MAPPING OF TRANSCRIPTIONAL START SITE AND THE IDENTIFICATION OF NOVEL sRNA IN *Mycobacterium tuberculosis* H37Rv

LEE LI PIN

UNIVERSITI SAINS MALAYSIA

2016

MAPPING OF TRANSCRIPTIONAL START SITE AND THE IDENTIFICATION OF NOVEL sRNA IN *Mycobacterium tuberculosis* H37Rv

by

LEE LI PIN

Thesis submitted in fulfillment of the requirements
for the degree of
Doctor of Philosophy

September 2016

ACKNOWLEDGEMENTS

First of all, I wish to express my sincere and deepest gratitude to my supervisor, Prof. Tang Thean Hock for his constant motivation, patience and support all these years. I am very lucky to have Prof. Tang as my mentor. His wide academic interests and enthusiasm has motivated me to take on challenges and work in a few small projects, where it was thrilling to learn new knowledge and lessons from situations.

Throughout my study, I received a lot of supports from the Institute of Experimental Pathology/Molecular Neurobiology, University of Muenster, Germany. To my field supervisor, Dr. Timofey Rozhdestvensky, thank you for your guidance and inspirational advices. To Dr. Carsten Raabe, your immense knowledge and insightful comments are very much appreciated. I am also grateful to Prof. Juergen Brosius for supporting my short attachment at the Institute of Experimental Pathology/Molecular Neurobiology.

My utmost appreciation would go to the Advanced Medical and Dental Institute (AMDI) and the Ministry of Education for providing financial assistance through the Graduate Research Assistant Scheme and the MyPhD scholarship, respectively. The financial support from the Institute of Postgraduate Studies, Universiti Sains Malaysia, for my research attachment at Germany is acknowledged. Besides, I also appreciate the Student Research Fund granted by AMDI for conducting my research project.

There are a few members from my RNA-Bio Research Group that need thanking: Dr. Hoe Chee Hock, for guiding me the basic techniques in RNA molecular work; Dr. Citartan Marimuthu, for helping me to review my research articles; Madam Siti Aminah Ahmed, for assisting me in the administrative works and plotting the growth curve of *Mycobacterium tuberculosis* H37Rv; Madam Ang Kai Cheen, for showing me the precautions in culturing the mycobacteria; and Mr. Cheah Hong Leong, for the inspiring ideas and opinions in bioinformatics analysis. I am really thankful to all my fellow group members, where I learned the importance of teamwork and interpersonal skills from them.

Last but not least, my most sincere appreciation dedicated to two very precious persons in my life - my father, Mr. Lee Chen Huat and my mother, Madam Doi Chiow Moi. I am really thankful for all the encouragement and unconditional supports given by them throughout my study. Their moral supports and love are my motivations to strive in works. Without the both of them, I wouldn't have reached this far.

This thesis is dedicated to my beloved parents who have supported me all the way since the beginning

TABLE OF CONTENTS

PAGE

LIST OF TABLES

PAGE

LIST OF FIGURES

PAGE

LIST OF SYMBOLS AND ABBREVIATIONS

%	:	percentage
°C	:	degree Celcius
(p)ppGpp	:	hyperphosphorylated guanine
µg/µl	:	microgram per microliter
µg/ml	:	microgram per milliliter
µg	:	microgram
µl	:	microliter
µm	:	micrometer
µM	:	microMolar
ABC	:	ATP-binding domain
APS	:	ammonium persulfate
asTSS	:	Antisense TSS
ATP	:	Adenosine triphosphate
BC	:	Before Christ
BCG	:	bacille Calmette-Guérin
bp	:	base pair
ca.	:	circa (approximate)
$CaCl_2$	:	calcium chloride
cDNA	:	complementary DNA
Ci	:	curie
DNA	:	deoxyribonucleic acid
Dnase	:	deoxyribonuclease
dNTP	:	dideoxynucleotide triphosphate
dRNA-seq	:	differential RNA sequencing
e.g.	:	exempli gratia (for example)
EDTA	:	ethylenediaminetetraacetic acid
Fe-	:	Mtb in iron stress, TEX-
Fe+	:	Mtb in iron stress, TEX+
$FeCl_3$	:	iron (III) chloride
g	:	gram
x *g*	:	relative centrifugal force (RCF)
G+C	:	guanine and cytosine
GO	:	Gene Ontology
HIV	:	human immunodeficiency virus
i.e.	:	id est (that is)
IGV	:	Integrative Genomics Viewer
Iso-	:	Mtb in isoniazid stress, TEX-
Iso+	:	Mtb in isoniazid stress, TEX+
iTSS	:	Internal TSS
J/cm^2	:	joule per square centimeter

Kan-	:	Mtb in kanamycin stress, TEX-
Kan+	:	Mtb in kanamycin stress, TEX+
L	:	Liter
Log-	:	Mtb in log phase, TEX-
Log+	:	Mtb in log phase, TEX+
M	:	Molar
mA	:	milliampere
mCi	:	millicurie
mg/ml	:	milligram per milliliter
$MgCl_2$	:	magnesium chloride
min	:	minute
miRNA	:	microRNA
miscRNA	:	miscellaneous RNA
ml	:	milliliter
mm	:	millimeter
mM	:	milliMolar
mmol	:	millimole
mRNA	:	messenger RNA
Mtb	:	*Mycobacterium tuberculosis*
NaCl	:	sodium chloride
NCBI	:	National Center for Biotechnology Information
nm	:	nanometer
npcRNA	:	non-protein-coding RNA
nt	:	nucleotide
OD_{600}	:	optical density at 600 nm wavelength
OMP	:	outer membrane protein
ORF	:	open reading frame
oTSS	:	Orphan TSS
PAGE	:	polyacrylamide gel electrophoresis
PBS	:	phosphate-buffered saline
PBS-	:	Mtb in nutrient starvation, TEX-
PBS+	:	Mtb in nutrient starvation, TEX+
PCR	:	polymerase chain reaction
PE	:	proline-glutamic acid
pH	:	potential of Hydrogen
PPE	:	proline-proline-glutamic acid
pTSS	:	Primary TSS
RBS	:	ribosome-binding site
RNA	:	ribonucleic acid
RNase	:	ribonuclease
RNA-seq	:	RNA sequencing
ROSE	:	Repression Of the heat Shock gene Expression element
RPKM	:	Reads Per Kilobase of transcript per Million mapped reads
rpm	:	revolutions per minute

rRNA	:	ribosomal RNA
s	:	second
SDS	:	sodium dodecyl sulfate
SDS-	:	Mtb in detergent stress, TEX-
SDS+	:	Mtb in detergent stress, TEX+
S-Fe	:	size-selected cDNA library of Mtb in iron stress
S-Iso	:	size-selected cDNA library of Mtb in isoniazid stress
S-Kan	:	size-selected cDNA library of Mtb in kanamycin stress
S-Log	:	size-selected cDNA library of Mtb in log phase
S-PBS	:	size-selected cDNA library of Mtb in nutrient starvation
sRNA	:	small regulatory RNA
S-SDS	:	size-selected cDNA library of Mtb in detergent stress
sTSS	:	Secondary TSS
TAE	:	Tris-Acetate-EDTA
TAP	:	tobacco acid pyrophosphatase
TB	:	tuberculosis
TBE	:	Tris-Borate-EDTA
TEMED	:	tetramethylethylenediamine
TEX	:	Terminator^{TM} 5'-phosphate-dependent exonuclease
TEX-	:	TEX-untreated cDNA library
TEX+	:	TEX-treated cDNA library
tRNA	:	transfer RNA
TSS	:	transcriptional start site
U	:	unit
UCSC	:	University of California Santa Cruz
UTR	:	untranslated regions
V	:	volt
v/v	:	volume/volume percent
w/v	:	weight/volume percent
WHO	:	World Health Organization

