GENOME ASSEMBLY OF NEXT-GENERATION SEQUENCING DATA FOR THE *ORYX BACILLUS*: SPECIES OF THE *MYCOBACTERIUM TUBERCULOSIS* COMPLEX

UNIVERSITY *of the*
WESTERN CAPE

Name: **Mmakamohelo Direko**

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at the South African National Bioinformatics Institute, Department of Biotechnology, Faculty of Science, University of the Western Cape

Date submitted for examination: December 2011

Supervisor: **Professor Alan Christoffels**

Co-supervisor: **Dr Junaid Gamieldien**

KEYWORDS

Mycobacterium tuberculosis

Oryx bacillus

Next generation sequencing

ABI SOLiD system

Genome assembly

Annotation pipeline

Comparative genomics

Single nucleotide polymorphism

Annotation

Phylogeny.

ABSTRACT

Next generation sequencing (NGS) technology platforms have accelerated ability to produce completed genome assemblies. Recently, collaborators at Tygerberg Medical School outsourced the sequencing of *Oryx bacillus*, a member of the *Mycobacterium tuberculosis* complex (MTC). A total of 31,271,059 short reads were generated and required filtering, assembly and annotation using bioinformatics algorithms. In this project, an NGS assembly pipeline was implemented, tailored specifically for SOLiD sequence data.

The raw reads were aligned to seven fully sequenced and annotated MTC members, namely, *Mycobacterium tuberculosis* H37Rv, H37Ra, CDC1551, F11, KZN 1435, *Mycobacterium bovis* AF2122/97 and *Mycobacterium bovis* BCG str. Pasteur 1173P2 using NovoalignCS. Depth and breadth of sequence coverage across each base of the reference genome was calculated using BEDTools, and structural variation. Structural variation at the nucleotide level including deletions, insertions and single nucleotide polymorphisms (SNPs) were called using three tools, GATK, SAMtools and Nesoni. These variations were further filtered using in-house PERL scripts. Putative functional roles for the alterations at the DNA level were extrapolated from the overlap with essential genes present in annotated MTC members.

Approximately 20,730,631 short reads (59.78%) out of a total of 31,271,059 reads aligned to the seven reference genomes. The per base sequence coverage calculations revealed an average of 1,243 unaligned regions. These unaligned regions overlapped with mycobacterial regions of difference (RD) and genetic phage elements acquired by the MTC through horizontal gene transfer and are genes prevalent in the clinical isolates of *M. tuberculosis*. A total of 2,680 genetic variations were identified and categorised into 845 synonymous and 1,724 non-synonymous SNPs together with 44 insertions and 67 deletions. Some of the variant alleles overlapped known genes to be involved in TB drug resistance. While the biological significance of our findings remain to be elucidated, it nonetheless deserves further attention, because SNPs have the potential to impact on strain phenotype by gene disruption. Therefore, any hypotheses generated from these large-scale analyses will be tested by our collaborators at Tygerberg medical school.

DECLARATION

I declare that "Genome assembly of next-generation sequencing data for the *Oryx bacillus*: species of the mycobacterium complex" is my own work, that it has not been submitted for degree or examination at any other university, and that all the resources I have used or quoted, and all work which was the result of joint effort, have been indicated and acknowledged by complete references.

Mmakamohelo Direko **December 2011**

ACKNOWLEDGEMENTS

This work was carried out at the South African National Bioinformatics Institute (SANBI), University of the Western Cape. The *Oryx bacillus*, *Mycobacterium tuberculosis* complex member, was isolated by the Tygerberg Hospital medical group, University of Stellenbosch and was sequenced by Co-factor Genomics, USA. Financial assistance for this MSc research was provided by the National research Foundation (NRF). With great pleasure, I would humbly like to express my sincere gratitude to:

The Almighty God for giving me an opportunity of this magnitude. Through it all, I have learned to trust and depend on God.

Supervisor and project leader - Professor Alan Christoffels for his warm welcome into his multicultural and diverse research group, SANBI. I am sincerely grateful to him for believing in me, his patience, and support and for grooming and allowing me to learn to the level, I am today.

The TB research team at Stellenbosch Tygerberg medical campus, Professor Nicolaas Claudius Gey van Pittius whose lab provided us with the data for this MSc thesis. The TB research team at SANBI, for being the good people they are, it would have not been the same without them. Through them I have learned so much, from an academic point of view, to a personal level. Knowing them was a lifetime experience, and as long as we have memories, yesterday remains.

My good friend and colleague, Dr. Mark Wamalwa, for being there always when I needed him most and for making the work seem easy when I felt the world was crushing down on me.

The technical team SANBI, Mario, Peter and Dale through every computational problem, software crashes, laptop deaths and script writing advise. They have never closed their doors on me.

A special thanks to my family and friends, Mmatshiu Direko, Dikgang Direko, Tlotlisang, Teboho, Ncebakazi Galada, Thandiwe Matube, Ziyanda Silevu, Kholeka Jamela, Anelisa Jaca, Kristen Wolfenden, Aasiyah Chafekar, Salome Meso, Fridah, Mpho and Kabelo Meso. Thank you for being there for me even on days when my goals in life did not make any sense.

