

**Computational characterisation of DNA
Methylomes in *Mycobacterium tuberculosis* Beijing
hyper- and hypo-virulent strains**



Thesis presented in fulfillment of the requirements for the Degree of *Doctor
Philosophiae* in Bioinformatics at the South African National Bioinformatics
Institute, University of the Western Cape

Advisor: Prof. Alan Christoffels

Co-advisor: Prof. Nico Gey van Pittius

December 2014

KEYWORDS

DNA Methylation

Methyltransferase

Next generation sequencing

Mycobacterium tuberculosis

PacBio single molecule sequencing

Virulence

Hyper-virulent strain

Hypo-virulent strain

Motif



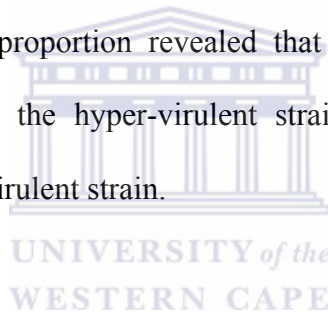
ABSTRACT

Mycobacterium tuberculosis, the causative agent of tuberculosis, is estimated to infect approximately one-third of the world's population and is responsible for around 2 million deaths per year. The disease is endemic in South Africa which has one of the world's highest tuberculosis incidence and death rates. The *M. tuberculosis* Beijing genotype are characterised by having an enhanced virulence capability over other *M. tuberculosis* strains and are the predominant strain observed in the Western Cape of South Africa. DNA methylation is a largely untapped area of research in *M.tuberculosis* and has been poorly described in the literature especially given its connection to virulence despite it being well characterised along with its role in virulence in other pathogenic bacteria such as *E.coli*. The overall aim was to characterise a global DNA methylation profile for two *M. tuberculosis* Beijing strains, hyper-virulent and hypo-virulent, using single molecule real time sequencing data technology. Moreover, to determine if adenine methylation in promoter regions has a possible functional role.

This study identified and characterised the DNA methylation profile at the single nucleotide resolution in these strains using Pacific Biosciences single molecule real time sequencing data. A computational approach was used to discern DNA methylation patterns between the hyper and hypo-virulent strains with a view of understanding virulence in the hyper-virulent strain. Methylated motifs, which belong to known Restriction Modification (RM) systems of the H37Rv reference genome were also identified. N6-methyladenine (m6A) and N4-methylcytosine

(m4C) loci were identified in both strains. m6A were identified in both strains occurring within the following sequence motifs CACGCAG (Type II RM system), GATNNNNRTAC/GTAYNNNNATC (Type I RM system), while the CTGGAGGA motif was found to be uniquely methylated in the hyper-virulent strain.

Interestingly, the CACGCAG motif was significantly methylated ($p = 9.9 \times 10^{-63}$) at a higher proportion in intergenic regions (~70%) as opposed to genic regions in both the hyper-virulent and hypo-virulent strains suggesting a role in gene regulation. There appeared to be a higher proportion of m6A occurring in intergenic regions compared to within genes for hyper-virulent (61%) and hypo-virulent (62%) strains. The genic proportion revealed that 35% of total m6A occurred uniquely within genes for the hyper-virulent strain while 27.9% for uniquely methylated genes in hypo-virulent strain.



A functional enrichment of the genes that were uniquely methylated in the hypo-virulent and hyper-virulent strains, revealed the following COG categories as significant: 1) Secondary metabolites biosynthesis, transport and catabolism was significant in the hyper-virulent strain ($p = 0.006$), 2) Energy production and conversion ($p = 0.015$) and 3) Cell wall/membrane/envelope biogenesis ($p = 0.047$) in the hypo-virulent strain. For the shared methylated genes between the two strains, no functional categories were statistically significant.

The DNA methylation patterns revealed that the majority (79%) of the m6A loci were shared between the hyper- and hypo-virulent strains. Despite the high proportion of shared methylated regions between hyper-virulent and hypo-virulent

strains we still observe strain specific pathways and genes. Analysis of methylated promoter associated genes reveal that the functional categories significantly enriched uniquely to the hyper-virulent strain were GO categories ‘response to acid’ and ‘energy conversion’. The enriched categories might suggest that these genes respond to stress conditions or adaptation to survival in the hyper-virulent strain.

This is the first study to our knowledge that identified and characterised m6A on a single nucleotide level using PacBio sequencing for *M.tuberculosis*. The DNA methylation analysis provides a starting point into understanding the role of DNA methylation and virulence strategies for *M. tuberculosis*.



DECLARATION

I declare that “*Computational characterisation of DNA Methylomes in Mycobacterium tuberculosis Beijing hyper- and hypo-virulent strains*” is my own work, that it has not been submitted for any degree or examination in any other university, and that all the resources I have or quoted have been indicated and acknowledged by complete references.

Alecia Geraldine Naidu



December 2014

Signed

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Prof Alan Christoffels, for giving me the opportunity to pursue my PhD studies and for his guidance and support. I would also like to extend my gratitude to my co-supervisor, Prof Nico Gey van Pittius for giving me the opportunity to pursue this project and to Prof Rob Warren for his valuable input and suggestions on my thesis. Much appreciation goes to Maryam and Ferial at SANBI for the admin support during my studies. I am grateful to The National Research Foundation and the Medical Research Council for financial support.

My deepest gratitude goes out to my parents for their loving support and selfless sacrifices throughout my studies. Thank you to all my friends and church friends for all the support, love and encouragement.

Last but not least, THANK YOU to my Lord and Saviour, Jesus Christ, whom through Him, this thesis would not be possible. Amen

TABLE OF CONTENTS

KEYWORDS	i
ABSTRACT	ii
DECLARATION	v
ACKNOWLEDGEMENTS	vi
TABLE OF CONTENTS	vii
LIST OF FIGURES	xi
LIST OF TABLES	xii
ABBREVIATIONS	xiii
CHAPTER 1: LITERATURE REVIEW	1
1.1 The <i>Mycobacterium tuberculosis</i> genome.....	1
1.1.1 Mycobacterium and the <i>M. tuberculosis</i> complex (MTC)	1
1.1.2 Tuberculosis pathogenesis in humans.....	2
1.1.3 Virulence of <i>M. tuberculosis</i>	3
1.1.4 The Beijing Genotype.....	5
1.2 Epigenetics.....	6
1.2.1 DNA Methylation in Bacteria.....	7
1.2.1.1 An overview of DNA methylation.....	7
1.2.1.2 DNA methylation recognition sites and DNA Methyltransferases	9
1.2.1.3 The role of DNA methylation in bacteria	10
1.2.1.4 Methylation and virulence in pathogenic bacteria	12
1.3 DNA Sequencing Technology	13
1.3.1 First and second generation DNA sequencing technologies	13

1.3.2 Third generation sequencing technology.....	17
1.3.3 Application of DNA sequencing to bacterial genomics	19
1.3.4 Whole genome bisulphite sequencing	21
1.4 Bioinformatics and NGS analysis.....	23
1.4.1 Data generation and format.....	23
1.4.2 Data analysis and algorithms	24
1.5 Rationale	26
CHAPTER 2: MATERIALS AND METHODS	29
2.1 Single Molecule Real Time DNA Sequencing.....	29
2.2 Computational Analysis of the PacBio sequencing data	31
2.2.1 Quality filtering of Raw Reads	31
2.2.2 Mapping to the H37Rv reference genome.....	32
2.2.3 Base Modification calling.....	32
2.3 Genome-wide methylome characterisation	34
2.3.1 Identification and characterisation of methylated recognition motifs and methyltransferases.....	34
2.3.2 Genic methylation Analysis and characterisation.....	35
2.3.3 Intergenic DNA methylation Analysis	36
2.3.3.1 Analysis of methylated promoter regions.....	36
2.3.3.1.1 Operon prediction and organisation.....	36
2.3.3.1.2 Comparative analysis of methylated promoters between hyper and hypo-virulent strains	37
2.3.3.2 Functional enrichment analysis of the methylated promoter operons.....	37
2.3.3.3 Metabolic Pathway Analysis	38

CHAPTER 3: RESULTS AND DISCUSSION	39
3.1 Description of methylation patterns.....	40
3.1.1 Mapping hyper-virulent and hypo-virulent <i>M. tuberculosis</i> Beijing genotype strain genomes to the reference <i>M. tuberculosis</i> H37Rv genome	40
3.1.2 Identification and characterisation of methylomes in hyper-virulent and hypo-virulent strains	41
3.1.2.1 Pinpointing the genomic locations of methylation	44
3.1.3 Summary.....	47
3.2 Methylated sequence motifs	48
3.3 Downstream analysis of methylation loci.....	54
3.3.1 Genic Methylation	54
3.3.1.2 Uniquely methylated genes in hyper-virulent and hypo-virulent strains.....	56
3.3.2 Intergenic methylation.....	59
3.3.2.1 Methylation in promoter regions	59
3.3.2.1.1 Functional enrichment of methylated promoters of operons	61
3.3.2.1.2 Metabolic pathways affected by methylated operon associated promoters.....	66
3.3.2.2 Summary	68
CHAPTER 4: CONCLUSIONS AND FUTURE WORK.....	72
4.1 Key findings.....	73
4.2 Novel aspects and impact on TB research	74
4.3 Limitations	75
4.4 Future work.....	76

BIBLIOGRAPHY 77

APPENDICES..... 99

Appendix I: Uniquely methylated promoter-associated genes for the Hyper-
virulent strain 99

Appendix II: Uniquely methylated promoter-associated genes for the Hypo-
virulent strain 101

Appendix III: Shared methylated promoter-associated genes in the Hyper- and
Hypo-virulent strains 103



LIST OF FIGURES

Figure 1.1 Graphical representation of methylated DNA bases. 6-methyl adenine (left); 5-methyl cytosine; 4-methylcytosine. The methyl group is highlighted in red. (Davis et al. 2013).....	8
Figure 1.2. Detection of DNA modifications using PacBio SMRT sequencing. a represents the sequencing reaction with the thymine nucleotide incorporated into the growing strand corresponding to a methylated (top) and unmethylated adenine (bottom). b represents the SMRT fluorescence real time detection. The x-axis shows the time for the nucleotide incorporation and the y-axis represents the fluorescence intensity. The interpulse duration (IPD) represented by the dashed line shows a delayed incorporation of the thymine due to the methylated adenine (top) compared to a shorter time for an unmethylated adenine (bottom) (Flusberg et al. 2010).....	18
Figure 2.1 Overview of data analysis. A) Represents the overview of the methodology for the identification of methylated DNA bases from the PacBio reads. B) Following identification of methylation sites, characterisation of the genic positions of methylated loci in uniquely and shared methylated positions between the hyper and hypo-virulent strains. C) Represents the promoter region methylation analysis and subsequent enrichment of unique and shared methylated promoter associated genes.	30
Figure 3.1 An illustration of the IPD ratio of 6.20 (purple block) for a methylated adenine within the predicted motif CACGCAG in the forward strand. The methylated adenine also occurs within the <i>OmpA</i> gene.	42
Figure 3.2 Frequency of genic methylation sites in hyper-virulent and hypo-virulent strains for both m6A and m4C sites.	45
Figure 3.3 Frequency of intergenic methylation sites in hyper-virulent and hypo-virulent strains for both m6A and m4C sites.	45

LIST OF TABLES

Table 3.1 A summary of the mapping statistics for hyper-virulent and hypo-virulent genomes	40
Table 3.2 The composition of base modifications identified in hyper-virulent and hypo-virulent strains. The percentages are indicated in parentheses.....	43
Table 3.3 A summary of the coverage and ModQv of m6A and m4C in hyper-virulent and hypo-virulent strains.....	44
Table 3.4 Stranded methylation in hyper-virulent and hypo-virulent strains.....	46
Table 3.5 Summary of methylated sequence motifs in the hyper-virulent strain.	49
Table 3.6 Summary of methylated sequence motifs in the hypo-virulent strain.	49
Table 3.7 Frequency of motifs categorized by the genomic locations in the hyper-virulent strain	50
Table 3.8 Frequency of motifs categorized by the genomic locations in the hypo-virulent strain	50
Table 3.9 Summary of methylation sites for intergenic and genic regions, per strand, in hyper-virulent and hypo-virulent strains.....	55
Table 3.10 Frequency of unique and shared genic and intergenic m6A sites in hyper-virulent and hypo-virulent strains and the frequency of methylated genes	55
Table 3.11 Significantly enriched functional categories of uniquely methylated operon-associated genes in hyper-virulent strain.....	62
Table 3.12 Significantly enriched functional categories of uniquely methylated operon-associated genes in hypo-virulent strain.....	62
Table 3.13 Significantly enriched functional categories of shared methylated operon-associated genes in hyper- and hypo-virulent strains.....	64
Table 3.14. Strain specific KEGG pathways in hyper-virulent strain	67
Table 3.15. Strain specific KEGG pathways in hypo-virulent strain	67

ABBREVIATIONS

ABC	ATP-binding cassette
ATP	Adenosine triphosphate
BCG	Bacillus Calmette–Guérin
BLASR	Basic Local Alignment with Successive Refinement
BP	Base pairs
CDS	Coding Sequence
CFU	Colony Forming Unit
COG	Clusters of Orthologous Groups
DAM	DNA adenine methyltransferase
DNTPS	deoxynucleosidetriphosphates
DDNTPS	di-deoxynucleosidetriphosphates
DNA	Deoxyribonucleic acid
DOOR	Database of PrOkaryote OpeRons
DOSR	Dormancy Survival Regulon
GFF	General Feature Format
GO	Gene Ontology
HDF	Hierarchical Data Format
HIV	Human Immuno-deficiency Virus
IC	Integrated Circuit
IFN-γ	Interferon gamma
IPD	interpulse duration
KEGG	Kyoto Encyclopedia of Genes and Genomes
M4C	4-methylcytosine
M5C	5-methylcytosine
M6A	6-methyladenine
MTASES	Methyltransferases
MTC	<i>Mycobacterium tuberculosis</i> Complex
NCBI	National Center for Biotechnology Information
NGS	Next generation sequencing
ORF	Open reading frame

PACBIO	Pacific Biosciences
PCR	Polymerase Chain Reaction
RM	Restriction Modification
SDH	Succinate Dehydrogenase
SMRT	Single Molecule Real Time
SNP	Single Nucleotide Polymorphism
TB	Tuberculosis
TBDB	Tuberculosis Database
TCA	Tricarboxylic acid
WHO	World Health Organization
XDR-TB	Extensively drug resistant TB
ZMW	Zero-mode waveguides



CHAPTER 1: LITERATURE REVIEW

1.1 The *Mycobacterium tuberculosis* genome

1.1.1 *Mycobacterium* and the *M. tuberculosis* complex (MTC)

Mycobacterium tuberculosis is a bacterium that belongs to the family *Mycobacteriaceae*, the suborder *Corynebacterineae*, the order *Actinomycetales* and the phylum *Actinobacteria*. Members of the suborder *Corynebacterineae* are acid-fast, gram-positive, high-GC bacteria, with a thick peptidoglycan layer in their cell wall that encloses their cell membranes (Silhavy et al. 2010). Prominent pathogens in the *Mycobacterium* genus are the *Mycobacterium tuberculosis* Complex (MTC) cluster, which include both human and non-human pathogens. These are the *M. tuberculosis*, whose target host are the primates; *M. bovis*, which target cattle but may also infect other hosts including human; *M. africanum*, which is the primary cause of human tuberculosis in West Africa and *M. microti*, which is rare but traditionally infects voles. Two additional species have recently been added to the MTC complex, and these include *M. canettii*, which was isolated from a Somali-born patient (van Soolingen et al. 1997) and *M. caprae*, a species that primarily targets goats (Cousins et al. 2003)

Both biochemical and molecular genomics techniques have been used in the isolation and characterisation of the MTC pathogens, since members of this group are genetically closely related, and may be considered as a subspecies of *M. tuberculosis*. Some of the widely applied classification techniques include: 1) DNA-

DNA hybridization, with a sequence similarity threshold greater than 95%, 2) multiple-locus enzyme electrophoresis, 3) 16s ribosomal RNA gene sequencing and 4) interrogation of the repetitive insertion sequence IS6110 and the direct repeat (DR) (van Soolingen et al. 1997).

1.1.2 Tuberculosis pathogenesis in humans

Tuberculosis is an infectious disease, which is caused by the pathogenic bacteria, *Mycobacterium tuberculosis* and affects the pulmonary system in humans and animals. It typically affects the lungs (pulmonary TB) but can affect other sites as well (extrapulmonary TB). The *M. tuberculosis* pathogen was first discovered by Robert Koch in 1882, and communicated to the Berlin Society of Physiology as the aetiological agent for the disease tuberculosis (Cambau & Drancourt 2014). The World Health Organization (WHO), declared TB a global health emergency in 1993, and estimates that there were 8.6 million incident cases of active TB disease globally in 2012 and 1.3 million deaths (WHO| Global Tuberculosis Report, 2013). Around a third of the human population is infected with TB and 90% of the cases remain latent. South Africa has the highest incidence rate in the world with 1000 cases per 100 000 population in some regions of South Africa.

Transmission of *M. tuberculosis* occurs when the human subject inhales droplets containing the bacilli. In the lung alveolar, macrophages engulf and kill the bacilli through phagocytosis, but those that survive are transported into the lung tissue where they remain within the macrophage structure. The pathogen engages in

different mechanisms to evade lysis, such as inhibiting the fusion of the phagosome with the lysosome. The presence of the bacilli in the phagosome leads to the formation of a granuloma as the immune system responds in an attempt to isolate the infected tissue. The most prominent cells in the fight against *M. tuberculosis* are the T cells, and particularly the Th1 effector CD4⁺ T cells, whose activation leads to the secretion of Interferon gamma (IFN- γ) and Tumor necrosis factor alpha (TNF- α). IFN- γ and TNF- α triggers the oxidative burst, which generates Nitric oxide (NO), as well as other reactive oxygen and nitrogen intermediates. However, the presence of these compounds *in vivo* only serves to control the bacillus as it fails to eradicate it (Kursar et al. 2007; Forrellad et al. 2013)

Further exacerbating TB is the co-infection with Human Immuno-deficiency Virus (HIV). There are approximately 1.1 million HIV positive new TB cases in 2012 globally of which 75 % of infected individuals live in sub-Saharan Africa. Furthermore TB is the leading cause of death among HIV infected patients and it presents emerging cases of multi-drug resistant (MDR-TB) and extensively drug resistant TB (XDR-TB). (WHO| HIV-Associated Tuberculosis, 2013).

1.1.3 Virulence of *M. tuberculosis*

M. tuberculosis does not have any classical virulence factors, such as pathogenicity islands or secretion of toxins that are prominent in other pathogenic bacteria (Forrellad et al. 2013). Currently, there are two models that are applied in quantifying TB virulence. These are the animal model, which uses non-human

primates and other subjects, and the cellular model that utilizes *in vitro* assay. When using animal subjects, colony forming units (CFU) of the *M. tuberculosis* bacilli, in the form of an aerosol, are administered intranasally to mimic natural infection (Smith 2003). Once infected, the different parameters is observed in the subject, such as its ability to cause secondary infections to induce extrapulmonary infection (Nicol & Wilkinson 2008).

In virulence quantification using the cellular model, the focus is Mycobacterial infection in the macrophage, dendritic, alveolar and the cells of the adipose tissue. Once infection has been established, three measures can be taken to assess virulence and disease progression and include 1) observation based on fluorescent microscopy, 2) the ability to interfere phagosome maturation, and 3) analysis of the cytokine profiles, and primary Tumor Necrosis Factor alpha (TNF α), Interleukin 2 (IL-2), Interleukin 12 (IL-12), and Interferon gamma (IFN γ) (Smith 2003). Despite having genetic similarity of 99% in *M. tuberculosis* strains (Brosch et al. 2002; Fleischmann et al. 2002) there are differences in virulence and immunogenicity in experimental infection models associated with certain strains. However, a link has not been established between observed virulence phenotypes, and the molecular profile of any given strain (Coscolla & Gagneux 2010).

