COMPARATIVE GENOMIC ANALYSIS OF *Mycobacterium tuberculosis* IN LUNG AND MENINGEAL TUBERCULOSIS

SYAKIRAH NURIZZATI BINTI MOHAMAD HOOD

FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

COMPARATIVE GENOMIC ANALYSIS OF *Mycobacterium tuberculosis* IN LUNG AND MENINGEAL TUBERCULOSIS

SYAKIRAH NURIZZATI BINTI MOHAMAD HOOD

DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF BIOTECHNOLOGY

INSTITUTE OF BIOLOGICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: SYAKIRAH NURIZZATI BINTI MOHAMAD HOOD

Registration/Matric No: SGF 140006

Name of Degree: MASTER OF BIOTECHNOLOGY

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):
COMPARATIVE GENOMIC ANALYSIS OF *Mycobacterium tuberculosis* IN LUNG AND MENINGEAL TUBERCULOSIS

Field of Study: BIOINFORMATICS

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

COMPARATIVE GENOMIC ANALYSIS OF *Mycobacterium tuberculosis* IN LUNG AND MENINGEAL TUBERCULOSIS

ABSTRACT

Tuberculosis is a serious lung disease that has infected a large human population throughout the world. Most cases of tuberculosis in humans are primarily caused by *Mycobacterium tuberculosis*. Tuberculosis cases worldwide mainly affect the lungs, but *Mycobacterium tuberculosis* also causes death by invading the central nervous system, or also known as meningeal tuberculosis. Although the percentage of infection is low, this disease has raised concerns as it is more lethal than lung tuberculosis. The purpose of this study is to identify the genomic differences between the strains that have been isolated from the lung and the cerebrospinal fluid. The dataset of the DNA of both samples that have been previously sequenced is obtained from online database. The lung strain that is used for this study is H37Rv (Genbank ID: NC_000962), which is known as the most common *Mycobacterium tuberculosis* strain that infects patients. The *Mycobacterium tuberculosis* strain from the brain used for this study is strain TKK_04_0158 (Genbank ID: KK324934). The result of this analysis has shown the differences of the genes presented in both strains, which has led to the discovery of the unique gene of each strain. The findings of this study will benefit the medical treatment of meningeal tuberculosis as it will provide an opportunity to develop or improve drugs and vaccine.

Keywords: *Mycobacterium tuberculosis*, pulmonary tuberculosis, meningeal tuberculosis, comparative genome analysis, pathogenicity.

ANALISIS GENOMIK BANDINGAN ANTARA STRAIN *Mycobacterium tuberculosis* PARU-PARU DAN MENIGEAL

ABSTRAK

Tuberkulosis (Batuk kering) ataupun lebih dikenali sebagai TB ialah penyakit paru-paru yang serius yang telah menjangkiti populasi masyarakat dunia yang besar. Kebanyakan kes TB yang menjangkiti manusia adalah disebabkan oleh *Mycobacterium tuberculosis*. Kes tuberkulosis di seluruh dunia adalah secara umumnya menjangkiti paru-paru, tetapi *Mycobacterium tuberculosis* juga boleh menyebabkan kematian dengan menjangkiti system saraf pusat, atau dikenali sebagai tuberculosis meningeal. Walaupun peratusan jangkitan adalah rendah, penyakit ini telah menaikkan kebimbangan oleh masyarakat kerana risiko penyakit ini untuk membawa maut terhadap penghidapnya adalah lebih tinggi berbanding dengan tuberculosis pada paru-paru. Tujuan kajian ini adalah untuk mengenalpasti perbezaan genomik antara strain yang diasingkan daripada paru-paru dan cecair serebrospina. Set data DNA kedua-dua strain yang sudah diaturkan telah diambil dari pangkalan data secara atas talian. Strain paru-paru yang yang digunakan untuk kajian ini adalah H37Rv (ID Genbank: NC_000962), strain ini telah dikenal pasti sebagai strain yang mencatatkan kebiasaan dalam menjangkiti pesakit. Strain *Mycobacterium tuberculosis* dari cecair serebrospina yang digunakan di dalam kajian ini adalah strain TKK_04_0158 (ID Genbank: KK324934). Keputusan daripada analisa ini telah menunjukkan perbezaan gen yang terdapat didalam kedua-dua strain, yang telah membawa kepada penemuan gen yang unik pada setiap strain. Penemuan kajian ini memberi manfaat kepada rawatan perubatan tuberkulosis meningeal kerana dapat memberi peluang mencipta dan menambah baik ubat-ubatan dan vaksin.

Kata kunci: *Mycobacterium tuberculosis*, batuk kering, tuberkulosis meningeal, analisis genomic perbandingan, sifat patogen.

ACKNOWLEDGEMENTS

All praise is due to Allah, the most gracious and most merciful for the strengths and His blessing in writing this dissertation. Completing this dissertation has leaved a great impact on me. I would like to reflect on the people who have helped and supported me so much throughout this period.

Special appreciation goes to my supervisor, Dr. Saharuddin Mohamad for his supervision and constant support. His guidance helped me in all the time of research and writing of this dissertation.

I would like to express my gratitude to the University of Malaya Post Graduate Research Fund, for contributing in financial support and to the Institute of Biological Science for providing all the necessaries and facilities for the completion of the research. My sincere appreciation also goes to all Bioinformatics laboratory staffs for their never-ending help and assistance.

My utmost appreciation goes to my friends for their great help in completing my research works. I would especially like to thank Hamidah, Yousri and Rosniyati for being the best friends and support system I could ever ask for. My research would not have been possible without their helps.

This work is dedicated to the parents for always believed in my ability to be successful in academic. Both of you have always believed in me that has made this journey possible. Most importantly, this dissertation could not have happened without my family. They always there, stood by me through my ups and downs. This dissertation stands as evidence to your unconditional love and inspiration.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

α	:	Alpha
β	:	Beta
γ	:	Gamma
ABC	:	ATP-binding cassette
AIDS	:	Acquired Immune Deficiency Syndrome
ATP	:	Adenosine Tri-Phosphate
BBB	:	Blood Brain Barrier
BLAST	:	Basic Local Alignment Search Tool
bp	:	Base Pair
BRIG	:	BLAST Ring Image Generator
cas	:	CRISPR Associated
cDNA	:	Complementary DNA
CNS	:	Central Nervous System
CRISPR	:	Clustered Regularly Interspaced Short Palindromic Repeats
DNA	:	Deoxyribonucleic Acid
GC	:	Guanine-Cytosine
HIV	:	Human Immunodeficiency Virus
HMM	:	Hidden Markov Model
IFN	:	Interferon
IL	:	Interleukin
Indel	:	Insertion and deletion
iNOS	:	Isoform Nitric Oxide Synthetase
KEGG	:	Kyoto Encyclopedia of Genes and Genomes
LAM	:	Lipoarabinomannan
LM	:	Lipomannan
LSP	:	Large Sequence Polymorphism
LTBI	:	Latent Tuberculosis Infection
ManLAM	:	Mannosyl Lipoarabinomannan
MHC	:	Major Histocompatibility Complex
MMR	:	Mismatch Repair
mRNA	:	Messenger Ribonucleic Acid
MTB	:	*Mycobacterium tuberculosis*
NADH	:	Nicotinamide Adenine Dinucleotide

NCBI	:	National Center for Biotechnology Information
NGS	:	Next Generation Sequencing
PAMP	:	Pathogen-Associated Molecular Pattern
PCR	:	Polymerase Chain Reaction
PE	:	Pro-Glu
PFGE	:	Pulsed-Field Gel Electrophoresis
PGRS	:	Polymorphic GC-rich Repetitive Sequence
PHAST	:	Phage Search Tool
PILAM	:	Phosphoinositide Residue Lipoarabinomannan
PPE	:	Pro-Pro-Glu
PRR	:	Pattern-Recognition Receptor
RAST	:	Rapid Annotation using Subsystem Technology
RFLP	:	Restriction Fragment Length Polymorphism
RNA	:	Ribonucleic Acid
RT-PCR	:	Reverse Transcription Polymerase Chain Reaction
SNP	:	Single Nucleic Polymorphism
TB	:	Tuberculosis
TBM	:	Tuberculosis Meningitis
TDM	:	Trehalose Dimycolates
TLR	:	Toll-Like Receptor
TMM	:	Trehalose Monomycolates
TNF	:	Tumor Necrosis Factor
tRNA	:	Transfer Ribonucleic Acid

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Overview

Mycobacterium tuberculosis is the primary source of worldwide spread disease of tuberculosis. This bacterium was discovered by Robert Koch in 1882. Tuberculosis has always been a serious disease phenomenon worldwide and normally human with a deficient immune system will have a high probability to develop tuberculosis (Finer, 2003). Tuberculosis cases worldwide mainly affect the lungs, but *Mycobacterium tuberculosis* infections have also been causing death by invading the central nervous system (Thwaites et al., 2008). Tuberculosis is one of the highest mortality-causing diseases in the world, whereby, 1.3 million cases have been reported relating to death caused by *Mycobacterium tuberculosis* in the year 2017 alone (WHO, 2018). Some cases of tuberculosis are known to be latent, where the patients are asymptomatic and able to show symptoms later in life. The common symptoms shown by patients include weight loss, fever and persistent cough.

Although *Mycobacterium tuberculosis* is well known to attack the lungs, it could also attack other organs. Meningeal tuberculosis, also known as tuberculosis meningitis is a disease caused by the infection of *Mycobacterium tuberculosis* to the central nervous system, where the bacteria invade the meninges of the host's brain. Although the reported cases are low compared to pulmonary tuberculosis, this disease is highly lethal.

1.2 Problem Statement

Completed and published genomic sequences of *Mycobacterium tuberculosis* that caused meningeal tuberculosis is very limited compared to the one that caused lungs tuberculosis. Due to that, there is currently a big gap in knowledge especially regarding genomic features between these two bacteria strains. The main question of this study is to identify and understand the genomic differences between the *Mycobacterium*

tuberculosis that attack the human lungs (lungs strain) and meninges (meningeal strain). In this study, it is hypothesized that there are unique genes of meningeal strains that became the causative agents allowing the penetration of the meningeal strain into the patient's meninges and thus leading to the meningeal tuberculosis symptoms.

The results of this study would provide more information on the genomic makeup for both bacteria strains. The information would also lead to ideas on how to design a more effective drug targeting these bacteria especially related to meningeal tuberculosis.

1.3 Objectives

In order to identify and understand the genomics differences between these two *Mycobacterium tuberculosis* strains, the objective of the study are:

1. To conduct a comparative genomic analysis of *Mycobacterium tuberculosis* in the lungs and meningeal strains.
2. To identify significant genomic different between both strains that leads to pathogenic properties of *Mycobacterium tuberculosis*.

1.4 Report Organization

Chapter 2 of this thesis is made purposely for the background of the study that has been done for this research. The chapter comprises of a detailed literature review of both disease background for respiratory and CNS tuberculosis, the genomic details and characteristics of the bacteria, the immune response of the host, why discrimination between strains need to be done and lastly the virulence factors of the bacteria.

The methodology conducted in this study has been explained in chapter 3. Chapter 4 focuses both on the results and discussion for this study, containing the comparative genomic blueprint which comprises figures and tables with explanations. The chapter is divided by sections that follow each method done in this study. Chapter 6 concludes the whole research. Raw data are inserted in the appendix.

CHAPTER 2: LITERATURE REVIEW

2.1 Tuberculosis

Caused by intracellular pathogens of *Mycobacterium* genus (Figure 2.1), tuberculosis or widely known as TB is a contagious and potentially life-threating disease. Tuberculosis has been recorded as one of the top 10 diseases in the world that has a high mortality rate that came from single causative agents. Most cases of tuberculosis in humans are caused primarily by *Mycobacterium tuberculosis*. This bacterium is usually transmitted from human to human mainly via air droplets. Another bacterium from this genus, *Mycobacterium bovis* causes transmission of the bacilli that causes tuberculosis in animals such as cows and later can infect human through milk consumption (Brosch et al., 2000). Although the disease is preventable such as by BCG vaccination and treatable by antibiotics, the increasing cases of multi-drug resistance *Mycobacterium tuberculosis,* high susceptibility of immunocompromised patients to the disease and protection limitation of BCG vaccine are warning signs of the spreading of the disease. The standard conventional method of detection and differentiation of the mycobacteria are by acid-fast staining and biochemical methods. To increase the disease detection efficiency, molecular biology techniques such as nucleic acid amplification tests (NAATs) has been developed to identify DNA markers as part of diagnosing process for tuberculosis patients (Chin et al., 2018).

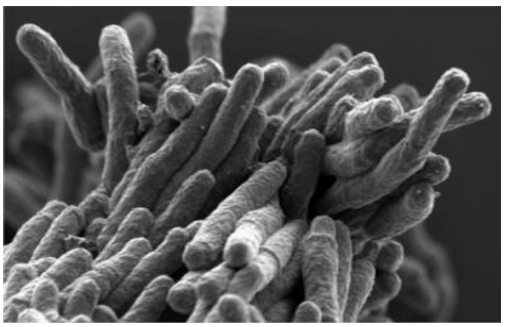


Figure 2.1: *Mycobacterium tuberculosis*. *(*Koch & Mizrahi, 2018), Citation graphic derived from Cell Press Reviews, Copyright Elsevier)

2.2 Tuberculosis Demographics

Tuberculosis is a concerning disease that have infected a large human population throughout the world. From numerous recorded data and reports, there have been 8.8 million of new cases has been reported in 2010 alone. African region shows the highest incidence rate while India and China also show a concerning rate where one in three of world's new Tuberculosis cases is reported from the two countries (WHO, 2011). In Malaysia, the number of cases detected shows an increasing high incidence for every year. Since 1989, Malaysian thoracic society has detected on the average 11,500 to 12,000 cases per year. In 1995, they have recorded 6,500 of total 11,778 detected cases are to be infectious. (Kuppusamy et al., 2018).

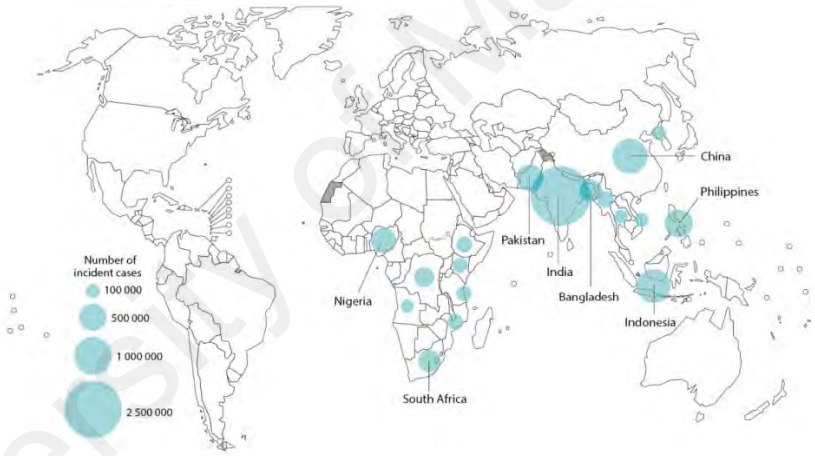


Figure 2.2: WHO reports on the estimated tuberculosis incidence in 2017. Citation report graphic is derived from Global Tuberculosis Report (WHO, 2018). Copyright World Health Organization Publications.

Recorded as one of the top 10 disease in the world that has a high mortality rate that came from single causative agents, tuberculosis incidence case is even higher than HIV/AIDS. Every year, millions of people are dead from tuberculosis. In 2017 alone, there are up to 1.3 million deaths from tuberculosis patients. Figure 2.2 shows that the distribution of tuberculosis in a high incidence countries (WHO, 2018). Although Malaysia is not listed as one of the top 30 countries of high tuberculosis prevalence, the total of infected Malaysian patients raises concern as the multi-drug resistant tuberculosis occurrence increases.

The increasing number of foreign workers entering Malaysia potentially causes a new wave of tuberculosis dissemination in the country. Malaysia receives foreign workers mainly from Indonesia, Bangladesh, India, Myanmar and the Philippines. Most of these countries are enlisted in the top 30 countries of highest tuberculosis incidence. India and Indonesia are in the top 3 mortality rate in 2017 (India: 421,000 death, Indonesia 116,000 deaths) (WHO, 2018). A study conducted on immigrant's health reported that, immigrants are relatively and less-likely to suffer non-transferable diseases, but are more prone to transmittable diseases, majority by tuberculosis and HIV/AIDS (Aldridge et al., 2018). Foreign workers are required to undergo tuberculosis screening upon entering new countries, but it must be noted that there are high percentage of tuberculosis are known to be latent, or known as latent tuberculosis infection (LTBI). A person with LTBI is observed to have a persistent immune response to *Mycobacterium tuberculosis* without showing or experiencing any clinically manifested evidence of active tuberculosis disease (WHO, 2018). Local Malaysian news reported that tuberculosis is the most prevalent among foreign workers (The Star, 2014) and currently the Malaysian immigration reported that there are 107,584 foreign workers are screened with contagious diseases including tuberculosis (The Star, 2018). Recent study on the prevalence of meningitis tuberculosis in Kota Kinabalu, Sabah, Malaysia reported that the disease is a causing a high mortality and morbidity among adults with central nervous system infections. 48.8% of samples collected between 2013 to 2013 are screened with tuberculosis meningitis (Lee et al., 2016).

2.3 Lung Tuberculosis

Most of people infected with *Mycobacterium tuberculosis* are asymptomatic known as latent tuberculosis infection. People with latent tuberculosis infection would have a 5–15% lifetime risk to progress to active tuberculosis disease. However, persons with compromised immune systems, such as people living with HIV, malnutrition or

diabetes, or people who use tobacco, have a much higher risk of falling ill. About 90% of active tuberculosis cases involves lungs with symptoms includes weight loss, fever, chest pain and a prolonged sputum producing cough (Finer, 2003) known as lung tuberculosis (also known as pulmonary tuberculosis or respiratory tuberculosis). When the *Mycobacterium tuberculosis* reach the lung alveolar air sacs, macrophages recognize the foreign entity and conduct phagocytosis in attempt to eliminate the bacteria. In the macrophage, the bacteria are in a membrane-bound vesicle known as phagosome which later transformed to phagolysosome by fusion with lysosome. The macrophage will destroy the foreign bacteria using reactive oxygen species and acidic pH within the phagolysosome. However, the thick and waxy mycolic acid capsule on the surface of *Mycobacterium tuberculosis* protects the bacteria from being destroyed. Furthermore, *Mycobacterium tuberculosis* are able to reproduce inside the macrophage. The bacteria will remain in the phagocytic cell, where it will be brought to the lymph node where the bacteria will be destroyed or continue to spread. At a certain point, the bacteria may be able to disrupt the phagocytic cell of the host and thus will be able to spread to other part of the body. The diseased stage is when the tubercle burst and thus releasing the bacteria into the bloodstream or brochi, the infectious material will be spread to the whole body and directly into the lungs.

2.4 Tuberculosis Meningitis

Tuberculosis for central nervous system only account for 1% of total tuberculosis cases (Thwaites, 2017). Despite its rarity, it is a big problem which they contribute to many death cases compared to other tuberculosis cases. Central nervous system (CNS) tuberculosis can causes two diseases which are tuberculosis meningitis and cerebral tuberculoma. Tuberculosis meningitis will cause death or severe neurological sequelae to half of total cases and it was considered as the most dangerous tuberculosis. Tuberculosis meningitis also can cause progressive confusion and coma that usually

will lead to death if not treated. Meanwhile cerebral tuberculoma is in contrast with tuberculosis meningitis where it is not directly life-threatening. Cerebral tuberculoma will form space-occupying lesions that is produced anywhere in the central nervous region, including the spinal cord (Thwaites, 2017).

From the investigation of Rich and McCordock in 1920s and 1930s, miliary tuberculosis will normally exist in bacteria that are still alive that will cause the occurrence of tuberculosis meningitis after the invasion by blood borne tuberculosis at meninges.