TABLE OF CONTENTS

Keywords ii
Abstract iii
Declaration iv
Acknowledgements v
Table of contents vi
Chapter 1 1
Introduction and literature review 1
1.1 History of tuberculosis 1
1.2 *Mycobacterium tuberculosis* complex (MTC) 2
1.3 Tuberculosis infection 2
1.3.1 Mode of Transmission 3
1.3.2 Epidemiology of TB 3
1.3.3 Treatment 4
1.3.4 Vaccination 5
1.4 Evolution of *Mycobacterium tuberculosis* 5
1.5 Emerging *Mycobacterium tuberculosis* strain-*Oryx bacillus* 11
1.6 Next generation sequencing 11
1.7 ABI SOLiD sequencing 14
1.8 Genome alignment and assembly 17
1.9 Comparative genomics analysis of the MTC 18
1.10 Gene duplication and multigene families 19
1.11 Genetic variation and annotation 20
1.11.1 Single nucleotide polymorphism 20
1.11.2 Insertion sequences and prophages 21
1.12 Thesis Rationale 23
1.13 Aims of the study 23
1.14 Thesis outline 23
Chapter 2 25
Materials and methods 25
2.1 Data Sources 25
2.1.1 ABI SOLiD file formats 27
2.1.2 Quality assessment and quality control 27
2.1.3 Read alignment and mapping 30
2.1.4 Alignment statistics 31
2.1.5 SAM file validation 31
2.1.6 Converting SAM to BAM 32
2.1.7 Post alignment processing of BAM files 32
2.1.8 Per base coverage calculations 33
2.1.9 Per gene coverage calculations 34
2.1.10 Zero coverage genes 34
2.2 *De novo* assembly 34
2.2.1 Sequence similarity search for unmapped contigs 36
2.2.2 Functional annotation 36
2.3 SNP calling 37
2.4 Short Insertion/Deletion (Indel) Calling 39
2.5 Indels and SNP annotation/filtering 39
2.6 Validation of genetic variants 40

Chapter 3 41
3.1 Quality assessment and quality control: basic Statistics 41
3.1.1 Sequence duplication levels 45
3.2 Genome assembly of *Oryx bacillus* 46
3.2.1 Read alignment mapping 46
3.2.2 Per base coverage 47
3.2.3 Significance of zero-covered regions 48
3.2.4 Deletions specific to *Oryx bacillus* 51
3.2.5 Visualization 53
3.2.6 Unmapped read assembly 55
3.2.7 Annotation of unmapped contigs 56
3.2.8 Functional annotation: GO analysis 57
3.3 SNP and Indel calling 59
3.3.1 Computational predictions of genetic variations in *Oryx bacillus* 59
3.3.2 Sequence Alignment and mapping 60
3.3.3 Genetic variations between *Oryx bacillus* and *M. tuberculosis* strains 61
3.3.4 Functional analysis of non-synonymous SNPs 66
3.3.5 Mutations associated with transcription factors and global regulations 66
3.3.6 Mutations affecting genes associated with metabolism and respiration 67
3.3.7 Stress-response and general metabolic proteins 68
3.3.8 Mutations affecting cell envelope and virulence 69
3.3.9 Mutations related to drug resistance 72
Chapter 4 74
discussion 74
Chapter 5 81
Conclusions 81
5.1 Genome assembly of *Oryx bacillus* 82
5.2 Genetic variations 85
5.3 Limitations of this study 87
5.4 Future perspective 88
5.5 References 89

LIST OF ABBREVIATIONS

TB	Tuberculosis
MTC	*Mycobacterium tuberculosis* complex
XDR	extremely drug resistant
MDR	Multi drug resistant
HGT	horizontal gene transfer
NR	non-redundant
BLAST	basic local alignment search tool
Contig	consensus sequence
DNA	deoxyribonucleic acid
RD	region of difference
E-value	expectation value
KEGG	Kyoto encyclopedia of genes and genomes
ORF	open reading frame
BCG	Bacille Calmette-Guerin
HIV	Human immuno-deficiency virus
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
SNP	Single nucleotide polymorphism
ABI	Applied Biosystems, Inc.

