated genes were selected from the differentially expressed gene list generated by Cufdiff (Tranpell et al, 2012), and primers for each gene was designed specific for *M. kansasii* type I by primer blast. qPCR validation using SYBR GREEN Master Mix (KAPA Biosystems) was performed with sigA being the housekeeping gene. Table 5 shows the genes and primer sequences used in this experiment.
Table 5. Primers used in qPCR validation

<table>
<thead>
<tr>
<th>Gene</th>
<th>Primer name</th>
<th>Primer Sequence (5’-3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>esxN</td>
<td>esxN_Forward</td>
<td>GCTTGCCAGGAGTTCATCA</td>
</tr>
<tr>
<td></td>
<td>esxN_Reverse</td>
<td>GCCTGCTCGTAGATCACCT</td>
</tr>
<tr>
<td>VapB43</td>
<td>VapB43_Forward</td>
<td>GTTATCGAGTCCAGCCGATT</td>
</tr>
<tr>
<td></td>
<td>VapB43_Reverse</td>
<td>CGTTGACGACAGATCCA</td>
</tr>
<tr>
<td>WhiB4</td>
<td>WhiB4_Forward</td>
<td>CGGATCAACTGGGTATCAAA</td>
</tr>
<tr>
<td></td>
<td>WhiB4_Reverse</td>
<td>GACCTTGTTTGAAGTGCATT</td>
</tr>
<tr>
<td>WhiB6</td>
<td>WhiB6_Forward</td>
<td>GCCGACTCAGGAATTACCA</td>
</tr>
<tr>
<td></td>
<td>WhiB6_Reverse</td>
<td>GACAACGACACCCGACAC</td>
</tr>
<tr>
<td>DrrB</td>
<td>DrrB_Forward</td>
<td>GGAACCTCAAACTGCTGTTGC</td>
</tr>
<tr>
<td></td>
<td>Drrb_Reverse</td>
<td>ATGATCACCAGTGACGAAC</td>
</tr>
<tr>
<td>Pks</td>
<td>Pks_Forward</td>
<td>CACCGCTATAGCCAGAATCA</td>
</tr>
<tr>
<td></td>
<td>Pks_Reverse</td>
<td>ACGATATCCTCGTGCCTCTTT</td>
</tr>
<tr>
<td>sigA</td>
<td>sigA_Forward</td>
<td>GTCGATGACGAGGAGGAGAT</td>
</tr>
<tr>
<td></td>
<td>sigA_Reverse</td>
<td>GTCCCTCGTCCCATACGAAAT</td>
</tr>
</tbody>
</table>

3.2 Results and discussion

3.2.1 RNA sequencing profiles of *M. kansasii* types I-V

RNA was extracted from *M. kansasii* types I-V, each with 3 biological replicates, during the stationary phase. The RNA was then subjected to the removal of ribosomal RNA, and was then used to generate cDNA, which where then analyzed by the Illumina Hiseq machine, giving a total of 14 transcriptome profile, as type V had only 2 biological replicates. To study the gene expression of *M. kansasii* types I-V, the core genome of this bacterium, which were 5075 genes (identified by OrthoMCL analysis). On average, total number of reads generated by Hiseq 2000 was 30,099,579 for type I, 24,964,575 for type II, 16,097,772 for type III and 31,763,621 for type IV. Mapping the reads of type I-V to the reference (*M. kansasii* type I) was done using the alignment program Bowtie, results can be seen in table 6. All the types exist in three biological replicates, however, type V was very difficult to grow, therefore, we only managed to get two biological replicates, and one of the replicates (MK5_3) showed very poor mapping percentage to type I. Due
to this reason, type V was taken out of this part of the project and was not analyzed in terms of gene expression profiling.