Genomic variations, such as Single Nucleotide Polymorphisms (SNPs), have been observed in translating into phenotypic changes that confer drug resistance, and particularly in the Beijing strains (Ebrahimi Rad et al. 2003). Generally, strain-specific virulence mechanisms in *M. tuberculosis* can be identified by first isolating

strains with similar clinical manifestations, followed by cellular assays for specific differentiating traits. For example, strains that have been observed to cause infection as a cluster, implying recent disease transmission, tend to grow at a rapid rate than unique ones in a macrophage *in vitro* assay (Zhang et al. 1999; Nicol & Wilkinson 2008).

1.1.4 The Beijing Genotype

Of great interest in *M. tuberculosis* studies is the Beijing strain family, which belongs to the principal genetic group 1 and was first described by (van Soolingen et al. 1995) as the dominant genotype of *M. tuberculosis* in South East Asia. Recent studies have demonstrated that it is one of the world's most wide-spread *M. tuberculosis* genotype (Glynn et al. 2002). The Beijing strain family exhibits selective advantage over other lineages, and this is conferred through increased virulence and rate of transmission that has been demonstrated through *in vitro* assays and *in vivo* using animal models (Li et al. 2002; López et al. 2003; Manca et al. 2004). It is considered hyper-virulent (Parwati et al. 2010) and is associated with multi-drug resistance (Borrell & Gagneux 2011). A number of reports link the Beijing strains with extrapulmonary TB or with treatment failure and relapse (Kong et al. 2006; Sun et al. 2006). The Beijing clade grows faster in mice, has an increased propensity to cause disease, evades the protective effect of the BCG vaccination (Spuy et al. 2009), and acquires drug resistance more frequently due to SNPs in the mismatch repair genes (Ebrahimi Rad et al. 2003).

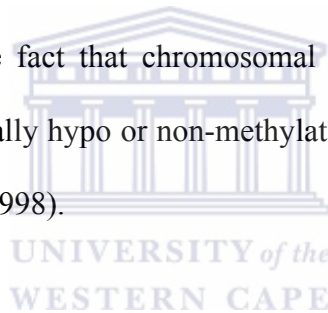
1.2 Epigenetics

Epigenetics is defined as the reversible changes that occur in the biochemical composition of DNA and its associated proteins. Although these changes do not modify the nucleotide sequence, they can be propagated through cell division and therefore may be heritable and could be associated with almost any known phenotype (Bird 2007). Epigenetic modifications play a functional role that connects an organism's genome to its environment. Therefore, the control of the mechanisms that regulate epigenetic changes may directly be linked to other physiological processes such as disease progression in a given environment. Conventional genetic modifications alter the coding DNA sequence within an organism whose expression leads into a structural change in the encoded protein sequence, and subsequently a change in the phenotype. In comparison, epigenetics involve alternative forms of direct DNA or protein sequence modifications, such as the attachments of methyl groups and other chemical signals. These changes, just like in conventional genetic modifications, are heritable and therefore can be passed from one generation to the next (Bock & Lengauer 2008).

One of the most commonly and extensively studied form of epigenetics is DNA methylation. This is a biochemical process that is mediated by the enzyme DNA methyltransferase, and involves the post-replicative transfer of a methyl group from S-adenosyl- methionine to the cytosine or adenine DNA nucleotides (Weber & Schübeler 2007). The resulting modified bases are the C5-methylcytosine, N6-methyl-adenine and N4-methylcytosine (Collier, 2009). In both prokaryotes and

eukaryotes, previous studies have demonstrated that DNA methylation could be playing a direct role in various physiological processes, such as the control of gene regulation, embryo development, as well as genomic imprinting and stability. Furthermore, DNA methylation constitutes chromatin structure, therefore enabling cell growth and differentiation (Bird 2007).

According to (Gardiner-Garden & Frommer 1987), two to seven percent of mammalian cytosine residues occur as 5-methylcytosine and are usually observed in regions referred to as the CpG islands and mostly are found near the transcription start sites. The observation that DNA methylation is involved in the control of gene expression is based on the fact that chromosomal regions that undergo through active transcription are usually hypo or non-methylated when compared to dormant regions (Bird 1992; Razin 1998).



1.2.1 DNA Methylation in Bacteria

1.2.1.1 An overview of DNA methylation

As recently reviewed by (Davis et al. 2013), DNA methylation is the most common form of epigenetic modification in prokaryotic organisms, and is driven by three families of methyltransferases (MTases) through a sequence-specific covalent attachment of methyl groups to either cytosine or adenine nucleotides. Two of these families interact with the exocyclic amino groups to form 6-methyladenine (6mA) and 4-methylcytosine (4MC) (Timinskas et al. 1995). The last family of MTases catalyzes the formation of 5-methylcytosine (5mC) by methylating the 5- carbon

pyrimidine ring of cytosine, and have close homologues in eukaryotes (Kumar et al. 1994).

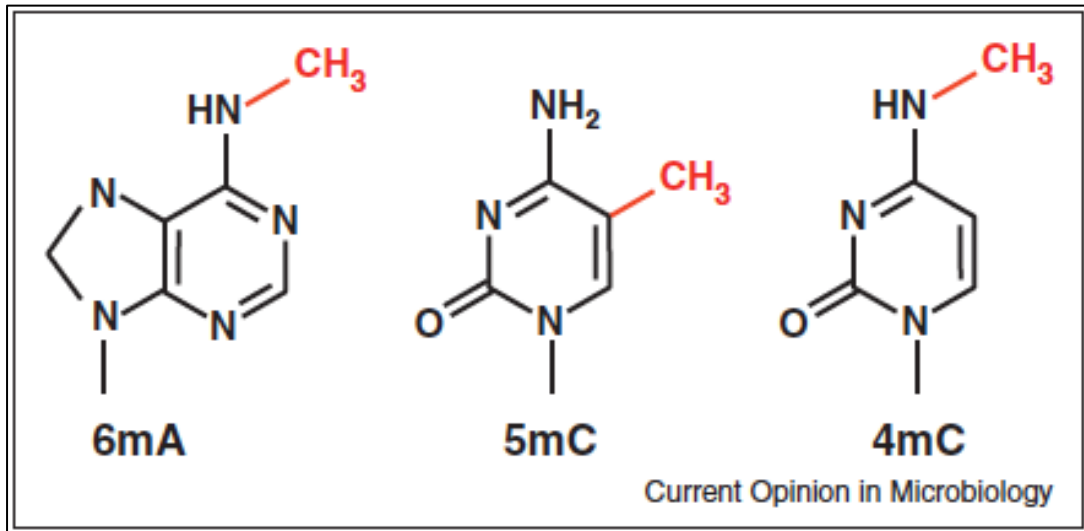
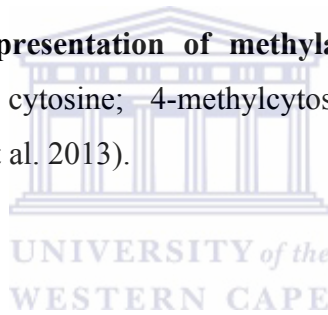


Figure 1.1 Graphical representation of methylated DNA bases. 6-methyl adenine (left); 5-methyl cytosine; 4-methylcytosine. The methyl group is highlighted in red. (Davis et al. 2013).



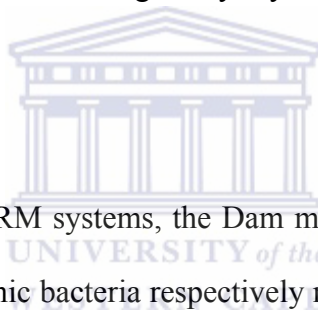
Bacteria exhibit flexibility in the number of MTases and their corresponding target sequence motifs, an observation that is common even in phylogenetically related species. However genomic studies have shown that a certain level of DNA methylation is conserved in prokaryotes, given that putative DNA MTases homologs have been found in at least 94% of over 5805 bacterial genomes sequences as of 2014 (Roberts et al. 2010). However, characterisation of these sequences hasn't been performed to establish the extent of this phenomenon. Similarly, there's lack of proper characterisation in other forms of DNA modifications such as phosphorothioation (Wang et al. 2011).

The two enzymes which play a critical role in prokaryotic methylation are the DNA adenine methyltransferase (Dam) and the cell cycle-regulated methyltransferase (CcrM), which are derived from the Gammaproteobacteria and the Alphaproteobacteria respectively. These two enzymes may have evolved from a common ancestral gene that lost its restriction domain and therefore adapted its function to epigenetic sequence modification (Marinus & Casadesus 2009). According to (Casadesús & Low 2006), the 6-methyladenine modification plays a role in prokaryotic gene expression by altering changes in DNA-protein interactions.

1.2.1.2 DNA methylation recognition sites and DNA Methyltransferases

In prokaryotes, MTases either constitute part of the host restriction-modification (RM) system or are independent of RM systems, for example Dam (Casadesús & Low 2006). RM systems occur in four main groups namely Types I, II, III and IV, with the first three classes being the most common. The Type I RM system are structurally composed of two R, two M and one S multisubunit proteins that function as a single complex. While the R subunit is primarily involved with the cleavage, the M subunit is what drives forward methylation. The S subunit determines the specificity for the recognition of the DNA sequence to be methylated. All studies have shown that the resultant methylation form is the m6A. Cleavage in Type I RM systems occurs at variable loci at far proximity from their recognition sites. In comparison, the Type II RM systems are able to identify specific recognition sites and the cleavage is performed always at constant loci

which are at close proximity to the recognition site, to yield the 5'-phosphates and 3'-hydroxyls. These systems may exist as monomers, dimers and occasionally as tetramers, utilizing the Mg^{2+} ions as the cofactor. Type III systems are made up of two protein coding genes, *mod* and *res*. Using Adenosine triphosphate (ATP) as an energy source, these two subunits are involved in DNA recognition, modification and restriction. On identification of the recognition site, cleavage occurs at a specific distance away from this locus. The final class is the Type IV RM systems, which have not been well characterised. They are made up of only one or two genes that cleave only modified DNA. Examples of such modified DNA include methylated, hydroxymethylated and glucosyl-hydroxymethylated bases (Roberts 2003).



In the case of independent RM systems, the Dam methylase and the CcrM in both Gamma- and Alpha-proteomic bacteria respectively mediate adenine methylation in prokaryotes. It involves the transfer of a methyl group to the adenosine moieties using S-adenosylmethionine as the donor. The DNA targets for this activity are the GATC and GANTC sequence motifs, for the Dam and CcrM respectively (Casadesús & Low 2006).

1.2.1.3 The role of DNA methylation in bacteria

Most DNA methylation studies have been performed on the model organism, *Escherichia coli*, whose DNA adenine methyltransferase (Dam) has been well studied (Barras & Marinus 1989; Marinus & Casadesus 2009; Casadesús & Low

2006). The specific mechanisms of methylation in *Mycobacterium tuberculosis* has not been fully established. According to (Jeltsch 2002), the molecular process of DNA methylation in bacteria primarily serves three functions which include 1) the ability to distinguish own and foreign derived DNA sequences, 2) assisting in DNA mismatch repair following replication and 3) guiding the replication machinery during the cell cycle. The ability to identify foreign DNA is performed through the RM systems and is at the core of defense and immunity in bacteria. The foreign DNA, which poses a threat, is identified by its methylation pattern subsequently cleaved off to eliminate the threat (Jeltsch 2002). In addition, previous studies have demonstrated that DNA adenine methyltransferases are involved in bacterial pathogenicity through the control of expression of virulence genes as well as the secretion of known virulence determinants (Low et al. 2001; Heithoff 1999). The mismatch repair system governs DNA replication fidelity. In principle, methylation does not occur on the newly synthesized strand, but rather mutations are checked on this strand and if found they are corrected on the nonmethylated template strand (Cooper et al. 1993).

Hemimethylation occurs when only one strand of a double-stranded DNA is methylated. In Dam methylation sites, this is informative as a direct indication of a completely replicated DNA molecule, and also as a label that marks the parental template. Given that certain bacterial gene promoters are activated only during the hemimethylation state, Dam methylation can be hypothesised to play a role in coupling DNA replication with the cell cycle (Barras & Marinus 1989). Due to the binding of the seqA protein, and therefore prevention of methylation, the origin of

DNA replication remains hemimethylated for the longest period. This provides a mechanism to control replication due to the inactive nature of the hemimethylated origins of replication (Jelsch 2002).

The role of DNA methylation in bacterial adaptation to a new environment has not been fully described. However, the adoption of different phenotypes through phase variation has been hypothesised as an outcome of gene expression that is controlled by RM associated MTases (Srikhanta et al. 2005; Srikhanta et al. 2011) .

1.2.1.4 Methylation and virulence in pathogenic bacteria

The association of DNA methylation with virulence control was initially established and described in *Salmonella enterica* serovar Typhimurium where the deletion of Dam methylase led to a decrease in the ability of the pathogen to colonize mice (Heithoff 1999). Since then, there has been extensive evidence that link adenine methylation to bacterial virulence. Studies based on model animals have shown virulence attenuation during the absence of Dam methylation in *Salmonella*, *Haemophilus*, and in some strains of *Yersinia pseudotuberculosis* (García-Del Portillo et al. 1999; Watson et al. 2004; Taylor et al. 2005). However, the association of virulence with Dam mutants is not a universal phenomenon as has been demonstrated in *Shigella flexneri* (Honma et al. 2004). Interestingly, other pathogens have shown the attenuation of virulence when there is an excess of Dam methylase (Julio et al. 2001). In *Haemophilus influenzae*, deletion of Dam leads to a decreased ability to invade host cells, as has been observed in *S. enterica* (Watson

et al. 2004). Furthermore, DNA methylation has been shown to influence the interaction of a bacterium with its host through regulation and modulation of the expression of surface proteins that are linked to virulence (Heusipp et al. 2007)

While bacterial virulence and methylation has been well characterised in major pathogens of medical importance, the same is lacking for Dam/Dcm methylation in *Mycobacterium tuberculosis*. A study by (Srivastava et al. 1981) has documented that there is no cytosine methylation in the avirulent *Mycobacterium tuberculosis* strain H37Ra, whereas methylation has been confirmed in the virulent H37Rv strain. In addition, (Hemavathy & Nagaraja 1995) have shown that DNA methylation is suspected to play a critical role in virulence since the level of methylcytosines in the virulent H37Rv was found to be higher than in the avirulent strain H37Ra.

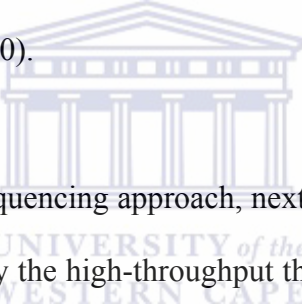
To date there has not been a rigorous study describing DNA methylation in *M. tuberculosis* and virulence using DNA sequencing technologies.

1.3 DNA Sequencing Technology

1.3.1 First and second generation DNA sequencing technologies

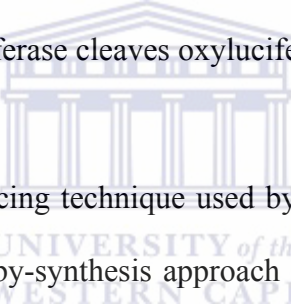
DNA sequencing is a molecular technique that is used to determine the order of nucleotides in a DNA molecule (Sanger et al. 1977). Over the course of its evolution, the technology has gone through three phases namely; first, second and the third generation. The first sequencing method, designated as the first generation, was developed by Frederick Sanger (1977) and was based on chain termination.

The method uses a single-stranded DNA template, a DNA primer, a DNA polymerase enzyme, normal deoxynucleosidetriphosphates (dNTPs), and modified di-deoxynucleosidetriphosphates (ddNTPs). The incorporation of ddNTPs by DNA polymerase terminates sequence elongation, based on the primer template, since they lack a 3'-OH group which is required for the formation of a phosphodiester bond between two nucleotides (Sanger et al. 1977) and the ddNTPs are usually radioactively or fluorescently labeled for rapid detection in automated sequencers. Based on its high accuracy and safety, Sanger sequencing was immediately adopted and commercialized, and still remains the gold standard for confirmatory *M. tuberculosis* (Merker et al. 2013; Kato-Maeda et al. 2013) and other mycobacteria sequencing (Wynne et al. 2010).



Unlike the first generation sequencing approach, next-generation sequencing (NGS) technology is characterised by the high-throughput that is achieved when massively parallel chemical reactions simultaneously amplify millions of DNA sequences. The key advantages of next-generation sequencing technology are the high throughput, which results in millions of short DNA fragments that range from 50 to 700 base pairs, and the subsequent reduction in sequencing cost (Niedringhaus et al. 2011). However, this compromises the quality of the obtained short-reads, when compared to Sanger sequencing (Niedringhaus et al. 2011). Currently, the main commercial vendors of NGS technology include Roche/454, Illumina and Life Technologies (Glenn 2011), and they all have the capacity to generate data of the order of giga basepairs (Gbp) per machine per day.

The 454 sequencing technology was implemented in 2006 by the 454 Life Sciences Corporation and subsequently applied in sequencing the Neanderthal genome (Noonan et al. 2006). The technology is based on the principle of pyrosequencing and has the capacity to generate up to 600 megabases per 10-hour run. In pyrosequencing, DNA fragments are nebulized and ligated to adapters, after which they are fixed onto beads and placed into a 29 micrometer well on a fibre optic chip referred to as a PicoTiterPlate. In this well, a mixture of enzymes such as luciferase, ATP sulfurylase and DNA polymerase are also added. The incorporation of a nucleotide by the DNA polymerase results in the release of a pyrophosphate molecule which in turn fuels a downstream set of chemical reactions and the production of light when luciferase cleaves oxyluciferin (Mardis 2008a).



Compared to the pyrosequencing technique used by 454, Illumina/Solexa Genome Analyzer uses a sequencing-by-synthesis approach (Metzker 2010), where all four fluorescently labelled nucleotides are added simultaneously into oligo-primed cluster fragments in flow-cell channels along with DNA polymerase. Fluorescent labelling is carried out with different colours for each of the four different bases. During each cycle in the sequencing protocol, the fluorescent colour which is specific to one of the four bases is automatically detected and read, allowing for sequence identification and the initialization of the next cycle until the entire DNA molecule is fully sequenced (Metzker 2010). Illumina sequencing offers several advantages. For instance, it is possible to sequence multiple strands in parallel due to the automated nature and hence massively boost the throughput. Furthermore, this technique employs only DNA polymerase as opposed to several, expensive

enzymes required by other sequencing techniques such as the 454 pyrosequencing (Metzker 2010). At present, the new Illumina HiSeq 2000 Genome Analyzer is capable of producing single reads of 2×100 basepairs (pair-end reads), and generates about 200 giga basepair (Gbp) of short sequences per run. The sequencing landscape is currently dominated by the Illumina sequencing technology due to its adaptability, superior read quality, relatively low cost and the ease of use (Zhang et al. 2011).

The principle behind the ABI-SOLiD sequencing technology differs from both 454 and Illumina in that oligonucleotides that are complementary to a series of bases in the template are ligated to a DNA molecule and the identity of the first two bases of the ligated oligonucleotide is specified by a flexible four color code with each color denoting four different dinucleotides. Given that each base in the template is interrogated twice in independent primer cycles, this technique offers some significant advantages in terms of accuracy. Furthermore, color reads can be converted into base reads if the first base of the sequence is known (Mardis, 2008a; Lui et al. 2012). Therefore, in resequencing applications, consideration of SOLiD sequencing data can facilitate differentiation between sequencing errors and biological artifacts such as SNPs. The drawback to this is that sequence errors result in major changes in downstream bases (Horner et al. 2010).

The Ion Torrent sequencing technique is based on the detection of the hydrogen ion that is released when a nucleotide is incorporated into a DNA sequence by the DNA polymerase (Niedringhaus et al. 2011). The release of hydrogen ions causes a

change in the pH within the microwell, which is subsequently detected using an ion-sensitive field-effect transistor. The Ion Torrent sequencing machine is the first commercial sequencing technology that does not use fluorescence and camera scanning. This results not only in smaller equipment, but also in an increase in sequencing speed, and low cost (Zhang et al 2011).