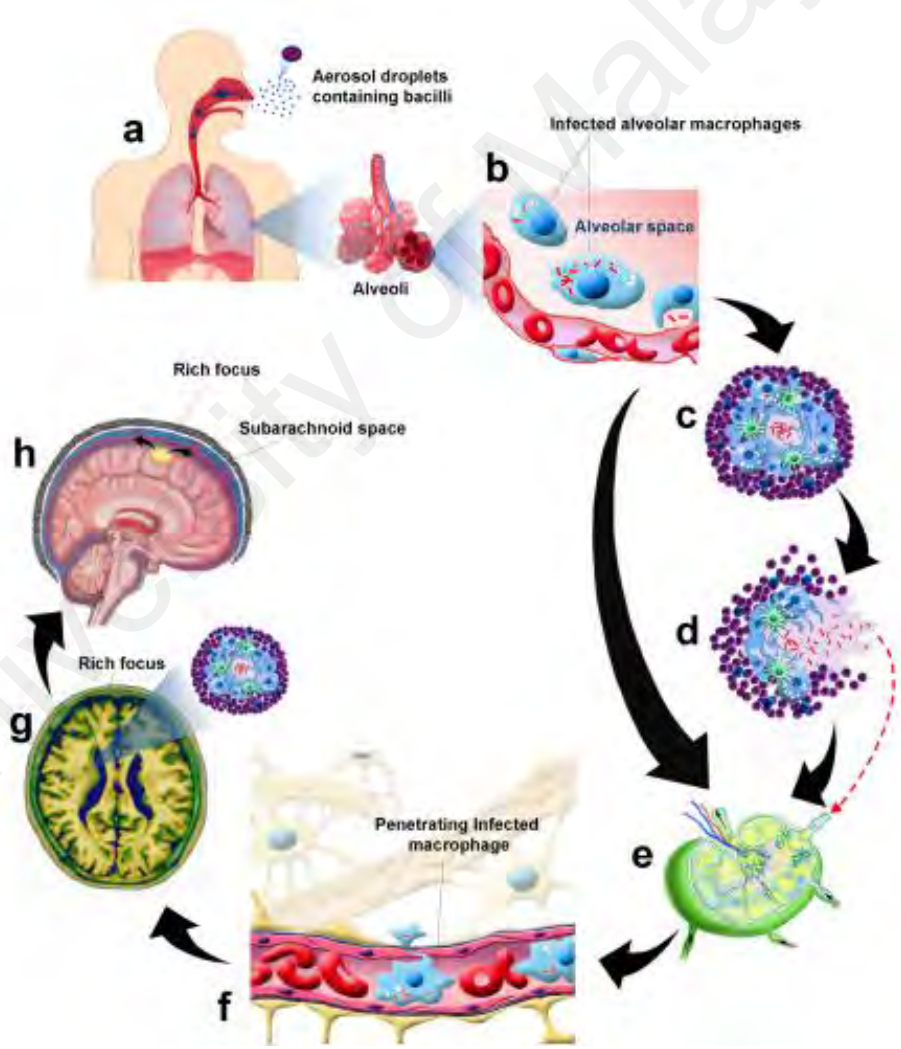


Figure 2.3: The neuropathology of *Mycobacterium tuberculosis* (MTB) (Faksri et al., 2012). Citation graphic is derived from Scientific Reports, Copyright Springer Nature.

A granuloma called Rich Foci has been found as the main element in the brain parenchyma rather than in meninges that was formed during prior bacteraemia. Tuberculosis meningitis is formed after mycobacterium contained in these lesions is freed into subarachnoid space, the occurrence that happens after few months of the first infection of the bacteria in the blood (Thwaites et al., 2008). Figure 2.3 shows the path taken by the bacteria, from a) Aerosol transmission of *Mycobacterium tuberculosis*, b) Phagocytosis of *Mycobacterium tuberculosis* by alveolar macrophages, c) Granuloma formation in the lung, e) *Mycobacterium tuberculosis* can cause meningeal tuberculosis by escalating from the lung or by secondary reactivation, f) *Mycobacterium tuberculosis* can enter the CNS through the blood brain barrier (BBB), bacilli seed to the meninges or the brain parenchyma forming Rich Foci, h) Rich foci increase in size, rupture and discharge into the subarachnoid space (Faksri et al., 2012).

The exudates of inflammation in the basal meninges of basal cisterns are affected, normal cerebrospinal fluid will be halted and thus resulted in hydrocephalus. Meanwhile, vasculitis and tuberculoma development are caused by inflammation of concentrated necrotizing granuloma, which often resulted with stroke syndromes. Basal ganglia and infarction of internal capsule occurred due to the perforation middle cerebral arteries, which is the most common affected area (Thwaites, 2017).

Physiological blood brain barrier protects the central nervous system against systemic circulatory system. Blood brain barrier (BBB) primarily consists of tightly connected microvascular endothelial cells of human brain. The components of blood brain barrier have the properties that support the barrier to be impermeable from various large, hydrophobic molecules and circulating pathogens. Spatial separation of circulatory system of circulatory system from cerebrospinal fluid is done by blood and cerebrospinal fluid barrier at choroid plexus. In spite of these properties, meningitis or

encephalitis can later be induced by viral pathogens that are able to cross the blood brain barrier (Be et al., 2015).

The molecular level of blood brain barrier is regulated by ectozymes, receptors and transporters that will control the traffic across the structure. Together with histological construction, a stable CNS surrounding are formed that is specific in 1. Ionic composition for neurones 2. Neurotransmitter pool 3. Low protein concentration in avoidance to cell proliferation, 4. Protection against systemic toxins to minimized neuronal damage, and 5. Low traffic of cells and molecules that causes inflammation (Varatharaj & Galea, 2017).

Tuberculosis meningitis is diagnosed by analysing the cerebrospinal fluid (CSF) obtained by performing lumbar puncture at every patient. There are few criteria that has been considered in the analysis of the cerebrospinal fluid which are the white blood cells with the mixture of neutrophils and lymphocyte, cerebrospinal protein, ratio of cerebrospinal fluid to the blood glucose and cerebrospinal fluid lactate (Thwaites et al., 2008). For the rapid diagnose of tuberculosis meningitis, acid fast bacilli is searched in the cerebrospinal fluid with staining.

2.4.1 Immune Response of Central Nervous System

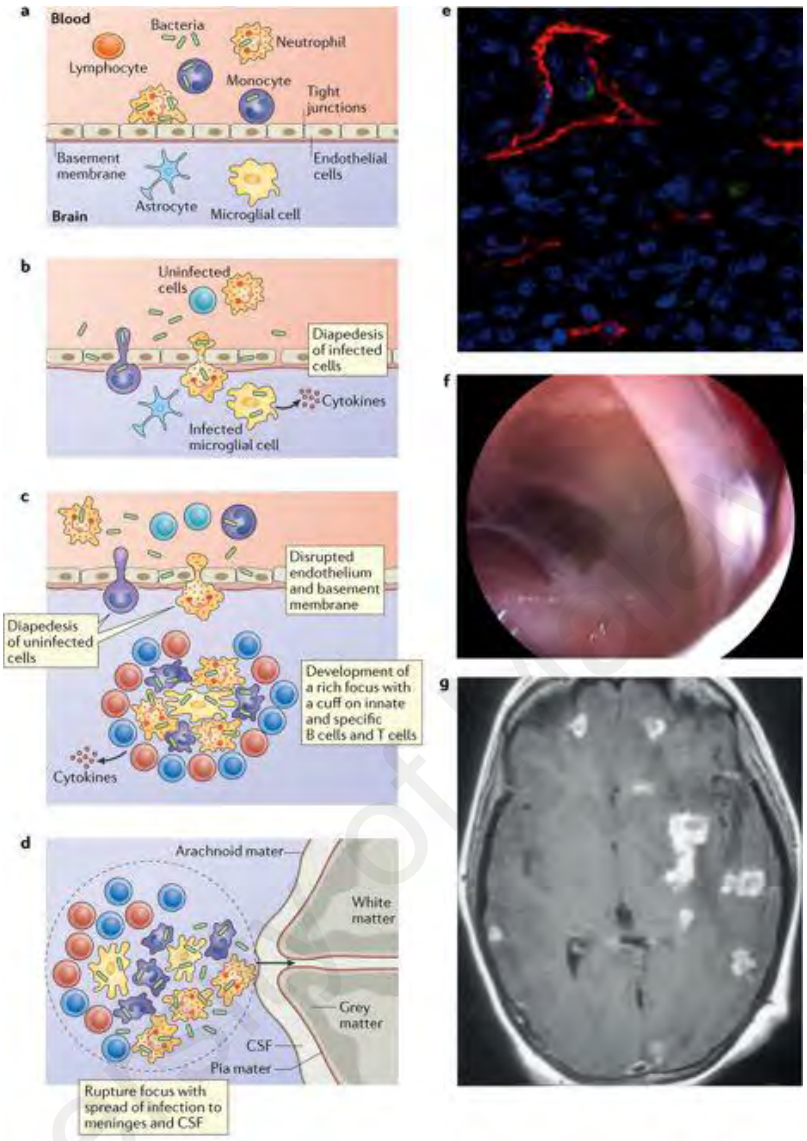


Figure 2.4: The immune response system of the blood brain barrier of human host upon the invasion of *Mycobacterium tuberculosis*. (Wilkinson et al., 2017). Citation graphic is derived from Nature Reviews, Copyright Springer Nature

Figure 2.4 shows the penetration of *Mycobacterium tuberculosis* through the blood brain barrier of a human host and how the immune system is responding. Strain of *Mycobacterium tuberculosis* is able to breach the blood brain barrier as a free organisms or by infecting neutrophils or monocytes. Formation of tuberculosis is established in brain parenchyma as such cellular traffic is restricted into central nervous system in the early infection. Ruptures of Rich Focus lead to invasion of *Mycobacterium tuberculosis* into subarachnoid space which will trigger a vigorous inflammatory T cell response

(Wilkinson et al., 2017). Increased level of TNF-α and IFN-γ has been recorded in previous studies of cerebrospinal fluid cytokine levels in patients with tuberculosis meningitis (Ransohoff et al., 2003).

The development of infection and severity of a disease is determined by the capability of the host to identify invading bacilli and to amount protective immune response against *Mycobacterium tuberculosis* (Joanne et al., 2000). *Mycobacterium tuberculosis* strain CDC1551 exhibits a high levels of TNF-α, interleukin (IL)-6, IL-10, IL-12 and interferon (IFN)- γ mRNA in the granuloma which later will be found in lungs (Manca et al., 1999). A model with a hyper-lethal *Mycobacterium tuberculosis* strain HN878 induced weak T cell proliferation and IFN-γ production by spleen and lymph nodes (Manca et al., 2001).

TNF-α can cause harm to the host when it is presented in excess and for a long time especially in the brain, although it is known that TNF-α has a protective characteristics against *Mycobacterium tuberculosis* infection. The tissue in brain can be damaged when TNF-α is associated with bacillary load, leukocytes and persistent TNF. This is because, the activation of TNF-α will trigger coagulation and thrombi formation which would eventually lead to vascular occlusion and necrosis (Tsenova et al., 1999).

The blood brain barrier is altered by proinflammatory cytokine in the CNS which will induce adhesion-molecule expression on the microvascular (Hickey, 1999). TNF can induce production of nitric oxide synthase, nitric oxide and nitric radical which will lead to endarteritis and tissue damage. Proinflammatory cytokine can be produced prior to the activation of the endothelial cell, with monocytes and T cell entering the tissue and further damaging the endothelial cell (Tsenova et al., 2005). Few of the non-disruptive blood brain barrier are mediated directly by cytokines. Some receptors such

as TNF- α, IL-1β and IL-6 are expressed on cerebral endothelium (Varatharaj & Galea, 2017).

Although generally mentioned as immunologically privileged, immune response in the brain parenchyma is better known for being selective and modified. This means that with the appearance of foreign antigen, the immune response is limited. This feature is fundamental to limit the internal brain inflammatory damage, as brain is very crucial and minimally regenerative (Galea et al., 2007).

The selective and modified immune reactivity of the central nervous system is moderated by several cellular and physiological factors. First; there are extremely limited antigen presences within the parenchyma. Next, the intact human brain has low level of MHCII molecules due to the restriction on dendritic cell (Galea et al., 2007). The third factor is the weak antigen presenting capabilities exhibited by microglia which is known as the macrophage of the brain (Ford et al., 1996). This will induce the apoptosis of the effector cells instead of proliferation. The last factor is the limitation of blood brain barrier and cell surface factors granted an access for systematic lymphocytes to brain parenchyma (Ransohoff et al., 2003).

The factors mentioned above contribute to inability of the brain to establish immediate inflammatory meningitis after the spreading of tuberculosis. Instead the inflammatory meningitis commonly to take place many months of the incidence of active disease (Be et al., 2015).

2.4.2 Contribution of Bioactive Lipid to Virulence

In an experiment conducted by past researchers state that they have demonstrated that tumor necrosis factor (TNF)-α as a determinant of immunity and pathogenicity of central nervous system. This has been achieved by inoculating *Mycobaterium tuberculosis* directly into the cisterna magna of rabbit models (Tsenova et al., 2005).

The activities of phenolic glycolipid hold a very crucial role in a hyper-virulence of *Mycobacterium tuberculosis* in the central nervous system. A rabbit model infected with a mutant strain that do not have this glycolipid displayed a minimal pathological and has a very low neurological signs although it appears to have a distinctive bacillary load in the brain. There are also a high TH1 immunity which then has increased the survival rate of the host. It also has been noted that the phenolic glycolipid mechanisms has not made any difference in the lungs (Reed et al., 2004).

After macrophage has been infected with *Mycobacterium strain* HN878 that possess phenolic glycolipid, the macrophage moves to the brain from the subarachnoid space with the aid from the bacilli. This macrophage has a high control in the bacillary growth at the brain. The macrophage is also able to breach the blood brain barrier more effectively than other strains like CDC1551 and H37Rv (Tsenova et al., 2005). As a part of the immune system, TNF-α is necessary for protection against *Mycobacterium tuberculosis*. This has been shown by the anti TNF monoclonal antibodies act of neutralization of TNF-α in rheumatoid arthritis patients that can induce latent tuberculosis infection. TNF-α is a part of TH1 protective response, alongside with other cytokine and chemokines that involves IL-2, chemokine receptor CCR-6, IL-6 and Fas-ligand; that will be induced during early interaction of virulent strains such as HN878 and CDC1551 (Cooper & Flynn, 1995).

A study on the protein globularity change where the enrichment of proteins that was involved in sex hormones from host could affect metabolism and virulence of bacteria. The hormones are degraded in order to be used for as carbon and energy sources. This mechanism could upset the unusual bacterial-host interaction to facilitate CNS invasion. Proteins of interests such as ABC transporters and glycophospholipids function to transfer molecules across cell membrane. They are also implicated in *Mycobacterium tuberculosis* transversal across the blood brain barrier. (Saw et al., 2016).

2.4.3 Tuberculosis of CNS in Children

Children who are infected by tuberculosis meningitis will usually shows similar complications. Hypothalamic injury or inflammation will cause hypnotremia, a syndrome where ADH production is in inappropriate amount. Hydrocephalus is another common complication that caused blockage to the cerebrospinal fluid flow. It is also noted that hydrocephalus is also a normal occurrence in common tuberculosis meningitis patients, not only children (Thwaites, 2017). The involvement of cranial nerve such as vasculitis of vessels in subarachnoid spaces will trigger total occlusion and thrombus formation. Neuritis is induced by nerve basal exudates entrapment. Vasculitis can also be a case of tuberculoma on the nerve within subarachnoid course (Muzumdar et al., 2018).

Tuberculosis meningitis may be secondary to tuberculosis in the lungs, lymph nodes, bones, urinary system and the rest of the body. A research conducted by (Huo et al., 2018) recorded that 65% of tuberculosis meningitis occurrence is among young adults. Most common clinical symptoms observed from the patients meningeal irritation, fever and headaches where the percentage of manifestation is more than 90%. An interesting case report from (Rahim & Ghazali, 2016) records that a Malay woman at the age of 19 years old was diagnosed with 1 year duration of psychosis, worsening before admission to the hospital. The patient also experiences mutism, stupor and in Glasgow coma scale of 11/15. Another research by Kumar shows similar case where the patient in study presented psychology symptoms such irrelevant talking, disorganized behaviour and physical symptoms such as constipation and reoccurring headaches.

2.5 The Genome of *Mycobacterium tuberculosis*

The most important aspects on the study of *Mycobacterium tuberculosis* will be in the aspects of determination of the genome. In 1998, it has been found that the most

used reference strain of *M. tuberculosis* contains 4,411,529 base pairs and has about 4000 genes. This genome is larger than most other bacteria (Brosch et al., 1998). For the chemical structure of the genome, it has found that *M. tuberculosis* is remarkably uniform with a high guanine and cytosine content which is 65.6% throughout, which indicates it has evolved with minimal incorporation of DNA from extraneous source (Grange, 2014). The mycobacterial genome is also unique in containing a large number of genes, up to 10% of the total coding potential, that code for two unrelated families, Pro-Glu and Pro-Pro-Glu, of acidic, glycine-rich proteins that contribute to the diversity in antigenic structure and virulence (Palucci et al., 2016).

Mycobacterium tuberculosis is one of the highly conserved bacterial groups. It also has very little phenotypic differences which have influence pathogenesis. Nevertheless, the genetic background of this bacterium is related to the difference in transmissibility and virulence among *Mycobacterium tuberculosis* strains. Certain strains of outbreaks are known to exhibit specific immune path and mortality rate (Lopez et al., 2003).

The method of genomic annotation has shown that *Mycobacterium tuberculosis* contains few unique characteristics. From the total genes, 6% of it or about 200 genes are annotated as encoding enzymes for fatty acid metabolism. The usage of this bacterium with fatty acid is possibly related to pathogenic ability to grow in the tissues of the infected host. This is because the fatty acid is the main source of carbon (Smith, 2003).

Mycobacterium tuberculosis genome also shows other significant feature, where there is the presence of the unrelated PE and PPE families of acidic, glycine-rich proteins. The families' names are adapted from the sequence of Pro-Glu (PE) and Pro-Pro-Glu (PPE) in the N-terminal region, where each of the families has about 110 and 180 amino acids long. Both of these families comprises for about 172 genes which is

4% of *Mycobacterium tuberculosis* total genes count (Brosch et al., 2000). PE-PGRS subfamily, a subdivided protein from PE family genes has conserved PE domains and extensions of C-terminal with Gly-Gly-Ala or Gly-Gly-Asn repeats in polymorphic GC-rich repetitive sequence (PGRS) domains. The PE-PGRS family formed alongside other proteins that are subdivided from 104 PE genes has unknown functions. But there are variations of the size of PE-PGRS family in clinical *Mycobacterium tuberculosis* strains, where most of these proteins are located in the cell wall and the cell membrane of the bacteria (Banu et al., 2002). This finding has suggested the antigenicity of the proteins and lead to the hypothesis that these proteins contributed to the antigenic variation of *Mycobacterium tuberculosis* during the course of infection in the host.

Chromosomal mutations which majorly by single nucleotide polymorphisms (SNP) can mainly cause the increasing of genetically encoded drug resistance in *Mycobacterium tuberculosis*. This has become the global health threat as the mutation that happened in the genome of *Mycobacterium tuberculosis* decreases the efficiency of drugs and can lead to undesirable side effects to the patients. There are number of factors to be considered in order to find the lead to the mutation of *Mycobacterium tuberculosis* during host infection. Evolution can be driven during DNA replication and repair through spontaneous error. Target novel agents can contribute a lot to a specific mechanism and pathway which can influence the mutation rate (Mcgrath et al., 2014).

There are several main types of genomic differences within *mycobacterium* complex that has been described before, which include single nucleotide polymorphism, long-sequence polymorphism, minisattelites and microsatellites (Palucci et al., 2016). The genome of *M. tuberculosis* also contains 'jumping genes' that associated to contribute to genetic variation and evolution. The genome members of *M. tuberculosis* also contains direct repair locus that consists of repetitive 36 base-pair units of DNA that is separated by non-repetitive 34-41 base pair spacer oligonucleotides. The direct repeat region of *M.*

tuberculosis is an example of a region present in all bacterial genomes and is termed clustered regularly interspaced short palindromic repeats (CRISPR). The function of this region is unknown, but it may be the bacterial analogue of the centromere found in eukaryocyte chromosomes. There are numerous possible combinations of spacer oligonucleotides, and these are very stable, providing a highly discriminative typing scheme known as spacer oligonucleotide typing, or 'spoligotyping' (Zhang et al., 2011).

2.5.1 CRISPR and its Relation to Virulence

Clustered regularly interspaced short palindrome repeat or shortened as CRISPR is known as a sequence that is capable to do task in an individual cell. CRISPR associated genes (cas) normally found alongside to CRISPR. Cas gene is important for encoding fundamental protein for immune response. CRISPR-cas system aims for DNA or RNA as a method of protection against viruses or any other mobile genetic elements (Hale et al., 2009). CRISPR locus existed for about 8% in archea and 45% in bacteria (Grissa et al., 2008), which also includes *mycobacterium tuberculosis*. Spacers with unique sequence separating an array of short repeated sequences. CRISPR is a part of anti-virus system as spacers are formed from nucleic acid of viruses and plasmids.

New viruses or other mobile genetic element can be identified by adding new spacers. This has made the spacer to be used as a recognition component in order to find a matching virus genome and therefore destroys them. Offspring of microbes will inherit any alteration made for protection, such as genome modification of CRISPR where spacer acquisition is done. The addition of the new spacer at the sides of CRISPR has made a record for microbial chronology that has been encountered by its ancestor (Rath at el., 2015).