LIST OF TABLES

	Title	Page
Table 1.1:	Sequenced MTC genomes …….....…….	8
Table 2.1:	The *Mycobacterium tuberculosis* reference strains used in this study …….	26
Table 3.1	FastQC generated statistics for the *Oryx bacillus* reads……	41
Table 3.2:	The sequence alignment statistics between *Oryx bacillus* and MTC reference genomes……...	47
Table 3.3:	Per base zero coverage regions on the reference genome considered in this study……..	48
Table 3.4:	Reference genes that overlap the zero coverage regions……	51
Table 3.5:	Reference genes that correspond to regions of difference specific to *Oryx bacillus*…….…….....	52
Table 3.6:	A summary of the assembly statistics of the unmapped reads……	55
Table 3.7:	A summary of sequence similarity search results for unmapped contigs…….…...	57
Table 3.8:	Summary of *Oryx bacillus* reads mapped to five reference genomes……..	60
Table 3.9:	Curated genetic variations between *Oryx bacillus* and *M. tuberculosis* strains…….	64
Table 3.10:	Functional categories of non-synonymous SNPs, identified in *Oryx bacillus* compared to five reference genomes……...	68
Table 3.11:	Mutations in genes implicated in invasion and persistence in the host macrophages…….....	70
Table 3.12:	Non-synonymous SNP containing genes relative to drug resistance in *Oryx bacillus*……	73

LIST OF FIGURES

	Title	Page
Figure 1.1:	Pathogenesis of tuberculosis…….……..	3
Figure 1.2:	Epidemiology of tuberculosis……	4
Figure 1.3:	Phylogenetic tree of members of the *Mycobacterium tuberculosis* complex……	7
Figure 1.4:	Phylogeny of *Mycobacterium tuberculosis* Complex	10
Figure 1.5:	The deBruijn graph algorithm for construction of contigs ……..…..…….	13
Figure 1.6:	Velvet assembler deBruijn approach.....	16
Figure 2.1:	ABI SOLiD input file format…….	27
Figure 2.2:	A computational workflow for the analysis of *Oryx bacillus* short reads …….....	29
Figure 2.3:	Workflow for post-processing analysis of the alignment BAM file……	33
Figure 3.1:	A plot showing quality scores across all bases……...	42
Figure 3.2:	Distribution of quality scores across all sequences in the library……..	42
Figure 3.3:	The sequence content across all bases in the library.....	43
Figure 3.4:	The mean % G+C content across all bases …….....	44
Figure 3.5:	Distribution of the mean % G+C content over all sequences…….....	45
Figure 3.6:	Sequence duplication levels……	46
Figure 3.7:	Genome coverage per base plots……	50
Figure 3.8 (a):	Zero coverage region overlaps MmpL14 gene encoded by CDC1551 reference genome……...	53
Figure 3.8 (b):	Zero coverage region overlaps phage-like element (phiRv1; RD3) in H37Rv reference genome…….	54
Figure 3.9:	The *PPE* gene coverage by *Oryx bacillus* reads……...	54
Figure 3.10:	A cumulative plot for the length distribution of the generated contigs……..	56
Figure 3.11:	Functional annotation of *Oryx bacillus* unmapped reads by GO terms…….	58

Figure 3.12:	Functional annotation of *Oryx bacillus* unmapped reads by GO Cellular component terms…….	59
Figure 3.13:	Venn diagram of the number of SNPs identified by three NGS SNP callers between *Oryx bacillus* and H37Rv…….	62
Figure 3.14:	Venn diagram of the number of SNPs identified by three NGS SNP callers between *Oryx bacillus* and H37Ra……..	62
Figure 3.15:	Venn diagram of the number of SNPs identified by three NGS SNP callers between Oryx bacillus and CDC1551…….	62
Figure 3.16:	Venn diagram of the number of SNPs identified by three NGS SNP callers between Oryx bacillus and F11	63
Figure 3.17:	Venn diagram of the number of SNPs identified by three NGS SNP callers between Oryx bacillus and KZN_1435……	63
Figure 3.18:	Genetic variation (Indels) between Oryx bacillus and MTC strains……	65
Figure 3.19:	Visualisation of nsSNP using IGV……	71
Figure 3.20:	Visualisation of indels using IGV……..	72

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 History of tuberculosis

Tuberculosis (TB) is caused by a group of closely related mycobacterial strains collectively known as the *Mycobacterium tuberculosis* complex (MTC). The disease kills approximately 2 million people yearly and there are about 9.27 million new infections annually (WHO, 2009; WHO, 2011). The mycobacteria genus consists of bacteria pathogenic to humans and animals. Together with other related genera such as the *Nocardia*, and *Corynebacteria*, the *Mycobacterium* genus is part of the *Actinobacteria*, a class that represents gram-positive bacteria with a high genomic G+C content (Runyon, 1959).

Mycobacteria are obligate aerobic, acid-fast rods that can be categorised into two sub-lineages (fast and slow growers). While *Mycobacterium abscessus*, which belongs to the fast growing mycobacteria, is considered incapable of causing harm, the majority of mycobacteria are slow growers and harmful. However, *Mycobacterium tuberculosis* is the most virulent of the mycobacteria (Pieters and Gatfield, 2002).

The discovery of *M. tuberculosis* as the causative agent of TB marked a key milestone in TB research (Koch, 1882). *Mycobacterium tuberculosis* was isolated from a tuberculous tissue and described as a rod-shaped bacillus and later, the isolate was experimentally verified using the Henle-Koch postulates (Koch, 1982). Tuberculosis has however, plagued human kind throughout its history and has possibly resulted in more deaths than any other microbial pathogen. Although other *M. tuberculosis*-like strains can cause TB in humans, these strains are generally restricted to their respective hosts, for example, *M. caprae* causes TB in goats, while the *Oryx bacillus* causes TB in antelopes.