1.3.2 Third generation sequencing technology

The single-molecule real-time (SMRT) technology was developed by Pacific Bioscience (Menlo Park, CA, USA), and is the first 3rd generation sequencing approach to directly observe a single molecule of DNA polymerase while it synthesizes a strand of DNA in real time (Eid et al. 2009). This technology directly leverages on the speed and processivity of the DNA Polymerase to address many of the shortcomings of second generation sequencing (Schadt et al. 2010). The technology is based on a SMRT cell, which consists of millions of zero-mode waveguides (ZMWs), embedded with one molecule of DNA polymerase and one molecule of DNA template which is recorded during the process. During the reaction, the DNA polymerase adds a nucleotide into the complementary strand and cleaves off the fluorescent dye previously attached to that nucleotide (Figure 1.1). Inside the machine, there is a camera that captures fluorescent signal in real time and records it in a movie format. This facilitates the derivation of the interpulse duration (IPD) from the signal differences, a feature that is paramount in predicting structural variance in the sequence, and is widely applied in epigenetic studies such as the identification of DNA methylation. The rate at which the DNA polymerase

molecule incorporates each nucleotide into the growing strand varies depending on the type of modifications present on each base (Clark et al. 2012).

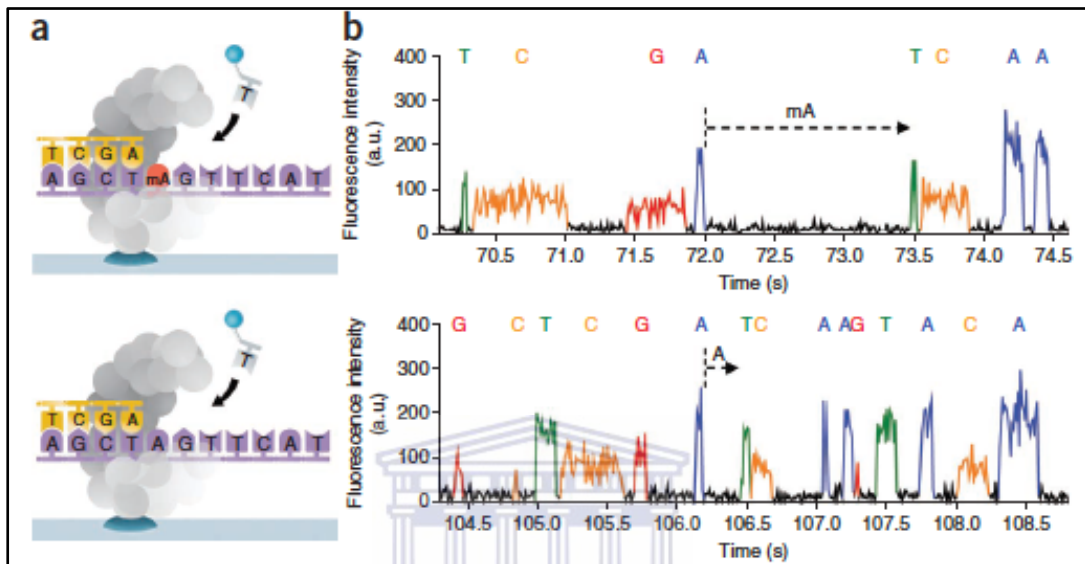
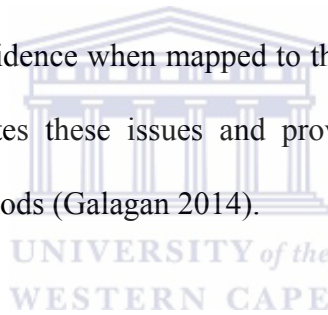


Figure 1.2. Detection of DNA modifications using PacBio SMRT sequencing. **a** represents the sequencing reaction with the thymine nucleotide incorporated into the growing strand corresponding to a methylated (top) and unmethylated adenine (bottom). **b** represents the SMRT fluorescence real time detection. The x-axis shows the time for the nucleotide incorporation and the y-axis represents the fluorescence intensity. The interpulse duration (IPD) represented by the dashed line shows a delayed incorporation of the thymine due to the methylated adenine (top) compared to a shorter time for an unmethylated adenine (bottom) (Flusberg et al. 2010).

PacBio SMRT sequencing offers four major advantages when compared to second generation sequencing; 1) rapid sample preparation which can be achieved in four to six hours instead of days, 2) elimination of errors and the bias that is associated with Polymerase Chain Reaction (PCR) amplification, 3) high turnover rate as the

runs can be completed within a day and 4) long average read length of up to 4000-6000 base pairs (bp), which is longer than any derived from the second-generation sequencers. Although the throughput of the PacBioRS is lower in comparison to the second-generation sequencers, the technology is widely applied in clinical laboratories, more so in microbiology research (Liu et al. 2012).

Second generation sequencing technology does not sequence long repeat regions effectively due to the size of the short reads (50 -250 base pairs). In *M. tuberculosis* insertion elements and highly repetitive regions such as the PE-PPE genes (Cole et al. 1998) are often not adequately sequenced due to the short reads resulting in gaps in coverage and lower confidence when mapped to the reference genome. The long reads from PacBio mitigates these issues and provides an advantage over 2nd generation sequencing methods (Galagan 2014).

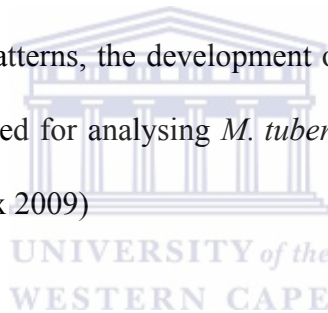


1.3.3 Application of DNA sequencing to bacterial genomics

Due to their small genomes, bacteria are easy and relatively cheap to sequence. Therefore, various studies that employ next-generation sequencing technology have been carried out ranging from outbreak patterns (Roetzer et al. 2013), drug resistance (Farhat et al. 2013), strain variation (Roetzer et al. 2013) and human microbiome metagenomics (Qin et al. 2010). Given that pathogens evolve continually driven by the acquisition of novel mutations, the sequencing of clinical isolates is of great importance especially if it can be analysed in the light of virulence markers, such as those conferring antibiotic susceptibility and resistance

(Mardis 2008b). For example, Nowrousian (2010) proposes that such information can be used to infer microbial evolution, identifying differentiating markers between pathogenic and nonpathogenic strains.

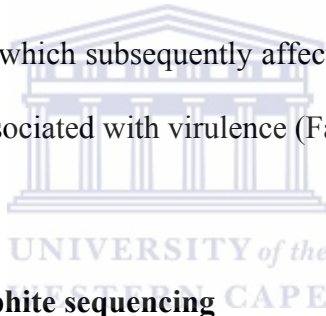
As of 2014, there are 23 complete and 43 draft *M. tuberculosis* genomes in the RefSeq database (Pruitt et al. 2007) at the National Center for Biotechnology Information (NCBI). Due to the success of *M. tuberculosis* and its ability to evade the immune system and also confer resistance to TB drug treatments, which presents an ongoing challenge, it is imperative to get a better understanding of this pathogen and identify its ability to persist with infection. In addition to studying emerging drug resistance patterns, the development of new drugs and discovery of biomarkers, prompts the need for analysing *M. tuberculosis* genomes (Mathema et al. 2006; Comas & Gagneux 2009)



The H37Rv laboratory reference strain was the first *M. tuberculosis* genome to be sequenced and annotated in 1998 by Cole and colleagues. To date, it is the most accurately annotated and well described strain of *M. tuberculosis* and is extensively used in various genomic studies as the reference strain of choice. The genome comprises of 4,411,532 base pairs, and approximately 4000 protein coding genes, and has a high GC content of around 65%. The H37Rv genomic and annotation information is stored in a public database TubercuList (<http://tuberculist.epfl.ch/>) (Lew et al. 2011). The sequencing of the whole genomic DNA of *M. tuberculosis* strain H37Rv in 1998 (Cole 1998) using Sanger sequencing was a turning point in

tuberculosis research, paving the way for the study of biology, metabolism and evolution of this pathogen (Metzker 2010).

In May 2011, during an ongoing outbreak of an exceptionally virulent shiga-toxin producing *Escherichia coli* O104:H4 in Germany, PacBio sequencing technology was employed not only in the genome assembly, but also to provide insight into the organism's methylation profile. More than 3000 people were infected with the new strain which was a hybrid of the entero-aggregative and the entero-hemorrhagic *E. coli*. The study revealed that the new strain had been infected with a bacteriophage leading to the acquisition of a shiga-toxin encoding gene and a new restriction modification (RM) system, which subsequently affected the transcription of several genes some of which are associated with virulence (Fang et al. 2012).



1.3.4 Whole genome bisulphite sequencing

Bisulphite genomic sequencing is the gold standard in methylation analysis. This method is used to identify methylated cytosines and has mostly been used to study cytosine, and specifically the CpG methylation patterns in eukaryotes. (Lister et al. 2008; Lister et al. 2009; Meissner et al. 2008). In *E. coli*, using whole genome bisulphite sequencing, methyl cytosines have been studied during different cell cycle stages and found to be regulators of stationary phase gene expression (Kahramanoglou et al. 2012). Currently, next-generation sequencing, coupled with sodium bisulphite modification of DNA, has proven to be a powerful method in the analysis of DNA methylation at single nucleotide resolution (Chatterjee et al. 2012).

In whole genome bisulphite sequencing, genomic DNA is treated with sodium bisulphite which converts unmethylated cytosines to uracil. The uracils are read as thymines by DNA polymerase, so amplifying bisulphite-treated DNA by PCR yields products where unmethylated cytosines appear as thymines. After subjecting the bisulphite-treated genomic DNA to next generation sequencing, the methylation state is determined from the sequence alignments by comparing the modified DNA with its original sequence (Krueger et al. 2012).

Bisulphite sequencing has some disadvantages in that 1) the treatment and sample preparation steps are time-consuming and costly, 2) DNA degradation can occur due to the bisulphite chemical treatment, 3) the complexity of the genome is reduced due to the conversion of all cytosines to thymines, which affects the mapping to the reference genomes, 4) methylcytosine and hydroxymethylcytosine cannot be discriminated and 5) this method is only restricted to identifying one type of nucleotide (cytosine) methylation (Flusberg et al. 2010). PacBio SMRT sequencing expedites this in that it directly sequences and identifies methylated bases without the need for prior bisulphite treatment, thereby providing a direct measurement.

1.4 Bioinformatics and NGS analysis

1.4.1 Data generation and format

Since its first description by Frederick Sanger in 1971, DNA sequencing (Sanger et al. 1977) has gone through three technological phases, also termed as generations, with each phase improving upon the previous one. This has led to huge amounts of data being generated for analysis and biological interpretation (Mardis 2008b). Recently, technical advancement in DNA sequencing has outpaced Moore's Law, a prevailing phenomenon that describes the trend observed in the development of computer hardware, whereby every two years the number of transistors that can be fitted on a single Integrated Circuit (IC) doubles (Dewitt et al. 2012). This calls for development of computational tools, algorithms and databases for analysing and managing the massive data which is currently posing a challenge for bioinformatics and computational biology (Zhang et al. 2011). Some of the areas of interest include 1) sequence alignment and mapping, 2) genome assembly 3) data storage and retrieval (Ansong 2009). The main disadvantages arising from the second generation sequencing platforms is the short read-length and decreased sequencing accuracy, compared to conventional Sanger sequencing. These pose significant algorithmic challenges for downstream analysis (Shendure & Ji 2008).

The base sequence in both first and second generation sequencing is as a result of decoded light signals through a process referred to as base-calling. Although this is dependent on the specific technology in use, each called base is associated with a probability score of accuracy referred to as the quality score. One of the widely

used quality scoring system is the phred score (Q), which is expressed mathematically as $Q = -10 \log_{10} P$, where P is the base-calling error probability (Ewing et al. 1998; Ewing & Green 1998). Therefore, a quality score of 20 corresponds to an error probability of 1 in 100. Different file formats have been adopted in the representation of both sequences and their corresponding quality scores, with the FASTQ being the most widely applied format by the next generation sequencing platforms. The first challenge encountered in the analysis of sequencing data is the file conversion as most software are designed to take data that is in a specific format (Nowrousian 2010).

Data obtained from the Pacific Biosciences (PacBio) sequencing technology (Eid et al. 2009), which is currently the representative platform for the third generation sequencing technology, differs significantly from the FASTQ formats of the first and the second (NGS) technologies in that 1) PacBio raw data in movie format and 2) sequencing information is stored in Hierarchical Data Format (HDF5) files, which contain information about the base calls, quality scores and meta data such as the polymerase kinetics.

1.4.2 Data analysis and algorithms

Other than the *de novo* assembly of sequenced genomes and genome fragments, genome mapping is a huge and often challenging area in the computational analysis of sequencing data. The overall objective is to align the sequenced reads to a known reference genome and determine all the regions of similarity and differences

between the two datasets (Ruffalo et al. 2011). One of the challenges in mapping is in the determination of the mismatches that can be allowed for any given read to be considered as aligned to a specific region on the reference genome, since mismatches may arise not only being biological artifacts, but rather as sequencing errors. The presence of repeats in the sequenced reads also poses a huge challenge in data analysis as they erode specificity and a single read can be mapped on to different regions of the reference (Nowrousian 2010). The principle underlying mapping software is the computational unique k-mers and their indices in either the reference genome or the reads. Several criteria have been used as a guide during the design of the mapping algorithm, for example the read quality scores (Nowrousian 2010).



Many of the mapping software are specially designed for the mapping of short reads generated by second generation sequencing technologies but not for long reads. Specialized mapping algorithms have been designed to deal with the long reads and movie format derived from the PacBio sequencing technology. The Basic Local Alignment with Successive Refinement (BLASR) algorithm is primarily used in the mapping of PacBio reads. The program searches for at least 10 kmers of size 15 using a Burrows-Wheeler transform algorithm and then uses them to anchor the read onto the reference sequence. Once the reads have been anchored, a dynamic alignment algorithm is employed to align the entire read (Chaisson & Tesler 2012). In addition to BLASR, PacBio has implemented a general-purpose data analysis platform, called SMRT Pipe, which performs a wide array of functions such as

logging and data management, parallel computation, error management, parameter optimization, and storage of temporary files.

1.5 Rationale

Molecular analysis of numerous *M. tuberculosis* clinical isolates has revealed that the species exhibits a high level of similarity at the genetic level with all strains showing greater than 99% genetic similarity (Hershberg et al. 2008). However, this level of genetic similarity does not necessarily translate to phenotypic homogeneity. Based on the numerous inter- and even intra-strain differences, previous studies have shown that some *M. tuberculosis* clones may undergo a virulence enhancing genetic alteration that potentially lead to occurrences of tuberculosis outbreaks (Hanekom et al. 2007; Nicol & Wilkinson 2008). In addition, van der Spuy et al (2009) has suggested that the Beijing strain family may possess an increased virulence capacity over other strains. Furthermore the Beijing genotype is the dominant strain in the Western cape, South Africa (Hanekom et al. 2007). In a study by de Souza et al. (2010) involving two strains that are closely related members of the Beijing strain sublineage 7 family, a virulence characterisation analysis highlighted significant differences between the two. While one strain was responsible for only a single case of disease, the other one was responsible for 147 cases. The strains have been designated the terms ‘hypo-virulent’ and ‘hyper-virulent’ respectively, and their virulence characteristics have been confirmed in a mouse model by (Souza et al. 2010).

DNA methylation is a largely untapped area of research in *M. tuberculosis* and has been poorly described in the literature especially given its connection to virulence. However, it has been extensively studied along with its role in virulence in other pathogenic bacteria such as *E.coli*. To date only three studies (Srivastava et al., 1981; Hemavathy and Nagaraja, 1995; Shell et al., 2013) have identified DNA methylation (either m6A and/or 5mC) in *M. tuberculosis* H37Rv, but none have elucidated genome-wide methylome characterisation for the Beijing strains which possess an enhanced virulence capacity over other *M. tuberculosis* strains.

Previous attempts to identify 5mC using bisulphite Illumina sequencing in both of these strains did not reveal any 5mC. Since 5mC was not detected in these strains, this led to the question if DNA methylation occurred on another nucleotide other than cytosine. With the recent availability of PacBio sequencing technology with applications in bacterial methylome analyses, the DNA methylation profiles of the *M. tuberculosis* hyper and hypo-virulent strains were investigated. The aims are as follows:

1. To characterise a global DNA methylation profile for hyper-virulent and hypo-virulent *Mycobacterium tuberculosis* Beijing strains from single molecule real time sequencing data by:

- 1.1 Identification and characterisation of DNA methylated bases viz. N4 methylcytosine and N6 methyladenine.

1.2 Identification of methylated sequence motifs and their subsequent methyltransferase enzymes and RM systems.

1.3 To compare the unique and shared methylated genomic loci of the hyper- and hypo-virulent strains with the aim of understanding the different disease phenotypes caused by these strains.

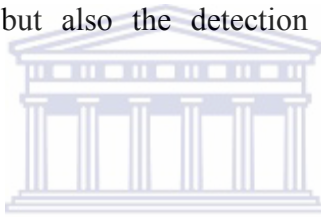
2. To computationally analyse adenine methylation in promoter regions.



CHAPTER 2: MATERIALS AND METHODS

2.1 Single Molecule Real Time DNA Sequencing

Two strains of *Mycobacterium tuberculosis* Beijing genotype sublineage 7; 1) hyper-virulent and 2) hypo-virulent strain, were sequenced using the Pacific Bioscience (PacBio) Single Molecule Real Time (SMRT) sequencing technology (Flusberg et al. 2010), in order to identify DNA methylation and other base modifications. The SMRT sequencing technology is ideal since it enables not only the sequencing of DNA, but also the detection of various other DNA base modifications.



The DNA samples were prepared as per the standard SMRTBell template protocol (Travers et al. 2010) generating a 1 kb insert size library. The samples were subjected to the C2 chemistry on the PacBio RS sequencing machine. This experiment was carried out at GATC Biotech in Germany. A total of 5 SMRT cells was used, yielding a sequencing coverage of 50x, 25x per strand of DNA with 2 sets of 45 minute movie protocol. The PacBio raw reads were downloaded from the GATC Biotech web portal and bioinformatics analysis was performed as outlined in the next section.