The fundamental process of CRISPR-Cas system involves three important steps. The first step is adaptation, where new spacer is inserted into CRISPR locus. There are two

types of adaptations. The first one is naïve spacer acquisition, where there is no previous information about the attacking virus stored in the bacterium. The second adaptation is known as prime acquisition that has involved the insertion of new spacer with a stored DNA information about the attacking virus. The adaptations are conducted with the help from cas1 and cas2 proteins. The second step of this system is expression. Here, the transcription of CRISPR locus and processing of CRISPR DNA take place. RNA that has been expressed with the virus DNA is transcripted. The last step is interference, is when the combined action of CRISPR RNA and cas protein identify and destroy the nucleic acid of the target (Rath et al., 2015).

By having a revolutionary genetic that consists of remnants and memories of previous bacteriophage attacks, CRISPR loci has the primary role to protect a bacterium against phage challenge. The discovery of this function has been the main aspect of CRISPR loci studies, but the functionality of CRISPR does not stop there. There are also important role that CRISPR possess such as the regulation of virulence, DNA repair and genome revolution of a bacteria.

CRISPR-cas system also has a distinctive role in the maintenance of bacteria that will effect on the survival inside a host. Bacterium such as *E.coli* has a Cas1 protein that is able to process the replication of single-stranded DNA, branched DNA and 5' end flaps. The interaction of cas1 protein with RecB, RecC and RuvB proteins suggests a vital role in DNA repairing. CRISPR-cas system also has a role in handling the accumulation of defective protein, where the system triggered by misfolded protein in the bacteria (Babu et al., 2011). CRISPR-Cas system also has a role in gene regulation of bacteria, like *Listeria monocytogenes*. The host chromosome is targeted by a CRISPR without cas gene leading to an increase of level of targeted RNA following stabilisation by CRISPR RNA (Perez-Rodriguez et al., 2011).

In the terms of virulence, cas9 protein is able to repress an endogenous lipoprotein gene for full virulence that has been seen in *F. novicida*. The repression is important for the virulence as the lipoprotein will trigger an innate immune system of the host. Expression of Cas9 protein that deficient in CRISPR loci has shown an increase in virulence, as recorded in *C. jejuni*.

A CRISPR study is still recent. Other functional roles that may be possessed by CRISPR are not yet well known and further studies need to be conducted for clarification.

2.6 Immune Response of Host

When *Mycobacterium tuberculosis* attacked a host, a course of infection triggers a chain of immune response. Wide range of probable clinical manifestation can take place in patients who cannot control the infection at any life stage. TH1 type cytokine promote macrophage activation which have the primary control of an infection. Inteferon-gamma (IFN-γ) and tumour necrosis factor (TNF-α) have a main control on the macrophage activation and the expression of isoform nitric oxide synthetase (iNOS).

This process will produce the important nitric oxide in order for intracellular mycobacteria to be killed. If this protective mechanism failed, TH2 type cytokine will be released for protective support (Chan et al., 1992).

In the early phase of infection, high level of TH1 cell cytokine will be produced alongside a high production of TNF-α and iNOS. This is crucial in order to temporarily control the infection (Wangoo et al., 2001). In this phase, the formation of granulomas will take place. After three weeks from the initial infection, TH1 cell cytokines will be expressed while the level of TNF-α and iNOS reduced. Granulomas are conquered by pneumonia which eventually can cause death.

Infected organism by *Mycobacterium tuberculosis* affects the ability of the bacterial pathogenicity, but it is independent of host factor. The extent of the severity of the pneumonia and mortality rate correlates with the rate of bacterial multiplication in lungs. Sustained TH1 cell response of the host successfully controls the *Mycobacterium tuberculosis* infection.

In an experiment conducted by Lopez *et al.,* (2013) course of infection in the brain is different for each type of virulent strain. The expression of TNF-α, iNOS and IFN-γ plays important role in controlling the infection. Beijing strain, which having a consistent accelerated bacterial multiplication, has a high level of TNF-α and iNOS expression in the early infection. During this phase, macrophages are rapidly activated. However, low level of IFN-γ expression is recorded as these macrophages are being deactivated and did not stimulate TH1 cells. This has led the inefficiency to arrest the bacillary multiplication. For Canetti strain that lead to a slow progressive disease, records a high and maintained level of TNF-α and iNOS gene expression that may limit the progression of the disease. This mechanism has prevented tissue damage and cause a low mortality to the host. By activating macrophage, high level of IFN-γ during early infection is not needed to control the infection (Lopez et al., 2003).

2.6.1 Pathogenic pattern recognition

Innate immunity is the first line of defense against any bacteria and other pathogens. Innate immunity is primarily consists of macrophages and dendritic cells. This immunity is not entirely unspecific. The system is able to differentiate between themselves and pathogens, by identifying microbes through pattern-recognition receptor (PRRs) (Akira et al., 2006).

The pattern recognition receptor has several roles in the immunity system. First, this receptor can identify microbial element in the form of pathogen-associated molecular

pattern (PAMPs). PAMPs is known to be fundamental for bacterial survival and therefore it is complicated for the bacteria to alter its structure. So the PAMPs of bacteria is highly conserved compared to other structures. The next importance of PRRs is that they are able to recognize pathogens in any state of its life cycle and they are expressed constitutively inside the host. PRRs are non-clonal, encoded in germline and they can be expressed in types of cells. One type of PRRs will react specifically with another type of PRRs, making them to be precisely particular. A distinctive expression pattern will activates a specific signaling pathway and will initiate particular anti-pathogen response (Akira et al., 2006).

One type of pattern-recognition receptors is toll-like receptor (TLR), which is a type 1 integral membrane glycoprotein that is classified by extracellular domains that contains a leucine-rich-repeats (LRR) motifs (Bowie & O'Neill, 2000). LRR is a protein with structural motifs that shapes α or β horseshoe fold, with a 20 to 30 amino acid stretch. Toll-like receptor can detect PAMPs. For example, TLR1, TLR2 and TLR6 are all recognize lipids, while TLR7, TLR8 and TLR9 recognize nucleic acids. Diverse type of immune cells such as macrophages, dendritic cell, B cells and T cells expresses toll-like receptors and they are expressed in different sites such as on cell surfaces and in intracellular components like nucleic acid and endosomes. An activation of TLR can induce a downstream signalling cascades and produces proinflammatory cytokines and chemokines.

Toll-like receptor 9 (TLR9) recognizes bacteria genomic DNA, which is an immunostimulant. This is due to the presence of unmethylated CpG-dinucleotides in particular base contact designated CpG-DNA (Hemmi et al., 2000). Although there are plenty of CpG motifs in bacterial genome, they were subdued and highly methylated in the mammalian genome. This has makes the mammalian immune cell not activated.

CpG-DNA display an intense activity of immunity stimulation that involves the production of inflammatory cytokine and TH1 immune response (Akira et al., 2006).

The pathogen-associated molecular patterns (PAMPs) in mycobacteria are able to survive in the macrophage of the host by several elaborated mechanisms. The cell wall of mycobacteria consists of thick and waxy combination which composed of lipid and polysaccharide. They also contain high amount of mycolic acid. Mycobacterial PAMPs activated a toll-like receptor 2 (TLR2).

2.6.2 Modulatory Immune Lipoglycans

Mycobacteria exhibit a strong modulatory immune lipoglycans which relates with lipomannan (LM) and lipoarabinomannan (LAM).

LAM is from LM that has been arabinosylated. The arabinan domain is capped by mannosyl (ManLAM) or by phosphoinositide residue (PILAM). ManLAM is known as an intense anti-inflammatory molecule that has been formed in a slow growing mycobacteria species, for example *Mycobacterium tuberculosis* and *Mycobacterium bovis*. Meanwhile PILAM is a powerful TLR2 stimulator, they were found in non-pathogenic and fast growing microbes. By examined the ratio of ManLAM : LM in the cell wall of a bacteria, it can determine response against virulent bacteria, which a mycobacteria has induced a cytokine in TLR2-dependent manner (Quesniaux et al., 2004). A cell wall associated lipoprotein which also a secreted antigen of *Mycobacterium tuberculosis* that can powerfully induce macrophages, are recognizable by TLR2 that has been associated with TLR1 (Thoma-Uszynski et al., 2001).

There are some bacteria and fungi are able to abuse TLR systems in order to evade the immune responses of the host. *Mycobacterium tuberculosis* can escape the death by macrophage by inhibits IFN-γ-mediated signalling that has been stimulated by TLR2.

Mycobacterium are able to evade T cell responses of the host by continuously signalling TLR2 and therefore are able to persist as a long-term infection (Quesniaux et al., 2004).

2.6.3 THP-1 System

The mortality and morbidity parameters are a way to determine the virulence of *Mycobacterium tuberculosis*. At the first 25 days of infection, the capacity of bacteria strain to multiply in livers, lungs and spleen is determined by observing its growth. It is also fundamental to determine the capability of the bacteria to produce a granulomatous response and cytokine mRNA. One way to verify the intracellular growth rates of *Mycobacterium tuberculosis* is by monitoring a primary culture of human macrophage and monocyte cell lines that has been transformed like THP-1 systems. (Theus et al., 2005).

In a study conducted by Zhang et al., *Mycobacterium tuberculosis* strain 210 that has caused for about 25% from overall cases in Los Angeles where the bacteria grew more quickly compared to a *Mycobacterium tuberculosis* that only affected one patient although the patient is highly exposed to unsanitary environment (Zhang et al., 1999). The relationship between the widely spread of a strain and the ability to for it multiply rapidly in the THP-1 cell system is important to identify on how *Mycobacterium tuberculosis* strain 210 is able to spread widely. It can also be discovered by identifying the virulence marker that has the ability to multiply quickly in human macrophage.

The persistent RFLP clusters of clinical strain have a remarkable quicker growth rates compared to the RFLP clusters from a unique clinical strain. Virulent strains such as *Mycobacterium tuberculosis* 210 is known to have a persistent RFLP clusters. This discovery proved that it is able to differentiate a strain that has evolved for disease transmission and establishment from other kind of strain. This differentiation is made able by the difference in activated THP-1 cell system phenotype. Factors like IL-10

rapid induction, rate of growth and TNF- α inhibition are considered to be the phenotypic markers that are relevant to epidemiologically related phenotype. The observation on gene *sigA* which is fundamental in the virulent gene expression supported this phenotype characteristic. This gene has been observed to be regulated in the *Mycobacterium tuberculosis* 210 and not in other strains (Wu et al., 2004).

The property of the bacterial strain growth phenotype is determined by the cytokine-inducing capacity of the bacteria. At the early phase of infection, TNF- α is secreted at the highest level. THP-1 cells of the infecting unique bacterial strain causes a notably high level of TNF- α to be induced as being compared to infection of a strain in clusters. The course of infection by bacterial strain with a slow-growth phenotype is mediated by the rapid TNF- α response. This will later restrict the replication of mycobacteria. Meanwhile, the rapid-growth bacterial strain has a suppressive TNF- α production. This is highly likely due to the high level of IL-10 production (Giacomini et al., 2001). The inhibitory effect of a macrophage is influenced by the concentration of IL-10. Not only IL-10 suppresses the type-1 cytokine in response to *Mycobacterium tuberculosis* infection, it also reduces the production of TNF- α from macrophage.

2.7 Virulence Factors of *Mycobacterium tuberculosis*

The preeminent source of mortality around the world is still remain to come from tuberculosis. Some of the factors that causes further spreading of this disease are from the rise of multi-drug resistance forms of tuberculosis, the widespread of HIV/AIDS and the degradation of the public health especially in developing countries (Forrellad et al., 2013).

To evade the viscous environment of a macrophage, the *Mycobacterium tuberculosis* species have evolved in few mechanisms. Such mechanisms include inhibiting the

phagosome lysosome fusion and to avoid from environments inside the phagolysosome (Meena & Rajni, 2010).

In recent years, it has been discovered that majority of the virulent genes encode the enzymes of cell surface proteins, proteins of signal transduction system, regulators and some lipid pathways. Mycobacteria are distinctively lack in prime virulence factors such as toxins that are common such other bacterial pathogens. Those multiple virulence genes in the *Mycobacterium tuberculosis* complex species are also conserved in non-pathogenic mycobacteria. This discovery suggest that with little acquisition of the exclusive virulence genes, the pathogenic species have to make use of their genomes to adapt from free lifestyle to the intracellular environment (Forrellad et al., 2013).

One of the condition to categorize a gene as a virulence factor is that its absence will weaken the virulence of the microorganism in an in vivo model. Still, this standard comprises a wide range of genes which also involved a housekeeping gene that poses a role in survival inside the host. These housekeeping genes are in part of the fundamental cellular metabolism and are not normally considered as virulence factors.

In the study by Forrelad (Forrellad et al., 2013) they have classified the virulence elements into several categories based on the role they possessed, molecular characteristics or cellular localization. Some of the important groups are: cell envelope proteins, proteins that restrict the antimicrobial effector of a macrophage, lipid and fatty acid metabolisms, gene expression regulators, PE and PE_PGRS that has the unknown protein function and several others. Another recent research that studies the mutant of H37Rv strains with inactivated prototype PE_PGRS protein shows that the mutant mycobacteria has a lowered efficiency during the entry to macrophage compared to non-mutant strains (Palucci et al., 2016).

Mycobacterium tuberculosis presents a large spectrum of complex lipids and lipoglycan on its cell surface. This has made them to be different among other bacterial pathogens. These unique cell wall lipids is well known to have a crucial part in pathogenicity; thus the genes are subjected for their biosynthesis, transport and degradation are highly possible as a virulence determinants that propose new targets for drug design.

The capability of *Mycobacterium tuberculosis* strains with the resistance inside the host can be contributed with number of factors aside from mutational rate, for instance like preferable advanced efflux system, higher replication rate and increasing potential adaptation to resistance of fitness loss.

2.7.1 Mycolic Acid

The first macromolecular layer after the peptidoglycan of mycobacteria is comprises of heteropolysaccharide arabinogalactant to which esterifies the mycolic acid. This mycolic acid is particularly identical in length but contrasting in structure. The structure is divided into three main sub-families which are clycopropanation, keto or methoxy. These three groups will later be esterified to become glycerol and trehalose. Trehalose will then be able to form either trehalose dimycolates (TDM) or trehalose monomycolates (TMM). Differently with most prokaryotes, mycobacteria uses the fatty acid synthase (FAS) system to make a long chains of fatty acid, compared to carbon-14 to carbon-18 range in most prokaryote (Barry, 2001).

In a mycobacteria system, mycolic acid methyl transferase is responsible for the classification of the three types of mycolic acid by adding methyl groups to the mycolic acid. The formation of these sub-families hold a vital information in the pathogenicity of the mycobacteria. An oxygenated mycolic acid is important in the infection process, as been shown before in an experiment that involves a mutant of *Mycobacterium*

tuberculosis tested in a mouse with inactivated *hma* (equal to *cmA* and *mmaA4*) gene. The altered gene show loss of oxygenated mycolic acids and an alteration in envelope permeability (Dubnau et al., 2000). Another gene that has gained attention throughout the years of research is *pcaA*. Deletion on *pcaA* gene can lead to modification of site-specific cyclopropyl of mycolic. This alteration has been a determinant for the interaction between *Mycobacterium tuberculosis* and the host it resides. One of the important virulence in this bacteria is caused by the alteration mentioned on the mycolates TDM as it also have a deep effect on the function of these lipids and not just modify the innate immune recognition of mycobacteria.

2.7.2 *kasB* Gene

A bacterial system known as FASII from FAS system, is encoded with *kasB* gene. This gene will causes a deletion that will affect the loss of acid-fast staining, alteration of the colony morphology and cording, which mean stunted normal serpentine growth. As done in experiment done by Bhatt *et al.* 2007, a mice with the disruption of *kasB* gene is able to survive without any disease or mortality. This result suggest that *kasB* gene is involved in the pathogenicity of *Mycobacterium tuberculosis* (Bhatt et al., 2007).

2.7.3 *mymA* Operon

A *mymA* operon has a fundamental function in the cell envelope of the bacteria as the operon system is involved in the mycolic acid export. It has been studied that the mutant of this operon has shown an altered and reduced content of mycolic acids also change in saturation of fatty acids, which later has been suggested that it has a connection in protecting the bacteria under any unfavourable conditions (Singh et al., 2005).

2.7.4 *pks* Gene

Another important set of genes that has known with its credibility on the virulence of mycobacterium complex is *pks*. A study by Reed et al. (2004) has concluded that a disruptions in the *pks1/15* gene which is a mutant, has caused a deficiency in the production of phenolic glycolipids. A production of phenolic glycolipid is highly associated to the hyper virulent phenotype as has been displayed by W-beijing family. The non-appearance of phenolic glycolipid will cause the inability by the bacteria to become a hyper-virulent phenotype, but without affecting the bacterial load during the disease period. A *pks10* gene also has been studied that a mutant strain with this gene shows the same phenotypes as *pks1/15* in the virulence attenuation. Meanwhile, a *pks5* gene functions in encodes a polyketide enzyme. By disrupting this gene, has displayed a severe growth defects in mouse although it show no difference in cell envelope lipid composition (Reed et al., 2004).

2.7.5 Mismatch Repair Enzyme

Mismatch repair (MMR) enzymes is a system presented in bacteria exclusively made for recognizing and repairing erroneous insertion and deletion throughout the process of DNA recombination and replication. Homologues of MMR enzymes however are not included in the DNA repair complement of *Mycobacterium tuberculosis* (Mizrahi & Andersen, 1998). Despite this absence, the lacking of MMR may applies a selective pressure on this bacteria that resulted in more stable genome and lower degree of polymorphism in simple sequence repeats or other known as microsatellite. Simple sequence repeat can affect in the polymerase slippage that causes frameshift mutation (Wanner et al., 2008). The nucleotides of *Mycobacterium tuberculosis* predictably are randomly distributed as they comprise of fewer long sequences. Based on a study conducted by Machowski et al., (2007) polymorphic GC-rich sequence (PGRS) repeats

are found among the PE_PGRS members of the highly conserved PE protein family. In this protein family, the existence of proline-glutamate residues of N-terminal region is not related to increasing mutational rate that associated with base substitutions (Machowski et al., 2007). These finding supports the fact that *Mycobacterium tuberculosis* genome is having limiting features of inherent variability that might be contributed by being under selective pressure.

Genetic variation in *Mycobacterium tuberculosis* can cause by the lacking of MMR. In other Mycobacteria species, MMR functions by forbidding recombination between non-identical sequences in the course of the repair of double strand-DNA breaks. The absence of MMR can raise the divergence of DNA sequence and therefore can lead to genome evolution (Springer et al., 2004).

2.7.6 Mutator Allele

A mutator strains is developed from spontaneous mutations inside a gene that codes for DNA metabolic proteins. This strain has a short-term selective advantage by having a capability to manufacture a lot of adaptive mutations. The relation of beneficial mutations with these mutator alleles provides bacterial maintenance to them. The beneficial relation only happened if the cost inherent in the rising risk of producing deleterious mutations can counterbalance or is not lesser than the fitness gains.

Some potential mutator allele has been studied in order to find the relation with the emergence of drug-resistance in *Mycobacterium tuberculosis*, for example, mutations in *mutT4* and *mutT2* in Beijing isolates. However there is no direct link between the mutator and DNA repair mechanism of the bacteria, but the monitored mutations gives other phenotype that indirectly influence the accession of drug resistance but not through DNA repair mechanism (Ebrahimi-Rad et al., 2003).

2.8 Discrimination Between Strains

Comparison between strains of *Mycobacterium tuberculosis* is very crucial. Studies has shown that this bacillus exist in several genotypes in term of lineage, families or clade. There are many approach has been discovered to define these genotypes such as spoligotyping and delineation of short nucleotide polymorphism (SNP), large sequence polymorphis (LSP), MIRU-VNTR and RFLP. For LSP variation, *Mycobacterium tuberculosis* was delineated into four major genotypes which are East Asian/Beijing, East African/Indian, Indo-Oceanic and Euro-American (Faksri et al., 2018). While spoligotyping, which more widely used has been divided into 62 sublineages. International database such as SpolDB4 has published thousands of *Mycobacterium tuberculosis* complex strain (Brudey et al., 2006).

Discrimination between strains is done to understand the diversity of bacterial pathogen. For *Mycobacterium tuberculosis* approach such as multilocus sequence typing is inappropriate because of the low phylogenetic resolution. While PFGE and RFLP approach comes with many drawbacks and is difficult to reproduce in laboratory (Comas et al., 2006). PCR-based genotyping possesses the best approaches. Clustered Regulatory Short Palindromic Repeats (CRISPR) are the regions of the bacterial genome that characterized by series of direct repeats by spacers which is a short unique region. Analysing CRISPR enable us to identify the encoded specialized defence mechanism against bacteriophages and phage susceptibility (Jain *et al*., 2006). Meanwhile Variable Number Tandem Repeats (VNTR) compares strains specific number of repeats of short DNA sequence at difference position of bacterial genome (Grissa *et al*., 2008). In *Mycobacterium tuberculosis* complex CRISPR is known as spoligotyping while VNTR as MIRU-VNTR (Comas, 2009).