PEMETAAN TITIK AWAL TRANSKRIPSI DAN PENGENALPASTIAN sRNA BARU BAGI *Mycobacterium tuberculosis* H37Rv

ABSTRAK

Penyesuaian *Mycobacterium tuberculosis* (Mtb) terhadap pelbagai perubahan persekitaran dalam sistem imun membolehkan bakteria tersebut untuk hidup berterusan dalam individu yang dijangkiti selama beberapa dekad. Pelbagai aspek dalam pengawalan ekspresi gen terhadap perubahan tekanan persekitaran telah dikaji secara meluas dalam Mtb. Namun, tidak banyak perhatian yang diberi terhadap fungsi pengawalan ekspresi gen oleh titik awal transkripsi (TSSs) dan RNA kecil (sRNAs) dalam Mtb, terutamanya dalam pelbagai rangsangan tekanan persekitaran. Ini menginspirasikan usaha dalam pengenalpastian TSSs dan sRNAs yang terlibat dalam pelbagai rangsangan tekanan persekitaran. Untuk mengenal pasti TSSs dalam Mtb, perpustakaan cDNA yang diperkaya dengan transkrip utama telah dibina untuk RNA sequencing, dan TSSpredator telah diguna untuk menganalisis TSSs bagi Mtb dalam pelbagai rangsangan tekanan persekitaran. Sebanyak 8,173 TSSs telah dikenal pasti dalam enam keadaan. Kebanyakan TSSs (3,904) terletak dalam gen (TSS dalaman) dan 1,932 TSSs primer yang terletak dalam lingkungan 300 nt daripada titik permulaan translasi. 2,861 TSSs merupakan permulaan bagi sRNAs antisense yang berpotensi (TSS antisense). Selain itu, 215 TSSs yang terletak dalam kawasan intergen merupakan permulaan yang berpotensi bagi sRNA berkod trans. Pengenalpastian TSSs membolehkan identifikasi transkrip alternatif dan kawasan tidak ditranslasi (UTR) pada sisi 5’ bagi gen, dengan itu membolehkan anotasi transkrip yang lebih tepat. Di samping itu, pengenalpastian TSSs juga mendapati

bahawa transkrip “leaderless” telah meningkat di Mtb yang berada dalam rangsangan. Ini mencadangkan bahawa transkrip “leaderless” terlibat dalam pengawalan fisiologi Mtb bagi penyesuaian terhadap pelbagai rangsangan. Untuk mengenal pasti sRNAs dalam Mtb, perpustakaan cDNA yang diperkaya dengan RNA yang bersaiz kecil (<120 nt) telah dibina untuk RNA sequencing. Program bioinformatik Rockhopper telah digunakan untuk identifikasi sRNAs Mtb dalam enam keadaan, di mana 865 sRNAs antisense dan 389 sRNAs berkod trans telah diramal. Di samping itu, miRNA-like sRNAs yang bersaiz kecil (<30 nt) juga dikenal pasti dalam kajian ini. Analisis pembezaan pengekspresan mendapati bahawa ekspresi kebanyakan sRNAs meningkat sewaktu dalam keadaan kekurangan nutrisi. Pengenalpastian sasaran gen bagi sRNAs tersebut menunjukkan potensi sRNAs tersebut dalam pengawalan ekspresi gen yang terlibat dalam proses metabolik bagi menjimatkan tenaga dalam sel. Kesimpulannya, kajian ini telah berjaya mengenal pasti TSSs dan sRNAs yang berpotensi dalam Mtb, dan membolehkan kajian selanjutnya dijalankan bagi pembinaan rangkaian pengawalan RNA secara berskala genom dan pengenalpastian fungsi-fungsi sRNAs yang berpotensi dalam Mtb.

MAPPING OF TRANSCRIPTIONAL START SITE AND THE IDENTIFICATION OF NOVEL sRNA IN *Mycobacterium tuberculosis* H37Rv

ABSTRACT

The adaptability of *Mycobacterium tuberculosis* (Mtb) to various environmental changes presented by host immune systems enables its persistence within infected hosts for decades. Various aspects of gene expression regulation in response to environmental stress have been extensively analyzed in Mtb. However, the roles of transcriptional start sites (TSSs) and the small regulatory RNAs (sRNAs) in regulating gene expression within Mtb are given very little attention, particularly under stress condition. This ignites the effort to map TSSs and identify sRNAs involved in different stress environments. In order to map the TSSs in Mtb, cDNA libraries that specifically enriched with primary transcripts were constructed for RNA sequencing, and TSSpredator was integrated for the analysis of TSSs in Mtb monitored in different stress environments. Here, a total of 8,173 TSSs were mapped in six conditions. The majority of TSSs (3,904) were detected within annotated genes (internal TSS), and 1,932 primary TSSs reside within 300 nt upstream of translation start site. 2,861 TSSs correspond to potential antisense sRNAs (antisense TSS). Finally, 215 novel TSSs are located within intergenic regions and represent potentially novel *trans*-encoded sRNAs. The mapping of TSSs allows identification of alternative transcripts and 5' untranslated regions (UTR), thus refining the existing annotation of transcriptional landscapes. In addition, mapping of TSSs also manifested an increasing number of leaderless transcripts under stress condition. This suggests the regulatory roles of leaderless transcripts in Mtb's physiology for stress

survival. To detect potential sRNAs in Mtb, cDNA libraries were constructed from size-fractionated total RNA to enrich the small sizes RNAs (<120 nt) for RNA-seq. The Rockhopper program identifies an abundance of novel transcripts in Mtb from all six conditions, including 865 and 389 potential *cis*-encoded antisense and *trans*-encoded sRNAs, respectively. Besides, miRNA-like sRNAs with small sizes (<30 nt) were also identified in this study. Differential expression analysis reveals that majority of the sRNA candidates are significantly up-regulated under nutrient starvation. Further sRNA target prediction and functional classification of the target genes showed the potential of the sRNA candidates to negatively regulate the expression of genes involved in the metabolic process for energy conservation. In conclusion, this study has successfully mapped the TSSs and identified a plethora of novel sRNA candidates in Mtb. The work set a stage for construction of genome-scale transcriptional regulatory networks and characterization of the potential sRNAs in Mtb.

CHAPTER 1

INTRODUCTION

1.1 Introduction

Tuberculosis (TB) still represents a major health threat nowadays and is particularly prevalent amongst poorer societies in South-East Asia and the Western Pacific Regions. The causative agent of TB, *Mycobacterium tuberculosis* (Mtb) primarily infects the host lungs (pulmonary TB). However, even infections of the brain, lymph nodes, pleura, genitourinary tract, skin, joints and bones are not uncommon (extrapulmonary TB).