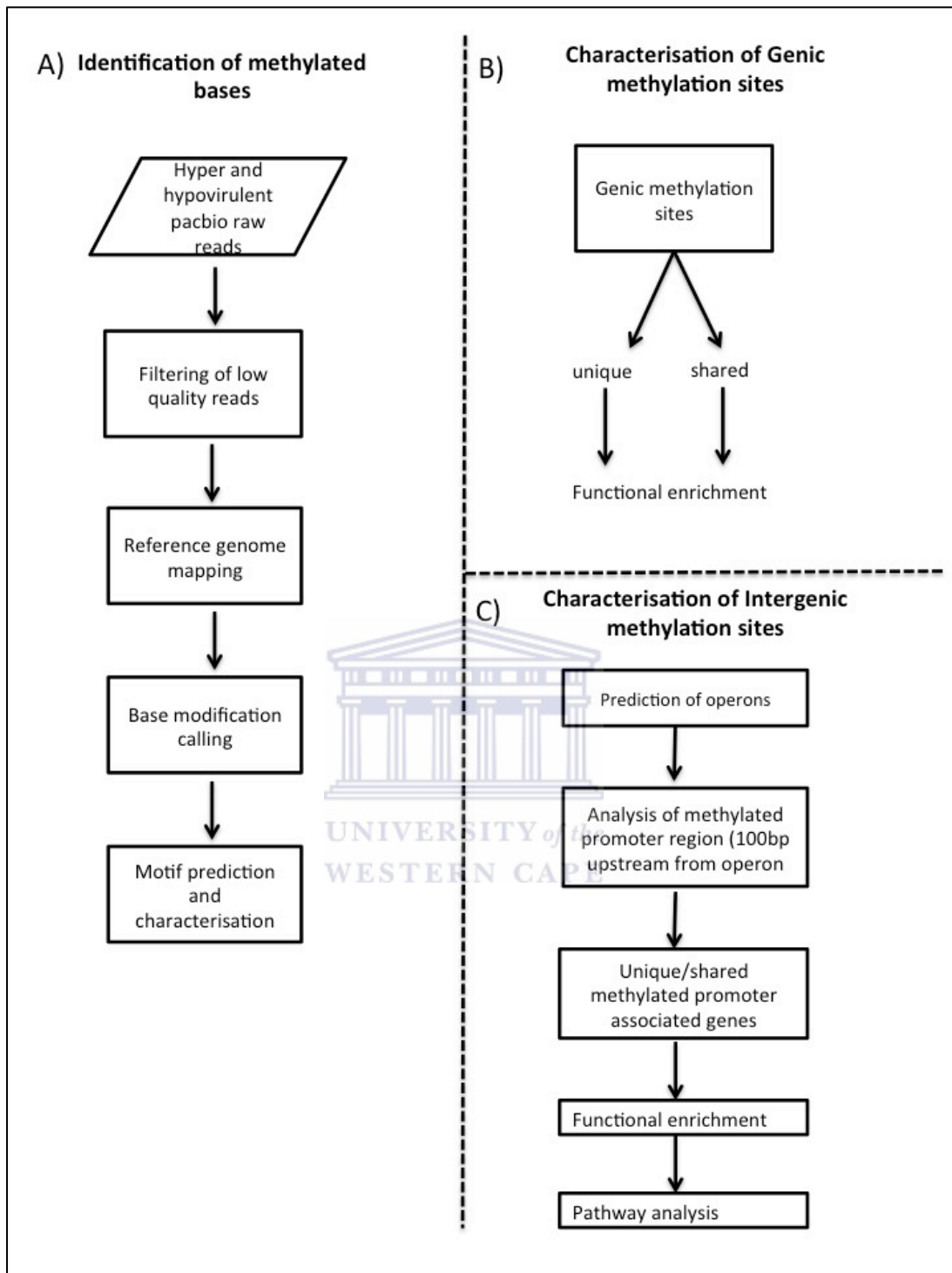
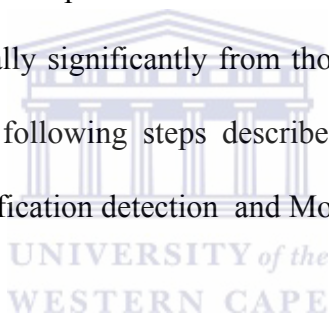


Figure 2.1 Overview of data analysis. A) Represents the overview of the methodology for the identification of methylated DNA bases from the PacBio reads. B) Following identification of methylation sites, characterisation of the genic positions of methylated loci in uniquely and shared methylated positions between the hyper and hypo-virulent strains. C) Represents the promoter region methylation analysis and subsequent enrichment of unique and shared methylated promoter associated genes.

2.2 Computational Analysis of the PacBio sequencing data

The SMRT analysis platform, version 1.3.3, was downloaded from the PacBio web portal (<https://github.com/PacificBiosciences/SMRT-Analysis/>). Data analysis was performed using The RS Modification and Motif Analysis protocol, with the default parameters using the graphical user interface. The SMRT analysis software is memory intensive and functions as a standalone tool thereby utilising all the memory and hard drive capacity on the server. The minimum disk space requirement is 250 GB per node.

The protocol analyses the interpulse distances to find bases with incorporation kinetics that differ statistically significantly from those of unmodified bases. This protocol comprises of the following steps described below: Quality filtering of reads; Mapping; Base Modification detection and Motif finding.

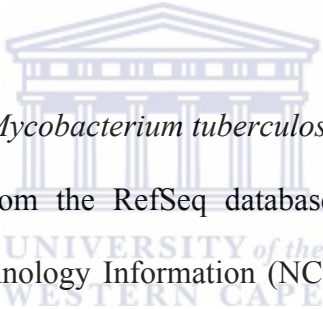


2.2.1 Quality filtering of Raw Reads

Quality filtering of PacBio Raw reads which focuses on; 1) adapter removal, 2) read-length settings, 3) subread length settings and 4) read quality scores. In this analysis, all the adapters were cleaved, while both the minimum read length and the subread length were set at 50. In the read quality assessment, the minimum cutoff score was set at 0.75.

2.2.2 Mapping to the H37Rv reference genome

Read mapping was performed using the Basic Local Alignment with Successive Refinement (BLASR) (Chaisson & Tesler 2012) software which is incorporated into the SMRT Analysis software portal. The BLASR algorithm is specially designed to cater for the long reads that are generated by the PacBio sequencing protocol. Prior to mapping the software filters out low quality reads which are usually characterised by relatively shorter lengths of less than or equal to 50 nucleotides, and are as a result of premature read termination during the sequencing.



The latest assembly of the *Mycobacterium tuberculosis* H37Rv genome (Cole et al. 1998), was downloaded from the RefSeq database (Pruitt et al. 2012) at the National Center for Biotechnology Information (NCBI) and used as the reference genome in read mapping. This procedure evaluates the maximum number of matches of each read to the reference sequence. The maximum divergence, which defines the allowed flexibility of the read in mapping to the reference, was set at 30%. The minimum anchor size was set at 12, implying that at least 12 base pairs in the read must match to the reference.

2.2.3 Base Modification calling

This step involves the identification of modified bases, as well as sequence motifs. In order to detect any base modification, the Mapping QV score, which represents

the confidence with which a read maps uniquely to the selected genomic interval, was set at 10 as the cutoff. Reads that fall within repeated genomic segments are automatically assigned low Mapping QV scores, and therefore subsequently removed from the analysis, so as to avoid erroneous base modification call as a result of mismatched reads.

The base modification and identification step uses an *in silico* kinetic reference and a *t*-test-based kinetic score to detect modified base positions within a genome. In this analysis, the log-transformed *p* value from the *t* test $-10 \log(p\text{-value})$, which is used as the kinetic score, was set at the threshold value of 100. Therefore, each reference base position that met this threshold was identified as methylated. In the identification of motifs, a minimum score of 40 was considered for assigning motifs to any given modified base. Based on the Inter-pulse Duration (IPD) ratio, and the specific kinetic signatures, different types of DNA base modifications were determined. These include; 1) 4-methylcytosine (4-mC), 2) 6-methyladenine (6-mA) and 3) modified base.


Once the methylation sites were predicted, further bioinformatics downstream analysis was carried out in order to characterise the methylome for the two strains and to determine their methylation patterns.

SMRTView® software from Pacific Biosciences was used to view the analysis and methylation sites and mapping results.

2.3 Genome-wide methylome characterisation

The methylation coordinates were separated by their strand orientation occurrence and analysed as described hereunder. The number of unique and shared m6A and m4C methylation sites between the hyper and hypo-virulent strains was determined. The methylation sites were also analysed for the genomic features they specifically overlap with.

2.3.1 Identification and characterisation of methylated recognition motifs and methyltransferases



Only confident methyl sites with a modification QV (confidence score) of 40 were selected and their sequence motifs further analysed. Subsequently, the predicted motifs sequences were searched for in the published Restriction Enzyme Database <http://tools.neb.com/~vincze/genomes> (REbase; Roberts et al. 2010), using the *M. tuberculosis* whole genome sequence as the query. REbase contains detailed information on prokaryote restriction enzymes, DNA methyltransferases and related proteins involved in the biological process of restriction-modification. This analysis yielded known methyltransferases and Restriction Modifications (RM) systems specific to the *Mycobacterium tuberculosis* H37Rv. In addition, genes encoding the methyltransferase enzymes in *M. tuberculosis* were searched for, in the REbase database.

2.3.2 Genic methylation Analysis and characterisation

Once the methylation sites were separated by their strand occurrence, the next step was to determine which genomic features they overlap in the H37Rv genome. The H37Rv annotation file, which contains all the genomic annotations in General Feature Format (GFF), was downloaded from the RefSeq database (Pruitt et al. 2012). In order to determine the genomic features where methylation sites overlap, *intersectBed* from the Bedtools sequence analysis software (Quinlan & Hall 2010) was used. The input files were the list of methylation genomic positions and the H37Rv genome annotation, both of which were in GFF format .

(intersectBed -wb -a input.txt -b annotation.gff >output.txt)

Once a list of genic methylation sites was determined for the hyper and hypo-virulent strains, a functional enrichment analysis using the Tuberculosis Database (TBDB) (Reddy et al. 2009), was performed in order to determine which functional categories were over represented. The Clusters of Orthologous Groups (COG) categories (Tatusov et al. 2003) were used as the functional references. This was run for the uniquely methylated genes in the hypo and hyper-virulent strains, as well as the shared methylated genes.

2.3.3 Intergenic DNA methylation Analysis

2.3.3.1 Analysis of methylated promoter regions

The overall aim was to computationally analyse DNA methylation occurring in promoter regions of operons.

2.3.3.1.1 Operon prediction and organisation

Genes in an operon are expressed together because they are co-transcribed or under the regulation of the promoter of the head gene in the operon (Mcguire et al. 2012). Hence it is important to consider operons and promoters of operons. The aim was to characterise the intergenic methylation sites, specifically in the promoter region of the operon, which may affect gene regulation.

The Database of Prokaryote Operons (DOOR2) online operon predictor software (Mao et al. 2009) <http://csbl.bmb.uga.edu/DOOR/annotate.php> was used to predict the operons for the *Mycobacterium tuberculosis* H37Rv genome.

The input files uploaded for the *M. tuberculosis* H37Rv genome were the gene location file in NCBI protein table format; the protein sequences and the genome sequence both in fasta format. The output is a tab-delimited file containing the list of predicted operons organised by their genomic positions and annotation information based on the COG category.

2.3.3.1.2 Comparative analysis of methylated promoters between hyper and hypo-virulent strains

The DOOR2 analysis, as described in section 2.3.3.1.1 , yields operons which are clusters of genes transcribed into a single polycistronic mRNA. The largest obtained operon was comprised of 6 genes. Promoter region's coordinates were identified for each operon by subtracting 100 base pairs upstream of the first gene in the operon. In order to identify the methylation sites falling within the predicted promoter regions, the m6A coordinates were compared to those of the target promoter region. If the methylation sites were i) greater than or equal to the promoter start coordinate or ii) less than or equal to the promoter end coordinate, they were assumed to be associated with the promoter region. Methylation sites that fell outside of this region were excluded. This was performed for both the hyper and the hypo-virulent *M. tuberculosis* strains to generate two lists of methylated promoter regions. Thereafter, the unique and the shared promoter regions, between the hyper and the hypo-virulent strains, were identified through direct comparison using the Unix 'cmp' command. The genes of these promoters were then further interrogated for their functional annotation using the Tuberculist web utility at <http://tuberculist.epfl.ch/> .

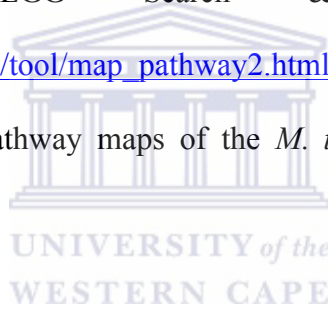
2.3.3.2 Functional enrichment analysis of the methylated promoter operons

Functional enrichment of these genes was carried out using the TBDB gene enrichment online tool, to determine which functional categories were

overrepresented. The TBDB web server also uses the Gene Ontology (GO) terms and Clusters of Orthologous Genes (COG) categories to augment the functional enrichment. A corrected p -value of 0.05 was used as the cutoff.

2.3.3.3 Metabolic Pathway Analysis

The uniquely methylated genes for the hyper and hypo-virulent strains, were mapped onto the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Ogata et al. 1999) in order to gain insight into their biological functions and relevance. The KEGG Search &Color Pathway tool http://www.genome.jp/kegg/tool/map_pathway2.html was used to map the list of genes on to the KEGG pathway maps of the *M. tuberculosis* H37Rv reference pathways.



The methylation genomic loci in GFF format were visualized in a local version Gbrowse genome browser and SMRT®View.

CHAPTER 3: RESULTS AND DISCUSSION

Overview

In this study, a comparative approach was employed in the analysis, and characterisation of the methylomes of two closely related *Mycobacterium tuberculosis* strains. The two strains, the hyper- and the hypo-virulent, belong to the same Beijing sublineage 7, and exhibit vastly different clinical phenotypes. The goal was to determine if there were any differences in DNA methylation patterns between the two strains that could explain the observed phenotypic differences. To date, only one study (Shell et al., 2013) has identified DNA methylation using NGS approaches (5mC and m6A) in *M. tuberculosis* despite its clinical importance as a pathogen. Furthermore, there have been no studies focused primarily on Beijing genotype strains which are predominant in the Western Cape region of South Africa. The data in this research was generated using the most recent single molecule sequencing technology available, Pacific Biosciences. A computational approach, as described in the methods chapter, was formulated and applied to enable the identification of the DNA methylation patterns, as well as their functional annotation.

This chapter presents the results, first by the prediction of the DNA methylation profiles for each of the two genomes, followed by the subsequent downstream analyses and characterisation of the identified methylated loci. Accordingly, this study provides a comprehensive genome-wide methylation analysis of the hyper- and hypo-virulent *M. tuberculosis* Beijing genotype strains.

3.1 Description of methylation patterns

3.1.1 Mapping hyper-virulent and hypo-virulent *M. tuberculosis* Beijing genotype strain genomes to the reference *M. tuberculosis* H37Rv genome

The SMRT sequencing platform generated reads with an average length of 2,461 base pairs (bp). The reads, which resulted from the adapter sequence trimming, are known as the ‘subreads’, which had an average length of 562 and 568 bp for hypo- and hyper-virulent strains respectively, were obtained (Table 3.1).

The H37Rv laboratory strain is the most accurately annotated *M. tuberculosis* genome to date and was appropriately selected since the downstream methylation analysis is dependent on accurate annotation and biological interpretation.

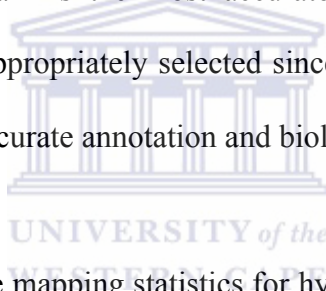
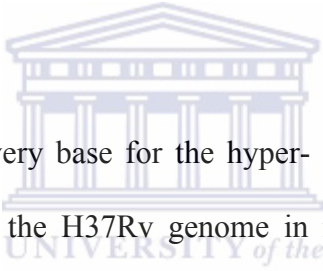


Table 3.1 A summary of the mapping statistics for hyper-virulent and hypo-virulent genomes

	Hyper-virulent	Hypo-virulent
Mean Mapped Subread Accuracy	87.54%	86.9 %
# of Reads	231363	135729
# of Mapped Reads	209668	119955
# of Mapped Subreads	779104	439319
Mean Mapped Readlength	2488 bp	2461 bp
Mean Mapped Subread Readlength	568 bp	562 bp
Mean Depth of coverage of H37Rv	96	53.69

The tabulated results (Table 3.1) show that the hypo-virulent genome was sequenced at a lower coverage, which resulted in a lower number of reads compared to the hyper-virulent strain. However, the same number of SMRT cells were used during the sequencing of both genomes. This implies that there was some loss in starting DNA material during the library preparation step (C König, personal communication).

3.1.2 Identification and characterisation of methylomes in hyper-virulent and hypo-virulent strains



The methylation state of every base for the hyper- and hypo-virulent strains was determined with respect to the H37Rv genome in real time by measurement of DNA polymerase kinetics during the sequencing reaction. This ensures the identification of DNA base modifications at the single base pair resolution.

The real time measurement of kinetics of the DNA polymerase enables the detection of modified DNA bases, which could be in the form of damaged DNA bases due to oxidative stress or epigenetic changes such as methylated DNA bases. The rate at which the polymerase incorporates each nucleotide is slowed down when encountering a modified base as opposed to an unmodified base. This difference in time between nucleotide incorporations is the interpulse duration (IPD) ratio, thus identifies a base as modified. Methylated bases such as N6 methyladenine and N4 methylcytosine have distinct kinetic signatures on the

polymerase within a 12 bp window and hence are detected accordingly to the *in silico* computational reference model (Figure 3.1). (Clark et al. 2012)

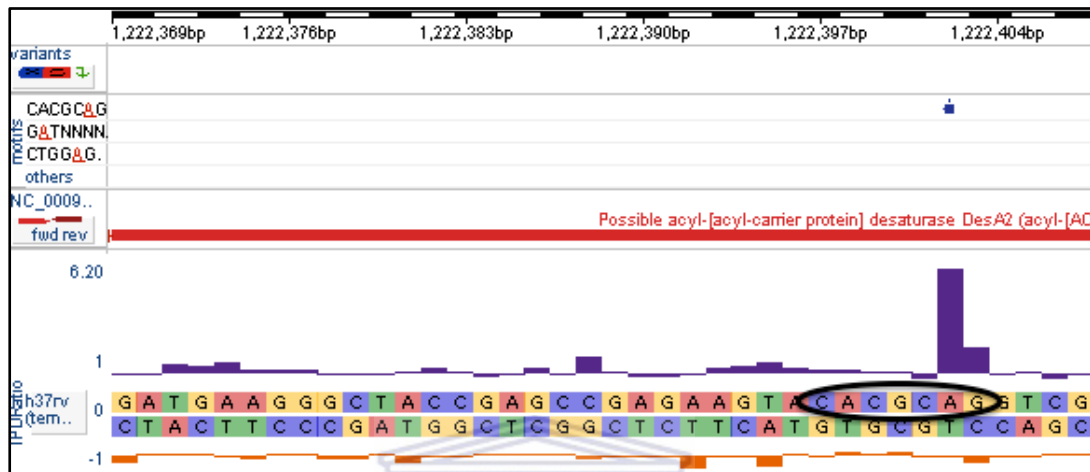


Figure 3.1 An illustration of the IPD ratio of 6.20 (purple block) for a methylated adenine within the predicted motif CACGCAG in the forward strand. The methylated adenine also occurs within the *OmpA* gene.

Base modifications, based on the kinetic signature of a specific modification type, were identified through comparison to an *in silico* control reference, for every position in the H37Rv genome. The PacBio sequencing detected 3 types of bases modifications namely, N6-methyladenine (m6A), N4-methylcytosine (m4C) and modified base, which refers to any unidentified bases other than the specific modified types mentioned above. These were excluded from the analysis due to their non-specific nature. The modification QV (ModQv) is the statistical phred-like score that quantifies the extent to which a base is correctly predicted as

modified. A score of 20 denotes a 99% chance that the base is modified and also corresponds to a p value of 0.01.

The IPD ratio for a methylated position is the average IPD ratio of all the molecules at that position which includes non-methylated and methylated strands. A heterogeneous sample containing a mixture of methylated and non-methylated strands of DNA hence will appear to have partial methylation at that site and will also have a lower modification Qv. (C Konig, personal communication).

The total number of methylated bases and their frequencies identified in hyper-virulent and hypo-virulent strains are summarized in Table 3.2.

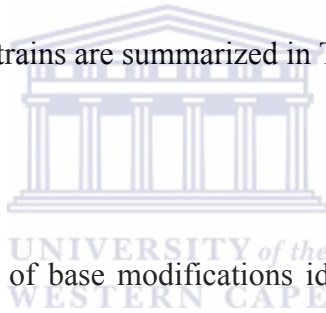


Table 3.2 The composition of base modifications identified in hyper-virulent and hypo-virulent strains. The percentages are indicated in parentheses.