There are few methods that have been used to analyse the genome of *Mycobacterium tuberculosis*. The methods are discovered for the purpose of quantifying the specific genes expression or to identify in a big scale and also to count the many RNA transcripts. One of the techniques is hybridization-based method. This technique is based on the cDNA production from RNA which the primer is from a known gene that later undergo PCR amplification. Reverse-transcriptase (RT-PCR) is a reliable method to measure individual transcripts. But the drawback of this method is that they are not unique due to contamination of the host RNA. One of the variations of RT-PCR, molecular beacon is developed to increase the reliability and accuracy. When *Mycobacterium tuberculosis* is grown under different stress, molecular beacon will measure the levels of several sigma factors gene mRNAs. It also counts mRNA of various *Mycobacterium tuberculosis* genes during the infections of human macrophage (Smith, 2003).

The rise of amplification-based hybridization techniques has made *Mycobacterium tuberculosis* transcripts able to be globally identified. The first variation of this technique is differential display which is used to identify few genes that are differentially expressed in *Mycobacterium tuberculosis* (RiveraMarrero et al., 1998). The second variation is cDNA method, also known as DECAL that eliminates abundant amount of RNAs which can disrupts PCR amplification specificity. This method also optimizes diverse parameters (Alland et al., 1998). The third variation is random cDNA synthesis. This method is achieved by subtractive hybridization and PCR amplification (Graham & Clark-Curtiss, 1999).

The emergence of DNA arrays development has enabled all *Mycobacterium tuberculosis* genes expression profiling. This development has become to be the fundamental benefit of *Mycobacterium tuberculosis* genome and its annotation. DNA arrays, or mainly known as DNA chips, is a breakthrough technology where the chip

has been probed with labelled cDNA complex mixtures. The important benefits of DNA arrays are; it allows an increase number of analysing genes and decreasing the significant amount of sample size requirement (Smith, 2003).

DNA arrays have been useful for *Mycobacterium tuberculosis* global gene analysis, especially where the sample obtained has been exposed to various conditions such as under anti-tubecular drug, acidic growth and heat shock. Mutations effects studies of *Mycobacterium tuberculosis* genes has been made possible, particularly the genes that codes for transcriptional regulatory proteins on global gene expression.

CHAPTER 3: METHODOLOGY

3.1 Whole Genome Sequencing of *M. tuberculosis*.

3.1.1 *Mycobacterium tuberculosis* Isolated from the Lungs

Whole genome sequencing projects for clinical *M. tuberculosis* isolates that have been sequenced and assembled before are acquired. Several sequenced whole genome from the isolate of pulmonary tuberculosis has been chosen for preliminary screening. Strain H37Rv from Sanger Institute has been chosen for final study, as it is one of the most used strains in the genomic studies. Genbank accession: NC_000962. (https://www.ncbi.nlm.nih.gov/nuccore/NC_000962)

3.1.2 *Mycobacterium tuberculosis* Isolated from Cerebrospinal Fluid (CSF).

Whole genome sequencing projects are acquired from the public database. The strain that has been obtained is strain TKK_04_0158, sequenced by Broad Institute. The strain was obtained in a set of scaffold consisting 68 contigs. Genbank accession: KK324934. (https://www.ncbi.nlm.nih.gov/nuccore/KK324934.1/)

3.2 De Novo Assembly

3.2.1 Whole Genome Sequence Draft Completion

GAPfiller v1.10 obtained from https://sourceforge.net/projects/gapfiller/ was used to fill the existing gaps of the cerebrospinal fluid (CSF) strain that have been obtained from the public database. Clean reads of forward and reverse sequences were used in .fastq files. Both reads was run together alongside the library file of the GAPfiller and the output files of the obtained scaffold. Artemis was then used to combine all the contigs of the scaffold, to become a single FASTA sequence.

3.2.2 Gene Prediction

PRODIGAL v2.60 downloaded from https://github.com/hyattpd/Prodigal was used to predict the number of genes that are presented in the of the *Mycobacterium tuberculosis* sample. Assembly files were uploaded to the prediction tool after scaffolding and gapfilling processes were completed. After the output file was generated in .gff format, the total number of genes were viewed and counted.

3.2.3 Gene Annotation

Both of the sequences that were obtained from the lungs and brain are uploaded into RAST annotation system at http://rast.nmpdr.org/. Each sequence in .fasta file was uploaded and computed separately. After the annotation process was completed, the annotation report was checked. The annotated genes were then browsed using the SEED viewer. Subsystem information chart distribution and features in subsystem were extracted. Complete recorded genes subsystem and features are presented in the Appendix A and B.

3.2.4 Genome Visualisation

DNAplotter was used to illustrate the genome of the strain isolated from the brain. Features such as coding DNA sequences, tRNA, mRNA, GC content and GC skews were visualised according to their respective tracks in circular pattern. Blast Ring Image Generator (BRIG) which was downloaded from http://brig.sourceforge.net/ ran alongside BLAST in the system. Strain H37Rv was selected to be uploaded as the reference and TKK_04_0158 as the query sequence. Each strain was colored accordingly and the visualized comparison ring was recorded.

3.3 Comparative Genomic Analysis

3.3.1 Genome Alignment

Both genomes from the lung and the brain that have been assembled into a single whole genome sequence are aligned together. Mauve software was used to generate a pairwise alignment of the newly assembled and annotated H37Rv and TKK_04_0158. Progressive Mauve alignment option was selected to compute the H37Rv sequence files that have been converted to GenBank file (A. E. Darling et al., 2010) .gbk previously by RAST. TKK_04_0158 sequence was added afterwards. After the alignment process has been completed, the visualization image that has been generated is studied and saved.

3.3.1.1 SNP Detection and Analysis

Mauve alignment was used to call SNPs from aligned reads of the sequenced genomes. By using Mauve progressive alignment program, the short reads will be aligned to the reference genome. The generated SNPs data that has produced is inserted in the appendix. The projected SNPs data was then classified according to its position in the chromosome. The chromosomal position of SNP density was classified in line with 10 000 bp intervals, later to be projected to a line graph to highlight the SNPs density peaks.

3.3.1.2 Insertion and Deletion Analysis

Mauve multiple alignment software and progressive alignment were used to perform multiple alignment of genomic sequences. The output file obtained by Mauve was parsed using custom Perl script to receive multiple align sequences for indel loci (Liu et al., 2014) The complete raw SNPs data that has been produced by Mauve is inserted in the Appendix D. The indel loci that has been obtained was separated by insertion and deletion, and each of them was clustered by their position, which is 40000 bp intervals.

3.3.2 CRISPR Locus Identification

CRISPRFinder (http://crispr.i2bc.paris-saclay.fr/Server/) was used to detect CRISPR loci from the draft TKK_04_0158 genome sequence and the published H37Rv genome sequence. The complete FASTA file was uploaded to the website query. BLAST was then used to detect the similarity searches between CRISPR spacer sequences and existing sequences in GenBank (Grissa et al., 2008).

3.3.3 Identification of Significant Genes

Significant genes were identified with the alignment from RAST server comparative tools, following the annotated procedure previously. The meningeal strain sequence was aligned with the sample from the lungs. The genes that are only presented in the brain and the genes that only presented in the lungs were identified.

3.3.4 Genomic Island Identification

Genbank (.gbk) file that has been converted from FASTA files for both *Mycobacterium tuberculosis* strain H37Rv and TKK_04_0158 were uploaded to http://www.pathogenomics.sfu.ca/islandviewer as input. The visualized output data was downloaded and all genomics islands were interpreted.

3.3.5 Pathogen Detection

Open web tool PHAST at http://phast.whirstlab.com/ was used to identify both incomplete and intact pathogens that were presented in the strain from the brain. The contigs in FASTA format of both strains were uploaded and run separately. The position was visualized in circular image, together with summary location table of the detected prophages.

CHAPTER 4: RESULTS AND DISCUSSION

In this study, high quality genomic datasets of *Mycobacterium tuberculosis* from lung tuberculosis strain H37Rv and meningeal tuberculosis TKK_04_0158 were assembled using *de novo* sequence assembly method in order to conduct a comparative genomic analysis of *Mycobacterium tuberculosis* in the lungs and meningeal strains. The notable information is further discussed afterward.

4.1 Genomes Data Sets

Table 4.1: Summary of genomic details

Feature	H37Rv (pulmonary)	TKK_04_0158 (meningeal)
Genomes size	4,411,532	4,425,690
GC content	65.6%	65.6%
No. of coding sequence	4312	4356
No. of RNAs	48	48
Contigs	1	1

Both working strain are obtained from genbank, where both of them reportedly were sequenced as a high quality isolate sequences. As shown in Table 4.1, both datasets were successfully assembled into a single continuing *de novo* sequence assembly method. The genomic size of strain TKK_04_0158 is larger than strain H37Rv, also contained higher amount of coding sequence. The percentage of GC content and total number of RNAs is the same.

4.2 Genome Visualisation and Annotation

The genomes of both organisms were visualized using DNAPlotter. Both genome sequences formatted in .gen files were added into the software. This tool provides a precise and clear visualized representation of both genomes in circular image. Projected image from this tool for both sequences is shown in Figure 4.1.

The image is evaluated by the genomic positions which are done by the coordinates tick marks. The outermost track of the circle which is in light pink represents the forward coding sequence, while the second outermost track which is in aqua blue represents the backward coding sequence. The third and fourth circular track plots for forward and reverse tRNA respectively, while the fifth and sixth track plots for forward and reverse rRNA. There are only few bands visible at the third track until the sixth track as it shows that tRNA and rRNA are not easily found throughout the whole genome positions. Second innermost track, the GC plot indicates the fragment of the bases which is either G or C. The innermost track shows the GC skew, which from the GC skew formula, a positive value represents a bias toward G (green spike), while negative value specify on a bias towards C (purple spike). The two innermost of GC track provides evidence that there is lateral gene transfer in few fragments of the genomes.

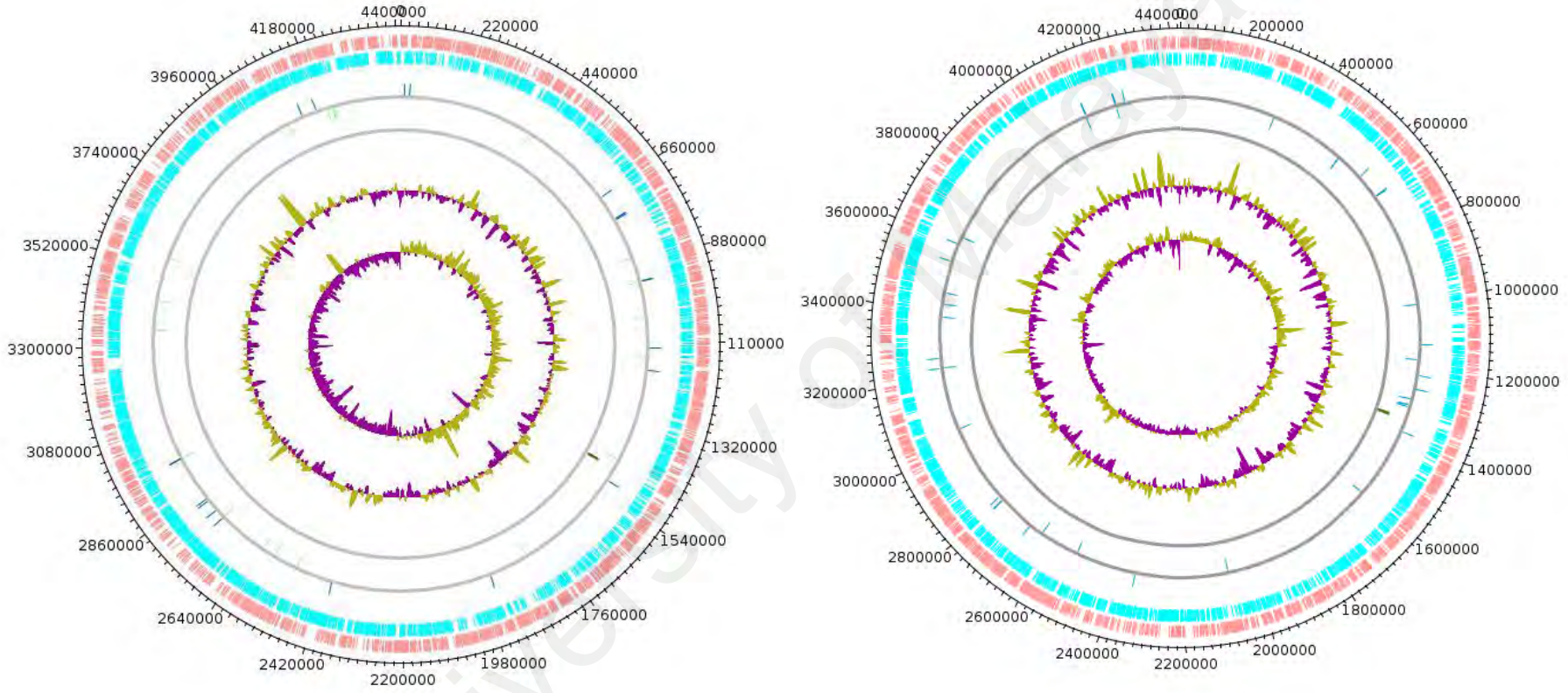


Figure 4.1: Circular representation of *Mycobacterium tuberculosis* strain H37Rv (lungs) compared to strain TKK_04_0158. From outermost to inner track each indicates forward sequence (pink), backward sequence (aqua), forward tRNA. Reverse tRNA, GC-plot (green, fraction base is G; purple, fraction base is C) and GC-skew (green, positive; purple, negative)

By comparing the circular representation of both strains, we were able to visualize the genome by whole chromosome; therefore enable us to identify the location of tRNA and rRNA positions.

GC skew occurred due to under or over copious guanine and cytosine at a certain region of a DNA. This happened particularly because of mutation occurrence that effectuates the nucleotide to be distributed randomly along the genome.

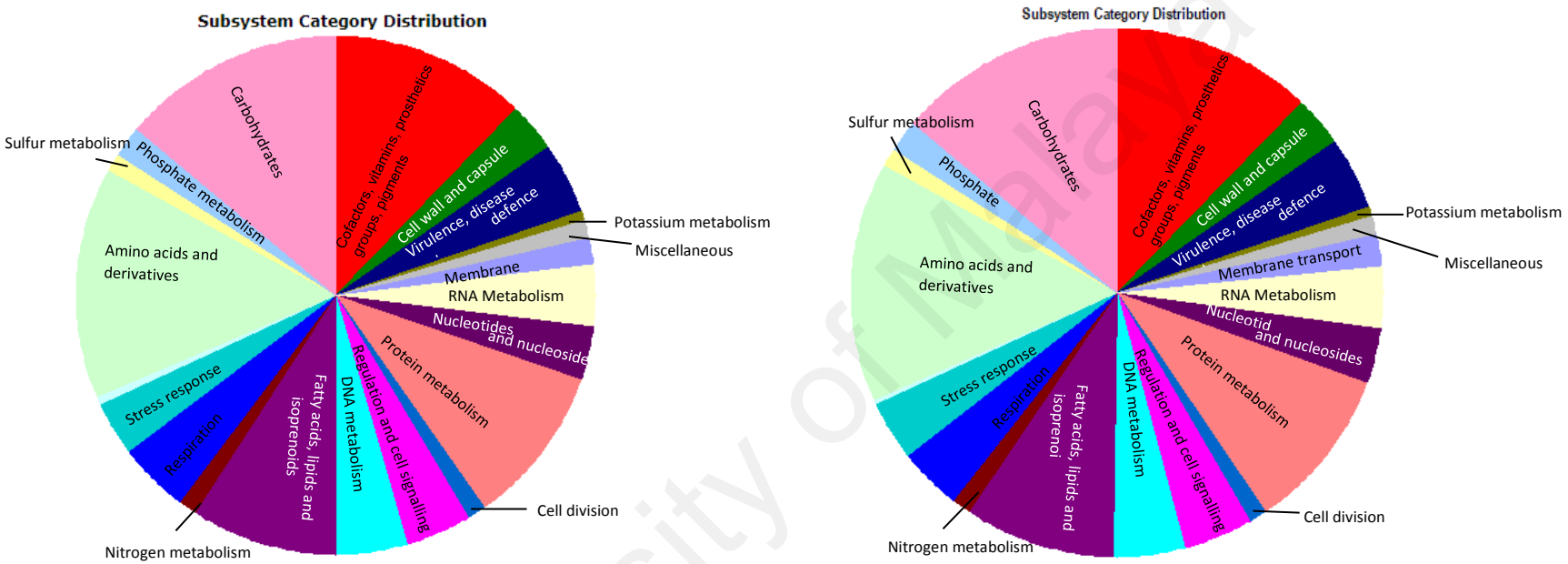
GC skew in the middle of the diagram (Figure 4.1) shows difference at the peak of the skew. Nevertheless, it is not considered as a significant difference, as the GC percentage does not show an apparent contrast as both strain records at 65.6%. The recorded value is presumed to be correct as multiple past researchers also records similar GC content percentage such as study from Brosch et al., 2000. By doing the GC skew analysis, the origin and the terminal of circular chromosome has been determined. The example of recent arrangement which are sequence inversion or integration of foreign DNA, are made able to be marked after the analysis of GC skew, and is beneficial for future research.

For gene annotation, both complete genomic sequence of brain and lung strains were uploaded into web-based annotation tool, Rapid Annotation Server (RAST) at https://rast.nmpdr.org/. The uploaded sequence is in FASTA format. The genes that have been calculated are assigned to 27 subsystem features. A total of 2459 genes from the meningeal tuberculosis strain and 2452 genes from pulmonary tuberculosis have been successfully sorted into respected categories (Table 4.2). The visualised subsystems by both genomes are shown side by side in Figure 4.2.

It is mandatory to execute gene annotation in this comparative study as this analysis provides detailed list of genes presented in each genome, with most of them are

identified according to classes of metabolism. The size and position of each gene in base pairs are also obtained.

All of the information obtained is applicable in the next analysis, where all the significant difference that has been recorded requires detailed information of genes. The details are referred by obtaining the exact position of interest, for example the peak of SNP density in SNP analysis will be evaluated with the type of genes that resides at the peak position.



H37Rv TKK_04_0158

Figure 4.2: Gene content analysis of *Mycobacterium tuberculosis* strain H37Rv (lungs) and TKK_04_0158 (meningeal) associated to subsystem according to functional categories. Both recorded the total of 27 subsystem feature, and each subsystem feature is labelled accordingly.

Both recorded subsystem feature counts has been simplified and compared in table 4.2 below. Full list of recorded subsystem and its subsequent genes has been inserted in the Appendix A and B. As highlighted in Table 4.2, there are two subsystem count that can be vital information in this genomic comparison analysis, which are Phages, Prophages, Transposable elements, Plasmids feature count and Virulence, Disease and Defense feature count. Nevertheless both of them showed similar value. Most of the categories do not shows significant difference in the gene number, with most difference between strains belongs to RNA metabolism and respiration.

Table 4.2: Comparison of subsystem feature count between *Mycobacterium tuberculosis* strain TKK_04_0158 and H37Rv

Category	**Total number of subsystem**	
	TKK_01_0548 (meningeal)	**H37Rv (lungs)**
Cofactors, Vitamins, Prosthetic Groups, Pigments	304	308
Cell Wall and Capsule	77	77
Virulence, Disease and Defense	**106**	**106**
Potassium metabolism	14	14
Miscellaneous	31	31
Phages, Prophages, Transposable elements, Plasmids	**2**	**6**
Membrane Transport	43	43
Iron acquisition and metabolism	2	2
RNA Metabolism	97	73
Nucleosides and Nucleotides	82	83
Protein Metabolism	247	245
Cell Division and Cell Cycle	26	25
Motility and Chemotaxis	2	2
Regulation and Cell signaling	105	108
Secondary Metabolism	2	2
DNA Metabolism	108	107
Fatty Acids, Lipids, and Isoprenoids	227	227
Nitrogen Metabolism	27	26
Dormancy and Sporulation	2	2
Respiration	97	108
Stress Response	86	85
Metabolism of Aromatic Compounds	10	10
Amino Acids and Derivatives	359	358
Sulfur Metabolism	30	30
Phosphorus Metabolism	49	49
Carbohydrates	327	325
Totals	2459	2452

Candidate genes identification has been the main aims of this study as it will provide lead information for the better understanding for the bacteria's virulence and its relativity. The list of genes and proteins that is listed by the tool are classified into subsystem feature, therefore the identification of genes has been specified from its function.