**Table 6. RNA sequencing profiles for cultures of *M. kansasii*.**

<table>
<thead>
<tr>
<th>Sample</th>
<th>Total reads</th>
<th>Reads Mapped</th>
<th>% of total</th>
<th>Reads PP**</th>
<th>% of total</th>
</tr>
</thead>
<tbody>
<tr>
<td>MK1_1</td>
<td>28,657,576</td>
<td>24,400,557</td>
<td>85.15</td>
<td>17,745,304</td>
<td>61.92</td>
</tr>
<tr>
<td>MK1_2</td>
<td>30,972,474</td>
<td>27,337,010</td>
<td>88.26</td>
<td>20,695,832</td>
<td>66.82</td>
</tr>
<tr>
<td>MK1_3</td>
<td>30,668,688</td>
<td>28,026,950</td>
<td>91.39</td>
<td>21,105,202</td>
<td>68.82</td>
</tr>
<tr>
<td>MK2_1</td>
<td>20,332,226</td>
<td>17,422,734</td>
<td>85.69</td>
<td>12,485,784</td>
<td>61.41</td>
</tr>
<tr>
<td>MK2_2</td>
<td>26,022,534</td>
<td>21,731,676</td>
<td>83.51</td>
<td>14,337,322</td>
<td>55.10</td>
</tr>
<tr>
<td>MK2_3</td>
<td>28,538,964</td>
<td>24,820,330</td>
<td>86.97</td>
<td>17,646,568</td>
<td>61.83</td>
</tr>
<tr>
<td>MK3_1</td>
<td>21,861,942</td>
<td>19,758,772</td>
<td>90.38</td>
<td>15,323,962</td>
<td>70.09</td>
</tr>
<tr>
<td>MK3_2</td>
<td>10,377,218</td>
<td>9,114,377</td>
<td>87.83</td>
<td>6,705,814</td>
<td>64.62</td>
</tr>
<tr>
<td>MK3_3</td>
<td>16,054,156</td>
<td>15,188,870</td>
<td>94.61</td>
<td>12,822,652</td>
<td>79.87</td>
</tr>
<tr>
<td>MK4_1</td>
<td>37,112,096</td>
<td>30,338,583</td>
<td>81.75</td>
<td>19,715,258</td>
<td>53.12</td>
</tr>
<tr>
<td>MK4_2</td>
<td>27,133,452</td>
<td>25,022,629</td>
<td>92.22</td>
<td>20,419,978</td>
<td>75.26</td>
</tr>
<tr>
<td>MK4_3</td>
<td>31,045,316</td>
<td>27,160,602</td>
<td>87.49</td>
<td>20,776,964</td>
<td>66.92</td>
</tr>
<tr>
<td>MK5_2</td>
<td>11,360,802</td>
<td>10,519,881</td>
<td>92.6</td>
<td>8,988,764</td>
<td>79.12</td>
</tr>
<tr>
<td>MK5_3</td>
<td>2,074</td>
<td>671</td>
<td>32.35</td>
<td>488</td>
<td>23.53</td>
</tr>
</tbody>
</table>

**Reads PP refer to properly paired reads**
To check for similarity and correlation of the RNA between the biological replicates of each studied type in this part of the thesis, figure 20 shows the correlation value (R) in red. If the R value is close to 1.0, this would refer to the high degree of similarity between the replicates. Types I, III and IV show similar R values (between 0.86 - 0.92), which suggests the high degree of similarity between the replicates of each type. Type II, however, shows a less degree of similarity between the replicates, as an R value of 0.78 was the highest (when replicate 3 was compared to 2). The reason behind this low correlation value between the replicates is unknown, as there were no clear observation during the culturing of the bacteria and library making that would suggest that they are not similar.

Figure 20. Correlation of the three *M. kansasii* types I-IV biological replicates. The red colour indicates the R value of the replicates compared to each other.
3.2.2 Comparative RNA seq of ESX-1 and 5

To limit the study of expression data, we focused on ESX-1 and ESX-5 operon that were expressed in *M. kansasii* types II-IV compared to type I (the reference) to complement the comparative genomics studies done in the previous chapter. Analysis of the expression data of the ESX-1 included the genes present within the operon and transcription factors that have some roles controlling the ESX-1 (WhiB transcription factors), especially Whib6 (Pang et al, 2013), and EspR, which is a transcriptional regulator that activates the *espA* operon (Pang et al, 2013). The heat map in figure 21 summarizes the expression values of these previously mentioned. Interestingly, type III shows many genes with high expression value, including the 2 secreted proteins of this system, esxA and esxB, however, the components of the *espA* operon, which is hypothesized to have a huge impact on the functionality of ESX-1, are down regulated in this type with an RPKM value of less than 5, which means it is deleted. As for type II and IV, they show very similar expression patterns, however, they differ in esxA and esxB expression, as type II seem to have higher expression of these two genes than type IV. However, *espA* operon (*espA*, C and D) seems to be down regulated in all, confirming the previous findings in the comparative genomics chapter, but EspR has a value of 1-4 log2 fold change in types II-IV, which is considered to be relatively higher than the *espA* operon. Studies on EspR show that it binds the *espA* promoter, where mutants of EspR showed decreased transcription of the *espA* operon, loss of ESAT-6 secretion, and reduced virulence (Raghavan et al, 2008). In our study, the *espA* operon are down regulated, however the EspR regulator gene shows a slight up regulation in all the studied types, which would suggest that it may not be directly involved in the *espA* operon regulation in *M. kansasii* types II-IV, however, further studies should be conducted to draw such a conclusion.
Figure 21. Heat map of ESX-1 and ESX-5 encoded genes expression patterns. A) shows ESX-1 encoded genes expression patterns of log2 fold change of each gene present in the ESX-1 operon and other genes affecting the functionality of ESX-1 present in types II-IV compared to the reference (type I). B) shows the ESX-5 operon expression pattern in types II-IV compared to type I. The dendogram on the side of each heat map was generated to cluster the genes according to the expression values.
As for ESX-5, all the studied types show similar expression patterns (figure 21b), as most of the genes within the operon seem to be up regulated, apart from the substrates of the operon (esxM and esxN), but esxM is only shown to be up-regulated in type II, which may suggest ESX-5 activation in that particular type, however, in types III and IV show down regulation of both secreted forms of the ESX-5 operons.