Unique characteristics of Mtb include lipid-rich cellular envelopes, high genetic homogeneity, dormancy within infected hosts, and intracellular pathogenesis (Cole et al., 1998). The mycobacterial cell envelope is extremely rich in long chain mycolic acids and forms the characteristic thick waxy lipid coat of mycobacteria. It also contains various surface proteins and in particular, adhesins that facilitate cell-cell contacts (Brennan and Nikaido, 1995). The envelope regulates cellular entry of nutrients and drugs and is majorly responsible for slow growth rates, antibiotic resistance and Mtb virulence (Kieser and Rubin, 2014). In summary, the cell envelope allows Mtb to cope with various stress related to the extremely hostile environment. Apart from several sigma factors, cellular stress leads to the activation of various genes involved in the regulation of anaerobic respiration, the production of cell wall components, siderophore synthesis and iron scavenge, respectively (Schnappinger et al., 2003). Although multiple genes associated with the Mtb's persistence are clearly identified to date (Geiman et al., 2006, Rodriguez et al.,

2002), mechanistic aspects of transcriptional regulation underlying intracellular survival are still poorly understood (Miotto et al., 2012).

Bacterial transcriptomes were generally considered to be less complex than their eukaryotic counterparts. Deep sequencing and RNA tiling array, however, established many additional layers of transcriptional regulation in prokaryotes. For instance, bacterial genes frequently harbor alternative promoter structures, which convey gene expression in response to environmental stimuli, and growth stage-specific usage of defined promoter subsets was reported. Hence, alternative promoters offer routes for more complex regulation of cellular transcription. This is particular beneficial for the smaller, rather size-streamlined genomes of bacteria. For instance, *Mycoplasma pneumoniae* grown under 173 different conditions revealed an extensive usage of alternative transcriptional start sites (TSSs) for stress survival (Güell et al., 2009).

The genome-wide mapping of TSSs via deep sequencing enables the identification of bacterial promoters and untranslated regions (UTRs). Terminator™ 5'-phosphate-dependent exonuclease (TEX) specifically digests transcripts with 5' monophosphorylated termini (processed transcripts) but leaves triphosphorylated RNAs (primary transcripts) unaltered. Enrichment of primary transcripts in TEX-treated versus untreated samples enables the detection of native RNA 5' ends and establishes genome-wide maps of primary transcription. So far, maps of TSSs and their corresponding changes in various stress conditions were not scrutinized in detail for Mtb. This study describes the mapping of primary TSSs for Mtb H37Rv within six different conditions (i.e. log phase, two antibiotic stresses, detergent stress, iron

stress, and nutrient starvation). Individual stress conditions were selected to most closely represent the endogenous environments of macrophages *in vitro* (Geiman et al., 2006).

Non-protein-coding RNAs (npcRNAs) do not provide templates to protein biosynthesis but contribute important regulatory functions. The regulatory RNAs often act based on structural features or in complex with proteins; they usually are differentiated according to size and processing pattern. npcRNAs with sizes of 50 - 500 nt commonly represent small regulatory RNAs (sRNAs) in bacteria. The contribution of sRNAs to the regulation of contemporary genomes is more relevant than anticipated earlier. *Cis*- and *trans*-encoded sRNAs control mRNA transcription and translation and thus directly participate in the regulation of many regulatory circuits. The involvement of various sRNAs in stress response is well documented. To complement the genome-wide analysis of TSSs and to generate more complete pictures of transcriptional regulation, this study investigated the mycobacterial sRNA transcriptome in the same conditions. This analysis sets the stage to understand mechanisms for Mtb's persistence and survival within macrophages.

1.2 Research Objectives

The main objectives of this study are:

1. To map the TSSs and identify alternative TSSs of Mtb in response to stress conditions.
2. To identify the potential sRNAs involved in Mtb's adaptation to various stress conditions.

CHAPTER 2

LITERATURE REVIEW

2.1 Tuberculosis (TB)

2.1.1 History of TB

The earliest evidence of human suffering TB was dated to the Neolithic period (ca. 5000 BC), where the fragments from the hump of a deformed spinal column showing signs of extrapulmonary tuberculous spondylitis (Pott's disease) (Kaufmann et al., 2008). In addition, scripture evidences for tuberculosis were also discovered in the Bible and Chinese literature (ca. 4000 BC) (Kaufmann et al., 2008). Besides, several Egyptian mummies (ca. 2400 BC) were found to be suffered from Pott's disease and in a few cases showing evidence of pulmonary tuberculosis (Crubezy et al., 1998, Zink et al., 2003).

From the Hippocrates era (460 – 370 BC) to the 18th century, TB was also known as phthisis, consumption, and scrofula. According to the *Corpus Hippocraticum*, phthisis was first believed to be a hereditary disease (Herzog, 1998). It was until the beginning of the Renaissance (1478 – 1553) that Girolamo Fracastoro proposed the transmission of phthisis occurs through a contagion (Kaufmann et al., 2008). Richard Morton (1637 – 1698) later expanded on Fracastoro's idea and claimed that the formation of tuberculous lesions precedes phthisis (Kaufmann et al., 2008). The severity of TB at that time also described by Morton as "the Consumption of Young Men, that are in the Flower of their Age" (Morton, 1720). Moreover, the Kings of France and England were believed to possess the power of healing scrofulosis (TB of the lymphatic glands) by performing the royal touch at that time

(Hunt, 2011). During the coronation of King Louis XIII in 1610, 800 scrofulosis patients received the royal touch, thus causing King Louis XIII to suffer from phthisis later (Kaufmann et al., 2008).

By the 18th and 19th century, TB was epidemic in Europe and North America where the annual mortality rates were 800 – 1000 per 100,000 per year (Daniel, 2006). Most TB cases were observed in the working classes whose living and working conditions were poor and overcrowded, where millions of deaths were reported over those two centuries (Frith, 2014). Strangely at that time, TB was romanticised by the society and thought to be fashionable. For example, French author Alexander Dumas wrote "It was the fashion to suffer from the lungs; everybody was consumptive, poets especially; it was good form to spit blood after each emotion that was at all sensational, and to die before reaching the age of thirty." (David, 2002).

There were several major breakthroughs of TB in the 19th century. In 1865, Jeanne-Antoine Villemin showed that TB could be transmitted from human to rabbits, proving the disease was infectious (Frith, 2014). The second major breakthrough is the discovery of tubercle bacillus (later renamed as *Mycobacterium tuberculosis*) as the cause for TB by Robert Koch in 1882 (Sakula, 1983). Following that, Koch introduced tuberculin as a TB diagnostic tool in 1890 and finally Charles Mantoux successfully developed intracutaneous tuberculin reaction in 1908, which is still used as a screening tool for TB today (Kaufmann et al., 2008). The next breakthrough is the introduction of bacille Calmette-Guérin (BCG) (attenuated strain

of *Mycobacterium bovis*) as a vaccine for newborn infants in 1924, where it is effective in protecting toddlers against TB (Hawgood, 2007).

2.1.2 The worldwide TB epidemic

Even with the scientific breakthroughs in identifying the causative agent and vaccine for the disease, TB is still a major global health problem. In 2014, an estimated of 9.6 million new TB cases and 1.5 million deaths have been reported, where majority (58%) of the new cases were in the South-East Asia and the Western Pacific Region (Figure 2.1) (WHO, 2015). Globally, WHO reported a decrease in TB incidence rate by 18% in 2014 compared to 2000, whereas the TB mortality rate has fallen 47% compared to 1990 (Figure 2.2) (WHO, 2015). Despite the reduction in the morbidity and mortality rate, TB still remains one of the world's biggest threats where it was ranked as the second leading cause of death from infectious disease, after the human immunodeficiency virus (HIV) (WHO, 2015).