As described in chapter 2, vast research has discovered a high number of pathogenic genes. From the genes data presented in Table 4.2, some of the key genes have been located in both strains, and some remain undetected. Nevertheless, the data have provided ample explanation on the biochemical metabolism during the course infection by the mycobacteria, and those related genes will be emphasized here.

A notable finding, where Rv0931c, one of a gene from a set of genes in relation to heparin binding hemaglutinin are present from the list of the annotated genes (Appendix A and B). These set of genes have been mentioned repeatedly by many researchers (For example; Rv0980c, Rv0987, Rv0989, Rv1801, Rv0311, Rv0805, Rv0931c). These are the genes that control the activity in the human brain microvascular endothelial cell during the pathogenic invasion. This is an important finding as the endothelial cells are vital for protection of central nervous system from the systemic circulation of blood brain barrier (Forrellad et al., 2013). Rv0931c was found to have similar in size recorded at 1995 bp, located at 1039914-1037920 bp and 1619546-1617552 bp for strain H37Rv and TKK_04_0158 respectively. It is also noted that the gene location does not falls within any of the SNP or indel density peaks. Rv0931c is known to be related to the production of gene *pknD*, a novel gene that is attenuated in brain and not lungs, and help in the survival of *Mycobacterium tuberculosis* (Be et al., 2012).

As been listed in Appendix A (brain) and Appendix B (lungs) few family of proteins that has been identified as a virulence factor such as PE-PGRS, PPE, *pks*, FASII and

mutT has been recorded similar for both strain. For PE-PGRS, H37Rv recorded 39 genes found while TKK_04_0158 was 40, where from the total virulence factor associated protein detected was 5 and 4 respectively for each strain. PPE protein in H37Rv recorded 66 genes in total while the TKK_04_0158 was 71. For *pks* and *mutT* both recorded similar number, only varies in length where it was longer in TKK_04_0158.

From the research done by Smith (2003), about 6% (200 genes) of total genes are bound to encode for fatty acid metabolism. From this study, a similar number gene for the category is recorded; which are 227 genes. The number has shown similarity for both TKK_04_0158 strain and H37Rv strain. Therefore, it is known here that pathogens ability to grow in the tissue of infected host under the activity of fatty acid metabolism does not contribute to the virulence difference between the two strains (Smith, 2003).

4.3 Genome Alignment

A Java-based tool for multiple alignment of whole genome called Mauve was downloaded and installed using the Linux operating system. The source of the tool is from darlinglab.org/mauve/download.html (Darling et al., 2004). This tool enabled the contigs uploaded to be ordered and oriented against another assembly. The input sequence file for the software is in FASTA format, which has been assembled without any gap. For this study, both sequences were uploaded simultaneously, with *Mycobacterium tuberculosis* H37Rv lined as the reference. The projected alignment is shown in Figure 4.3.

A set of genome assemblies of both studied strains were taken up by Mauve and pair-wise genome alignment is generated. As seen in Figure 4.3, this tool recognizes the sequence homology blocks which each is assigned with different colours. The visualisation was made easier as the sequence of all the coloured blocks are visualized

as a single whole genome. This has enabled for the conserved regions to be identified, as well as unique regions.

A visualisation tool named BLAST ring image generator (BRIG) (Alikhan et al., 2011) that also has been downloaded and installed is used to visualize both of *Mycobacterium tuberculosis* genomes in a circular genome map when they were aligned next to each other. Figure 4.4 shows the alignment.

It is crucial to conduct pair-wise alignment as the overall differences of genetic components such as insertion and deletion (indels) and SNPs can be detected by this alignment method.

It is worth to note that Mauve is mainly used for generating SNP and indels from the alignment and the generated image on Figure 4.3 represents the alignment by homology blocks, not by percentage of identity. The aligned data from Mauve is then used in BRIG in order to generate a simpler and concise visualization by percentage of identity, as presented in Figure 4.4.

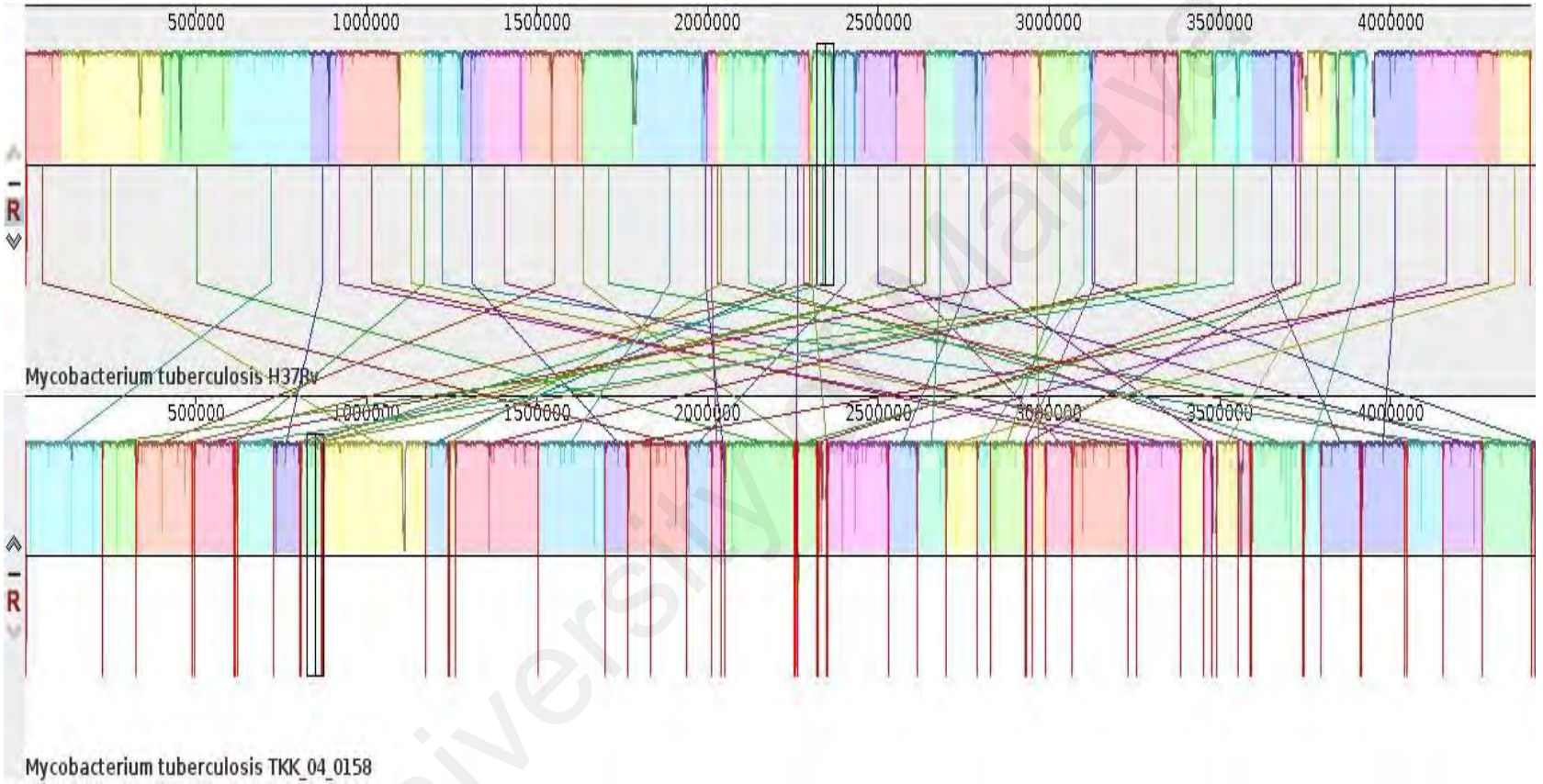


Figure 4.3: Both *Mycobacterium tuberculosis* H37Rv and TKK_04_0158 genomes are aligned to each other. Coloured blocks outline genome sequence that aligned to other part of genome and was conclude to be homologous and internally free from genome rearrangement (LCBs). White regions are sequences that were not aligned in the reverse complement orientation.

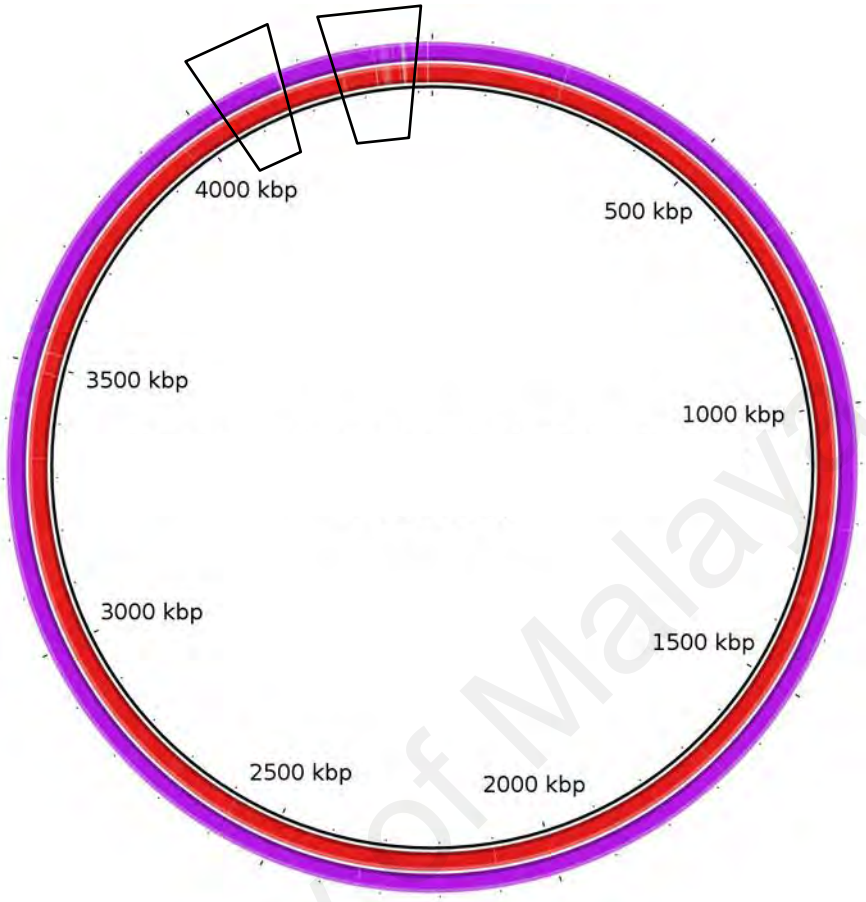


Figure 4.4: *Mycobacterium tuberculosis* strain TKK_04_0158 (red circle) aligned to strain H37Rv as a reference strain (Purple strain). Any discoloration formed in the circle indicates the difference in the genome sequence.

The projected alignment shows a high symmetrical identity. As presented in the Figure 4.4, the genetic content when aligned is highly similar. Main information extracted from the Mauve alignment has been utilized in analysing the SNPs and indels content, which will later be discussed.

The absence of a complete genome of *Mycobacterium tuberculosis* genome strains that are extracted from the brain in the genomic database has caused the inability to execute multiple alignments between brain strains, to allow further investigation in the conservation pattern for resistance of the bacteria. The absence of complete genome also poses a challenge for an assured assembly and comparative analysis.

The genomic content of the scaffolds that has been assessed by BRIG, which the scaffolds has been made as the query sequences, were investigated further into notable regions. Regions that show a high level of variability when compared to the reference genome were narrowed down into the identification of gene content throughout the region. Although there is difference recorded in the Figure 4.4, the genes that reside at the pinpointed regions does not classified as highly significant. It is noted that the region contains the predictable CRISPR genes such as cas1, cas2, csm6, csm5 and csm2. These genes are related to the results obtained in CRISPR analysis, which will further discussed be in section 4.6. Upon acquiring the comparative circular blast alignment for the plasmid sequence that has been visualized in Figure 4.4, it is safe to conclude that the homology between both strains is highly similar.

4.4 SNP Analysis

After both *Mycobacterium tuberculosis* strain TKK_04_0158 and H37Rv has been aligned to each other by using Mauve genome alignment tool, single nucleotide polymorphisms (SNP) analysis were executed. By using a specific PERL script, as shown in Appendix C, a detailed list of SNP has been generated, according to the base positions. Further analysis has been done, which involves a simplification of the SNP by classifying each of them according to groups of genomic positions per 10000 base pairs.

From the data extracted from the list of the SNPs as shown in Table 4.3, a SNP graph is drawn in order to have a clear view of the SNPs frequency according to the base positions.

As seen in Table 4.3 and Figure 4.5, there are significant peaks of SNP concentration that can be observed at position between 340000-350000 bp, 1100000-1110000 bp, 1640000-1350000 bp, 2000000-2010000 bp and 3740000-3750000 bp. All of these regions recorded values above 100 SNPs, a benchmark point to indicate a high density

of SNP in a region. The highest SNPs recorded at 269. Every virulence gene candidates that has been residing the high density SNPs regions are identified and has been listed in Table 4.4.

Table 4.3: SNP density and its position. Bolded data indicates a high density of SNP occurence

Position	SNP density	Position	SNP density	Position	SNP density	Position	SNP density	Position	SNP density
<10000	8	<510000	3	<1010000	1	<1510000	0	<2010000	1
<20000	5	<520000	1	<1020000	0	<1520000	2	<2020000	0
<30000	4	<530000	0	<1030000	0	<1530000	1	<2030000	8
<40000	6	<540000	2	<1040000	2	<1540000	4	<2040000	1
<50000	2	<550000	1	<1050000	2	<1550000	3	<2050000	4
<60000	1	<560000	3	<1060000	1	<1560000	1	<2060000	6
<70000	3	<570000	2	<1070000	2	<1570000	0	<2070000	0
<80000	4	<580000	2	<1080000	7	<1580000	1	<2080000	3
<90000	2	<590000	2	<1090000	4	<1590000	1	<2090000	0
<100000	1	<600000	3	<1100000	**244**	<1600000	1	<2100000	2
<110000	6	<610000	0	<1110000	6	<1610000	1	<2110000	2
<120000	2	<620000	1	<1120000	0	<1620000	2	<2120000	1
<130000	2	<630000	2	<1130000	2	<1630000	1	<2130000	4
<140000	3	<640000	4	<1140000	0	<1640000	**159**	<2140000	2
<150000	3	<650000	1	<1150000	3	<1650000	2	<2150000	3
<160000	3	<660000	0	<1160000	1	<1660000	2	<2160000	1
<170000	1	<670000	3	<1170000	2	<1670000	0	<2170000	2
<180000	2	<680000	3	<1180000	3	<1680000	3	<2180000	2
<190000	2	<690000	4	<1190000	1	<1690000	1	<2190000	0
<200000	3	<700000	2	<1200000	3	<1700000	4	<2200000	0
<210000	3	<710000	1	<1210000	1	<1710000	1	<2210000	1
<220000	1	<720000	0	<1220000	2	<1720000	1	<2220000	4
<230000	3	<730000	2	<1230000	2	<1730000	1	<2230000	4
<240000	2	<740000	1	<1240000	0	<1740000	1	<2240000	1
<250000	0	<750000	0	<1250000	1	<1750000	1	<2250000	1
<260000	0	<760000	4	<1260000	0	<1760000	2	<2260000	2
<270000	2	<770000	2	<1270000	1	<1770000	3	<2270000	5
<280000	1	<780000	2	<1280000	1	<1780000	71	<2280000	1
<290000	3	<790000	1	<1290000	1	<1790000	8	<2290000	5
<300000	1	<800000	1	<1300000	2	<1800000	4	<2300000	0
<310000	1	<810000	0	<1310000	1	<1810000	3	<2310000	58
<320000	3	<820000	2	<1320000	3	<1820000	2	<2320000	1
<330000	1	<830000	1	<1330000	4	<1830000	0	<2330000	1
<340000	**215**	<840000	14	<1340000	0	<1840000	2	<2340000	3
<350000	2	<850000	2	<1350000	7	<1850000	2	<2350000	5
<360000	1	<860000	2	<1360000	1	<1860000	2	<2360000	0
<370000	0	<870000	3	<1370000	1	<1870000	2	<2370000	1
<380000	1	<880000	1	<1380000	3	<1880000	0	<2380000	2
<390000	3	<890000	5	<1390000	2	<1890000	1	<2390000	2
<400000	1	<900000	2	<1400000	4	<1900000	2	<2400000	2
<410000	5	<910000	4	<1410000	2	<1910000	2	<2410000	1
<420000	3	<920000	1	<1420000	3	<1920000	1	<2420000	3
<430000	1	<930000	4	<1430000	1	<1930000	1	<2430000	3
<440000	2	<940000	0	<1440000	0	<1940000	2	<2440000	1
<450000	1	<950000	5	<1450000	4	<1950000	5	<2450000	5
<460000	2	<960000	1	<1460000	2	<1960000	2	<2460000	0
<470000	9	<970000	0	<1470000	0	<1970000	4	<2470000	2
<480000	4	<980000	1	<1480000	2	<1980000	1	<2480000	0
<490000	2	<990000	1	<1490000	6	<1990000	7	<2490000	2
<500000	1	<1000000	3	<1500000	1	<2000000	**134**	<2500000	1

Table 4.3, continued.

Position	SNP density	Position	SNP density	Position	SNP density	Position	SNP density
<2510000	4	<3010000	3	<3510000	1	<4010000	1
<2520000	1	<3020000	1	<3520000	2	<4020000	1
<2530000	2	<3030000	1	<3530000	0	<4030000	2
<2540000	1	<3040000	0	<3540000	2	<4040000	3
<2550000	0	<3050000	0	<3550000	2	<4050000	0
<2560000	1	<3060000	1	<3560000	1	<4060000	2
<2570000	1	<3070000	3	<3570000	1	<4070000	5
<2580000	2	<3080000	2	<3580000	0	<4080000	3
<2590000	2	<3090000	2	<3590000	3	<4090000	1
<2600000	1	<3100000	0	<3600000	3	<4100000	2
<2610000	1	<3110000	4	<3610000	1	<4110000	1
<2620000	1	<3120000	2	<3620000	2	<4120000	1
<2630000	1	<3130000	2	<3630000	3	<4130000	0
<2640000	6	<3140000	4	<3640000	2	<4140000	2
<2650000	1	<3150000	0	<3650000	2	<4150000	1
<2660000	2	<3160000	0	<3660000	0	<4160000	4
<2670000	2	<3170000	3	<3670000	0	<4170000	3
<2680000	1	<3180000	3	<3680000	0	<4180000	2
<2690000	1	<3190000	1	<3690000	0	<4190000	3
<2700000	1	<3200000	1	<3700000	1	<4200000	0
<2710000	0	<3210000	0	<3710000	1	<4210000	1
<2720000	2	<3220000	1	<3720000	3	<4220000	1
<2730000	0	<3230000	4	<3730000	1	<4230000	4
<2740000	2	<3240000	3	<3740000	**269**	<4240000	0
<2750000	2	<3250000	15	<3750000	0	<4250000	2
<2760000	2	<3260000	1	<3760000	0	<4260000	3
<2770000	2	<3270000	3	<3770000	0	<4270000	2
<2780000	1	<3280000	2	<3780000	1	<4280000	0
<2790000	2	<3290000	0	<3790000	0	<4290000	0
<2800000	1	<3300000	1	<3800000	2	<4300000	2
<2810000	4	<3310000	1	<3810000	0	<4310000	3
<2820000	3	<3320000	1	<3820000	1	<4320000	2
<2830000	6	<3330000	0	<3830000	4	<4330000	2
<2840000	1	<3340000	3	<3840000	3	<4340000	1
<2850000	0	<3350000	0	<3850000	1	<4350000	1
<2860000	1	<3360000	3	<3860000	4	<4360000	3
<2870000	2	<3370000	2	<3870000	1	<4370000	1
<2880000	1	<3380000	0	<3880000	2	<4380000	4
<2890000	5	<3390000	0	<3890000	8	<4390000	3
<2900000	4	<3400000	0	<3900000	5	<4400000	3
<2910000	0	<3410000	1	<3910000	1	<4410000	2
<2920000	3	<3420000	2	<3920000	0	<4420000	0
<2930000	3	<3430000	6	<3930000	1		
<2940000	2	<3440000	0	<3940000	0		
<2950000	3	<3450000	3	<3950000	0		
<2960000	1	<3460000	2	<3960000	4		
<2970000	3	<3470000	2	<3970000	0		
<2980000	3	<3480000	0	<3980000	1		
<2990000	1	<3490000	2	<3990000	2		
<3000000	3	<3500000	0	<4000000	1		