3.2.3 RNA seq validation (qRT-PCR) of selected differentially expressed genes

Cufdiff was used to generate the list of differentially expressed genes, which investigates the expression in a given number of samples and tests the statistical significance of each observed change in expression between them (Trapnell et al, 2012). When analyzing expression patterns of type II compared to type I, the differentially down and up regulated genes were 963 and 514, respectively, out of a total of 5075 genes (core genome from OrthoMCL analysis). Type III had 1188 differentially down regulated genes and 1240 of differentially up regulated genes, and type IV differentially expressed genes were 1399 down regulated and 1443 up regulated. For qPCR validation of the differential expressed genes, table 7 shows the selected genes for qPCR analysis, which represent genes of interest in terms of their functional category, and their initial log2 fold change when compared to type I.
Table 7. Selected down and up regulated genes (in log2fold change) for qPCR validation

<table>
<thead>
<tr>
<th>Down/ Up regulated</th>
<th>Gene</th>
<th>Functional Category</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Down regulated</td>
<td>esxN</td>
<td>cell wall and cell processes</td>
<td>-6</td>
<td>-7</td>
<td>-4.82</td>
</tr>
<tr>
<td></td>
<td>VapB43</td>
<td>virulence, detoxification, adaptation</td>
<td>-5.07</td>
<td>-5.07</td>
<td>-5.07</td>
</tr>
<tr>
<td></td>
<td>WhiB4</td>
<td>regulatory proteins</td>
<td>-0.92</td>
<td>-4.9</td>
<td>-5.51</td>
</tr>
<tr>
<td>Up regulated</td>
<td>WhiB6</td>
<td>regulatory proteins</td>
<td>2.21</td>
<td>6.46</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>DrrB</td>
<td>cell wall and cell processes</td>
<td>1.34</td>
<td>3.96</td>
<td>1.24</td>
</tr>
<tr>
<td></td>
<td>Pks</td>
<td>lipid metabolism</td>
<td>0.58</td>
<td>4.38</td>
<td>2.28</td>
</tr>
</tbody>
</table>

From the qPCR results, all the selected down regulated genes were validated except for WhiB4 in type III (figure 22, A). This could not be explained, as the primary expression analysis shown in table 7 gives a log2 fold change value of -4.9. All the up regulated genes that were validated by qPCR confirmed the values given in table 7.
Figure 22. Representation of qPCR results of selected down regulated (A) and up regulated (B) genes.

3.3 Future work

For this part of the project, many aspects should be enhanced. Firstly, include *M. kansasii* type V in the analysis to give a complete view of the transcriptome of this bacterium and its types and to gain more insights of the transcriptome of type V. Repeat the qPCR analysis to assess if WhiB4 in type III is down or up regulated. To understand the function(s) of the differentially expressed genes, mapping to the KEGG ontology to identify genes involved in known metabolic or several signalling pathways. It would also be useful to add more conditions to the bacteria, such as: stress, oxygen depletions or during log phase of their growth, in order to get a view of the change, in the gene expression profile, if any, and compare between exponential and stationary phase, as opposed to the aforementioned conditions.
This project was limited to the study of ESX-1 and 5, it would have been interesting to investigate the expression patterns of ESX 2, 3 and 4 as well. Also, studying the transcription profile of the previously studied functional groups (PE/PPE genes, phospholipase and \textit{mce} operons). If the knock-in approach would be done for the \textit{espA} operon (previously mentioned in chapter 2) in types II-V, it would be interesting to study the gene expression profile of the genome as a whole and in the ESX operons (1-5).
CONCLUSIONS

From the data obtained from the sequenced genome of *M. kansasii* types I-V and the comparative genomics approaches, preliminary data gave us some justification as to why *M. kansasii* type I is considered pathogenic, while the remaining types are known to be environmental. This was proved by the lack of the *espA* operon in the genome of types II-V, and the effect of the operon on ESX1 when the expression data from the types were compared to type I. Other findings that can help support the notion that type I is pathogenic and the other types are not is the presence of a stop codon in one components of the *plc* operon, *plcB*, in types III and IV, which would render it inactive. This could reduce the pathogenicity, as *plc* has a role in host-pathogen interactions. As for analysis of *mce* operons, *mce-2* was absent in all the studied genomes, similar to closely related species *M. marinum*. On average, *mce-1* operon is duplicated, and *mce-3* and *mce-4* is present in one copy in types I-V, however type III lacked *mce-3* operon. RNA profiles of types I-IV were analysed, generated the list of differentially expressed genes, and selected down and up regulated genes were validated by qPCR, confirming the expression level status.
BIBLIOGRAPHY