The syndemic interaction between HIV and TB is one of the factors that magnified the burden of TB disease, where the risk of HIV-positive patients in developing TB is estimated to be 26 – 31 times greater compared to healthy individuals (WHO, 2015). HIV not only increases the risk of progression to active TB during the initial period of infection, but also accelerates the reactivation of latent TB infection (Kwan and Ernst, 2011). This was probably due to the HIV-infected patients are incapable to mount long-lasting protective immune responses against the pathogen (Glynn et al., 2010). Among the 9.6 million of new cases, an estimated of 12% (1.2 million) were HIV co-infected (WHO, 2015). The HIV-associated TB

affects the African Region greatly, where it constituted 75% of the total HIV-positive TB cases (WHO, 2013).

Besides, the emergence of multidrug-resistant TB (TB that resistant to two first-line drugs: isoniazid and rifampicin) and extensively drug-resistant TB (multidrug-resistant TB plus resistance to any fluoroquinolone and at least one of the second-line drugs) also increases the burden of the disease and threatens the global TB control. Although more TB patients were tested for drug resistance nowadays (globally, 58% of previous cases and 12% of new cases were tested), however the case detection gaps are still large, for example the Western Pacific Region is the worst with only 19% of the estimated multidrug-resistant TB cases being detected (WHO, 2015).

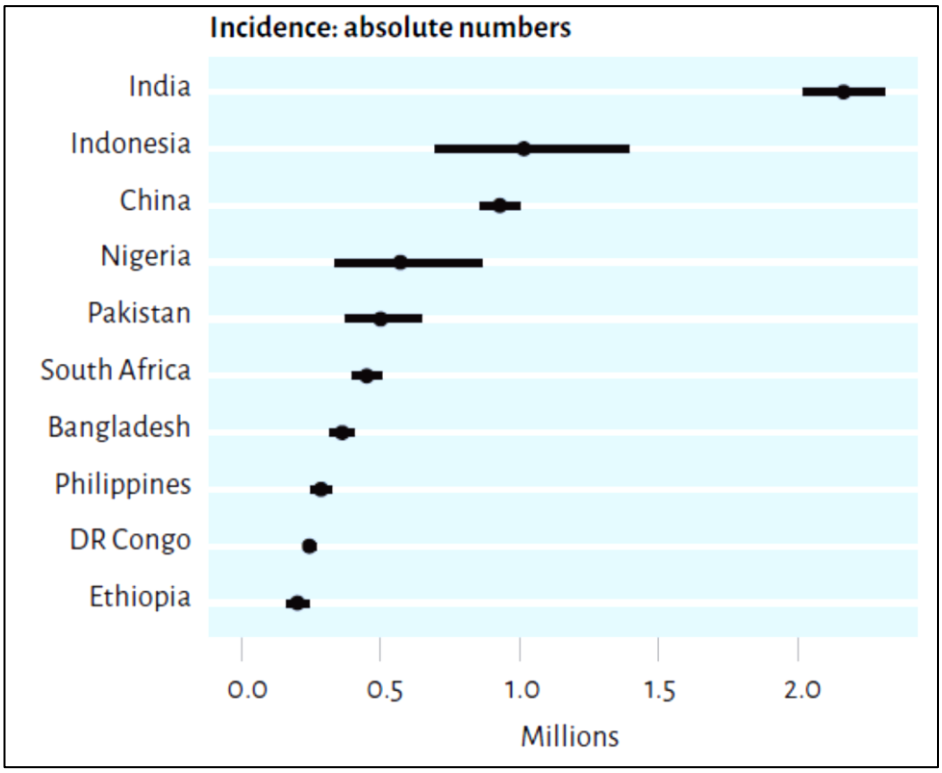

Figure 2.1: The estimated number of TB incidence in top-ten countries, 2014 (adapted from WHO Global Tuberculosis Report 2015).

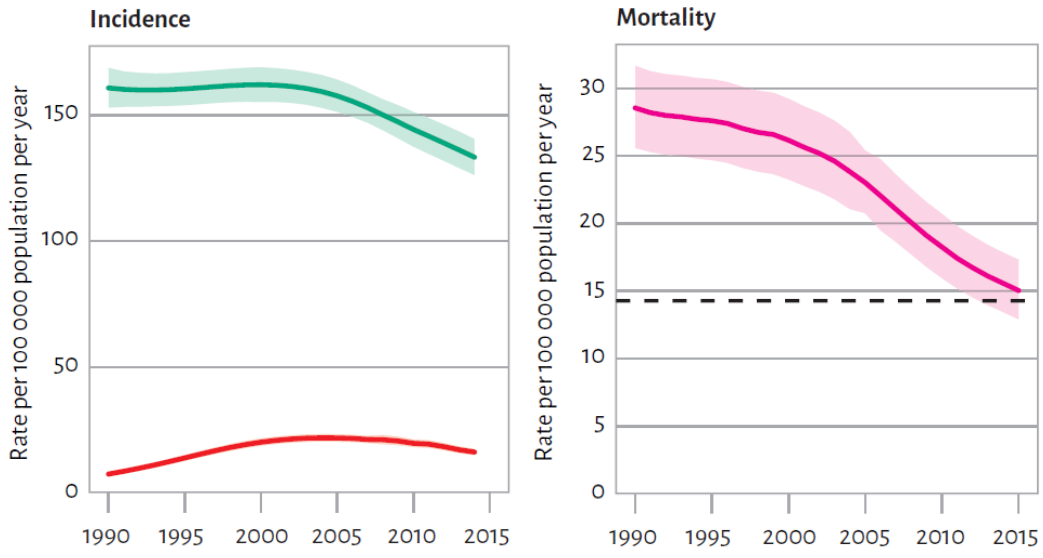

Figure 2.2: The global estimated rates of TB incidence and mortality.
Left: Estimated incidence rate including HIV co-infected TB (green) and estimated incidence rate of HIV co-infected TB only (red). Right: Horizontal dashed line represents the Stop TB Partnership's target of mortality rates by 2015 (adapted from WHO Global Tuberculosis Report 2015).

2.1.3 *Mycobacterium tuberculosis* (Mtb)

Mtb is the causative agent of TB, which infects the lungs (pulmonary TB) and other organs as well, such as brain, lymph nodes, pleura, genitourinary tract, skin, joints and bones (extrapulmonary TB). Individuals with pulmonary TB usually experienced chest pain, prolonged cough with sputum, night sweats, fatigue, fever, weight loss and loss of appetite. On the other hand, the symptoms of extrapulmonary TB may vary depending on the infected site.

Mtb is a rod-shaped bacillus with the size of 1 – 4 µm in length and 0.2 – 0.5 µm in width (Figure 2.3), which can be visualised by Ziehl-Neelsen (acid-fast) staining (Brennan and Nikaido, 1995). It is a slow-growing member of the Mycobacteriaceae family, where the generation time in synthetic medium or infected animals is around 24 hours (Cole et al., 1998). Besides the slow growth rate, other characteristics of Mtb include extensively lipids-rich cell envelope, high genetic homogeneity, the state of dormancy within infected host, and intracellular pathogenesis (Cole et al., 1998, Pérez-Lago et al., 2011).

The mycobacterial cell envelope has unique characteristics and is impermeable towards many compounds (Hett and Rubin, 2008). The Mtb cell envelope consists of: (1) the plasma membrane, (2) a thin, inner peptidoglycan layer, (3) a layer of arabinogalactan linked to both the peptidoglycan layer and the mycolic acids outer membrane, (4) an outer membrane rich in long-carbon-chain mycolic acids that forms the thick waxy lipid coat of mycobacteria, and (5) the surface proteins that functions as enzymes or adhesins that enable the adherent to host cells (Figure 2.4) (Brennan and Nikaido, 1995, Kieser and Rubin, 2014). This complex

cell envelope structure impedes the entry of nutrients and drugs into the cell, causing the slow growth rate, antibiotics resistance and virulence of Mtb (Kieser and Rubin, 2014).