Currently available treatment options for TB include first line drugs (Isoniazid, Rifampicin,Streptomycin), second-line drugs (Fluoroquinolones, Ethionamide) and third-line drugs (Amoxicillin-clavulanate, clofazimine). Despite the availability of these drugs, the microbe has the ability to evade current treatments, which require prolonged use by patients. Inefficient diagnostic methods coupled to poor drug regiment have resulted into emergence of drug-resistant TB strains.

1.2 *Mycobacterium tuberculosis* complex (MTC)

The members of MTC are genotypically closely related with a 99.9% similarity at the nucleotide level, however, they are phenotypically distinct, with different levels of virulence and a wide host range. MTC strains lack interstrain genetic diversity, hence tuberculosis occurs as a result of a number of genetically related species (Cole, 2002). However, despite their relatedness at the genomic level, these members differ remarkably with respect to their host range and pathogenicity, for example, voles are infected by *M. microti*, dassies are infected by *Dassie bacillus*, seals are infected by *M. pinnipedii*, antelopes are infected by *Oryx bacillu*s while cattle and goats are infected by *M. bovis* and *M. caprae* respectively (Greth et al., 1994; Behr et al., 1999; Veyrier et al., 2011).

1.3 Tuberculosis infection

Tuberculosis is an airborne disease primarily characterised by a small patch of caseous bronchopneumonia to the lower lobes of the lung(s). Once an active infection has been established, it usually stabilizes after a certain period, and later goes into latency. In some cases, the disease manifests as chronic pulmonary or extra-pulmonary TB. Active TB disease either remains localised, to the lungs or spreads to other parts of the body (Figure 1.1). *M. tuberculosis* can cause infection anywhere within the host body but the primary focus is mainly within the lungs (Crevel et al., 2002). Depending on the infected host, the disease can continue for months to years and can be fatal if untreated. Tuberculosis is a highly infectious disease and can be transmitted between species; this phenomenon was introduced in 1865, when Villemin transmitted the disease from human autopsy material to a rabbit (Daniel, 2006; http://www.britannica.com/EBchecked/topic/629218/Jean-Antoine-Villemin).

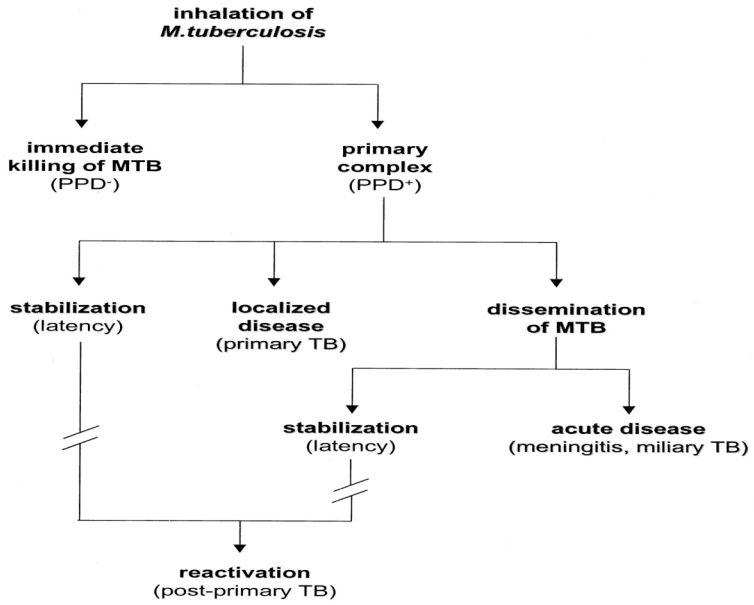


Figure 1.1. Pathogenesis of tuberculosis.

Disease progression after inhalation of *M. tuberculosis* (Crevel et al., 2002).

1.3.1 Mode of Transmission

Tuberculosis is an airborne disease, transmitted mainly through inhalation when hosts with active TB cough, sneeze or spit. Transmission can be contained by isolation of patients with active TB and commencing anti-tuberculosis therapy.

1.3.2 Epidemiology of TB

Tuberculosis is a world-wide epidemic with approximately 9.27 million new cases and two million deaths annually (WHO; report 2009). Despite the availability of chemotherapy and vaccines, about one third of the world population is latently infected, 10% of which develop disease during their lifetime (Figure 1.2). The rise of HIV infections and abuse of TB control programs have contributed to resurgence of TB infections.