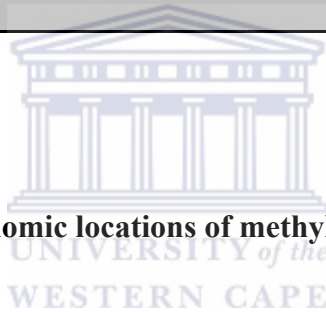
	Hyper-virulent	Hypo-virulent
m6A	1946 (2.3%)	1825 (4.6%)
m4C	3049 (3.7%)	2011 (5.4%)
Modified base	81119 (94%)	35779 (90%)
Total	86187	39672

As seen in Table 3.3, the m4C has a low methylation signal based on the ModQv. The m6A sites that do not occur within predicted motifs, have a lower Modification

Qv compared to the sites within motifs. The low modification score could be due to the low sequencing coverage or sequencing errors in these regions.

Table 3.3 A summary of the coverage and ModQv of m6A and m4C in hyper-virulent and hypo-virulent strains.

	Hyper-virulent	Hypo-virulent
Mean coverage m4C	47	27
Mean ModQv m4C	31	24
Mean coverage m6A	47	26
Mean ModQv m6A	40	29
Mean ModQv m6A with motif	80	50



3.1.2.1 Pinpointing the genomic locations of methylation

Figure 3.2 summarises the frequency of methylation sites occurring in genomic features. Out of the total methyladenines identified, 38.2% and 40% of these sites occur within genes for hypo-virulent and hyper-virulent strains respectively. Both strains have 44.8% methylcytosines out of a total of all methylcytosines occurring in genes (Figure 3.2). The remainder, which is the majority of the methylation, occurs in intergenic regions as shown in Figure 3.3.

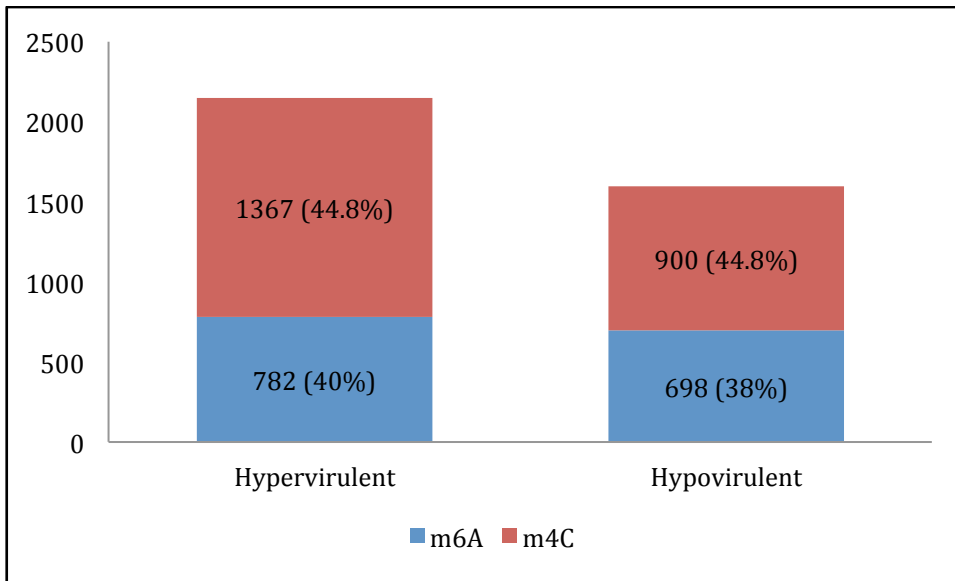


Figure 3.2 Frequency of genic methylation sites in hyper-virulent and hypo-virulent strains for both m6A and m4C sites.

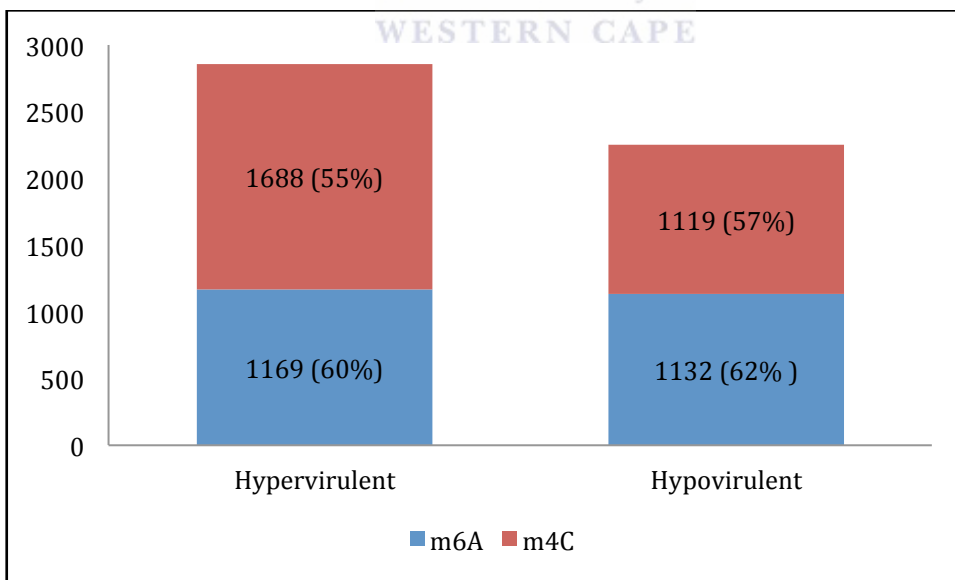


Figure 3.3 Frequency of intergenic methylation sites in hyper-virulent and hypo-virulent strains for both m6A and m4C sites.

The majority of the intergenic methyl adenines have 897 shared positions between hyper-virulent and hypo-virulent strains, with 272 occurring uniquely in the hyper-virulent strain and 235 unique sites for the hypo-virulent strain.

A total of 980 and 939 methyl adenines were detected by the software in the negative strand for the hyper-virulent and hypo-virulent respectively (Table 3.4). There were 965 and 885 methyl adenines detected in the positive strand for hyper-virulent and hypo-virulent strains respectively. A Fishers statistical test did not show any overrepresentation of strandedness for total adenine methylation, in both the hyper- and hypo-virulent strains.

Since this is a comparative analysis between two strains, a unique and shared methylation approach between the hyper- and hypo-virulent strains, was used in the following reported analyses. As a general trend, the shared methylation sites occur mostly within the predicted motifs. Most of the m6A methylation sites are shared between the hyper- and hypo-virulent with only a small percentage of the total m6A sites being unique to either strain.

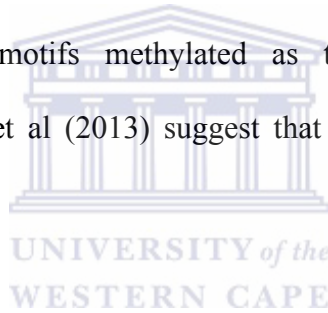
Table 3.4 Stranded methylation in hyper-virulent and hypo-virulent strains

	Hyper-virulent		Hypoviulent	
	Neg strand	Pos strand	Neg strand	Pos strand
m6A genic	407 (42%)	375 (39%)	360 (38%)	338 (38%)
m6A intergenic	577 (58%)	593 (61%)	582 (62%)	550 (62%)
Total	980	965	939	885

3.1.3 Summary

N4-methylcytosine is usually targeted by methyltransferases which belong to restriction modification systems. Although less understood than m6A RM systems, m4C methylated bases can be detected using the PacBio sequencing technology.

The m4C had low Modification Qv scores, which represents the confidence that a base is modified (Table 3.3), despite having a consistent coverage as that of the m6A. It can be deduced that the m4C methylation signal is weaker in comparison to that of m6A signal in the hyper- and hypo-virulent genomes. Murray et al (2012) have seen a similar result with the detection of m4C in *Bacillus cereus* having less than half of the m4C motifs methylated as targeted by their respective methyltransferases. Davis et al (2013) suggest that this might be due to a weak intensity of the m4C signal.



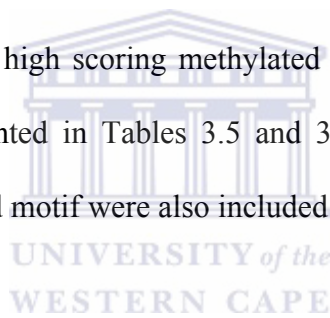
m4C bases was identified in the both the strains, albeit at a low confidence. This had not been previously described in *M. tuberculosis*. However, we did not proceed with further downstream analysis due to the low signal, which could be interpreted as false positives (C Konig, personal communication).

The following sections focus on an in-depth m6A characterisation and annotation between the hyper-virulent and hypo-virulent strains.

3.2 Methylated sequence motifs

DNA methylation in bacteria is carried out by an enzyme known as a methyltransferase which confers a methyl group to the nucleotide base. These methyltransferases identify a recognition sequence or motif in which the respective base is methylated. Methyltransferases belong to a Restriction Modification (RM) system along with its cognate restriction enzyme which functions as the bacteria's defence against foreign viruses among other roles. Hence it is imperative to identify these motifs, which provide information on the type of restriction modification system involved.

The 4 predicted motifs for high scoring methylated positions with a Modification Qv cutoff of 35 are presented in Tables 3.5 and 3.6. The methylation sites not identified within a predicted motif were also included in the downstream analysis.



There was a higher percentage of motifs that were methylated in the hyper-virulent genome compared to hypo-virulent genome, (Tables 3.5 and 3.6). Out of 1945 m6A sites, 1441 occur in predicted motifs for the hyper-virulent strain, while the other 505 (26%) are independent of the predicted motifs. On the other hand, 1291 m6A sites in the hypo-virulent strain occur in predicted motifs, out of a total of 1824 m6A sites. The other 535 sites, which translates to 42%, are not associated with any of the predicted motifs. Therefore, the observed hypo-virulent methylation signal is weaker given the low coverage and the low modification scores, when compared to the hyper-virulent strain (Tables 3.5 and 3.6).

Table 3.5 Summary of methylated sequence motifs in the hyper-virulent strain.

Motif	Modified	%	# of	# of	Mean	Mean	Partner Motif
	Position	Motifs	Motifs	Motifs	Modification	Motif	
		Detected	Detected	in	QV	Coverage	
				Genome			
CACGCAG	6	95.24	781	820	82.1	45.3	
GATNNNNRTAC	2	87.05	316	363	84.1	47.4	GTAYNNNNATC
GTAYNNNNATC	3	84.57	307	363	78.9	47.8	GATNNNNRTAC
CTGGAGGA	5	24.11	27	112	52.2	56.0	

Methylated base highlighted in bold

Table 3.6 Summary of methylated sequence motifs in the hypo-virulent strain.

Motif	Modified	%	# of	# of	Mean	Mean	Partner Motif
	Position	Motifs	Motifs	Motifs in	Modification	Motif	
		Detected	Detected	Genome	QV	Coverage	
CACGCAG	6	64.27	527	820	58.3	29.2	
GTAYNNNNATC	3	58.95	214	363	55.2	28.5	GATNNNNRTAC
GATNNNNRTAC	2	57.02	207	363	58.1	28.9	GTAYNNNNATC

Methylated base highlighted in bold

From Tables 3.7 and 3.8, the CACGCAG motif is significantly ($p = 9.9 \times 10^{-63}$) methylated at a higher proportion in intergenic regions (~70%) as opposed to genic regions in both the hyper-virulent and hypo-virulent strains.

Table 3.7 Frequency of motifs categorized by the genomic locations in the hyper-virulent strain

	Genic*	Intergenic
CTGGAGGA	29	8
GATNNNNRTAC	120	201
GTAYNNNNATC	146	146
CACGCAG	231	563

*Genes include repeat regions and pseudogenes

Table 3.8 Frequency of motifs categorized by the genomic locations in the hypo-virulent strain

	Genic*	Intergenic
GATNNNNRTAC	113	195
GTAYNNNNATC	109	117
CACGCAG	222	542

*Genes include repeat regions and pseudogenes

Two out of three predicted motifs are known methyltransferase targets in Type I and Type II H37Rv RM systems found in REbase, recently identified using PacBio sequencing (Roberts et al. 2010). It is important to classify the m6A based on its specific sequence context as each motif is recognised by a methyltransferase enzyme which belong to specific classes of RM systems as described below:

The CACGCAG motif

This motif was not found in REbase for H37Rv RM systems, however it was found in the *M. bovis* AN5 genome Type II RM system although the actual cleavage site for the restriction enzyme has not been determined. This motif is non-palindromically methylated. It has also been identified in *M. tuberculosis* HN878 Beijing strain as part of the Type II RM system although the actual methylated base within the CACGCAG motif was not annotated.

The GATNNNNRTAC/ GTAYNNNNATC motif

The GATNNNNRTAC/ GTAYNNNNATC motif is the target of a possible Type I: subtype Gamma methyltransferase. The cognate restriction enzyme has not been identified for this RM system. Type I RM systems consists of a complex comprised of three polypeptides which are coded for by the *HsdM* gene M subunit, DNA methyltransferase, the S protein specificity determinant *hsdS.1* and the *hsdS* protein (Lew et al. 2011). The S subunit has the same recognition motif as the M subunit. These genes have been annotated in Tuberculist as forming the M and S subunits which are constituents of the methyltransferase, and methylate the adenine in this bi-partite motif in both strands.

The CTGGAGGA motif

The motif, CTGGAGGA, was identified as unique to the hyper-virulent strain. The m6A sites within this motif do not have high confident m6A sites as seen in Table 3.5. Furthermore, it was identical to the H37Rv type II Restriction Modification

system motif, CTGGAG, which was found in REbase, but without the two extra bases at the 3' end.

Shell et al. (2013) identified the similar motif (CTGGAG) to be methylated at the same adenine in *M. tuberculosis* H37RV and Euro-American strains. This site was methylated by a methyltransferase encoded for by gene *mamA* (Shell et al 2013) and was shown to be unmethylated at the motif for Beijing HN878 strain due to a point mutation in the gene. Their results reveal a 3-fold reduction in adenine methylation and attribute the methylation to *mamA*-independent methylation. Our results differ, in that the CTGGAGGA motif is methylated in very few bases only in the hyper-virulent strain. Soolingen et al. 1996, also reveal that this motif was not methylated in the Beijing lineage.

There were no significant difference in the motifs identified for m4C in both the hyper- and hypo-virulent strains.

Motifs characterised in two RM systems according to the REbase database were identified. According to REbase these motifs are the target for enzymes from a Type II RM system whose principle function in the bacteria immune system is to digest foreign DNA. In contrast orphan methyltransferases such as Dam in *E.coli* are not part of RM systems and have functions such as the regulation of virulence genes (Davis et al. 2013; Kumar & Rao 2013; Fang et al. 2012). Given that the accompanying restriction enzymes were not characterised for *M. tuberculosis* and are not found in REbase, suggesting that the role of methyltransferases in *M.*

tuberculosis might be other than host immunity through restriction modification. The methylation sites not occurring within a predicted motif could be due to the inaccurate motif prediction from the software. These non-specific m6A sites were not identified within a predicted motif due to the ModQv score of 30 which was set as the cut off.

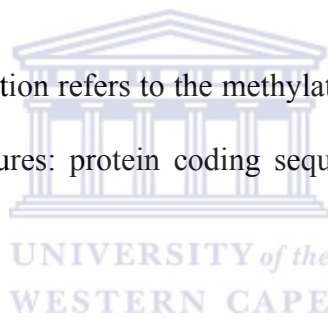


3.3 Downstream analysis of methylation loci

The downstream analysis of m6A by their genomic regions occurrence has been broken down into methylation sites occurring within genes (henceforth termed ‘genic methylation’) and intergenic regions. In this section, the unique and shared approach between hyper-virulent and hypo-virulent strains was employed in comparatively analysing the uniquely methylated genomic regions.

3.3.1 Genic Methylation

In this study, genic methylation refers to the methylation positions occurring within the following genome features: protein coding sequences, repeat regions, mobile elements and pseudogenes.



There is evidence to suggest that genic methylation for m6A and 5mC methylation has an effect on gene expression in prokaryotes and eukaryotes but the exact mechanism remains unknown (Shenker & Flanagan 2012; Dayeh et al. 2014; Kahramanoglou et al. 2012; Furuta et al. 2014). Additionally, there is evidence to show that gene expression is mostly affected by methylation in promoter regions and noncoding regions (Heithoff 1999; Low et al. 2001). However seeing that a considerable number of methyl adenines, 41% and 38% (Table 3.9), for hyper-virulent and hypo-virulent strains respectively are present in genic regions, it is worth discussing them bearing in mind that this would warrant further investigation into the exact mechanisms and functional consequences therein and/or thereof.

Table 3.9 Summary of methylation sites for intergenic and genic regions, per strand, in hyper-virulent and hypo-virulent strains.

	Hyper-virulent		Hypo-virulent	
	Neg strand	Pos strand	Neg strand	Pos strand
m6A genic	407 (42%)	375 (39%)	360 (38%)	338 (38%)
m6A intergenic	577 (58%)	593 (61%)	582 (62%)	550 (62%)
Total	980	965	939	885

Interestingly, the shared sites mostly occur within the predicted motifs while this is not the case for the unique sites.

For the m6A positions, a total of 775 and 692 sites overlap in genomic features which include CDS and repeat regions and mobile elements for hyper-virulent and hypo-virulent strains, respectively. The number of genic m6A positions that are uniquely found in the hyper-virulent are 278 (35%) and 195 (27.9%) compared to the hypo-virulent strain while 497 sites were shared as outlined in Table 3.10.

Table 3.10 Frequency of unique and shared genic and intergenic m6A sites in hyper-virulent and hypo-virulent strains and the frequency of methylated genes

	Genic	Methylated genes only	Intergenic
Shared	497	461	897
Unique hypervirulent	278	213	273
Unique hypovirulent	195	145	235

There were 461 shared methylated genes. In addition, there are 213 and 145 (Table 3.10) uniquely methylated genes in hyper-virulent and hypo-virulent strains, respectively. Given that the number of shared genic positions, at 497, is slightly higher than the number of methylated genes (461), multiple methylated sites per gene could be as a result of gene overlap.

From the 278 unique genic methylation sites in the hyper-virulent strain, only 83 (29%) occur within a predicted motif. 51 of these sites are in the GATNNNNRTAC/GTAYNNNNATC motif, 17 from the CTGGAGGA motif and 15 from the CACGCAG motif.

For the hypo-virulent strain, out of 195 methylated genic sites, only 7 (3.6%) occur in motifs, all 7 of which are the GATNNNNRTAC/GTAYNNNNATC motif.

3.3.1.2 Uniquely methylated genes in hyper-virulent and hypo-virulent strains

A functional enrichment of the genes that were uniquely methylated in the hypo-virulent and hyper-virulent strains, revealed the following COG categories as significant: 1) Secondary metabolites biosynthesis, transport and catabolism was significantly associated with m6A methylation in the hyper-virulent strain ($p = 0.006$), 2) Energy production and conversion ($p = 0.015$) and 3) Cell

wall/membrane/envelope biogenesis ($p = 0.047$) in the hypo-virulent strain. For the shared methylated genes between the two strains, no functional categories were statistically significant. The enrichment results depicts that methylation has a preference for genes of a certain function but does not necessarily mean the non-enriched genes should be excluded.

Despite the little knowledge of the effects of intragenic methylation and their functional consequence, the genic methylation signal reported here is a starting point for further investigation into the functional effects in *M. tuberculosis*.

One plausible explanation for the genic methylation occurrence, is that promoters of the adjacent gene could overlap in an open reading frame (ORF) in the previous gene as evidenced by (Gonzalez-y-Merchand et al. 1999). Hence, in actual fact it is a promoter that is methylated. For example, the outer membrane protein A gene *OmpA*, a methyladenine occurs very close to the end of the gene [position 1003616]. This is 189 base pairs away from the start of the adjacent gene *Rv0901* [position 1003805] and is close to the end of *OmpA* [position 1003792] implying that the promoter regions overlaps with *OmpA* gene (figure 3.1). This is expected seeing that bacterial genomes are tightly packed so promoters of genes will overlap in an ORF.