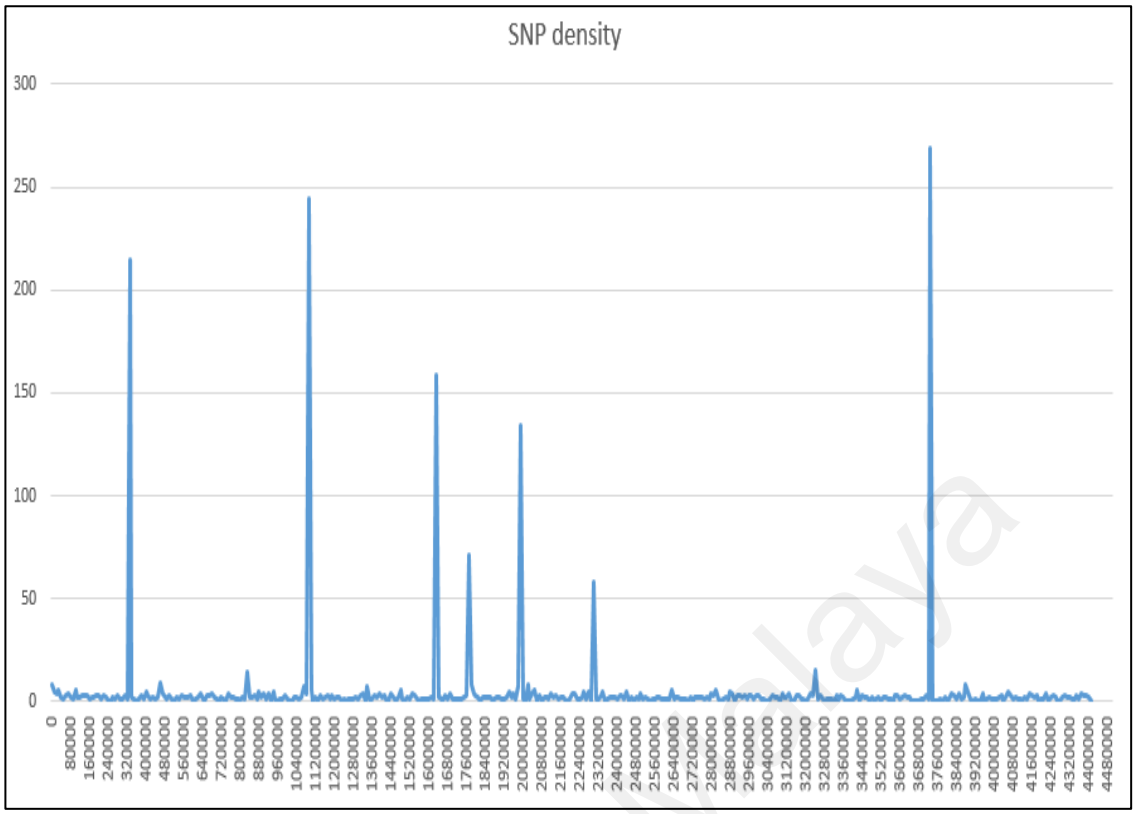


Figure 4.5: SNP density plotted according to the base pair position

Table 4.4: Notable features according to SNP density region. The features were extracted from list of annotated genes, according to the position

High SNP density region	Notable features	Chromosomal position	Length
340000-349999	PPE family protein	339247-340974bp	1728bp
1100000-1199999	DNA repair protein RecN	1101464-1103227bp	1764bp
	PPE family protein	1114435-1113278bp	1158bp
1780000-1789999	PPE family protein	1794692-1790493bp	4200bp
2000000-2009999	PE family protein	2012150-2010462bp	1689bp
3740000-3749999	PE-PGRS family protein	3740617-3742131bp	1515bp

From the SNP density distribution that has been sorted according to the position, it is observed that the SNP distribution is not random. There is some significance in diversifying selections in some of the genes. As been studied in past research, the genes such as *mutT* structure and mechanism has a notable build-up of SNPs in excess

(Ebrahimi-Rad et al., 2003). Most of the genes that are located at the high density position are relatable to host-pathogen interaction can thus became the source of virulence aspects uniquely for the each strain. The highest SNP build-up recorded is at position 3740617-3742131 where the recorded length was 1515 bp. The notable protein found in this region is PE-PGRS family protein. The GC content predicted is 65.29% which confirms the pathogenicity.

Past study by Saw et al. (2016) states that SNPs and rearrangement that includes indel in their cerebrospinal fluid (CSF) sample, does not presented in pulmonary *Mycobacterium tuberculosis* genome. They also include that PE/PPE genes, transcriptional and membrane proteins were found abundant in the rearranged regions. The finding was comparable to CNS strains which the genes are also found to code for PE/PPE proteins.

Lateral gene transfer can also contribute to high SNP density at some region that could be one of the causes for some inconsistence in the distribution. This is due to the deficiency of virulence information that can be differentiated by each strain. On the consistent part, the result from past studies has found to be related.

A study by Saw et al. (2016) also reports that there is abundant of gene that encode for PE/PPE, transcriptional and membrane protein is observed in the rearranged segments. The non-synonymous SNPs of the CSF strains were commonly found to have PE/PPE coding genes (Saw et al., 2016).

4.5 Insertion and Deletion Analysis

Similar to SNP analysis, insertion and deletion analysis has been done after genome alignment of both *Mycobacterium tuberculosis* strain TKK_01_0158 and H37Rv by using a specific PERL script. Each insertion and deletion of base is also classified according to the base pair positions. The complete raw data output has been inserted in the Appendix D.

The full list of insertions and deletions that has been obtained is sorted according to their genomic positions in base pairs. In order to make the visualised data to be simpler and compact, both insertions and deletions are grouped to 40 000 bp intervals. Any grouped base pairs regions that record the occurrence of both insertions and deletions above 1000 occurrence; they are marked as high density region, and is shown as peaks in Figure 4.6.

As what have been done in SNPs analysis, each of the high peak positions are studied to track the kind of genes of interest that resides in them. The list of notable features is presented in Table 4.6.

In the study by Saw et al. (2016) strains that are affected are found to induce deletions that has resulted truncation of gene, one of which was found to be PPE57. The gene is known to be associated in lipid metabolism and degradation, also the oxidation of fatty acid (Saw et al., 2016).

Table 4.5: Insertion and deletion according to its position. Bolded data indicates high insertion and deletion occurrence

Position	Insertion	Deletion	Position	Insertion	Deletion	Position	Insertion	Deletion
<40000	4	88	<1560000	0	0	<3080000	4	66
<80000	9	0	<1600000	0	10	<3120000	0	1
<120000	4	537	<1640000	5	71	<3160000	3	0
<160000	5	166	<1680000	**1321**	160	<3200000	388	3
<200000	0	0	<1720000	65	4	<3240000	9	282
<240000	75	1	<1760000	0	0	<3280000	0	1
<280000	0	77	<1800000	0	0	<3320000	0	0
<320000	1	0	<1840000	0	1	<3360000	60	3
<360000	**1362**	5	<1880000	12	40	<3400000	0	0
<400000	1	0	<1920000	7	61	<3440000	0	0
<440000	64	28	<1960000	66	0	<3480000	9	0
<480000	3	30	<2000000	**1362**	1	<3520000	**1364**	**5003**
<520000	1	0	<2040000	543	893	<3560000	0	0
<560000	0	0	<2080000	**2610**	65	<3600000	60	4
<600000	0	1	<2120000	0	0	<3640000	6	1
<640000	0	0	<2160000	0	0	<3680000	295	152
<680000	**2722**	6	<2200000	28	4	<3720000	**1361**	3
<720000	0	**203**	<2240000	45	0	<3760000	105	55
<760000	**1374**	0	<2280000	**4933**	0	<3800000	0	0
<800000	0	0	<2320000	121	467	<3840000	0	0
<840000	**1453**	88	<2360000	195	175	<3880000	0	0
<880000	0	2	<2400000	6	61	<3920000	0	0
<920000	**3514**	0	<2440000	1	1	<3960000	3	0
<960000	0	2	<2480000	0	0	<4000000	0	0
<1000000	1	0	<2520000	0	0	<4040000	0	0
<1040000	1	0	<2560000	**1365**	4	<4080000	373	0
<1080000	**3072**	**5999**	<2600000	136	4	<4120000	62	0
<1120000	5	34	<2640000	0	0	<4160000	61	3
<1160000	1	0	<2680000	4	28	<4200000	0	3
<1200000	0	0	<2720000	5	157	<4240000	**1369**	115
<1240000	0	0	<2760000	135	5	<4280000	0	0
<1280000	1	0	<2800000	79	60	<4320000	0	0
<1320000	4	395	<2840000	2	104	<4360000	176	1
<1360000	0	0	<2880000	0	0	<4400000	0	120
<1400000	1	0	<2920000	0	0	<4440000	0	0
<1440000	84	7	<2960000	**1361**	3			
<1480000	0	0	<3000000	0	0			
<1520000	0	0	<3040000	0	1			

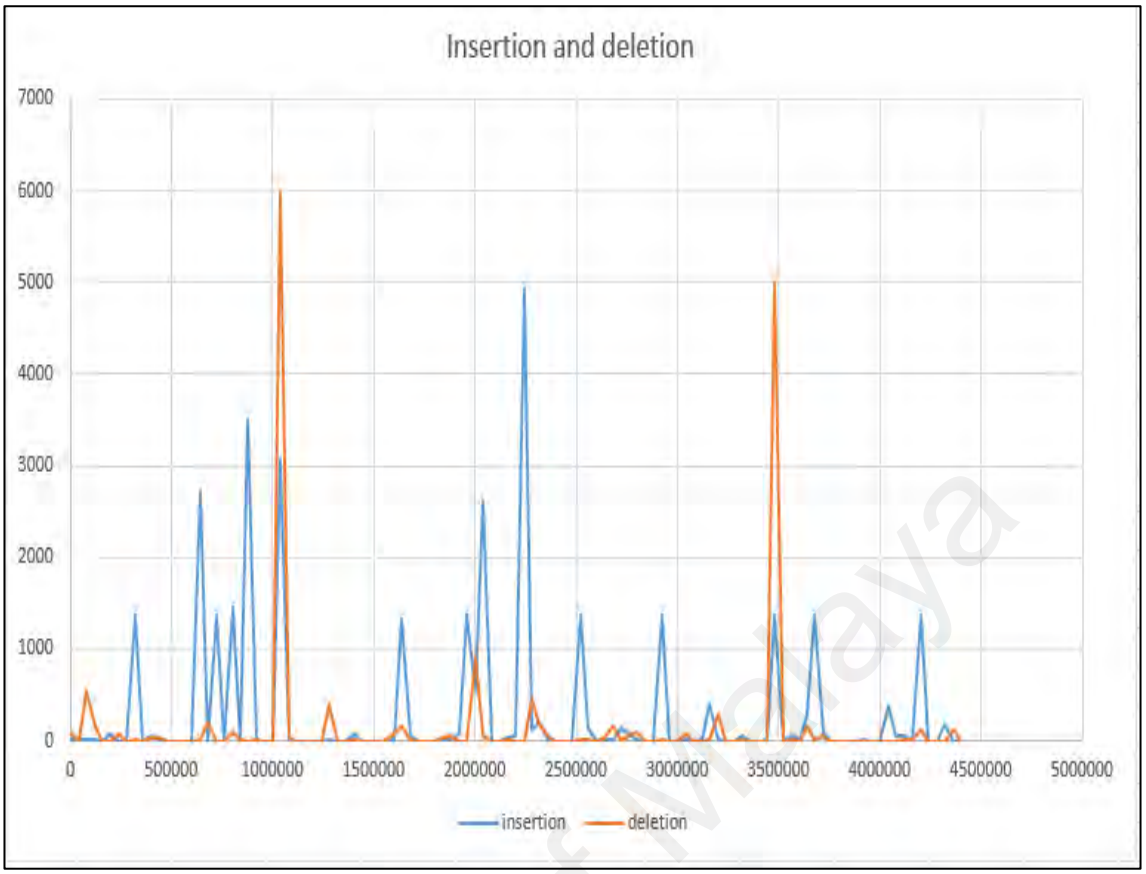


Figure 4.6: Insertion (blue lines) and deletion (orange line) plotted according to the base pair positions

From the analysis, it is observed that there are more insertions of bases recorded compared to deletions. As shown in Figure 4.6, the regions which records high insertions are 320000-360000 bp (1362 bp), 640000-680000 bp (2722 bp), 720000-760000 bp (1374 bp), 800000-840000 bp (1453 bp), 880000-920000 bp (3514 bp), 1040000-1080000 bp (3072 bp), 1640000-1680000 bp (1321 bp), 1960000-2000000 bp (1362 bp), 2040000-2080000 bp (2610 bp) , 2240000-2280000 bp (4933 bp), 2520000-2560000 bp (1365 bp), 2920000-2960000 bp (1361 bp), 3480000-3520000 bp (1364 bp), 3680000-3720000bp (1361 bp) and 4200000-4240000 bp (1369 bp).

Although the deletions do not occur in various regions, they show a significantly high occurrence at 2 positions, at 1040000-1080000 bp (5999 bp) and 3480000-3520000 bp (5003 bp). This has makes them to form the highest peaks among others.

Table 4.6: Notable features according to indel density region. The features were extracted from list of annotated genes, according to the position

High indel density region	Notable features	Position	Length
440000-479999	*pks*2	446381-44046bp	1728bp
840000-879999	Possible mutator protein *mutT3*	862934-863530bp	596bp
1080000-1119999	PE family protein	1037535-1038467	932bp
	Polyketide synthase	1069797-1072850	3053bp
	DNA-3-methyldeline glycosylase II	1094760-1095371	812bp
	DNA repair RecN	1101464-1103227	1764bp
	Polyketide synthase *pks7*	1063475-1063455	4809bp
	PE-PGRS family protein	1047153-1044118	3036bp
	Polyketide synthase *pks7*	1057075-1063455	6381bp
	PPE family protein	1115659-1114475	1185bp
1680000-1719999	PPE gene clusters	1678834-1680240	1407bp
	NADH dehydrogenase	1682911-1684620	1709bp
	PE-PGRS family protein	1690217-1688730	1487bp
	PE family protein	1717014-1715467	
2000000-2039999	PE family protein	2012150-2010462	1689bp
2080000-2119999	PPE gene clusters	2113546-2114766	1221bp
2280000-2319999	PPE gene clusters	2578803-2577628	1176bp
	PE-PGRS family protein	2571082-2570690	393bp
2560000-2599999	PPE gene clusters	2588486-2578803	1143bp
2960000-2999999	PE family protein	2960944-2962521	1578bp
	PE family protein	2986015-2985188	828bp
	PPE gene clusters	2985066-2983924	1143bp
	PPE gene clusters	2987445-2986339	1107bp
3520000-3559999	PPE gene clusters	3524062-3524247	186bp
	PE-PGRS family protein	3529727-3528957	771bp
3720000-3759999	Probable *pks* associate protein	3725192-3719991	5209bp
	PE-PGRS family protein	3738802-3740463	1661bp
	pks2	3763558-3757271	6288bp
	PPE gene clusters	3751373-375198	609bp
	Possible conserved *pks* associate protein PapA2	3749679-3748465	1215bp
	pks associate protein Pap1	3757097-3755685	1413bp
4240000-4279999	PE-PGRS family protein	4241068-4238420	2649bp

Insertion and deletion are the main cause for variability of pathogens activity. Pathogens such as *Mycobacterium tuberculosis* gains benefit for their virulent action during infection and transmission of the host. Possible candidates for resistance and virulence related component can be specified by investigating the indel loci. This has been proven by the deletion of *pks* gene, which has been studied to be homologous to mycocerosic acid synthase whereby the product of the gene will influence the secretion of multimethylated branched lipids. From the past research (Reed et al., 2004), the mutant of the gene will show intense defects in growth of the bacteria.

Another common set of protein that has been discussed in many studies before (Brosch et al., 2000); (Banu et al., 2002); PE and PPE-PGRS family protein has been found in position 1047153, 1690217, 20125150, 2560000, 3524062, 3738802 and 4241068. These are the position where the high peaks of indels are observed as seen in Figure 4.6. In the aspect of positions, few positions have found to be notable due to presence of high numbers important genes that are interconnected to host pathogenicity. Position 1080000 to 1120000 has elevated occurrence of both insertion and deletion unlike other regions where only either one is high. This particular position has recorded the manifestation of virulent affiliated genes which are PE, PE-PGRS and PPE family, polyketide synthase (*pks*) genes, DNA-3-methyladeline glycosylase and DNA-repair ReCN. Another region, position 1680000, exhibit high insertion where the occupancy of PPE, PE-PGRS and PPE family is significant with addition of NADH dehydrogenase. Position 840000 shows where gene *mutT* resides and position 360000 and 3720000 contains *pks* genes.

The presence of these genes that arisen from indel has proven them to be the main cause of the antigenic variability of *Mycobacterium tuberculosis*. Therefore the result of the protein found in the indel loci is comparable and consistent to past research.

4.6 CRISPR Locus Identification

A web based tool called CRISPRfinder at http://crispr.i2bc.paris-saclay.fr/Server/ has been used to identify any CRISPR locus presented in both *Mycobacterium tuberculosis* TKK_04_0158 and H37Rv. FASTA formatted file is used as an input which were uploaded to the website. The analysed data comprises the list of confirmed and suspected CRISPR.

This analysis was conducted in order to obtain the information that are available in the CRISPR loci. The same analysis was conducted in past research as they presumably affect the plasticity of *Mycobacterium tuberculosis* genomes (Liu et al., 2014).

From Table 4.7 to Table 4.10 of CRISPR display, the highlighted bases sequence in yellow represent the repeats, they are all arranged on the left side vertically to indicate the total number of direct repeat consensus. While the bases sequences on the right side which are highlighted in multiple colours represents the spacers that follows the repeating consensus. All CRISPR loci sequences are marked with base pair position at the beginning and the end.

In contrast to *Mycobacterium tuberculosis* strain H37Rv which contain two confirmed CRISPR loci, strain TKK_04_0158 was predicted to contain three CRISPR loci.