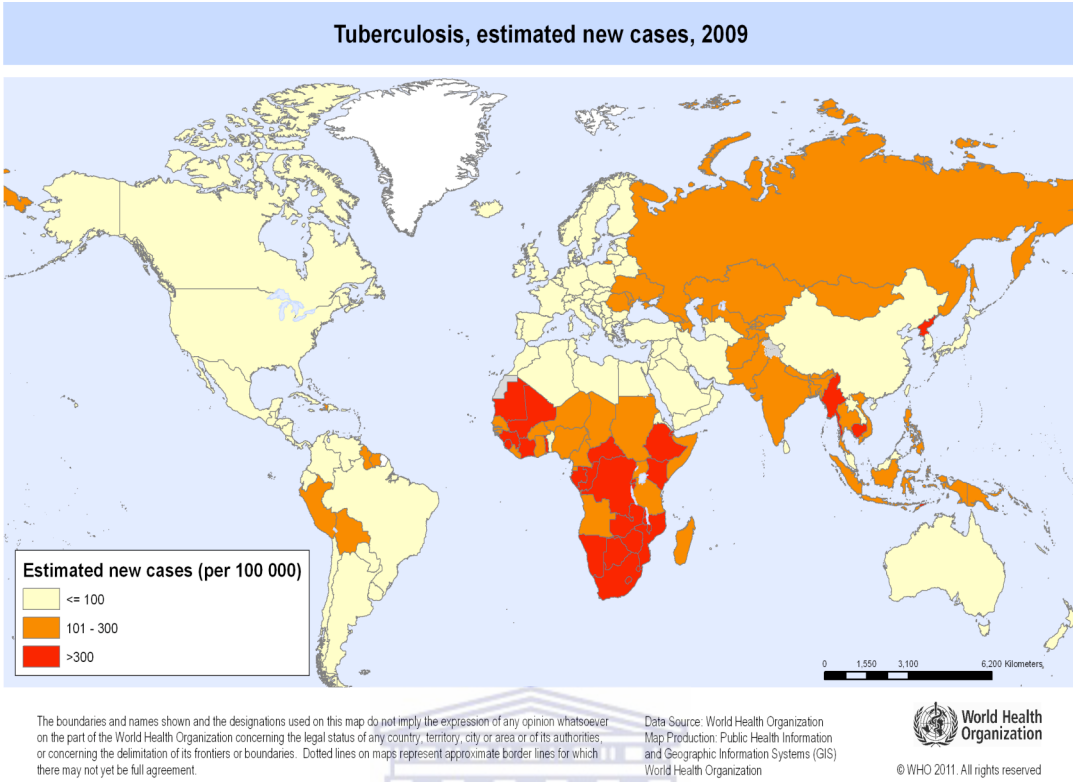


Figure 1.2. Epidemiology of tuberculosis.
Global distribution of the disease (WHO, 2009).

1.3.3 Treatment

The discovery of the tuberculin skin-test paved a way into TB research leading to the development of therapeutic agents against the disease. Tuberculosis can be treated by antibiotics for example, streptomycin (Jones et al., 1944; Schatz et al., 1944). Despite the use of antibiotics, multi-drug resistant (MDR) and extensively drug-resistant (XDR) strains of *M. tuberculosis* have emerged, due to deficiencies in antibiotic use at the level of the clinic or at the level of the TB management program. Overcoming drug resistance requires a combination of drug regimes. Isoniazid (INH) (Steenken et al., 1952) and rifampicin (RMP) (Furesz et al., 1963) are the two most common first line drugs, with pyrazinamide (PZA) (McDermott et al., 1954) and ethambutol (EMB) (Forbes et al., 1962) also commonly used with a 6-9 months treatment regime. Patient compliance and non-adherence to drug regimen has led to the emergence of MDR and XDR-TB (Pablos-Mendez et al., 1998). It is estimated that 55 countries had globally reported at least one extensively-drug resistant TB (XDR-TB) case with 0.5 million

cases of multi-drug resistant TB (MDR-TB) (Figure 1.2) (WHO, 2009), hence there is a great need for new antimicrobial compounds against *M. tuberculosis*.

1.3.4 Vaccination

Despite drug resistance, vaccination against TB remains the most viable public health intervention that is likely to affect both the incidence and the prevalence of the disease (WHO, 2009). As a result, a worldwide campaign against TB was initiated with the involvement of the World Health Organisation, UNICEF and the Red Cross which was based on the use of tuberculin skin test followed by Bacille Calmette-Guerin (BCG) vaccination, a live attenuated bacterium (Comstock, 1994). Current approved prophylactic TB vaccines (BCG and its derivatives) are of variable efficiency in adult protection against pulmonary TB (0%–80%), and are directed against the early phases of infection (Calmette, 1924; Calmette, 1928).

1.4 Evolution of *Mycobacterium tuberculosis*

Members of the *Mycobacterium* genus are dichotomously classified and they belong to the *Mycobaterium tuberculosis* complex (Veyrier et al., 2011). The availability of genome sequences of various strains of the *M. tuberculosis* complex suggest that modern strains namely *Mycobacterium tuberculosis*, *Mycobacterium bovis*, *Mycobacterium africanum* (subtype 1a and 1b), *Mycobacterium microti*, *Mycobacterium pinnipedii, Oryx bacillus* as well as the *Dassie bacillus*, had a common African ancestor, approximately 15-35 million years ago (Sreevatsan et al., 1997; Brosch et al., 2002). *M. tuberculosis* is the first member to have found its origin as a member of the MTC (Figure 1.3). It originated from a prototuberculosis as a result of an evolutionary bottleneck (Gutierrez et al., 2005).