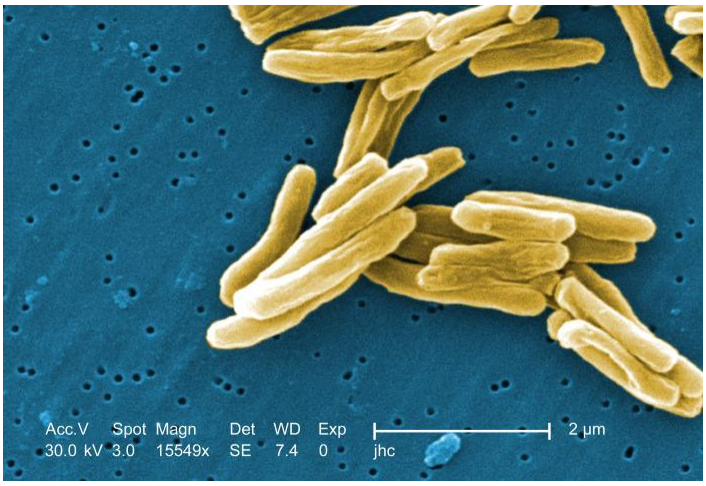

Figure 2.3: Image of Mtb under scanning electron microscope with 15,549X magnification (adapted from Centers for Disease Control and Prevention – Division of Tuberculosis Elimination).

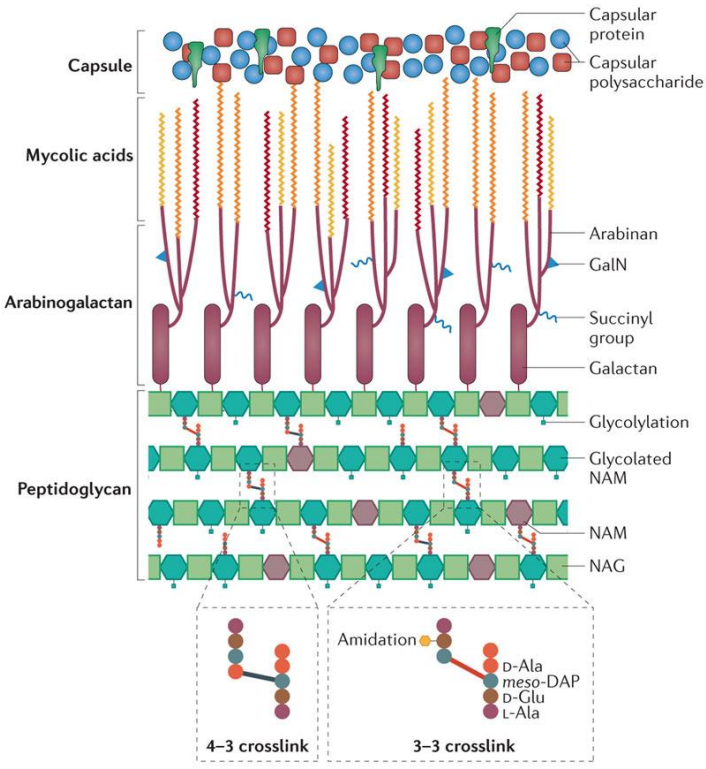

Figure 2.4: Schematic representation of the complex Mtb cell wall (adapted from Kieser and Rubin, 2014).

2.1.4 Transmission and pathogenesis

The transmission of TB between individuals occurs via infectious droplet nuclei containing Mtb where the particles are generated by patients with pulmonary or laryngeal TB during coughing or sneezing (Gengenbacher and Kaufmann, 2012). These infectious droplets are transmitted through the air and remain suspended in the air for several hours (McNerney et al., 2012). Upon inhalation of the droplets, Mtb can traverse the mouth or nasal passages, reaches the lung airways and later phagacytosed by the alveolar macrophages (Figure 2.5) (Gengenbacher and Kaufmann, 2012). The infected macrophages induce a localised proinflammatory response, which attracts the mononuclear cells such as macrophages, lymphocytes and dendritic cells from neighbouring blood vessels, causing the formation of granuloma (Silva Miranda et al., 2012). At this stage, Mtb is able to enter into a dormant state by shutting down its central metabolism activities and terminating replication process (Gengenbacher and Kaufmann, 2012). Reactivation of Mtb would provoke the death of the infected macrophages, leading to the formation of a necrotic zone in the centre of the granuloma (Silva Miranda et al., 2012). The granuloma that loses its solidity due to the formation of cavities in the lung would eventually disintegrate and allows the release of Mtb, which will then spread to other parts of the lungs or released into the airways as contagious droplet nuclei (Silva Miranda et al., 2012). Among the 50 million individuals estimated to be infected by Mtb per year, ~2 billion of them remain latently infected with no symptoms of the disease (Gengenbacher and Kaufmann, 2012). Only less than 10% of the latently infected individuals would progress into the reactivation state, due to factors such as HIV infection, antitumor necrosis factor therapy, alcoholism or malnutrition (Silva Miranda et al., 2012).