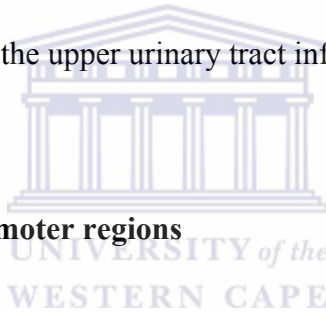
In comparison to a study by (Furuta et al. 2014), methylation due to Type I M and S genes in *H. pylori* repressed the expression of genes which contained adenine methylation. However, the exact mechanism is yet to be determined. They also

identified the CATG motif (Type II RM system) methylated within ORFs and found hypermethylation within genes, in particular an RNA polymerase gene amongst others whose biological significance could not be determined. Furthermore (Lluch-Senar et al. 2013) revealed that adenine methylation occurring within genes did not have a significant effect on gene expression at different growth phases in *Mycoplasma pneumoniae*. Their results also reveal the occurrence of unmethylated adenines in the stop codon and hypothesise that these unmethylated sites are protected from methylation by an interacting protein.

As a functional consequence of genic methylation, it could possibly be that DNA methylation interferes with binding of RNA polymerase since the methyl group of adenine lowers the thermodynamic stability of DNA, affecting the curvature and hence interfering with protein-DNA interactions (Kumar & Rao 2013). The actual role of genic methylation is still unclear, hence its biological relevance cannot be explained and will have to be tested in conjunction with experimental transcriptomic data to unravel their functional consequences.

3.3.2 Intergenic methylation

In this study the intergenic region was defined as any genomic region between gene features. The results show that more m6A sites occur within intergenic regions which might suggest gene regulation activities of DNA methylation. The hypothesis is that DNA methylation occurring in promoters of operons could affect initiation of transcription and thereby gene regulation. Therefore, these promoter region methyladenines were further examined. Adenine methylation in promoters of operons and non-operon genes has been observed in *E.coli* where the expression of pyelone-phritis associated pili operon is regulated by *Dam* methyltransferase and is a critical virulence factor in the upper urinary tract infections (Heithoff 1999).



3.3.2.1 Methylation in promoter regions

A computational approach was employed to examine the methylated promoter domains upstream of operons. The hypothesis is that methylation sites overlapping within possible promoter binding proteins positions may affect gene expression. Genes in an operon are expressed together and are under the same control of one head promoter, hence methylation overlap within a promoter region could possibly affect transcription and gene expression of these operon genes. To reiterate, the promoter region is defined as a 100bp domain upstream from the start of the head gene of the operon.

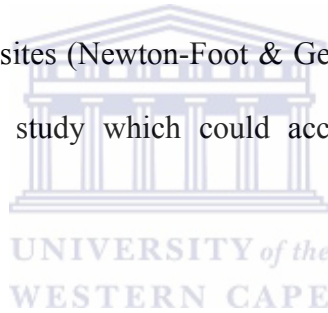
As reviewed by (Newton-Foot & Gey van Pittius 2013), the expression of most genes in Mycobacterial species is mediated through regulatory regions situated within the upstream intergenic region. Consequently, methylation can lead to up or down regulation of a gene (Marinus & Casadesus 2009). Operons form the most basic unit of organization in bacterial genomes, and are fundamental in understanding transcriptional regulation (Roback et al. 2007). Hence operons should be taken into account when analysing the promoter regions.

The DOOR software predicted a total of 2341 operons for the H37Rv genome. A total of 65 operons in the hyper-virulent strain, whose promoter regions are methylated, and 68 operons for the hypo-virulent strain. A total of 20 operon promoter regions were found to be uniquely methylated in the hyper-virulent strain and 23 for the hypo-virulent strain, while 45 shared methylated operon promoters between them. Interestingly, most of the uniquely methylated promoters do not have methylation sites occurring in predicted regulatory motifs while the shared methylated promoters have predicted regulatory motifs. From the 20 uniquely methylated promoters only six have predicted motifs for the hyper-virulent strain. While only one promoter contains a predicted motif out of the 23 uniquely methylated promoters for the hypo-virulent strain.

Intergenic methylation sites occurring in operon promoter regions

In the hypo-virulent strain, only 68 (10%) of 1132 intergenic sites occur within the promoter regions while 67 sites (10%) out of 1170 intergenic sites for the hyper-virulent strain. The remaining 90% of intergenic sites occur in spacer regions between genes of an operon and beyond the defined 100bp promoter region.

The 90% of intergenic methylation not occurring in promoter domains could be due to its typical methylation role such as the restriction modification processes or maybe playing a role other than gene regulation. Another explanation could be that in some cases there could be more complex promoter regions extending up to 1 kb from the transcription start sites (Newton-Foot & Gey van Pittius 2013) other than the defined 100bp in this study which could account for the 90% intergenic methylation.



After identifying the uniquely methylated promoter-associated operon regions within the hyper-virulent and hypo-virulent strains, functional enrichment was carried out on the operon-associated genes.

3.3.2.1.1 Functional enrichment of methylated promoters of operons

All genes belonging to a given methylated promoter-associated operon were included in the gene list for the functional enrichment.

A total of 36 uniquely methylated promoter associated genes were discovered in hyper-virulent (Appendix I) and 32 in hypo-virulent strain (Appendix II) and 68 shared methylated promoter associated genes (Appendix III). The GO terms and COG categories enriched for uniquely and shared methylated operon-associated genes for the hyper- and hypo-virulent strains are presented in Tables 3.11 - 3.13.

Table 3.11 Significantly enriched functional categories of uniquely methylated operon-associated genes in hyper-virulent strain

Functional category	<i>p</i> value
GO-Growth of symbiont in host cell	0.0
GO-Response to acid	0.0
GO-Succinate dehydrogenase activity	0.013
COG-Energy production and conversion	0.001



Table 3.12 Significantly enriched functional categories of uniquely methylated operon-associated genes in hypo-virulent strain

Functional category	<i>p</i> value
GO- Response to zinc ion	0.005
COG-Function unknown	0.038

Response to acid

The functional categories in Table 3.11 for hyper-virulent strain possibly suggests it's adaptation to host environment and stress and subsequent survival of *M.*

tuberculosis in the host. The macrophage phagolysosome provides an acidic environment hence creating stress on the bacilli and as a means to evade the acidic nature of the phagosome *M. tuberculosis* adjusts its intracellular pH levels to adapt. This is also demonstrated in *in vitro* studies on mutants of acid sensitive *M. tuberculosis* which are attenuated in animal models of infection (Vandal et al. 2009)

Furthermore the deletion of the *mymA* (Rv3083-Rv3089) operon, whose promoter region is found to be methylated in the hyper-virulent strain, is required for growth in macrophages and is activated in response to acidic conditions when inside macrophages (Cheruvu et al. 2007). Singh et al (2003) have shown that when exposed to a low pH the promoter of the *mymA* operon is induced and the loss of this operon resulted in *M. tuberculosis* failing to persist in spleens of infected guinea pigs and showed increased drug sensitivity as well as resulting in death of *M. tuberculosis* due to activated macrophages. This operon promoter was found to be uniquely methylated in the hyper-virulent strain suggesting that it plays a key role in the adaptation of *M. tuberculosis* to acidic stress and highlights the possible role of DNA methylation in regulating promoter activity in virulence. Additionally, the Rv3089 (*fadD13*) protein was found to be differentially abundant in the hyper-virulent strain based on proteomic experiments (S. Fortuin personal communication). Therefore, the hypothesis could be that methylation in this promoter region may have increased transcription of the *fadD19* gene. This may explain the strategy that *M. tuberculosis* uses to adapt to acidic stress and pathogenesis thereof.

Succinate dehydrogenase activity

The m6A site occurring in the promoter of the *Rv0247c-Rv0249c* operon does not occur in any of the predicted motifs. These genes also belong the tricarboxylic acid (TCA) cycle pathway. Interestingly, *Rv2048c* protein was differentially abundant in the hyper-virulent strain (S. Fortuin, personal communication). The GO term succinate dehydrogenase (SDH) activity has been significantly enriched in genes uniquely methylated in the hyper-virulent strain. According to (Black et al. 2014), SDH plays a crucial role in physiological processes such as membrane maintenance, hypoxia and ATP synthesis, and could therefore be a potential drug target. In *M. tuberculosis*, the sustained metabolism of succinate through SDH is a critical adaptation to hypoxia (Eoh & Rhee 2013). The occurrence of methylation in the promoter region, might suggest that these genes are differentially regulated in the hyper-virulent strain to possibly ensure survival in hypoxic conditions.

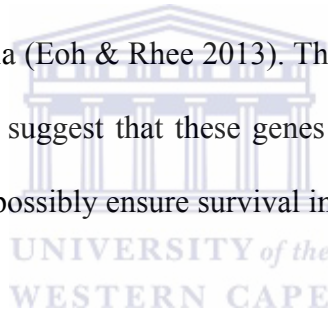


Table 3.13 Significantly enriched functional categories of shared methylated operon-associated genes in hyper- and hypo-virulent strains

Functional category	<i>p</i> value
GO- DNA topoisomerase (ATP-hydrolyzing) activity	0.013
GO- Protein heterodimerization activity	0.002

The shared methylated operon-associated genes were enriched for the GO term DNA topoisomerase (ATP-hydrolyzing) activity (Table 3.13) with these genes playing a role in DNA replication, more specifically in the DNA gyrase operon. The

promoter region of this operon is methylated in the Type I motif in both the hyper- and hypo-virulent strains which suggests that methylation plays a role in DNA replication in *M. tuberculosis*, similar results have been seen in other bacteria such as *E.coli* and *Caulobacter crescentus*. Interestingly, the gyrase genes in *M. tuberculosis* are also known drug targets (Mdluli & Ma 2007).

Selected operons of interest in the hyper-virulent strain

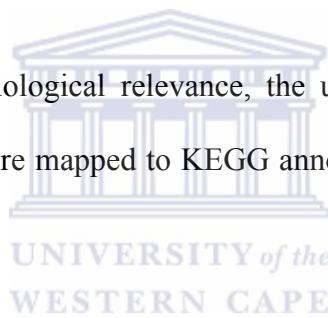
Of interest, is the gene *Rv2280*, coding a possible dehydrogenase involved in oxidoreduction, whose promoter region is methylated uniquely within the CACGCAG motif. In a study by (Safi et al. 2004), it was demonstrated that in H37Rv this gene also has an *IS6110* element insert present in the promoter region, and has increased gene expression levels. Currently, it is unknown whether methylation regulates *Rv2280* expression.

The *Rv1216c-Rv1219c* predicted operon is of special interest. *Rv1219c* is a transcriptional regulator for the other genes within this operon and the methylation site overlaps in between the -10 and -35 region upstream of this gene. Our results reveal that the GATNNNNRTAC motif, being part of the Type I RM system, is methylated upstream of the *Rv1219c* gene in between the -10 and -35 region. This implies that methylation might be directly interfering with transcriptional binding proteins, for example an RNA polymerase. *Rv1217c* and *Rv1218c* form part of the multidrug efflux system ATP-binding cassette (ABC) transporter family and their gene expression is controlled by the *Rv1219c* protein. Furthermore, *Rv1217c-*

Rv1218c has been shown to be involved in multidrug antibiotic resistance (Balganesh et al. 2010; Wang et al. 2013). Methylation could affect the expression of the transcriptional regulator (*Rv1219c*) which in turn regulated the expression of the efflux pump operon. This suggests a possible link between methylation and drug resistance, seeing that this gene is uniquely methylated in the hyper-virulent strain.

3.3.2.1.2 Metabolic pathways affected by methylated operon associated promoters

To further elucidate the biological relevance, the uniquely methylated promoter operon associated genes were mapped to KEGG annotated pathways of the H37Rv genome.



A total of 36 uniquely methylated promoter-associated genes were mapped to 19 pathways in the hyper-virulent strain, and to 15 pathways in the hypo-virulent strain. Of these pathways, 10 were unique to the hyper-virulent (Table 3.14) and 5 to the hypo-virulent strain (Table 3.15).

Table 3.14. Strain specific KEGG pathways in hyper-virulent strain

KO Number	Pathway name
mtu00363	Bisphenol degradation
mtu00650	Butanoate metabolism
mtu00190	Oxidative phosphorylation
mtu00020	Citrate cycle (TCA cycle)
mtu00071	Fatty acid degradation
mtu00626	Naphthalene degradation
mtu00564	Glycerophospholipid metabolism
mtu00625	Chloroalkane and chloroalkene degradation
mtu01220	Degradation of aromatic compounds
mtu03420	Nucleotide excision repair

Table 3.15. Strain specific KEGG pathways in hypo-virulent strain

KO number	Pathway Name
mtu00260	Glycine, serine and threonine metabolism
mtu00760	Nicotinate and nicotinamide metabolism
mtu00770	Pantothenate and CoA biosynthesis
mtu03010	Ribosome
mtu05152	Tuberculosis

Energy Conversion and Oxidative phosphorylation

Oxidative phosphorylation produces ATP which is essential for certain processes in replicating persistent bacteria or to promote virulence and survival (Berney & Cook

2014; Black et al. 2014). The unique metabolic pathways are involved in the hyper-virulent strain revealed the oxidative phosphorylation pathway that mycobacteria require ATP for replicating in host cells. Along with the fact that energy conversion COG category is also enriched might suggest that the hyper-virulent strain may require extra ATP and DNA methylation is probably playing a role in driving this process. Furthermore, in a study done by our collaborators (S. Fortuin, PhD thesis, 2013) the hyper-virulent strain showed differentially abundant proteins involved in energy metabolism which suggested these genes are up-regulated due to higher ATP requirements by the hyper-virulent strain and maybe required for survival and adaptation in the host.



3.3.2.2 Summary

The uniquely methylated regions (21%) in the hyper- and hypo-virulent strains may provide insights into their difference in disease phenotypes. Most of the methylation positions are shared (79%) between the hyper- and hypo-virulent strains and occur in predicted motifs with very few unique methylation positions. The shared methylated loci are possibly the regions involved in the typical RM functions of bacterial DNA methylation. However, in the case of the *bcp* gene, the promoter region is methylated in both hyper- and hypo-virulent strains and this gene belongs to the virulence, detoxification and adaptation category according to Tuberculist

(Lew et al. 2011). This implies that DNA methylation might also play a common role in expression of certain virulence genes in both the strains.

Promoter DNA methylation in bacteria has been linked to virulence and pathogenesis as seen in *E. coli*, *H influenza*, *Salmonella enterica* (Low et al. 2001). Our results suggest that the methylation of promoters of certain genes might explain the adaptation of the hyper-virulent strain to stress, allowing adaptation to the environment and not virulence *per se*. These genes may possibly be up-regulated or down-regulated due to the overlapping of a methylation site with a transcriptional regulatory protein site in the promoter region. In some cases, the promoter regions in which some methylation sites occur, is within -10 and -35 promoter sequence or sigma factor sequences. This may provide the understanding into the mechanism of gene expression due to the overlap of a methylated adenine with a sigma factor binding site occurring upstream region of *Rv0213-Rv0214*, as well as the promoter region of operon *Rv1216c-Rv1219c*.

A minor caveat is the absence of annotated transcription start sites for the hyper- and hypo-virulent strains with only coordinates of the gene which is not necessarily the ORF from which the promoter regions were extrapolated. However about 26% of CDSs have leaderless ORFs which contain no UTRs so the gene start is the absolute start position of the ORF (Cortes et al. 2013).

A study by (Hemavathy & Nagaraja 1995) showed fairly similar levels of N6 methyladenine between the virulent H37Rv and avirulent H37Ra *M. tuberculosis*

strains. These findings were based on mass spectrometry but the exact genomic positions were not pinpointed and only the methyladenines levels were quantified. In parallel, our results generated through the third generation sequencing technology reveal that majority of the methyl adenine loci are shared between hyper- and hypo-virulent strains. However, the subtle differences in unique methylated loci in the hyper-virulent strain might suggest adaptation to survival in the host.

Shell et al (2013) reveal that the H37Rv strain had the CTGGAG motif methylated in the -10 Sigma factor binding site promoter region, and in response to hypoxia certain genes were overexpressed. Similarly, our results indicate that the promoter region of certain genes does overlap with adenine methylation sites. The methylation sites in promoter regions in *M. tuberculosis* Beijing strains have not been previously characterised using genome-wide single base pair approach with PacBio technology.

DNA methylation of certain regulatory regions is thought to play a role in phase variation, a remarkable strategy employed by bacteria to rapidly regulate transcription in response to environmental stimuli. This allows bacterial pathogens such as *E. coli* to propagate their virulence and to adapt to environmental stimuli in the host. Phase variation results in bacterial subpopulations controlled by epigenetic mechanisms which result in phenotypic diversity (Casadesús & Low 2013). There could be a possibility that *M. tuberculosis* uses a similar approach to adapt to the

environmental stresses in the granuloma, although this hasn't been experimentally confirmed.

Given the different pathogen characteristics and disease phenotypes seen in the Beijing family lineages (Aguilar et al. 2010) there is a possibility that DNA methylation, specifically adenine methylation, could be a virulence and/or an adaptation strategy of *M. tuberculosis* Beijing sublineage 7 strains in response to stress conditions in order to persist infection especially in the hyper-virulent strain.



CHAPTER 4: CONCLUSIONS AND FUTURE WORK

The overall aim of this study was to characterise the methylomes of two closely related Beijing hyper and hypo-virulent strains with the view of understanding virulence. The comparative analysis between hyper and hypo-virulent strains was determined by examining the shared and unique methylated genomic regions, which might explain their different disease phenotypes.

Using the latest single molecule sequencing technology available and with computational analysis, the methylomes for the Beijing sublineage 7 hyper- and hypo-virulent strains were identified and characterised. To our knowledge this has not been characterised before on a single base pair resolution for *M. tuberculosis*.

The exact location of methylated DNA bases was pinpointed, namely N6 methyladenine and N4 methylcytosine, which even though was described for *M. tuberculosis* using non-sequencing methods, but not its specific location throughout the genome. The goals of the thesis have been met whereby the methylation profiles and specific positions as well methylated sequence motifs were described for the hyper- and hypo-virulent strains. By interrogating the uniquely methylated genomic regions in the hyper-virulent strain, intergenic methylated domains of operons were identified that may suggest persistence and survival in the host. In doing so these results may provide a starting point into the possible gene regulation functional effects that DNA methylation might have in the survival and persistence of infection in the hyper-virulent strain.

4.1 Key findings

The repertoire of adenine methylation (m6A) was examined within these two strains using a comparative approach. Knowledge of these identified methylated genomic regions could be used as a molecular typing and/or virulence marker. m6A was identified in both strains occurring within the following sequence motifs: CACGCAG (Type II RM system), GATNNNNRTAC/GTAYNNNNTC (Type I RM system), while the CTGGAGGA motif was found to be uniquely methylated in the hyper-virulent strain. Interestingly, the CACGCAG motif was significantly methylated at a higher proportion in intergenic regions in both the hyper- and hypo-virulent strains suggesting a role in gene regulation.

The majority (79%) of the methylated positions were shared between the hyper-virulent and hypo-virulent strains, with very few strain-specific methylated regions. The results in this study indicate that some operons in the hyper-virulent strain, for example the *mymA* operon (response to acid) and the *Rv1217c- Rv1218c* operon, which is a known drug target, have methylated promoter regions, appear to have roles in response to stress conditions and not virulence *per se*. These have implications in the adaptations of *M. tuberculosis* to stress by possibly using methylation occurring in promoter regions as a mechanism to alter gene regulation and could be a potential area to investigate since the exact mechanism on gene expression has yet to be explored.

The PacBio SMRT sequencing employed in this study was proven to be a reliable method for DNA methylation detection in *M. tuberculosis* compared to previous unsuccessful attempts using Illumina bisulphite sequencing for detecting 5 methylcytosine.

4.2 Novel aspects and impact on TB research

A plethora of bacterial methylomes have been recently identified owing to the PacBio SMRT technology but to our knowledge none have been determined for the medically important *M. tuberculosis*. DNA methylation of the *M. tuberculosis* genome will change the way we approach the understanding of this pathogen and its use as a possible survival mechanism and virulence factor in the host. There are several genetic variation studies being undertaken to understand virulence mechanisms and pathogenesis, but not DNA methylation-based, with the exception of Shell et al (2013), for *M. tuberculosis*. The computational approach used in this thesis opens the doors to understanding *M. tuberculosis* methylomes.