Mycobacterium tuberculosis complex genomic deletions (denoted as RD or regions of difference) are irreversible evolutionary events used as important biomarkers (Cole et al., 1998). Eleven of the fourteen RD's (RD1, RD4, RD5, RD6, RD7, RD9, RD10, RD11, RD12, RD13 and RD15), which encompasses >80kb genomic DNA in *M. tuberculosis*, are absent from the vaccine strains of *M. bovis* (Mustafa and Al-Attiyah, 2009). RD3 is present in *M. tuberculosis* and all virulent isolates of *M. bovis* but is absent in the *M. bovis* BCG (Bibb and Hatfull, 2002). This region is situated between

two genes *bioD* and *bioB* where the REP13E12 family serves as an attachment for the phiRv1 phage (Bibb and Hatfull, 2002). It encodes 14 complete open reading frames (ORFs) of which five are predicted to encode proteins similar to putative phage proteins, namely, the capsid subunit, prohead protease, terminase, promase/helicase and an intergrase (Bibb and Hatfull, 2002; Bibb et al., 2005). Due to the few ORFs and the limited subset of phage genes represented in the RD3, this region is not an intact prophage capable of independently generating infectious particles. Hence, it is referred to as a prophage-like element (phiRv1) (Hendrix, 1999; Bibb and Hatfull, 2002). The reference genome H37Rv carries an additional copy of a prophage-like element (phiRv2) which is absent in *M. bovis* BCG strain. PhiRv2 prophage-like element is specific to *M. tuberculosis*, however, it occupies the RD11 and has similar characteristics to phiRv1 (Cole et al., 1998; Gutierrez et al., 2005). Relative to other MTC members, H37Rv genome encodes deletions (RvD1-5) as well as RD1, an *M. tuberculosis* specific deletion (TbD1) (Brosch et al., 2002). Previous studies identified regions of difference RvD2, RvD3, RvD4, RD5, RD7, RD8, RD9, RD10 and RD13 specific to *Oryx bacillus* (Brosch et al., 2002; Marmiesse, 2004; Mostowy et al., 2005).

Based on the presence or absence of an *M. tuberculosis* specific RD regions, *M. tuberculosis* strains can be divided into ancestral and "modern" strains, the latter comprising representatives of major epidemics like the CDC1551, H37Rv, H37Ra, XDR (KZN 605), MDR (KZN 1435), DS (KZN 4207), F11, Haarlem, and African *M. tuberculosis* clusters (Brosch et al., 2002). The TbD1 region however is absent in the laboratory strain H37Rv.

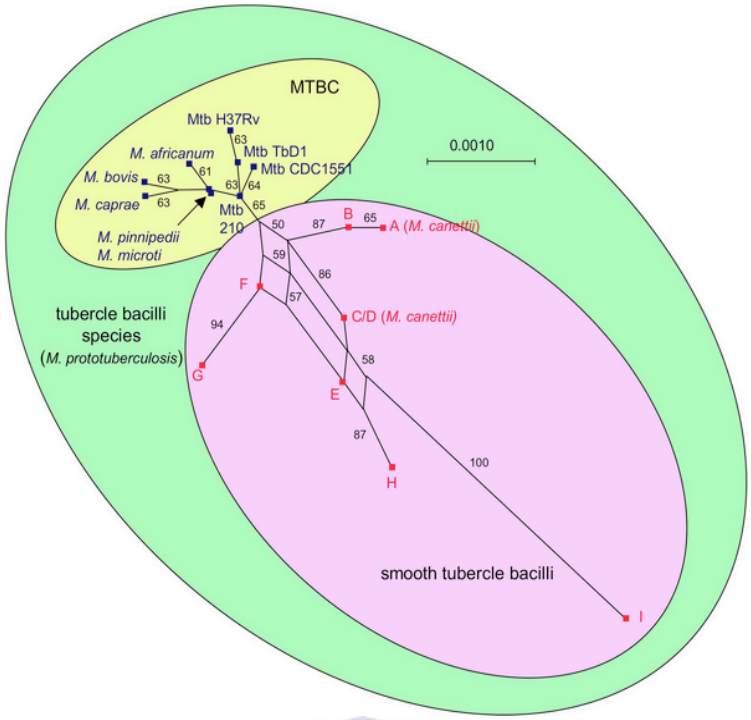


Figure 1.3. Phylogenetic tree of members of the *Mycobacterium tuberculosis* complex.

The ancestral nodes (red color) of the tree represent the smooth tubercle bacilli, while blue nodes represent MTC members (Gutierrez et al., 2005).

Oryx bacillus was first isolated and characterised from antelopes (*Oryx*) and is a subtype of *M. bovis*, a strain with a very broad host spectrum infecting many mammalian species, including man (Brosch et al., 2002). As an *M. bovis* isolate and a member of the MTC, *Oryx bacillus* is not only infectious to antelopes but it has the potential to cause infection in other mammalian species including man.