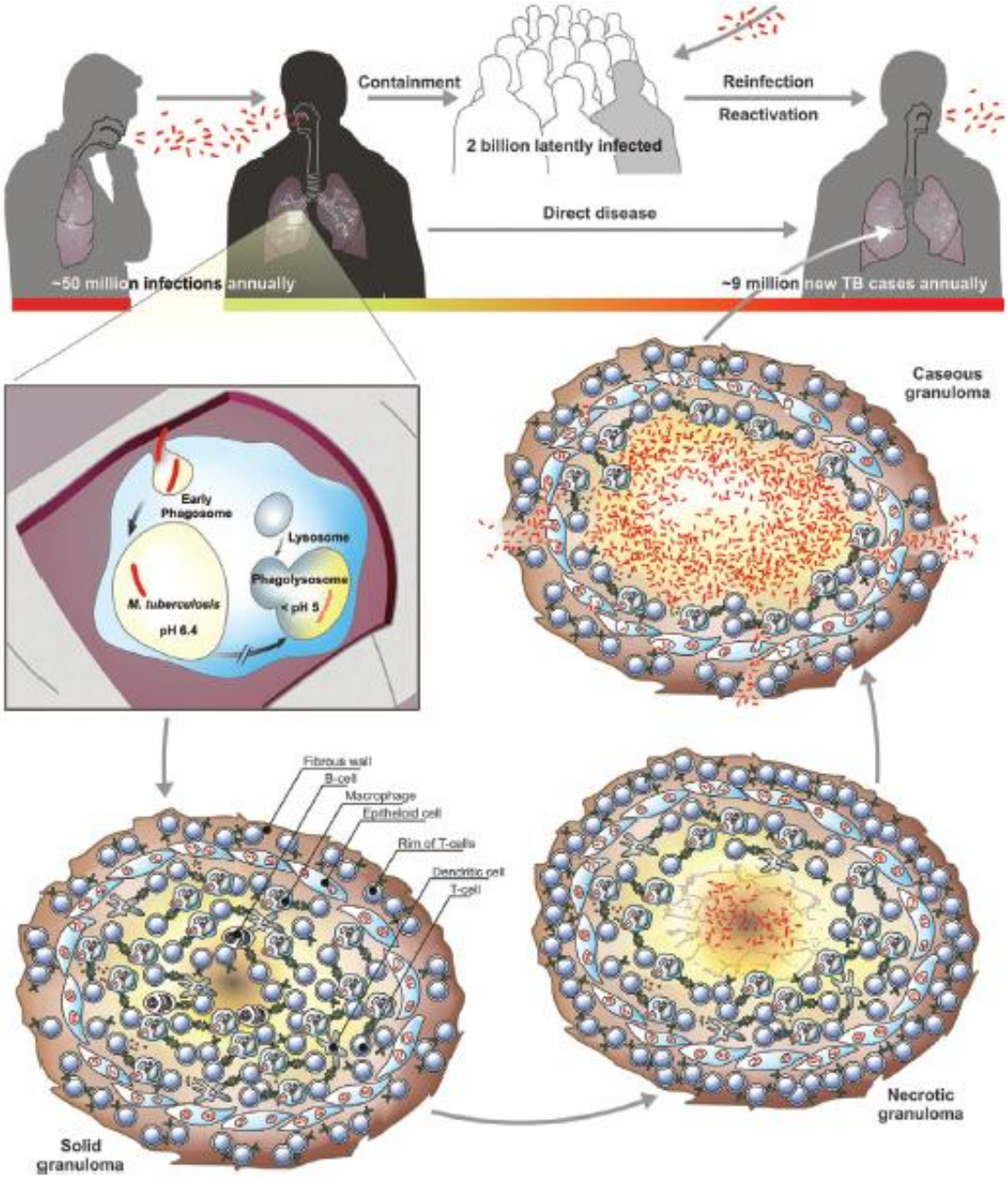

Figure 2.5: Summary of the transmission and pathogenesis of TB (adapted from Gengenbacher and Kaufmann, 2012).

2.1.5 Mtb's survival within macrophages – the key to the pathogenesis of Mtb

After the phagocytic uptake by macrophages, Mtb resides within a subcellular compartment called phagosome. One of the major factors that enables successful persistence of Mtb within macrophages is the prevention of phagosome maturation into acidic and hydrolase-rich phagolysosomes (Katti et al., 2008, Russell, 2001, Russell et al., 2009, Schnappinger et al., 2003, Vergne et al., 2003, Vergne et al., 2005, Walburger et al., 2004). Thus, the phagosomes are retained at a stage where pH is close to neutral (pH 6.4) and the access to extracellular compartment is maintained. This allows the mycobacteria to acquire nutrients and avoid the host cell killing mechanisms (de Chastellier, 2009, Pieters, 2008, Russell, 2001).

Mtb that failed to prevent phagosome maturation were found to be viable by switching into a dormant state within the granulomas (Via et al., 1998). The conditions within granulomas are often hypoxic, acidic, nutrient limited and with the presence of immune effectors such as nitric oxide (Rustad et al., 2009). In response to the stress conditions, mycobacterial genes involved in stress responses, anaerobic respiration, and production of cell wall components, diverse sigma factor, siderophore and scavenging of iron were found to be upregulated (Beste et al., 2007, Schnappinger et al., 2003, Ward et al., 2010).

2.1.6 Persistence within macrophages upregulates several gene and protein expressions

Genes encoding the RNA polymerase sigma factors were amongst the first regulator genes discovered to be involved in stress responses (DeMaio et al., 1996, Gomez et al., 1997). For example, a sigma factor gene, *sigF* were found to be highly expressed during stationary phase, and other stress conditions such as nitrogen depletion, oxidative stress, anaerobic, alcohol shock and cold shock (DeMaio et al., 1996). However, sigF was undetectable during the exponential growth phase (DeMaio et al., 1996). The essential role of sigF for Mtb to adapt to host defenses and persist within host was further confirmed when a *sigF*-deletion strain failed to survive in the later stages of infection (Chen et al., 2000).

Under the carbohydrates-limited environment, the lipid metabolism of Mtb was found to be elevated (Kim et al., 2010). Upon infection, the host genes encoding cholesterol, cholesteryl esters, triacylglycerols and lactosylceramide were highly expressed (Kim et al., 2010). These findings suggested that when Mtb switches to lipids as preferred carbon source within granulomas, the mycobacteria also manipulated the host lipid metabolism for its survival (Kim et al., 2010). On the other hand, the hyperphosphorylated guanine, (p)ppGpp was also utilized as a signalling molecule to regulate expression of genes involved in Mtb survival during starvation (Primm et al., 2000). In response to nutrient deprivation, (p)ppGpp was produced by Rel(Mtb), a protein homolog to the RelA and SpoT proteins identified in Gram-negative bacteria (Primm et al., 2000). The rel(Mtb)-deletion mutant strain was significantly less able to survive during *in vitro* nutrient starvation model,

thereby suggesting that the Rel(Mtb) protein is essential for the persistence within host (Primm et al., 2000).

Besides, a transcriptional regulator of Mtb, DosR is also involved in controlling the mycobacterial dormancy survival within granulomas (Boon and Dick, 2002). DosR is a member of the DosS/DosR two-component regulatory system induced by hypoxia and nitric oxide exposure (Converse et al., 2009). The two-component system is generally consists of a heme-sensor kinase, DosS that autophosphorylates in the presence of external stimulus. The phosphate group was then transferred to its cognate response regulator, DosR where it binds to the promoter regions and regulates the transcription of its target genes (Sivaramakrishnan and Ortiz de Montellano, 2013). The upregulation of 48 genes were reported to be DosR-dependent following the exposure of Mtb H37Rv to hypoxia and nitric oxide (Park et al., 2003, Voskuil et al., 2003). These genes were scattered around the chromosome and this arrangement was postulated to facilitate a more rapid and coordinated response towards the stress stimuli (Park et al., 2003, Voskuil et al., 2003).

Although multiple genes associated with the Mtb's persistence have been identified through differential gene expression studies (Geiman et al., 2006, Rodriguez et al., 2002), the mechanisms of transcriptional regulations for the mycobacterial intracellular survival are still poorly understood (Miotto et al., 2012). RNA-based regulation and the alternative operon structures also have been shown to play a role in regulating the bacterial transcription (Cho et al., 2013, Sorek and Cossart, 2010).

2.2 The bacterial transcriptome

The total transcriptome of a bacteria is defined as the complete set of cellular RNA transcripts, which can be generally divided into two categories (Brosius and Tiedge, 2004): (1) the protein-coding RNAs that are transcribed and subsequently translated into amino acids, which is the messenger RNAs (mRNAs), and (2) the non-protein-coding RNAs that are transcribed but not translated, which included the housekeeping RNAs and small regulatory RNAs (sRNAs) (Wagner and Vogel, 2003, Wassarman et al., 1999). Housekeeping RNAs are constitutively expressed and provide general regulatory functions, such as processing of primary transcripts, and protein translation (e.g. ribosomal RNAs, transfer RNAs, transfer-messenger RNAs, RNase P RNA, signal recognition particles-RNA and etc.) (Szymanski et al., 2003). On the other hand, sRNAs are involved in control of gene expression at the levels of transcription, transcript stability, and translation (Szymanski et al., 2003, Wagner and Vogel, 2003).