DNA methylation has not been an extensive area of research in *M. tuberculosis* and has been poorly described in the literature despite being well characterised along with its role in virulence in other pathogenic bacteria such as *E.coli*. With the findings presented in this thesis, we now have a better understanding of DNA methylation patterns in *M. tuberculosis* and provides a starting point for in depth methylation characterisation of the RM systems in *M. tuberculosis*. To our

knowledge this is the first time *M. tuberculosis* has been characterised using SMRT technology in Beijing strains.

4.3 Limitations

This study used the *M. tuberculosis* H37Rv genome as the reference strain for mapping and identifying the methylated positions hence genomic regions unique to the Beijing strains with respect to H37Rv were missed and may possibly contain methylated loci. Hence there may be absent methylation loci uniquely occurring in the Beijing strains.

These *in vitro* isolates are snapshots of the hyper and hypo-virulent strains and represents a once off representation and not the entire repertoire of DNA methylation in every Beijing strain. Thus, additional strains could be sequenced in future investigations to achieve a more robust DNA methylation profile and to account for the low sequencing coverage used. Though not the aim of this study, what would need to be considered to comprehensively link DNA methylation in promoter regions with virulence, is for adenine methylation to be measured over a timescale in combination with the corresponding transcriptome and proteome for these strains. Despite these limitations this study has definitely revealed novel insights into DNA methylation in *M. tuberculosis*, which has not been well characterised previously more especially in Beijing strains given their enhanced virulence capacity.

One short-coming of this study was the low confidence m4C positions which resulted in not proceeding with further downstream analysis. A deeper sequencing coverage is therefore recommended to have a clearer picture of the m4C repertoire. The m4C signal in the hyper-virulent and hypo-virulent strains appeared to be low whereas the m6A signal is more distinct as evidenced by their predicted motifs. These could be false positives and deeper resequencing would be required to get highly confident m4C predictions.

4.4 Future work

Comparing the transcriptomes for the two strains to measure the effects of gene expression with the overlapping adenine methylation sites predicted in the promoter regions in this study could be explored. Additionally, another approach could be to measure the virulence attenuation in the identified operons whose promoters were methylated.

In conclusion, bacterial methylome discoveries has been a fairly new area of research using the latest single molecule sequencing technology. We are only now beginning to understand the role of DNA methylation after characterising their genomic positions and will require further extensive investigations to understand their functional consequences. This thesis provides the basis for the identification and characterisation of DNA methylation in *M. tuberculosis* Beijing strains and the novel results presented herein contributes to the largely untapped area of mycobacterial epigenetics.

BIBLIOGRAPHY

- Aguilar, D.L., Hanekom, M., Mata, D., Gey van Pittius, N.C., van Helden, P.D., Warren, R.M. & Hernandez-Pando, R., 2010. Mycobacterium tuberculosis strains with the Beijing genotype demonstrate variability in virulence associated with transmission. *Tuberculosis (Edinburgh, Scotland)*, 90(5), pp.319–25.
- Ansorge, W.J., 2009. Next-generation DNA sequencing techniques. *New biotechnology*, 25(4), pp.195–203.
- Balganesh, M., Kuruppath, S., Marcel, N., Sharma, S., Nair, A. & Sharma, U., 2010. Rv1218c, an ABC transporter of Mycobacterium tuberculosis with implications in drug discovery. *Antimicrobial agents and chemotherapy*, 54(12), pp.5167–72.
- Barras, F. & Marinus, M.G., 1989. The great GATC: DNA methylation in E. coli. *Trends in genetics : TIG*, 5(5), pp.139–43.
- Berney, M. & Cook, G.M., 2014. Respiration and Oxidative Phosphorylation in Mycobacteria. In M. F. Hohmann-Marriott, ed. *The Structural Basis of Biological Energy Generation*. Springer Netherlands, pp.277–293.
- Bird, A., 2007. Perceptions of epigenetics. *Nature*, 447(7143), pp.396–8.
- Bird, A., 1992. The essentials of DNA methylation. *Cell*, 70(1), pp.5–8.

- Black, P. a, Warren, R.M., Louw, G.E., van Helden, P.D., Victor, T.C. & Kana, B.D., 2014. Energy metabolism and drug efflux in *Mycobacterium tuberculosis*. *Antimicrobial agents and chemotherapy*, 58(5), pp.2491–503.
- Bock, C. & Lengauer, T., 2008. Genome analysis Computational epigenetics. *Bioinformatics*, 24(1), pp.1–10.
- Borrell, S. & Gagneux, S., 2011. Strain diversity, epistasis and the evolution of drug resistance in *Mycobacterium tuberculosis*. *Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 17(6), pp.815–20.
- Brosch, R., Gordon, S. V, Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., Garnier, T., Gutierrez, C., Hewinson, G., Kremer, K., Parsons, L.M., Pym, A.S., Samper, S., van Soolingen, D. & Cole, S.T., 2002. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proceedings of the National Academy of Sciences of the United States of America*, 99(6), pp.3684–9.
- Cambau, E. & Drancourt, M., 2014. Steps towards the discovery of *Mycobacterium tuberculosis* by Robert Koch, 1882. *Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 20(3), pp.196–201.
- Casadesús, J. & Low, D., 2006. Epigenetic Gene Regulation in the Bacterial World
Epigenetic Gene Regulation in the Bacterial World. , 70(3).

- Casadesús, J. & Low, D. a, 2013. Programmed heterogeneity: epigenetic mechanisms in bacteria. *The Journal of biological chemistry*, 288(20), pp.13929–35.
- Chaisson, M.J. & Tesler, G., 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC bioinformatics*, 13(1), p.238.
- Chatterjee, A., Stockwell, P.A., Rodger, E.J. & Morison, I.M., 2012. Comparison of alignment software for genome-wide bisulphite sequence data. *Nucleic Acids Research*, pp.1–8.
- Cheruvu, M., Plikaytis, B.B. & Shinnick, T.M., 2007. The acid-induced operon Rv3083-Rv3089 is required for growth of *Mycobacterium tuberculosis* in macrophages. *Tuberculosis (Edinburgh, Scotland)*, 87(1), pp.12–20.
- Clark, T.A., Murray, I.A., Morgan, R.D., Kislyuk, A.O., Spittle, K.E., Boitano, M., Fomenkov, A., Roberts, R.J., Korlach, J., 2012. Characterisation of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic acids research*, 40(4), pp.1–12.
- Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V, Eiglmeier, K., Gas, S., Barry, C.E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M.A., Rajandream, M.A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R.,

- Squares, S., Sulston, J.E., Taylor, K., Whitehead, S. & Barrell, B.G., 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, 393(6685), pp.537–44.
- Comas, I. & Gagneux, S., 2009. The past and future of tuberculosis research. *PLoS pathogens*, 5(10), p.e1000600.
- Cooper, D.L., Lahue, R.S. & Modrich, P., 1993. Methyl-directed mismatch repair is bidirectional. *The Journal of biological chemistry*, 268(16), pp.11823–9.
- Cortes, T., Schubert, O.T., Rose, G., Arnvig, K.B., Comas, I., Aebersold, R. & Young, D.B., 2013. Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in *Mycobacterium tuberculosis*. *Cell reports*, 5(4), pp.1121–31.
- Coscolla, M. & Gagneux, S., 2010. Does *M. tuberculosis* genomic diversity explain disease diversity? *Drug discovery today. Disease mechanisms*, 7(1), pp.e43–e59.
- Cousins, D. V, Bastida, R., Cataldi, A., Quse, V., Redrobe, S., Dow, S., Duignan, P., Murray, A., Dupont, C., Ahmed, N., Collins, D.M., Butler, W.R., Dawson, D., Rodríguez, D., Loureiro, J., Romano, M.I., Alito, A., Zumarraga, M. & Bernardelli, A., 2003. Tuberculosis in seals caused by a novel member of the *Mycobacterium tuberculosis* complex: *Mycobacterium pinnipedii* sp. nov. *International journal of systematic and evolutionary microbiology*, 53(Pt 5), pp.1305–14.

- Davis, B.M., Chao, M.C. & Waldor, M.K., 2013. Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. *Current opinion in microbiology*, 16(2), pp.192–8.
- Dayeh, T., Volkov, P., Salö, S., Hall, E., Nilsson, E., Olsson, A.H., Kirkpatrick, C.L., Wollheim, C.B., Eliasson, L., Rönn, T., Bacos, K. & Ling, C., 2014. Genome-wide DNA methylation analysis of human pancreatic islets from type 2 diabetic and non-diabetic donors identifies candidate genes that influence insulin secretion. *PLoS genetics*, 10(3), p.e1004160.
- Dewitt, N.D., Yaffe, M.P. & Trounson, A., 2012. Building stem-cell genomics in California and beyond. *Nature biotechnology*, 30(1), pp.20–5.
- Ebrahimi Rad, M., Bifani, P., Martin, C., Kremer, K., Samper, S., Rauzier, J., Kreiswirth, B., Blazquez, J., Jouan, M., van Soolingen, D. & Gicquel, B., 2003. Mutations in putative mutator genes of Mycobacterium tuberculosis strains of the W-Beijing family. *Emerging infectious diseases*, 9(7), pp.838–45.
- Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J.,

- Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J. & Turner, S., 2009. Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)*, 323(5910), pp.133–8.
- Eoh, H. & Rhee, K.Y., 2013. Multifunctional essentiality of succinate metabolism in adaptation to hypoxia in *Mycobacterium tuberculosis*. , 2013.
- Ewing, B. & Green, P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research*, 8(3), pp.186–94.
- Ewing, B., Hillier, L., Wendl, M.C. & Green, P., 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome research*, 8(3), pp.175–85.
- Fang, G., Munera, D., Friedman, D.I., Mandlik, A., Chao, M.C., Banerjee, O., Feng, Z., Losic, B., Mahajan, M.C., Jabado, O.J., Deikus, G., Clark, T. a, Luong, K., Murray, I. a, Davis, B.M., Keren-Paz, A., Chess, A., Roberts, R.J., Korlach, J., Turner, S.W., Kumar, V., Waldor, M.K. & Schadt, E.E., 2012. Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. *Nature Biotechnology*, 30(12), pp.1232–9.
- Farhat, M.R., Shapiro, B.J., Kieser, K.J., Sultana, R., Jacobson, K.R., Victor, T.C., Warren, R.M., Streicher, E.M., Calver, A., Sloutsky, A., Kaur, D., Posey, J.E., Plikaytis, B., Oggioni, M.R., Gardy, J.L., Johnston, J.C., Rodrigues, M., Tang, P.K.C., Kato-Maeda, M., Borowsky, M.L., Muddukrishna, B., Kreiswirth, B.N., Kurepina, N., Galagan, J., Gagneux, S., Birren, B., Rubin, E.J., Lander,

- E.S., Sabeti, P.C. & Murray, M., 2013. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nature genetics*, 45(10), pp.1183–9.
- Fleischmann, R.D., Alland, D., Eisen, J.A., Carpenter, L., White, O., Peterson, J., DeBoy, R., Dodson, R., Gwinn, M., Haft, D., Hickey, E., Kolonay, J.F., Nelson, W.C., Umayam, L.A., Ermolaeva, M., Salzberg, S.L., Delcher, A., Utterback, T., Weidman, J., Khouri, H., Gill, J., Mikula, A., Bishai, W., Jacobs Jr, W.R., Venter, J.C. & Fraser, C.M., 2002. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *Journal of bacteriology*, 184(19), pp.5479–90.
- Flusberg, B. a, Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T. a, Korlach, J. & Turner, S.W., 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature methods*, 7(6), pp.461–5.
- Forrellad, M.A., Klepp, L.I., Gioffré, A., Sabio, J., Morbidoni, H.R., De, M., Santangelo, P., Cataldi, A.A. & Bigi, F., 2013. Virulence factors of the *Mycobacterium tuberculosis* complex. *Virulence*, 4(1), pp.1–64.
- Furuta, Y., Namba-Fukuyo, H., Shibata, T.F., Nishiyama, T., Shigenobu, S., Suzuki, Y., Sugano, S., Hasebe, M. & Kobayashi, I., 2014. Methylome Diversification through Changes in DNA Methyltransferase Sequence Specificity. *PLoS genetics*, 10(4), p.e1004272.

- Galagan, J.E., 2014. Genomic insights into tuberculosis. *Nature Reviews Genetics*, 15(5), pp.307–320.
- García-Del Portillo, F., Pucciarelli, M.G. & Casadesús, J., 1999. DNA adenine methylase mutants of *Salmonella typhimurium* show defects in protein secretion, cell invasion, and M cell cytotoxicity. *Proceedings of the National Academy of Sciences of the United States of America*, 96(20), pp.11578–83.
- Gardiner-Garden, M. & Frommer, M., 1987. CpG islands in vertebrate genomes. *Journal of molecular biology*, 196(2), pp.261–82.
- Glenn, T.C., 2011. Field guide to next-generation DNA sequencers. *Molecular ecology resources*, 11(5), pp.759–69.
- Glynn, J.R., Whiteley, J., Bifani, P.J., Kremer, K. & van Soolingen, D., 2002. Worldwide occurrence of Beijing/W strains of *Mycobacterium tuberculosis*: a systematic review. *Emerging infectious diseases*, 8(8), pp.843–9.
- Gonzalez-y-Merchand, J.A., Colston, M.J. & Cox, R.A., 1999. Effects of growth conditions on expression of mycobacterial *murA* and *tyrS* genes and contributions of their transcripts to precursor rRNA synthesis. *Journal of bacteriology*, 181(15), pp.4617–27.
- Hanekom, M., van der Spuy, G.D., Streicher, E., Ndabambi, S.L., McEvoy, C.R.E., Kidd, M., Beyers, N., Victor, T.C., van Helden, P.D. & Warren, R.M., 2007. A recently evolved sublineage of the *Mycobacterium tuberculosis* Beijing strain

- family is associated with an increased ability to spread and cause disease. *Journal of clinical microbiology*, 45(5), pp.1483–90.
- Heithoff, D.M., 1999. An Essential Role for DNA Adenine Methylation in Bacterial Virulence. *Science*, 284(5416), pp.967-970.
- Hemavathy, K.C. & Nagaraja, V., 1995. DNA methylation in mycobacteria : Absence of methylation at GATC (Dam) and CCA / TGG (Dcm) sequences. *FEMS Immunol Med Microbiol.*, 11(4), pp.291–6.
- Hershberg, R., Lipatov, M., Small, P.M., Sheffer, H., Niemann, S. & Homolka, S., 2008. High Functional Diversity in Mycobacterium tuberculosis Driven by Genetic Drift and Human Demography. *PloS Biology*, 6(12) pp. 2658-2671.
- Heusipp, G., Fälker, S. & Schmidt, M.A., 2007. DNA adenine methylation and bacterial pathogenesis. *International journal of medical microbiology : IJMM*, 297(1), pp.1–7.
- Honma, Y., Fernández, R.E. & Maurelli, A.T., 2004. A DNA adenine methylase mutant of *Shigella flexneri* shows no significant attenuation of virulence. *Microbiology (Reading, England)*, 150(Pt 4), pp.1073–8.
- Horner, D.S., Pavesi, G., Castrignano, T., Meo, P.D.O. De, Liuni, S., Sammeth, M., Picardi, E. & Pesole, G., 2010. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Briefings in Bioinformatics*, 11(2), pp.181–97.

- Jeltsch, A., 2002. Beyond Watson and Crick: DNA methylation and molecular enzymology of DNA methyltransferases. *Chembiochem : a European journal of chemical biology*, 3(4), pp.274–93.
- Julio, S.M., Heithoff, D.M., Provenzano, D., Klose, K.E., Sinsheimer, R.L., Low, D.A. & Mahan, M.J., 2001. DNA adenine methylase is essential for viability and plays a role in the pathogenesis of *Yersinia pseudotuberculosis* and *Vibrio cholerae*. *Infection and immunity*, 69(12), pp.7610–5.
- Kahramanoglou, C., Prieto, A.I., Khedkar, S., Haase, B., Gupta, A., Benes, V., Fraser, G.M., Luscombe, N.M. & Seshasayee, A.S.N., 2012. Genomics of DNA cytosine methylation in *Escherichia coli* reveals its role in stationary phase transcription. *Nature communications*, 3, p.886.
- Kato-Maeda, M., Ho, C., Passarelli, B., Banaei, N., Grinsdale, J., Flores, L., Anderson, J., Murray, M., Rose, G., Kawamura, L.M., Pourmand, N., Tariq, M. a, Gagneux, S. & Hopewell, P.C., 2013. Use of whole genome sequencing to determine the microevolution of *Mycobacterium tuberculosis* during an outbreak. *PloS one*, 8(3), p.e58235.
- Kong, Y., Cave, M.D., Zhang, L., Foxman, B., Marrs, C.F., Bates, J.H. & Yang, Z.H., 2006. Population-based study of deletions in five different genomic regions of *Mycobacterium tuberculosis* and possible clinical relevance of the deletions. *Journal of clinical microbiology*, 44(11), pp.3940–6.
- Krueger, F., Kreck, B., Franke, A. & Andrews, S.R., 2012. DNA methylome analysis using short bisulphite sequencing data. *Bioinformatics*, 9(2).

- Kumar, R. & Rao, D.N., 2013. Role of DNA methyltransferases in epigenetic regulation in bacteria. *Sub-cellular biochemistry*, 61, pp.81–102.
- Kumar, S., Cheng, X., Klimasauskas, S., Mi, S., Posfai, J., Roberts, R.J. & Wilson, G.G., 1994. The DNA (cytosine-5) methyltransferases. *Nucleic acids research*, 22(1), pp.1–10.
- Kursar, M., Koch, M., Mittrücker, H.W., Nouailles, G., Bonhagen, K., Kamradt, T. & Kaufmann, S.H.E., 2007. Cutting Edge: Regulatory T cells prevent efficient clearance of *Mycobacterium tuberculosis*. *Journal of immunology (Baltimore, Md. : 1950)*, 178(5), pp.2661–5.
- Lew, J.M., Kapopoulou, A., Jones, L.M. & Cole, S.T., 2011. TubercuList--10 years after. *Tuberculosis (Edinburgh, Scotland)*, 91(1), pp.1–7.
- Li, Q., Whalen, C.C., Albert, J.M., Larkin, R., Zukowski, L., Cave, M.D. & Silver, R.F., 2002. Differences in rate and variability of intracellular growth of a panel of *Mycobacterium tuberculosis* clinical isolates within a human monocyte model. *Infection and immunity*, 70(11), pp.6489–93.
- Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, a H. & Ecker, J.R., 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell*, 133(3), pp.523–36.
- Lister, R., Pelizzola, M., Downen, R.H., Hawkins, R.D., Hon, G., Tonti-filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q., Edsall, L., Antosiewicz-bourget, J., Stewart, R., Ruotti, V., Millar, A.H., Thomson, J.A., Ren, B. & Ecker, J.R.,

2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271), pp.315–322.
- Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. & Law, M., 2012. Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology*, 2012, p.251364.
- Lluch-Senar, M., Luong, K., Spittle, K., Clark, T.A., Schadt, E., Turner, S.W., Korlach, J. & Serrano, L., 2013. Comprehensive Methylome Characterisation of *Mycoplasma genitalium* and *Mycoplasma pneumoniae* at Single-Base Resolution. , 9(1), pp.1–12.
- López, B., Aguilar, D., Orozco, H., Burger, M., Espitia, C., Ritacco, V., Barrera, L., Kremer, K., Hernandez-Pando, R., Huygen, K. & van Soolingen, D., 2003. A marked difference in pathogenesis and immune response induced by different *Mycobacterium tuberculosis* genotypes. *Clinical and experimental immunology*, 133(1), pp.30–7.
- Low, D.A., Weyand, N.J. & Mahan, M.J., 2001. Roles of DNA Adenine Methylation in Regulating Bacterial Gene Expression and Virulence. *Infection and Immunity*, 69(12), pp.7197-7201.
- Manca, C., Reed, M.B., Freeman, S., Mathema, B., Kreiswirth, B., Barry, C.E. & Kaplan, G., 2004. Differential monocyte activation underlies strain-specific *Mycobacterium tuberculosis* pathogenesis. *Infection and immunity*, 72(9), pp.5511–4.