Table 4.7: List of Confirmed TKK_04_0158 (Meningeal) CRISPR

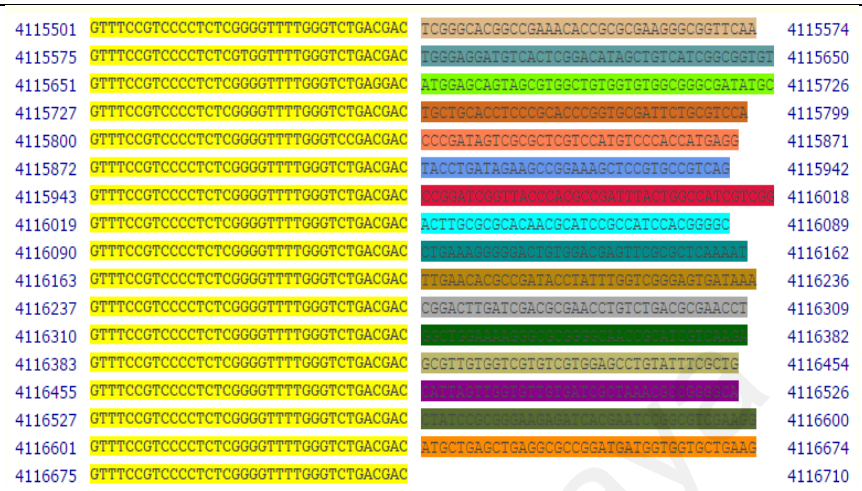
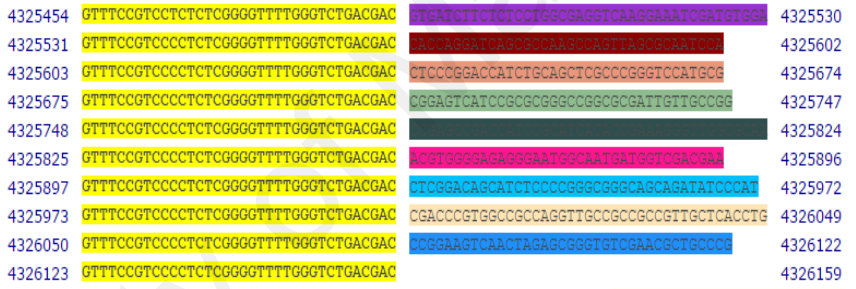
CRISPR details	CRISPR display
Start position: 4115501 End position: 4116710 Length: 1209 Direct repeat length: 36 Spacers: 16	 Direct repeat consensus: GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC
Start position: 4325454 End position: 4326159 Length: 705 Direct repeat length: 36 Spacers: 9	 Direct repeat consensus: GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC
Start position: 4326454 End position: 4328021 Length: 1567 Direct repeat length: 36 Spacers: 21	 Direct repeat consensus: GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC

Table 4.8: List of Questionable TKK_04_0158 (Meningeal) CRISPR

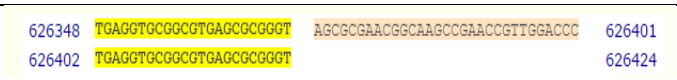
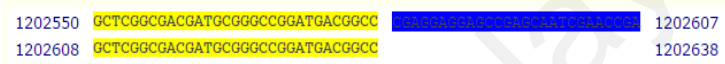
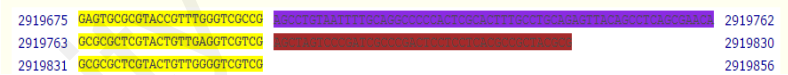
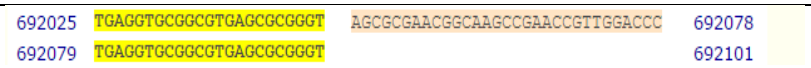
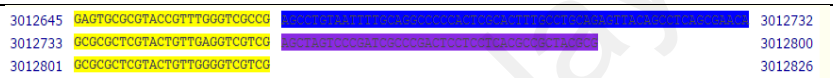
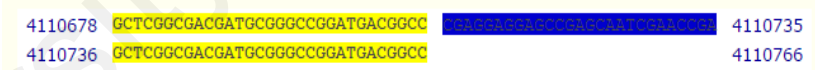
CRISPR details	CRISPR display
Start position: 626348 End position: 626424 Length: 76 Direct repeat length: 23 Spacers: 1	 Direct repeat consensus: TGAGGTGCGGCGTGAGCGCGGG
Start position: 1202550 End position: 1202638 Length: 88 Direct repeat length: 31 Spacers: 1	 Direct repeat consensus: GCTCGGCGACGATGCGGGCCGGATGACGGCC
Start position: 2919675 End position: 2919856 Length: 181 Direct repeat length: 26 Spacers: 2	 Direct repeat consensus: GCGCGCTCGTACTGTTGAGGTCGTCG

Table 4.9: List of Confirmed H37Rv (Lungs) CRISPR

CRISPR details	CRISPR display
Start position: 3119185 End position: 3120468 Length: 1283 Direct repeat length: 36 Spacers: 17	 Direct repeat consensus: GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC
Start position: 3121862 End position: 3123576 Length: 1714 Direct repeat length: 36 Spacers: 23	 Direct repeat consensus: GTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC

Table 4.10: List of Questionable H37Rv (Lungs) CRISPR

CRISPR details	CRISPR display
Start position: 692025 End position: 692101 Length: 76 Direct repeat length: 23 Spacers: 1	 Direct repeat consensus: TGAGGTGCGGCGTGAGCGCGGGT
Start position: 3012645 End position: 3012826 Length: 181 Direct repeat length: 26 Spacers: 2	 Direct repeat consensus: GCGCGCTCGTACTGTTGAGGTCGTCG
Start position: 4110678 End position: 4110766 Length: 88 Direct repeat length: 31 Spacers: 1	 Direct repeat consensus: GCTCGGCGACGATGCGGGCCGGATGACGGCC

The direct repeat regions that have been predicted in the *Mycobacterium tuberculosis* strain TKK_04_0158 has the length of 36 bp for all three confirmed CRISPR. With total 9 and 16 spacers in between, all three CRISPR region starts at the position between 4115501 bp and ends at 4326159 bp. In comparison, strain H37Rv has two CRISPR

repeat region with the same recorded length of 36 bp. The CRISPR region of strain H37Rv in contrast was found located at 3119185 bp of the chromosome.

The CRISPR region that includes the length of the spacer sequence of both *Mycobacterium tuberculosis* strain H37Rv is concluded similar to strain TKK_04_0158. From that note, it can be a strong evidence of their evolutionary relevance and hence predict that CRISPR region in *Mycobacterium tuberculosis* is conserved. Similarly compared to past study, the CRISPR founded in the chromosome has the attributes to interfere the resistance of the host immunity and thus are able to add repairing properties to aid in resistance (Babu et al., 2011). Nevertheless, the CRISPR protein founded are not a strong evidence to properly claim that they majorly contribute in pathogenicity, as no virulent CRISPR protein such as cas9 was found in this investigation.

It is rational to conclude that there are only few CRISPR which contains little spacers in contrast to other bacteria that are not pathogenic.

4.7 Significant Genes Identifications

Upon completing gene annotation, the web-based tool RAST was used execute a gene comparison analysis. The comparison tool analysed both *Mycobacterium tuberculosis* strains and provides a list of genes that each of the bacteria strains are unique to each other, which will give a detailed comparison on a notable genes that can be significant. Table 4.11 below shows the listed notable genes.

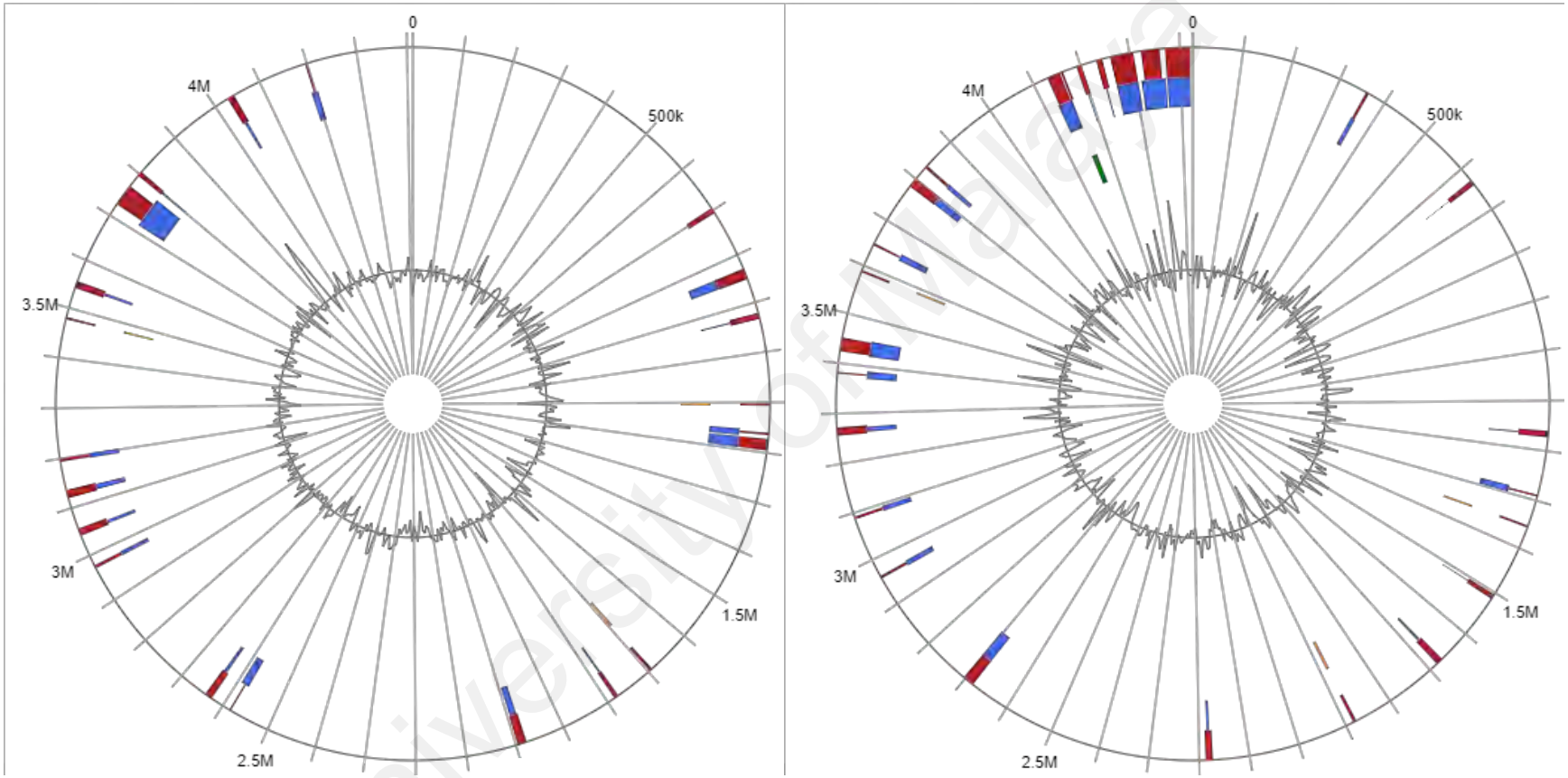
Table 4.11: List of Notable Genes Found.

No	TKK_04_0158 (brain)	H37Rv (Lungs)	Category	Role
1	✓	✘	Clustering-based system	tRNA uridine 5-carboxymethylaminomethyl modification enzyme GidA
2	✓	✘	Protein Metabolism	Cys-tRNA(Pro) deacylase YbaK
3	✓	✘	Stress Response	transcriptional regulator, Crp/Fnr family
4	✘	✓	Amino Acids and Derivatives	Kynurenine formamidase, bacterial (EC 3.5.1.9)
5	✘	✓	Clustering-based subsystems	Transcription regulator in CO-DH cluster
6	✘	✓	Phages, Prophages, Transposable elements, Plasmids	DNA primase/helicase, phage-associated

Some of the genes that are highlighted are found to be related to the virulence of the bacteria, but not much is associated with *Mycobacterium tuberculosis*. In a study conducted in *Streptococcus* pyogenes, it is found that the mutation in GidA will weaken the translation effectiveness. This will then lower the level of virulence factors (Cho & Caparon, 2008). Although H37Rv strain in this study was unable to detect any Ybak gene that is only found exclusive to the meningeal strain, other research has found this gene to be discovered in H37Rv. From there, Ybak is known to be related to defence against low oxygen stress, starvation and extreme temperature conditions (Akhter et al., 2007).

4.8 Genomic Island Identification

To identify a genomic island, a web-based island viewer tool is used, from http://www.pathogenomics.sfu.ca/islandviewer. The input file used for this tool is in genbank (.gbk) file. The output of the analysis is a visualisation in circular image representing the whole chromosome of the bacteria, and where the genomic island lies in the chromosomal position. The Figure 4.7 shows the genomic island for both of the studied strains (Bertelli et al., 2017).



H37Rv TKK_04_0158

Figure 4.7: The genomic island for *Mycobacterium tuberculosis* strain H37Rv (lungs) compared to strain TKK_04_0158 (meningeal), projected by IslandViewer. The colored bands inside the circular images indicates the predicted genomic islands according to their position

The tool projected the genomic island by using IslandPath-DIMOB(blue band), SIGI-HMM (orange band), IslandPick(green band) and integrated method (red band). There is no curated virulence factors recorded, as well as homologous virulence factors and curated resistance factors. For strain H37Rv, there are total of 21 predicted possible genomic islands. The largest island that has been discovered for the strain has a size of 43,917 bp, while the smallest size recorded is at 2,565. The predicted total genomes for strain TKK_04_0158 is slightly higher, which records for 26 genomic islands, with the largest size at 49,882 bp while the smallest at 3010 bp.

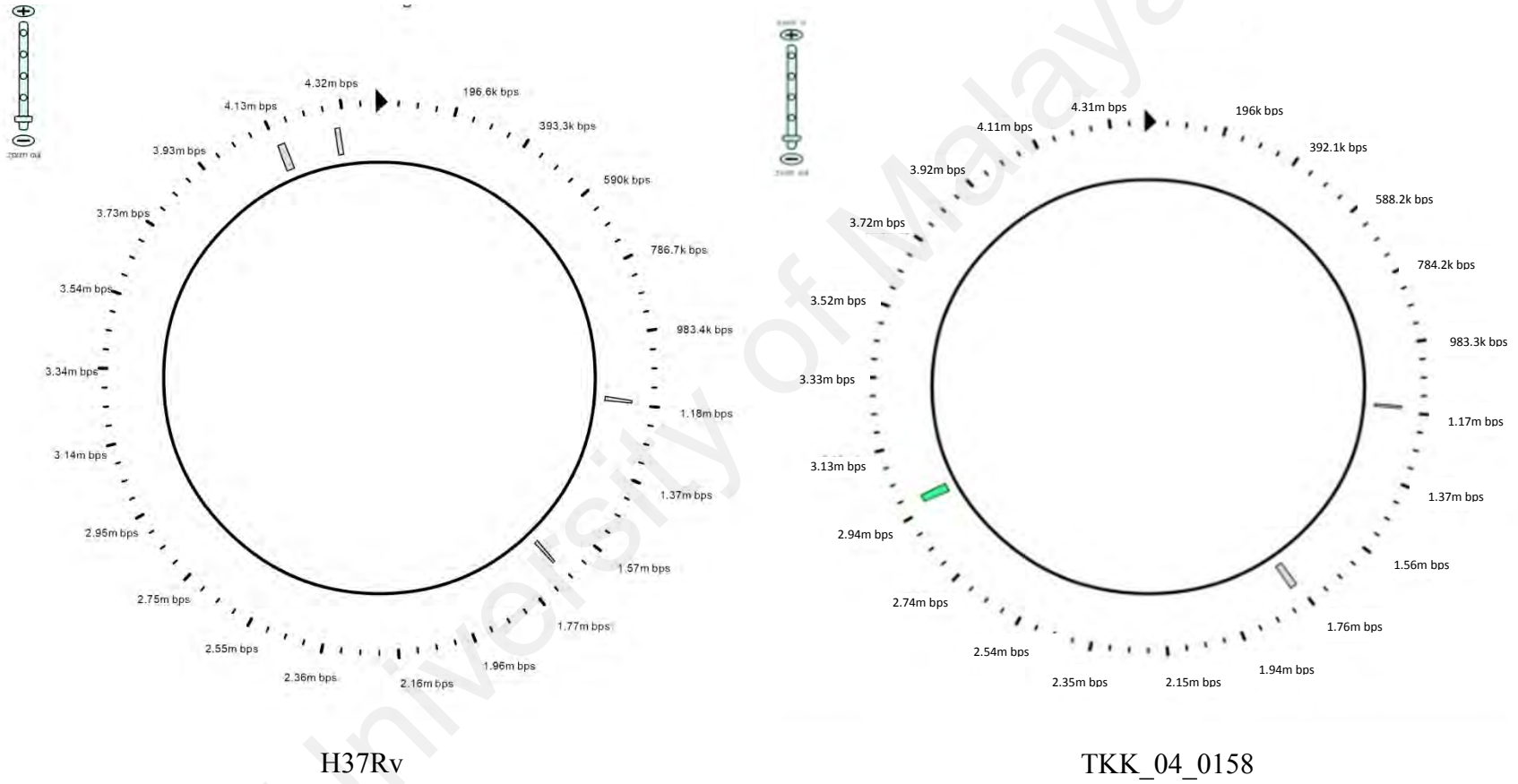
It is noted that there are accumulation of genomic islands at the end of the chromosome ring for strain TKK_04_0158. One of the regions comprises of CRISPR associated proteins, and therefore supports the earlier claim of CRISPR build up area executed by CRISPR finder. By executing the analysis of these genome island, it was found that the genes existed in these genomic island both strains are mainly encodes for PE_PGRS protein, mobile elements, probable phage protein (phiRv1 and phiRv2) and hypoethical proteins.

In a microarray profiling study that has been conducted by Kain et al., (2006) there are inflated level of up-regulation of 33 *Mycobacterium tuberculosis* genes in the course of early invasion of the bacteria to the blood brain barrier of the host. Among them, there are 18 genes known to belonged to genomic island (Rv0960-Rv1001). The mutant for the high regulation of genes Rv0980c, Rv0987, Rv0989c, and Rv1801 are discovered to be weak to able to invade the blood brain barrier model (Jain, Paul-Satyaseela, Lamichhane, Kim, & Bishai, 2006).

4.9 Pathogen Candidate Gene Detection

To identify the phage sequences from both *Mycobacterium tuberculosis* strands, a typing tool PHAST is used (Zhou, Liang, Lynch, Dennis, & Wishart, 2011). This tool is

a web-based addressed at phast.whshartlab.com. The input of this tool is in contigs in FASTA format. The output of this analysis is circular genome map that shows the locations of prophages within the genome and a summary table that indicates the location of prophage sequences. Figure 4.8 shows the circular genome map for the studied strains.



H37Rv TKK_04_0158

Figure 4.8: Genomic circular map that plots the location of both incomplete phage (grey band) and questionable phage (green band)

Table 4.12: Summary of Prophage of meningeal strain (TKK_04_0158)

Region	Region length	Completeness	Score	#CDS	Region position	GC %
1	12.8kb	incomplete	50	14	1163451-1176330	64.06
2	11.6kb	incomplete	50	12	1670002-1681699	67.36
3	25kb	incomplete	50	17	4130028-4155124	66.26
4	15.9kb	Incomplete	50	15	4298473-4314386	66.55

Table 4.13: Summary of Prophage of pulmonary (H37Rv)

Region	Region length	Completeness	Score	#CDS	Region position	GC %
1	7.5Kb	incomplete	50	13	1157963-1165490	65.25
2	21.1Kb	incomplete	50	18	1766987-1788115	66.09
3	26.3Kb	questionable	90	20	2957504-2983874	66.17

It is essential to verify the candidates as they are diverse mobile genetic elements that are formed through horizontal gene transfer. This analysis has predicted 4 candidates in total for *Mycobacterium tuberculosis* strain TKK_04_0158 (Table 4.12). From the total of genes predicted, none of them is intact. Due to incompleteness the region lengths are rather short, as intact predicted prophages ought to possess over 100 kb.

In comparison, *Mycobacterium tuberculosis* strain H37Rv has 2 incomplete prophage region and one questionable prediction (Table 4.13). The predicted prophage region for H37Rv has different position compared to the other strain.

The identification of prophage in this study is only considered to be as pathogen gene candidates. Moreover no intact prophage has been detected. This indicates that more research and better analysis needed to be executed in order to finalize the actual prophages that are confirmed to affect the host-pathogen interaction.

4.10 Summary of Findings

Table 4.14: Summary and comparison of the overall findings

	H37Rv (Pulmonary)	**TKK_04_0158 (Meningeal)**
Annotation	Rv0931c (1995bp) PE-PGRS (39 genes) PPE (66 genes) *Pks* *mutT* FASII (227 genes)	Rv0931c (1995bp) PE-PGRS (40 genes) PPE (71 genes) *Pks* *mutT* FASII (227 genes)
Visualisation	All bands are presented, similar GC plot and GC skew, GC%:65.6%	
Alignment	BRIG: Region with high level variability: CRISPR genes (cas1,cas2,csm2,csm5,csm6)	
SNPs	Highest SNP: 1515bp: PE-PGRS family	
Insertion and deletions	Highest Insertion: 4933bp: PPE gene clusters Highest deletion: 5999bp: PE, PE-PGRS, PPE, DNA repair, *pks*	
CRISPR	3 confirmed CRISPR loci • longest: 1567bp • highest spacers:9 • direct repeat: 36	2 confirmed CRISPR loci • longest: 1714bp • highest spacers: 23 • direct repeat: 36
Genome Island	21 genomic islands • largest:43917bp	26 genomic islands • largest: 49882bp
Pathogen gene candidate	3 incomplete prophages	2 incomplete, 1 questionable prophages

As a general discussion on the results, it can be concluded the comparative genomics was executed successfully, and the objective is well achieved. Most of the outcome in each methods record high similarities. Most of gene of interest that contributes to pathogenicity exists in both strains, but in different sizes and residing at different locations (Table 4.14). Although there are few genes of interest such as the Rv0931c and *Pks* genes were found in the meningeal strain, they were also observed to exist in almost similar size in the pulmonary strain; with only difference is the residing location. Hence, more in depth studies need to be done to achieve this particular goal.

This outcome of this study suggests that studies like proteomics and transcriptomes can be done for the mentioned genes in this chapter. Therefore gene adaptions of human host particularly in blood brain barrier can be further understood. Multiple strains can be compared in the same time, not necessarily pair-wise alignment, as what have been done in this research that therefore can produce a better quality results. Metabolic pathway studies can be conducted by doing KEGG pathway analysis, and each protein can be studied in depth to understand their relationships with any virulent activity.