The whole-transcriptome study is a powerful tool for understanding gene structures and RNA-based regulation through mapping of transcripts to unique regions of the genome with single-base-pair resolution (Cho et al., 2013). The transcriptome-wide studies have been employed for characterization of fundamental regulatory mechanisms in eukaryotes for more than a decade (Sorek and Cossart, 2010). The transcriptomes of bacteria have not received attention until recently mainly due to the assumption that prokaryotic transcriptomes are simpler compared to the eukaryotic transcriptomes because their transcripts due to the lack of introns and thus the absence of alternative splicing events in the former (Sorek and Cossart, 2010). However, recent findings from *Burkholderia cenocepacia* (Yoder-Himes et

al., 2009), *Escherichia coli* (Thomason et al., 2015), *Helicobacter pylori* (Sharma et al., 2010), *Listeria monocytogenes* (Toledo-Arana et al., 2009), and *Salmonella* spp. (Perkins et al., 2009, Sittka et al., 2008) that grown in different conditions have uncovered many unexpected features including the alternative transcriptional start sites within operon structures, and the RNA-based regulation mediated by sRNAs in the bacterial transcriptomes. The transcriptomic studies which enable the mapping of transcriptional start sites (Sharma et al., 2010, Thomason et al., 2015) and discovery of sRNAs (Arnvig et al., 2011, Raabe et al., 2010), have provided insights into the functional genomic elements and their regulatory roles in bacteria, thus re-shaping our understanding of the complexity of the bacterial transcriptome.

2.2.1 Bacterial transcription initiation

Transcription is a process where the genetic information stored in DNA is converted into complementary RNA by enzymes called DNA-dependent RNA polymerases (Weiss and Gladstone, 1959). Unlike the transcription in eukaryotes that is initiated by three different RNA polymerases (Pol I, II, and III) (Roeder and Rutter, 1969, Kedinger et al., 1970), the bacterial transcription is initiated by a single type of RNA polymerase that binds to specific promoter sequences located upstream of the coding regions (Burgess et al., 1969). Prior to the transcription initiation, the bacterial RNA polymerase core complex ($\alpha_2\beta\beta'\omega$) interacts with regulatory subunit known as sigma factor and converts the core enzyme into the RNA polymerase holoenzyme ($\alpha_2\beta\beta'\omega\sigma$) that will only initiates transcription at specific promoter sequences (Burgess et al., 1969).

Sigma (σ) factors are responsible for the specificity of promoter recognition by RNA polymerase. They were first postulated to have regulatory roles in affecting the transcription patterns of lytic phages in different growth phases (Burgess et al., 1969). Among the many different σ factors that are present in the cell, most of them belong to a homologous family closely related to the σ^{70} from *E. coli* (Gribskov and Burgess, 1986). The σ^{70} is the primary σ factor that suffices the growth of most bacteria under nutrient-rich conditions. Structural analysis showed that the σ^{70} factor is comprised of four regions, i.e. σ1, σ2, σ3 and σ4, which are further divided into subregions (Figure 2.6) (Campbell et al., 2002, Malhotra et al., 1996). σ2.4 and σ4.2 are the subregions identified to recognize the -10 and -35 regions of the promoters, respectively (Campbell et al., 2002). The locations of σ2.4 and σ4.2 are consistent with their roles in promoter recognition where the distance between regions σ2.4 and σ4.2 is ~75 Å, which are corresponding to the 16 – 18 bp distance between the -10 and -35 regions (Murakami et al., 2002).

Although the bacterial promoters are expected to share common structural features for similar interaction with RNA polymerase, however their sequences are not identical (Pribnow, 1975). Therefore by analyzing the sequences of 112 well-defined promoters in *E. coli,* consensus sequences recognized by the RNA polymerase containing σ^{70} have been identified, i.e. 5'-TATAAT-3' at the -10 region and 5'-TTGACA-3' at the -35 region (Hawley and McClure, 1983). For mycobacteria, a hexameric consensus sequence (5'-TATAAT-3') that is similar to the consensus sequence in nearly all prokaryotic promoters has been identified at the -10 region (Bashyam et al., 1996). However, there is little homology between the sequences at the -35 region (Bashyam et al., 1996). The lack of conserved sequences

at the -35 region might due to the diversity of the sequences which are recognized by the different σ factors in mycobacteria (Newton-Foot and Gey van Pittius, 2013). The different associations of σ factors and promoter sequences could provide alterations in gene expression at the level of transcription, thus enabling changes in bacterial transcription patterns in response to different environemnts (Burgess et al., 1969). Therefore, in order to understand the complex transcriptional regulation of Mtb during infection and persistence within host, further investigations on the Mtb's promoters and their association with σ factors in response stress conditions are crucial (Newton-Foot and Gey van Pittius, 2013).

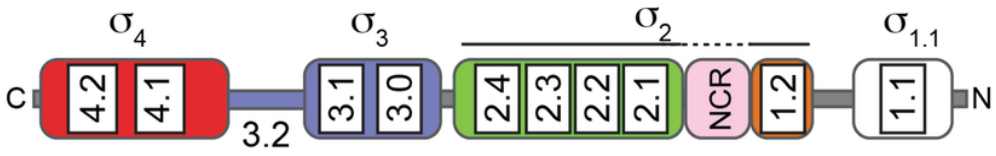

Figure 2.6: Structure of bacterial σ^{70} factor.
The schematic diagram shows the structure of σ^{70} factor which is consists of four conserved regions (σ1, σ2, σ3 and σ4). The regions are represented by numbered boxes and are color coded. NCR (in pink) represents the nonconserved region (adapted from Paget, 2015).

2.2.2 Mapping of transcriptional start sites (TSSs)

TSS refers to the location where RNA polymerase initiates the transcription of DNA into RNA. Previously, the TSSs were identified through primer extension and S1 nuclease protection mapping assay that are labour-intensive (Berk and Sharp, 1977). High-density tiling arrays also have been used for TSSs mapping (McGrath et al., 2007, Voorhies et al., 2011), however these studies did not map the TSSs directly due to a lack of discrimination between the primary (mRNAs and sRNAs with 5'-

triphosphate group) and processed (mature rRNAs and tRNAs with 5'-monophosphate or 5'-hydroxyl group) transcripts. With the advent of RNA sequencing (RNA-seq) coupled with relative enrichment of primary transcripts and depletion of processed RNAs during construction of cDNA libraries, a single-base-pair resolution map of the *H. pylori* transcriptome that allows the genome-wide mapping of TSSs was generated (Sharma et al., 2010). The purposes of mapping the genome-wide TSSs in bacterial genome are discussed in the following sections and illustrated in Figure 2.7 and 2.8.

2.2.2 (a) Discovery of new genes

The mapping of TSSs is beneficial for improving the bacterial genomic annotation. Most of the completely sequenced bacterial genomes were annotated using gene-prediction software. However, these computational predictions are usually error prone and failed to annotate genes encoding short peptides and sRNAs due to the much longer threshold size set for open reading frames (ORFs) identification (McHardy et al., 2004, Overbeek et al., 2007). With the mapping of TSSs and transcript coverage analysis, small genes that were previously failed to be annotated by the computational predictions are able to be identified (Figure 2.7A).

2.2.2 (b) Correction of gene annotations

TSS mapping can be used to screen for mis-annotated genes in the bacterial genome. The genes in bacterial genome are usually annotated by identifying the largest ORFs spanning between a candidate start codon and stop codon (DeJesus et al., 2013). The gene annotation by computational method usually have an accuracy rate of 99%, however previous study showed that the error rate of gene annotation could rise to

14% in the GC-rich bacterial genomes (Dunbar et al., 2011). Therefore, the identification of actual TSS that located downstream from the computational predicted one could indicate an incorrect start codon have been predicted initially and the gene is shorter than annotated (Figure 2.7B) (Wurtzel et al., 2010).