The laboratory strain *M. tuberculosis* H37Rv, was first isolated in 1905 and has been used as a reference genome in demonstrating differences between recent isolates and already existing genomes (Camus et al., 2002). H37Ra is an avirulent, attenuated progeny of the strain H37Rv and comprises of 3,989 genes (Kato-Maeda et al., 2001; Cubillos-Ruiz et al., 2008).

Mycobacterium tuberculosis CDC1551 genome was sequenced by the institute for genomic research (TIGR) and it encodes 4,189 genes (Table 1.1). CDC1551 is a highly infectious strain which was isolated following an outbreak in the United States (Garnier et al., 2003). The strain's virulence is associated with the Mycobacterial

membrane lipid Large (MmpL) which when compared to H37Rv, has an extra copy MmpL14, acquired through gene duplication (Domenech et al., 2005).

Mycobacterium tuberculosis F11 strain is an aerobic, chemoorganotroph, rod-shaped, non-motile, non-sporulating human pathogen isolated from TB patients in the Western Cape region of South Africa. The *M. tuberculosis* F11 genome was sequenced by The Broad Institute and has a total of 3,959 genes (Table 1.1). Victor and colleagues (2004) reported that isolates of F11 are not only a major contribution to the TB epidemic in South Africa but are also present in different continents and several other countries in the world.

Mycobacterium tuberculosis Haarlem is an aerobic human pathogen first discovered in Haarlem, The Netherlands, belongs to a widely distributed genotype with reduced virulence in murine models (Cubillos-Ruiz et al., 2008). It is multi-drug resistant (MDR) and its genome was sequenced by The Broad Institute. The genome of the Harlem strain encodes 3,866 genes.

Table 1.1. Sequenced MTC genomes used in this study

Genome	Size	Number of genes	Sequencing center
Mycobacterium tuberculosis H37Rv	4.41 Mb	3989 genes	TIGR
Mycobacterium tuberculosis H37Ra	4.42 Mb	4034 genes	Broad Institute
Mycobacterium tuberculosis F11	4.42 Mb	3959 genes	Broad Institute
Mycobacterium tuberculosis KZN1435	4.38 Mb	3851 genes	Broad Institute
Mycobacterium tuberculosis CDC1551	4.4 Mb	4189 genes	TIGR
Mycobacterium bovis BCG	4.3 Mb	3952 genes	Institut Pasteur
Mycobacterium bovis AF2122/97	4.3 Mb	3920 genes	Institut Pasteur

The KZN (Kwazulu-Natal) strain family of *M. tuberculosis* have recently caused an outbreak of extensively-drug resistant TB in the Kwazulu-Natal region of South Africa (Ioerger et al., 2009). Between 2005 and 2007, the KZN strain virulence had increased to 80% (Gandhi et al., 2006). This strain family constitutes three members; the extensively-drug resistant (XDR), the multi-drug resistant (MDR) and the drug-sensitive (DS) strains. The XDR strain may have emerged from a pre-existing F15/LAM4/KZN family strain through clonal expansion (Ioerger et al., 2009). The strain's mode of transmission is still unclear except the fact that it has mutations conferring drug resistance (Dheda et al., 2010). The XDR and MDR strains share a

130 bp (base pair) deletion in the glucose-inhibited division protein B (gidB). However, XDR is resistant to two more second-line TB-drugs (kanamycin and ofloxan) while the MDR is resistant to all three injectable first-line TB-drugs (isoniazid, rifampicin and the fluoroquinolines) (Gandhi et al., 2006). These three KZN isolates were sequenced by the Broad Institute. The XDR strain has a total of 4,024 genes while the MDR encodes 3,851 genes.

Mycobacterium africanum is a highly heterogeneous strain with characteristics similar to *M. tuberculosis* and *M. bovis*. It was isolated in Senegal, West Africa in 1968 (Castets et al., 1969). As a result of its geographical origin, *M. africanum* species are differentiated into two major subtypes, the West Africa subtype l closer to *M. bovis* and the East Africa subtype ll closer to *M. tuberculosis* (Niemann et al., 2002). In contrast to *M. bovis* and *M. tuberculosis*, the *M. africanum* strain shows high variation of phenotypic attributes (Niemann et al., 2002).

Bovine tuberculosis is caused by *M. bovis*, a mycobacterium strain that is highly similar to *M. tuberculosis* and the second most epidemiologically and historically notable sub-species which is a member of the MTC. The main host of *M. bovis* is cattle (*Bos Taurus*) and goats, however, it has been shown to affect humans (Karlson and Lessel, 1970). Sheep and horses are rarely infected by *M. bovis* (Long et al., 1999; Palomino et al., 2007). Variation between the *M. tuberculosis* and *M. bovis* resides either in large genomic RD regions or in SNPs (Figure 1.4). Brosch and co-workers (2002) demonstrated that the region of difference 9 (RD 9) was absent from all strains of the evolutionary lineage represented by *M. africanum* and *M. bovis*, but present in *M. tuberculosis*. Additionally, the TbD1 deleted region is absent in *M. tuberculosis* but present in *M. bovis* (Brosch et al., 2002). These findings demonstrated that *M. bovis* as well as the *M. tuberculosis* may have evolved from a common ancestor (Brosch et al., 2000).