CHAPTER 5: CONCLUSION

Tuberculosis is an infection caused by *Mycobacterium* tuberculosis and is one of the most fatal diseases if the patient does not receive treatment. Tuberculosis are commonly known to affect the human lungs, resulting in continuous severe cough. But patients worldwide recorded another invasion of this bacteria to other part of human host body, for instance to the central nervous system. Meningeal tuberculosis recorded a high mortality upon the diseased patients, primarily among children. As this disease has arises concern in public health, more understanding is needed in the mechanism and the cellular properties of this bacteria, especially on the virulence and pathogenicity aspects. This study tackles the problem by focusing on the genomic components of *Mycobacterium tuberculosis* and conducting comparative genomic studies between the pulmonary strain and meningeal strain. Both sequences were obtained from NCBI database, pulmonary strain (H37Rv) was obtained in completed whole genome sequence while the meningeal strain (TKK_04_0158) was obtained in scaffolds. The outcome of this study finds that there is the presence heparin binding haemoglobin producing gene, Rv0931c in both strains, where there is a possible activation in the meningeal strain and none in lungs. This is due to the fact that this gene is highly affiliated with brain micro vascular activity, specifically at the blood brain barrier. It is also a precursor for *pknD* genes, one of an essential gene for the bacterial virulence activity in the central nervous system. Other set of genes and proteins that were found to be consequential to the pathogenicity of both strains are PE-PEGRS, PPE, *pks*, FASII and *mutT*. The second level of the study, which is the comparative analysis, generates a fine genomic data and visualisation. SNP and indel graph of high density peaks indicates all the possible regions where genes of interest reside. As an additional procedure compared to past research on comparative genomics, CRISPR study has been implemented and the result is compatible with the visualised alignment comparison. To

simplify, this study has found similarities between strains and not much significance in the difference. To utilize the data and hence achieving the final goal of finding the novel gene that marks the pathogenicity of meningeal *Mycobacterium tuberculosis*, more research can be done to study the differences in depth, by performing proteomics and metabolisms of each genes of interest. As a final conclusion, this research has completed a successful comparative genomic blueprint, where the similarities and differences observed has been studied and beneficial for further investigation.

REFERENCES

Akhter, Y., Tundup, S., & Hasnain, S. E. (2007). Novel biochemical properties of a CRP / FNR family transcription factor from *Mycobacterium tuberculosis*, *International Journal of Medical Microbiology 297*(6), 451–457.

Akira, S., Uematsu, S., & Takeuchi, O. (2006). Pathogen recognition and innate immunity. *Cell*, *124*(4), 783–801.

Aldridge, R. W., Nellums, L. B., Bartlett, S., Barr, A. L., Patel, P., Burns, R., … Abubakar, I. (2018). Articles global patterns of mortality in international migrants : a systematic review and meta-analysis. *The Lancet*, *392*(10164), 2553-2566.

Alikhan, N. F., Petty, N. K., Ben Zakour, N. L., & Beatson, S. A. (2011). BLAST ring image generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics*, *12*(6), 402.

Alland, D., Kramnik, I., Weisbrod, T. R., Otsubo, L., Cerny, R., Miller, L. P., … Bloom, B. R. (1998). Identification of differentially expressed mRNA in prokaryotic organisms by customized amplification libraries (DECAL): the effect of isoniazid on gene expression in *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(22), 13227–13232.

Babu, M., Beloglazova, N., Flick, R., Graham, C., Skarina, T., Nocek, B., … Yakunin, A. F. (2011). A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. *Molecular Microbiology*, *79*(2), 484–502.

Banu, S., Honoré, N., Saint-Joanis, B., Philpott, D., Prévost, M. C., & Cole, S. T. (2002). Are the PE-PGRS proteins of *Mycobacterium tuberculosis* variable surface antigens? *Molecular Microbiology*, *44*(1), 9–19.

Barry, C. E. (2001). Interpreting cell wall "virulence factors" of *Mycobacterium tuberculosis*. *Trends in Microbiology*, *9*(5), 237–241.

Be, N. A., Bishai, W. R., & Jain, S. K. (2012). Role of *Mycobacterium tuberculosis* pknD in the pathogenesis of central nervous system tuberculosis. *BMC Microbiology*, *12*(1), 7.

Be, N. A., Kim, K. S., Bishai, W. R., & Jain, S. K. (2015). Pathogenesis of central nervous system tuberculosis. *Current Molecular Medicine, 33*(4), 395–401.

Bertelli, C., Laird, M. R., Williams, K. P., Lau, B. Y., Hoad, G., Winsor, G. L., & Brinkman, F. S. L. (2017). IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Research*, *45*(W1), W30–W35.

Bhatt, A., Fujiwara, N., Bhatt, K., Gurcha, S. S., Kremer, L., Chen, B., … Jacobs, W. R. (2007). Deletion of kasB in *Mycobacterium tuberculosis* causes loss of acid-fastness and subclinical latent tuberculosis in immunocompetent mice. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(12), 5157–5162.

Bowie, A., & O'Neill, L. A. (2000). The interleukin-1 receptor/toll-like receptor superfamily: signal generators for pro-inflammatory interleukins and microbial products. *Journal of Leukocyte Biology*, *67*(4), 508–514.

Brosch, R., Gordon, S. V, Pym, a, Eiglmeier, K., Garnier, T., & Cole, S. T. (2000). Comparative genomics of the mycobacteria. *International Journal of Medical Microbiology : IJMM*, *290*(2), 143–152.

Brudey, K., Driscoll, J. R., Rigouts, L., Prodinger, W. M., Gori, A., Al-Hajoj, S. A., ... & Binder, L. (2006). Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC microbiology*, *6*(1), 23.

Chan, J., Xing, Y., Magliozzo, R. S., & Bloom, B. R. (1992). Killing of virulent *Mycobacterium tuberculosis* by reactive nitrogen intermediates produced by activated murine macrophages. *Journal of Experimental Medicine, 175(4), 1111-1122.*

Chin, K. L., Sarmiento, M. E., & Acosta, A. (2018). DNA markers for tuberculosis diagnosis. *Tuberculosis*, *113*, 139-152.

Cho, K. H., & Caparon, M. G. (2008). tRNA modification by GidA / MnmE is necessary for *Streptococcus pyogenes* virulence : a new strategy to make live attenuated strains. *Infection and Immunity, 76*(7), 3176–3186.

Comas, I., Homolka, S., Niemann, S., & Gagneux, S. (2009). Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies. *PloS one*, *4*(11), e7815.

Cooper, a. M., & Flynn, J. L. (1995). The protective immune response to *Mycobacterium tuberculosis*. *Current Opinion in Immunology*, *7*(4), 512–516.

Darling, A. C. E., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve : multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, *14*(7), 1394-1403.

Darling, A. E., Mau, B., & Perna, N. T. (2010). Progressive mauve : multiple genome alignment with gene gain , loss and rearrangement. *PloS One, 5*(6), e11147.

Dubnau, E., Chan, J., Raynaud, C., Mohan, V. P., Lanéelle, M. A., Yu, K., … Daffé, M. (2000). Oxygenated mycolic acids are necessary for virulence of *Mycobacterium tuberculosis* in mice. *Molecular Microbiology*, *36*(3), 630–637.

Ebrahimi-Rad, M., Bifani, P., Martin, C., Kremer, K., Samper, S., Rauzier, J., … Gicquel, B. (2003). Mutations in putative mutator genes of *Mycobacterium tuberculosis* strains of the W-Beijing family. *Emerging Infectious Diseases*, *9*(7), 838–845.

Faksri, K., Prammananan, T., Leechawengwongs, M., & Chaiprasert, A. (2012). Molecular epidemiology and drug, resistance of tuberculous meningitis. *Meningitis*. IntechOpen. *4910*(8), 1-10.

Finer, K. R. (2003). Tuberculosis. *Deadly diseases and epidemic*. New York, NY: Infobase Publishing.

Flynn, J. L., & Ernst, J. D. (2000). Immune responses in tuberculosis. *Current Opinion in Immunology*, *12*(4), 432–436.

Flynn, J. L., Goldstein, M. M., Chan, J., Triebold, K. J., Pfeffer, K., Lowenstein, C. J., … Bloom, B. R. (1995). Tumor necrosis factor-α is required in the protective immune response against *Mycobacterium tuberculosis* in mice. *Immunity*, *2*(6), 561–572.

Ford, B. A. L., Foulcher, E., Lemckert, F. A., & Sedgwick, J. D. (1996). Microglia induce CD4 T lymphocyte final effector function and death. *Journal of Experimental Medicine*, *184*(5), 1737-1745.

Forrellad, M. A., Klepp, L. I., Gioffré, A., Sabio y García, J., Morbidoni, H. R., de la Paz Santangelo, M., … Bigi, F. (2013). Virulence factors of the *Mycobacterium tuberculosis* complex. *Virulence*, *4*(1), 3–66.

Galea, I., Bechmann, I., & Perry, V. H. (2007). What is immune privilege (not)?. *Trends in Immunology*, *28*(1), 12–18.

Giacomini, E., Iona, E., Ferroni, L., Miettinen, M., Fattorini, L., Orefici, G., … Coccia, E. M. (2001). Infection of human macrophages and dendritic cells with *Mycobacterium tuberculosis* induces a differential cytokine gene expression that modulates T cell response. *The Journal of Immunology*, 116(12), 7033-7041.

Graham, J. E., & Clark-Curtiss, J. E. (1999). Identification of *Mycobacterium tuberculosis* RNAs synthesized in response to phagocytosis by human macrophages by selective capture of transcribed sequences (SCOTS). *Proceedings of the National Academy of Sciences of the United States of America*, *96*(20), 11554–11549.

Grissa, I., Vergnaud, G., & Pourcel, C. (2008). CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research*, *36*(suppl_2), 52–57.

Hale, C. R., Zhao, P., Olson, S., Duff, M. O., Graveley, B. R., Wells, L., … Terns, M. P. (2009). RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell*, *139*(5), 945–956.

Hemmi, H., Takeuchi, O., Kawai, T., Kaisho, T., Sato, S., Sanjo, H., … Akira, S. (2000). A toll-like receptor recognizes bacterial DNA. *Nature*, *408*(6813), 740–745.

Hickey, W. F. (1999). Leukocyte traffic in the central nervous system: the participants and their roles. *Seminars in Immunology*, *11*(2), 125–137.

Hu, B., Xie, G., Lo, C. C., Starkenburg, S. R., & Chain, P. S. G. (2011). Pathogen comparative genomics in the next-generation sequencing era: genome alignments, pangenomics and metagenomics. *Briefings in Functional Genomics*, *10*(6), 322–333.

Huo, Y., Zhan, Y., Liu, G., & Wu, H. (2018). Tuberculosis meningitis: early diagnosis and treatment with clinical analysis of 180 patients. *Radiology of Infectious Diseases*, *6*(1), 21-25.

Jain, S. K., Paul-Satyaseela, M., Lamichhane, G., Kim, K. S., & Bishai, W. R. (2006). *Mycobacterium tuberculosis* invasion and traversal across an in vitro human blood-brain barrier as a pathogenic mechanism for central nervous system tuberculosis. *The Journal of Infectious Diseases*, *193*(9), 1287–1295.

Koch, A., & Mizrahi, V. (2018). *Mycobacterium tuberculosis*. *Trends in Microbiology*, *26*(6), 555–556.

Kuppusamy, I., Tan, M. H., Omar, A., Md. Zain, Z., Liam, C. K., Wong, W. K., … Balakrishnan. (2018). Guidelines on management of tuberculosis. Retrieved from http://www1.mts.org.my/EDUCATION/Guidelines/Management-of-Tuberculosis

Lee, H. G., William, T., Menon, J., Ralph, A. P., Ooi, E. E., Hou, Y., … Yeo, T. W. (2016). Tuberculous meningitis is a major cause of mortality and morbidity in adults with central nervous system infections in Kota Kinabalu, Sabah, Malaysia: An observational study. *BMC Infectious Diseases*, *16*(1), 1–8.

Liu, F., Hu, Y., Wang, Q., Li, H. M., Gao, G. F., Liu, C. H., & Zhu, B. (2014). Comparative genomic analysis of *Mycobacterium tuberculosis* clinical isolates. *BMC Genomics*, *15*(1), 469.

Lopez, B., Aguilar, D., Orozco, H., Burger, M., Espitia, C., Ritacco, V., … Soolingen, D. Van. (2003). A marked difference in pathogenesis and immune response induced by different. *Clinical and Experimental Immunology*, *133*(1), 30-37.

Machowski, E. E., Barichievy, S., Springer, B., Durbach, S. I., & Mizrahi, V. (2007). In vitro analysis of rates and spectra of mutations in a polymorphic region of the Rv0746 PE_PGRS gene of *Mycobacterium tuberculosis*. *Journal of Bacteriology*, *189*(5), 2190–2195.

Manca, C., Tsenova, L., Barry, C. E., Bergtold, A., Freeman, S., Haslett, P. A. J., … Kaplan, G. (1999). *Mycobacterium tuberculosis* CDC1551 induces a more vigorous host response in vivo and in vitro, but is not more virulent than other clinical isolates. *The Journal of Immunology*, *162*(11), 6740–6746.

Manca, C., Tsenova, L., Bergtold, A., Freeman, S., Tovey, M., Musser, J. M., … Kaplan, G. (2001). Virulence of a *Mycobacterium tuberculosis* clinical isolate in mice is determined by failure to induce Th1 type immunity and is associated with induction of IFN-alpha /beta. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(10), 5752–5757.

Mcgrath, M., Gey van pittius, N. C., Van helden, P. D., Warren, R. M., & Warner, D. F. (2014). Mutation rate and the emergence of drug resistance in *Mycobacterium tuberculosis*. *Journal of Antimicrobial Chemotherapy*, *69*(2), 292–302.

Meena, L. S., & Rajni, T. (2010). Survival mechanisms of pathogenic *Mycobacterium tuberculosis* H 37Rv. *FEBS Journal*, *277*(11), 2416–2427.

Mizrahi, V., & Andersen, S. J. (1998). DNA repair in *Mycobacterium tuberculosis*. What have we learnt from the genome sequence? *Molecular Microbiology*, *29*(6), 1331–1339.

Move to prevent immigrants with TB from entering Sabah. *The Star*. Retrieved 12, december, 2018, from https://www.thestar.com.my/news/nation/2018/11/15/move-to-prevent-immigrants-with-tb-from-entering-sabah/

Muzumdar, D., Vedantam, R., & Chandrashekhar, D. (2018). Tuberculosis of the central nervous system in children. *Child's Nervous System*, *34*(10) 1–11.

Palucci, I., Camassa, S., Cascioferro, A., Sali, M., Anoosheh, S., Zumbo, A., … Delogu, G. (2016). PE-PGRS33 contributes to *Mycobacterium tuberculosis* entry in macrophages through interaction with TLR2. *PLoS ONE*, *11*(3), 1–15.

Perez-Rodriguez, R., Haitjema, C., Huang, Q., Nam, K. H., Bernardis, S., Ke, A., & DeLisa, M. P. (2011). Envelope stress is a trigger of CRISPR RNA-mediated DNA silencing in *Escherichia coli*. *Molecular Microbiology*, *79*(3), 584–599.

Quesniaux, V., Fremond, C., Jacobs, M., Parida, S., Nicolle, D., Yeremeev, V., … Ryffel, B. (2004). Toll-like receptor pathways in the immune responses to mycobacteria. *Microbes and Infection*, *6*(10), 946–959.

Rahim, M. J. C., & Ghazali, W. S. W. (2016). Psychosis secondary to tuberculosis meningitis. *BMJ Case Reports*, *2016*, 1–3.

Ransohoff, R. M., Kivisakk, P., & Kidd, G. (2003). Three or more routes for leukocyte migration into the central nervous system. *Nature Reviews Immunology*, *3*(7), 569–581.

Rath, D., Amlinger, L., Rath, A., & Lundgren, M. (2015). The CRISPR-Cas immune system: Biology, mechanisms and applications. *Biochimie*, *117*, 119–128.

Reed, M. B., Domenech, P., Manca, C., Su, H., Barczak, A. K., Kreiswirth, B. N., … Barry, C. E. (2004). A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. *Nature*, *431*(7004), 84–87.

RiveraMarrero, C. A., Burroughs, M. A., Masse, R. A., Vannberg, F. O., Leimbach, D. L., Roman, J., & Murtagh, J. J. (1998). Identification of genes differentially expressed in *Mycobacterium tuberculosis* by differential display PCR. *Microbial Pathogenesis*, *25*(6), 307–316.

Saw, S. H., Tan, J. L., Chan, X. Y., Chan, K. G., & Ngeow, Y. F. (2016). Chromosomal rearrangements and protein globularity changes in *Mycobacterium tuberculosis* isolates from cerebrospinal fluid. *PeerJ*, *4*, e2484.

Singh, A., Vishwakarma, R. A., Narayanan, P. R., Paramasivan, C. N., Ramanathan, V. D., & Tyagi, A. K. (2005). Requirement of the mymA operon for appropriate cell wall ultrastructure and persistence of *Mycobacterium tuberculosis* in the spleens of guinea pigs. *Journal of Bacteriology*, *187*(12), 4173–4186.

Smith, I. (2003). *Mycobacterium tuberculosis* pathogenesis and molecular determinants of virulence. *Clinical Microbiology Reviews*, *16*(3), 463–496.

Springer, B., Sander, P., Sedlacek, L., Hardt, W.-D., Mizrahi, V., Schär, P., & Böttger, E. C. (2004). Lack of mismatch correction facilitates genome evolution in mycobacteria. *Molecular Microbiology*, *53*(6), 1601–1609.

Theus, S. A., Cave, M. D., & Eisenach, K. D. (2005). Intracellular macrophage growth rates and cytokine profiles of *Mycobacterium tuberculosis* strains with different transmission dynamics. *The Journal of Infectious Diseases*, 191(3), 453-460.

Thoma-Uszynski, S., Stenger, S., Takeuchi, O., Ochoa, M. T., Engele, M., Sieling, P. A., … Modlin, R. L. (2001). Induction of direct antimicrobial activity through mammalian toll-like receptors. *Science (New York, N.Y.)*, *291*(5508), 1544–1547.

Thwaites, G. (2017). Tuberculous meningitis. *Medicine (United Kingdom)*, *45*(11), 670–673.

Thwaites, G., Caws, M., Chau, T. T. H., D'Sa, A., Lan, N. T. N., Huyen, M. N. T., ... & Nhu, N. T. Q. (2008). Relationship between Mycobacterium tuberculosis genotype and the clinical phenotype of pulmonary and meningeal tuberculosis. *Journal of Clinical Microbiology*, *46*(4), 1363-1368.

Tsenova, L., Bergtold, A., Freedman, V. H., Young, R. A., & Kaplan, G. (1999). Tumor necrosis factor alpha is a determinant of pathogenesis and disease progression in mycobacterial infection in the central nervous system. *Proceedings of the National Academy of Sciences USA*, *96*(10), 5657–5662.

Tsenova, L., Ellison, E., Harbacheuski, R., Moreira, A. L., Kurepina, N., Reed, M. B., … Kaplan, G. (2005). Virulence of selected *Mycobacterium tuberculosis* clinical isolates in the rabbit model of meningitis is dependent on phenolic glycolipid produced by the bacilli. *The Journal of Infectious Diseases*, *192*(1), 98–106.

Tuberculosis the most prevalent disease among foreign workers. *The Star*. Retrieved 12, december, 2018, from https://www.thestar.com.my/news/nation/2014/02/13/tb-higherst-foreign-workers/

Varatharaj, A., & Galea, I. (2017). The blood-brain barrier in systemic inflammation. *Brain, Behavior, and Immunity*, *60*, 1–12.

World Health Organization. (2018). *Global tuberculosis report 2018*. Geneva: World Health Organization.

Wangoo, A., Sparer, T., Brown, I. N., Snewin, V. A., Janssen, R., Thole, J., … Young, D. B. (2001). Contribution of Th1 and Th2 Cells to Protection and Pathology in Experimental Models of Granulomatous Lung Disease. *The Journal of Immunology*, *166*(5), 3432–3439.

Wanner, R. M., Guethlein, C., Springer, B., Boettger, E. C., Ackermann, M., Cole, S., … Cebula, T. (2008). Stabilization of the genome of the mismatch repair deficient *Mycobacterium tuberculosis* by context-dependent codon choice. *BMC Genomics*, *9*(1), 249.

Wilkinson, R. J., Rohlwink, U., Misra, U. K., Crevel, R. Van, Thi, N., Mai, H., … Thwaites, G. E. (2017). Tuberculous meningitis. *Nature Publishing Group*, *13*(10), 581–598.

Wu, S., Howard, S. T., Lakey, D. L., Kipnis, A., Samten, B., Safi, H., … Barnes, P. F. (2004). The principal sigma factor sigA mediates enhanced growth of *Mycobacterium tuberculosis* in vivo. *Molecular Microbiology*, *51*(6), 1551–1562.

Zhang, M., Gong, J., Yang, Z., Samten, B., Cave, M. D., & Barnes, P. F. (1999). Enhanced capacity of a widespread strain of *Mycobacterium tuberculosis* to grow in human macrophages. *Journal of the Infectous Diseases*, *179*(5), 1213–1217.

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., & Wishart, D. S. (2011). PHAST: A Fast Phage Search Tool. *Nucleic Acids Research*, *39*(SUPPL. 2), 347–352.