2.2.2 (c) Definition of 5' untranslated regions (5' UTRs)

The 5' UTR of gene which also known as leader sequence, is the region from the transcriptional start site to the translational start codon. The mapping of TSSs can be used to determine the length of 5' UTRs (Figure 2.7C). Usually, the 5' UTRs in bacteria are shorter than 30 bp (Sorek and Cossart, 2010), however recent studies have identified 5' UTRs that are more than 100 bp in *Salmonella* Typhi (Perkins et al., 2009). The 5' UTRs in prokaryotes are known to contain riboregulators such as metabolite-sensing riboswitches (Vitreschak et al., 2004), RNA thermometers (Kortmann and Narberhaus, 2012), and pH responsive RNA element (Padan et al., 2005). In this regard, the mapping of TSSs and further bioinformatics analysis on the 5' UTRs could lead to the detection of putative riboregulators in Mtb H37Rv (Hammann and Westhof, 2007).

The determination of 5' UTRs also allows the identification of ribosome-binding site (RBS). For bacterial translation to occur, the initiation process involves the binding of 16S rRNA and initiator tRNA, $tRNA_f^{Met}$ to the RBS located within the 5' UTR of mRNA (Kozak, 1983). The RBS generally extends 20 nt on either side of the translational start codon (Gold, 1988). In order to facilitate mRNA anchoring and adaptation to the 30S ribosome, the RBS is complementary to the highly conserved sequence in the 3' end of 16S rRNA, which is one of the ribosomal RNA

components (Steitz and Jakes, 1975). For most bacteria, the core sequence near the 3' end of 16S rRNA is 5'-CCUCCU-3', therefore most of the prokaryotic RBS has a subset of the sequence 5'-AGGAGG-3' (Omotajo et al., 2015). For Mtb H37Rv, the RBS was identified by analyzing the sequence that is complementary to the 3' end of 16S rRNA (5'-ACCTCCTT-3') (sequence of 16S rRNA gene was obtained from NCBI Reference Sequence: NC_000962.3; May 14, 2015). The RBS in Mtb H37Rv has the sequence 5'-AAGGAGGT-3', which also contains a subset of the prokaryotic RBS core sequence (5'-AGGAGG-3') (Omotajo et al., 2015). Interestingly, leaderless transcripts that are lack of RBS were found to be actively translated in some bacteria (Shean and Gottesman, 1992). Recently, increased leaderless mRNAs were also identified in Mtb exposed to nutrient starvation (Cortes et al., 2013), indicating the functional potentialities of these leaderless mRNAs in response to stress conditions.

2.2.2 (d) Determination of operon structural arrangement

The genome-wide TSSs maps also identify the potential operon structures in bacteria (Figure 2.7D). In the prokaryotic genome, consecutive genes with functional association are often arranged in an operon, under the control of a single promoter. A transcriptome analysis of *Mycoplasma pneumoniae* under 173 different growth conditions revealed that there are extensive alternative TSSs located within the operon structures in response to different condition, suggesting that these alternative sub-operon structures could act in a manner that is analogous to the alternative splicing in eukaryotes (Güell et al., 2009).

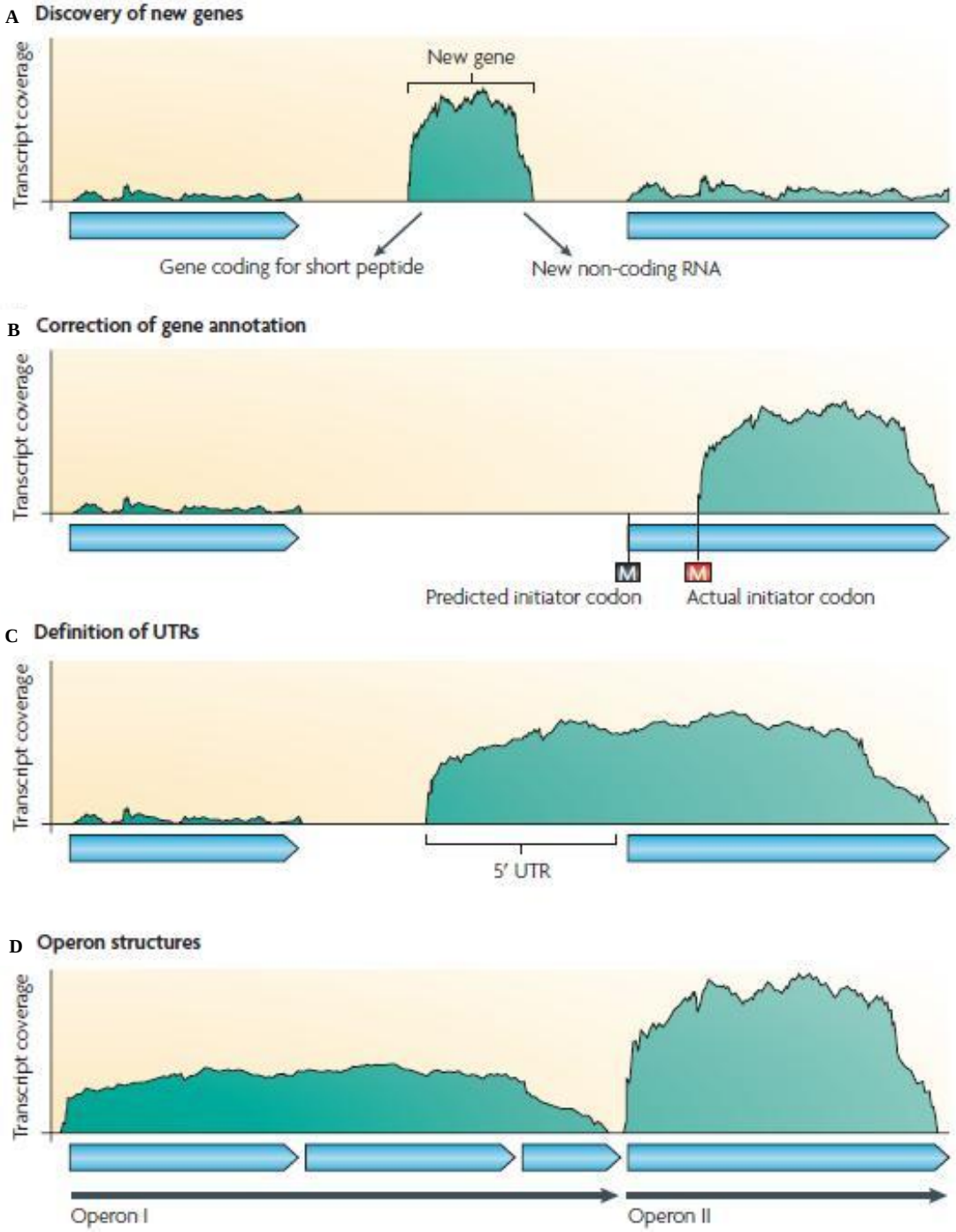

Figure 2.7: Contributions of TSSs mapping to improve bacterial genome annotations. In all panels, the y-axis represents the transcript coverage derived from RNA-seq or tiling array experiments; while the x-axis represents the schematic genomic regions. The arrows in light blue represent the annotated protein-coding genes. The transcriptomic information from TSSs mapping can be used to: A. discovery of novel genes encoding short peptides or sRNAs; B. correction of gene annotations where the black M (methionine) depicts the first predicted start codon and the red M depicts the start codon in corrected annotation; C. map the 5' UTRs across the entire genome; and D. determination of operon relationships (adapted from Sorek and Cossart, 2010).