- Mao, F., Dam, P., Chou, J., Olman, V. & Xu, Y., 2009. DOOR: a database for prokaryotic operons. *Nucleic acids research*, 37(Database issue), pp.D459–63.
- Mardis, E.R., 2008a. Next-generation DNA sequencing methods. *Annual review of genomics and human genetics*, 9, pp.387–402.
- Mardis, E.R., 2008b. The impact of next-generation sequencing technology on genetics. *Trends in genetics : TIG*, 24(3), pp.133–41.
- Marinus, M.G. & Casadesus, J., 2009. Roles of DNA adenine methylation in host-pathogen interactions: mismatch repair, transcriptional regulation, and more. *FEMS microbiology reviews*, 33(3), pp.488–503.
- Mathema, B., Kurepina, N.E., Bifani, P.J. & Kreiswirth, B.N., 2006. Molecular epidemiology of tuberculosis: current insights. *Clinical microbiology reviews*, 19(4), pp.658–85.
- Mcguire, A.M., Weiner, B., Park, S.T., Wapinski, I., Raman, S., Dolganov, G., Peterson, M., Riley, R., Zucker, J., Abeel, T., White, J., Sisk, P., Stolte, C., Koehrsen, M., Yamamoto, R.T., Iacobelli-martinez, M., Kidd, M.J., Maer, A.M., Schoolnik, G.K., Regev, A. & Galagan, J., 2012. Comparative analysis of Mycobacterium and related Actinomycetes yields insight into the evolution of Mycobacterium tuberculosis pathogenesis. *BMC Genomics*, 13(1), p.120.
- Mdluli, K. & Ma, Z., 2007. Mycobacterium tuberculosis DNA gyrase as a target for drug discovery. *Infectious disorders drug targets*, 7(2), pp.159–68.

- Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B., Gnirke, A., Jaenisch, R. & Lander, E.S., 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(August), pp.766–771.
- Merker, M., Kohl, T. a, Roetzer, A., Truebe, L., Richter, E., Rüsç-Gerdes, S., Fattorini, L., Oggioni, M.R., Cox, H., Varaine, F. & Niemann, S., 2013. Whole Genome Sequencing Reveals Complex Evolution Patterns of Multidrug-Resistant Mycobacterium tuberculosis Beijing Strains in Patients. *PloS one*, 8(12), p.e82551.
- Metzker, M.L., 2010. Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1), pp.31–46.
- Newton-Foot, M. & Gey van Pittius, N.C., 2013. The complex architecture of mycobacterial promoters. *Tuberculosis (Edinburgh, Scotland)*, 93(1), pp.60–74.
- Nicol, M.P. & Wilkinson, R.J., 2008. The clinical consequences of strain diversity in Mycobacterium tuberculosis. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 102(10), pp.955–65.
- Niedringhaus, T.P., Milanova, D., Kerby, M.B., Snyder, M.P. & Barron, A.E., 2011. Landscape of next-generation sequencing technologies. *Analytical chemistry*, 83(12), pp.4327–41.

- Noonan, J.P., Coop, G., Kudaravalli, S., Smith, D., Krause, J., Alessi, J., Chen, F., Platt, D., Pääbo, S., Pritchard, J.K. & Rubin, E.M., 2006. Sequencing and analysis of Neanderthal genomic DNA. *Science (New York, N.Y.)*, 314(5802), pp.1113–8.
- Nowrousian, M., 2010. Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. *Eukaryotic cell*, 9(9), pp.1300–10.
- Parwati, I., van Crevel, R. & van Soolingen, D., 2010. Possible underlying mechanisms for successful emergence of the Mycobacterium tuberculosis Beijing genotype strains. *The Lancet infectious diseases*, 10(2), pp.103–11.
- Pruitt, K.D., Tatusova, T., Brown, G.R. & Maglott, D.R., 2012. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic acids research*, 40(Database issue), pp.D130–5.
- Pruitt, K.D., Tatusova, T. & Maglott, D.R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 35(Database issue), pp.D61–5.
- Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D.R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H.B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J.,

- Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S.D. & Wang, J., 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285), pp.59–65.
- Quinlan, A.R. & Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6), pp.841–2.
- Razin, A., 1998. CpG methylation, chromatin structure and gene silencing—a three-way connection. *The EMBO journal*, 17(17), pp.4905–8.
- Reddy, T.B.K., Riley, R., Wymore, F., Montgomery, P., DeCaprio, D., Engels, R., Gellesch, M., Hubble, J., Jen, D., Jin, H., Koehrsen, M., Larson, L., Mao, M., Nitzberg, M., Sisk, P., Stolte, C., Weiner, B., White, J., Zachariah, Z.K., Sherlock, G., Galagan, J.E., Ball, C. a & Schoolnik, G.K., 2009. TB database: an integrated platform for tuberculosis research. *Nucleic acids research*, 37(Database issue), pp.D499–508.
- Roback, P., Beard, J., Baumann, D., Gille, C., Henry, K., Krohn, S., Wiste, H., Voskuil, M.I., Rainville, C. & Rutherford, R., 2007. A predicted operon map for *Mycobacterium tuberculosis*. *Nucleic acids research*, 35(15), pp.5085–95.
- Roberts, R.J., 2003. A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Research*, 31(7), pp.1805–1812.

- Roberts, R.J., Vincze, T., Posfai, J. & Macelis, D., 2010. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic acids research*, 38(Database issue), pp.D234–6.
- Roetzer, A., Diel, R., Kohl, T.A., Rückert, C., Nübel, U., Blom, J., Wirth, T., Jaenicke, S., Schuback, S., Rüsç-Gerdes, S., Supply, P., Kalinowski, J. & Niemann, S., 2013. Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. *PLoS medicine*, 10(2), p.e1001387.
- Ruffalo, M., LaFramboise, T. & Koyutürk, M., 2011. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics (Oxford, England)*, 27(20), pp.2790–6.
- Safi, H., Barnes, P.F., Lakey, D.L., Shams, H., Samten, B., Vankayalapati, R. & Howard, S.T., 2004. IS6110 functions as a mobile, monocyte-activated promoter in Mycobacterium tuberculosis. *Molecular microbiology*, 52(4), pp.999–1012.
- Sanger, F., Nicklen, S. & Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp.5463–7.
- Schadt, E.E., Turner, S. & Kasarskis, A., 2010. A window into third-generation sequencing. *Human molecular genetics*, 19(R2), pp.R227–40.

- Shell, S.S., Prestwich, E.G., Baek, S.-H., Shah, R.R., Sasseti, C.M., Dedon, P.C. & Fortune, S.M., 2013. DNA methylation impacts gene expression and ensures hypoxic survival of *Mycobacterium tuberculosis*. *PLoS pathogens*, 9(7), p.e1003419.
- Shendure, J. & Ji, H., 2008. Next-generation DNA sequencing. *Nature biotechnology*, 26(10), pp.1135–45.
- Shenker, N. & Flanagan, J.M., 2012. Intragenic DNA methylation: implications of this epigenetic mechanism for cancer research. *British journal of cancer*, 106(2), pp.248–53.
- Silhavy, T.J., Kahne, D. & Walker, S., 2010. The bacterial cell envelope. *Cold Spring Harbor perspectives in biology*, 2(5), p.a000414.
- Smith, I., 2003. *Mycobacterium tuberculosis* Pathogenesis and Molecular Determinants of Virulence. *Clinical microbiology reviews*, 16(3), pp.463-496.
- Van Soolingen, D., Qian, L., de Haas, P.E., Douglas, J.T., Traore, H., Portaels, F., Qing, H.Z., Enkhsaikan, D., Nymadawa, P. & van Embden, J.D., 1995. Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of east Asia. *Journal of clinical microbiology*, 33(12), pp.3234–8.
- Van Soolingen, D., de Haas, P.E., Blumenthal, R.M., Kremer, K., Sluijter, M., Pijnenburg, J.E., Schouls, L.M., Thole, J.E., Dessens-Kroon, M.W., van Embden, J.D. & Hermans, P.W., 1996. Host-mediated modification of PvuII

- restriction in *Mycobacterium tuberculosis*. *Journal of bacteriology*, 178(1), pp.78–84.
- Van Soolingen, D., Hoogenboezem, T., de Haas, P.E., Hermans, P.W., Koedam, M.A., Teppema, K.S., Brennan, P.J., Besra, G.S., Portaels, F., Top, J., Schouls, L.M. & van Embden, J.D., 1997. A novel pathogenic taxon of the *Mycobacterium tuberculosis* complex, Canetti: characterisation of an exceptional isolate from Africa. *International journal of systematic bacteriology*, 47(4), pp.1236–45.
- De Souza, G. a, Fortuin, S., Aguilar, D., Pando, R.H., McEvoy, C.R.E., van Helden, P.D., Koehler, C.J., Thiede, B., Warren, R.M. & Wiker, H.G., 2010. Using a label-free proteomics method to identify differentially abundant proteins in closely related hypo- and hyper-virulent clinical *Mycobacterium tuberculosis* Beijing isolates. *Molecular & cellular proteomics : MCP*, 9(11), pp.2414–23.
- Spuy, G.D. Van Der, Kremer, K., Ndabambi, S.L., Beyers, N., Dunbar, R., Marais, B.J., Helden, P.D. Van & Warren, R.M., 2009. Changing *Mycobacterium tuberculosis* population highlights clade-specific pathogenic characteristics. *Tuberculosis*, 89(2), pp.120–125.
- Srikhanta, Y.N., Gorrell, R.J., Steen, J. a, Gawthorne, J. a, Kwok, T., Grimmond, S.M., Robins-Browne, R.M. & Jennings, M.P., 2011. Phasevarion mediated epigenetic gene regulation in *Helicobacter pylori*. *PloS one*, 6(12), p.e27569.
- Srikhanta, Y.N., Maguire, T.L., Stacey, K.J., Grimmond, S.M. & Jennings, M.P., 2005. The phasevarion: a genetic system controlling coordinated, random

- switching of expression of multiple genes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(15), pp.5547–51.
- Srivastava, R., Gopinathan, K.P. & Ramakrishnan, T., 1981. Deoxyribonucleic acid methylation in mycobacteria. *Journal of bacteriology*, 148(2), pp.716–9.
- Sun, Y.-J., Lee, A.S.G., Wong, S.-Y. & Paton, N.I., 2006. Association of Mycobacterium tuberculosis Beijing genotype with tuberculosis relapse in Singapore. *Epidemiology and infection*, 134(2), pp.329–32.
- Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E. V, Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A. V, Vasudevan, S., Wolf, Y.I., Yin, J.J. & Natale, D. a, 2003. The COG database: an updated version includes eukaryotes. *BMC bioinformatics*, 4, p.41.
- Taylor, V.L., Titball, R.W. & Oyston, P.C.F., 2005. Oral immunization with a dam mutant of Yersinia pseudotuberculosis protects against plague. *Microbiology (Reading, England)*, 151(Pt 6), pp.1919–26.
- Timinskas, A., Butkus, V. & Janulaitis, A., 1995. Sequence motifs characteristic for DNA [cytosine-N4] and DNA [adenine-N6] methyltransferases. Classification of all DNA methyltransferases. *Gene*, 157(1-2), pp.3–11.
- Travers, K.J., Chin, C.-S., Rank, D.R., Eid, J.S. & Turner, S.W., 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic acids research*, 38(15), p.e159.

- Vandal, O.H., Nathan, C.F. & Ehrt, S., 2009. Acid resistance in *Mycobacterium tuberculosis*. *Journal of bacteriology*, 191(15), pp.4714–21.
- Wang, K., Pei, H., Huang, B., Zhu, X., Zhang, J., Zhou, B., Zhu, L., Zhang, Y. & Zhou, F.-F., 2013. The expression of ABC efflux pump, Rv1217c-Rv1218c, and its association with multidrug resistance of *Mycobacterium tuberculosis* in China. *Current microbiology*, 66(3), pp.222–6.
- Wang, L., Chen, S., Vergin, K.L., Giovannoni, S.J., Chan, S.W., DeMott, M.S., Taghizadeh, K., Cordero, O.X., Cutler, M., Timberlake, S., Alm, E.J., Polz, M.F., Pinhassi, J., Deng, Z. & Dedon, P.C., 2011. DNA phosphorothioation is widespread and quantized in bacterial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 108(7), pp.2963–8.
- Watson, M.E., Jarisch, J. & Smith, A.L., 2004. Inactivation of deoxyadenosine methyltransferase (dam) attenuates *Haemophilus influenzae* virulence. *Molecular microbiology*, 53(2), pp.651–64.
- Weber, M. & Schübeler, D., 2007. Genomic patterns of DNA methylation: targets and function of an epigenetic mark. *Current opinion in cell biology*, 19(3), pp.273–80.
- WHO. 2013. Global Tuberculosis Report 2013. World Health organisation.
- Wynne, J.W., Seemann, T., Bulach, D.M., Coutts, S.A., Talaat, A.M. & Michalski, W.P., 2010. Resequencing the *Mycobacterium avium* subsp. *paratuberculosis*

K10 genome: improved annotation and revised genome sequence. *Journal of bacteriology*, 192(23), pp.6319–20.

Zhang, J., Chiodini, R., Badr, A. & Zhang, G., 2011. The impact of next-generation sequencing on genomics. *Journal of genetics and genomics = Yi chuan xue bao*, 38(3), pp.95–109.

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. *The Journal of infectious diseases*, 179(5), pp.1213–7.



APPENDICES

Appendix I: Uniquely methylated promoter-associated genes for the Hyper-virulent strain

Methylated position	Motif methylated	Mod Qv	Strand	Genes in operon						
699838		49.0	+	Rv0603						
960158		26.0	-	Rv0861c						
3640526		69.0	+	Rv3261	Rv3262					
4212226		30.0	+	Rv3766						
301750		24.0	-	Rv0247c	Rv0248c	Rv0249c				
3014976		32.0	-	Rv2699c						
1442759		45.0	+	Rv1289						
2551532	CACGCAG	66.0	+	Rv2280						
881511	GTAYNNNNATC	65.0	-	Rv0786c						
2837595	GTAYNNNNATC	76.0	+	Rv2521						
256047	CTGGAGGA	44.0	+	Rv0214						

3355056		67.0	+	Rv2997									
3128995	GTAYNNNNATC	111.0	-	Rv2819c	Rv2820c	Rv2821c							
385988		37.0	-	Rv0317c									
256044		45.0	-	Rv0212c	Rv0213c								
1363387	GATNNNNRTAC	44.0	-	Rv1216c	Rv1217c	Rv1218c	Rv1219c						
1179374		28.0	+	Rv1057									
3448500		29.0	+	Rv3083	Rv3084	Rv3085	Rv3086	Rv3087	Rv3088	Rv3089			
4221070		40.0	+	Rv3776									
1951840		26.0	+	Rv1726	Rv1727								

Appendix II: Uniquely methylated promoter-associated genes for the Hypo-virulent strain

Methylated position	Motif methylated	Mod Qv	Strand	Genes in operon			
3312872		29.0	-	Rv2959c			
241392		35.0	-	Rv0201c	Rv0202c		
1417447		23.0	-	Rv1267c			
1833532		32.0	+	Rv1630	Rv1631		
1588519		31.0	-	Rv1410c	Rv1411c		
565716		25.0	+	Rv0475			
4130405		21.0	-	Rv3688c			
1512677		22.0	-	Rv1347c			
4062427	GTAYNNNNATC	39.0	+	Rv3623			
361326		25.0	+	Rv0297			
2291213		18.0	+	Rv2046			
1690366		23.0	-	Rv1498A			

2123102	28.0	+	Rv1873				
447063	17.0	-	Rv0368c	Rv0369c			
2718848	22.0	-	Rv2418c	Rv2419c	Rv2420c	Rv2421c	
2641203	20.0	+	Rv2358	Rv2359			
2740698	27.0	+	Rv2443				
4198230	42.0	-	Rv3749c				
608554	23.0	-	Rv0516c				
3017820	30.0	+	Rv2703	Rv2704			
831753	19.0	+	Rv0740				
2566768	27.0	+	Rv2295				
324610	26.0	-	Rv0269c				

Appendix III: Shared methylated promoter-associated genes in the Hyper- and Hypo-virulent strains

Methylated position	Motif methylated	Mod Qv	Strand	Genes in operon			
3621309	CACGCAG	158.0	-	Rv3241c			
4234602	CACGCAG	80.0	-	Rv3786c	Rv3787c		
3889393	GTAYNNNNATC	71.0	-	Rv3468c	Rv3469c	Rv3470c	Rv3471c
684351	CACGCAG	86.0	-	Rv0585c			
32038	GATNNNNRTAC	65.0	+	Rv0029			
1984790	CACGCAG	63.0	-	Rv1753c			
147780	CTGGAGGA	61.0	-	Rv0120c			
1399915	GATNNNNRTAC	76.0	-	Rv1251c	Rv1252c		
3926485	GATNNNNRTAC	62.0	+	Rv3507			
1852115	CACGCAG	105.0	-	Rv1638A	Rv1639c	Rv1640c	
2062683,206	GATNNNNRTAC	29.0,63.0	-	Rv1818c			
916361	CACGCAG	85.0	-	Rv0821c	Rv0822c		
2627066	GATNNNNRTAC	104.0	-	Rv2348c			
3492115	GATNNNNRTAC	122.0	+	Rv3127			

4350714	GATNNNNRTAC	98.0	+		Rv3872	Rv3873		
4056419		50.0	-		Rv3616c			
1717616	GATNNNNRTAC	52.0	+		Rv1523	Rv1524	Rv1525	
3701027	GTAYNNNNATC	127.0	-		Rv3312A			
131170	CACGCAG	90.0	-		Rv0108c			
2835424	GTAYNNNNATC	54.0	-		Rv2518c			
1717623	GTAYNNNNATC	74.0	-		Rv1522c			
3550341		82.0	+		Rv3182	Rv3183		
1302765	GATNNNNRTAC	78.0	-		Rv1172c			
4052925	CACGCAG	57.0	+		Rv3611			
3700356	GATNNNNRTAC	159.0	-		Rv3312c			
1374277	GTAYNNNNATC	68.0	-		Rv1229c	Rv1230c		
703208	GATNNNNRTAC	55.0	+		Rv0608	Rv0609	Rv0609A	
3059249	GTAYNNNNATC	101.0	-		Rv2746c			
2752259	CACGCAG	54.0	-		Rv2450c			
1736478	CACGCAG	90.0	+		Rv1536			
1277802	CACGCAG	101.0	+		Rv1149			

1231316	CACGCAG	86.0	-	Rv1102c	Rv1103c		
1902334	CACGCAG	79.0	+	Rv1678	Rv1679	Rv1680	Rv1681
1211498	GATNNNNRTAC	65.0	+	Rv1087			
210866	CACGCAG	68.0	-	Rv0179c			
3059242	GATNNNNRTAC	78.0	+	Rv2747			
1399908	GTAYNNNNATC	59.0	+	Rv1253	Rv1254		
1951779	CACGCAG	56.0	-	Rv1724c	Rv1725c		
468039	CACGCAG	72.0	-	Rv0387c	Rv0388c		
3133591	GTAYNNNNATC	73.0	-	Rv2825c			
302066	GTAYNNNNATC	80.0	-	Rv0250c			
1172186	CTGGAGGA	26.0	-	Rv1048c			
5163,5231	GATNNNNRTAC, GTAYNNNNATC	96.0,72.0	+	Rv0005	Rv0006		
2916222	GATNNNNRTAC	91.0	-	Rv2588c			
1410366	CACGCAG	66.0	+	Rv1263			