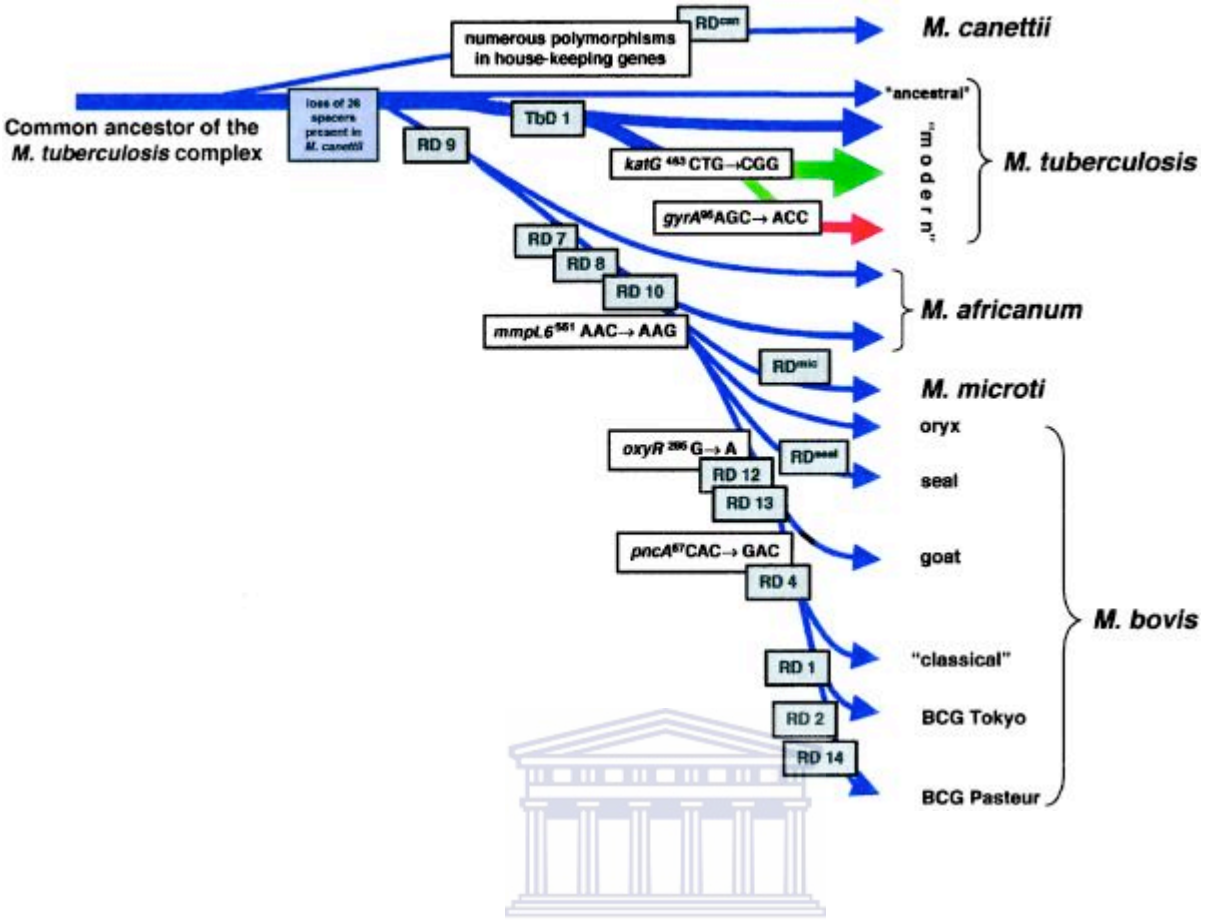


Figure 1.4. Phylogeny of *Mycobacterium tuberculosis* Complex

The proposed evolutionary pathway of the tubercle bacilli illustrating successive loss of DNA in certain lineages (grey boxes). Blue arrows indicate strains that are characterised by mutations of *katG*463 gene. Mutations involving *gyrA* (CTG (Leu), *gyrA*95 ACC (Thr)) are typical for group one mycobacterial strains. Green arrows indicate strains that belong to group two characterized by the following mutations of *katG*463: CGG (Arg), *gyrA*95 ACC (Thr). The red arrow indicates that strains belong to group three, characterised by mutations of *katG*463 involving the following changes: CGG (Arg), *gyrA*95 AGC (Ser)(Brosch et al., 2002).

1.5 Emerging *Mycobacterium tuberculosis* strain-*Oryx bacillus*

Oryx bacillus is an MTC member and a causative agent of tuberculosis in antelopes. It was first isolated from the Arabian Oryx and camelids in 1986 (Greth et al, 1994). The bacterium was recently isolated from a healthy buffalo on a farm known to have had imported buffalo from a zoo in Portugal (Van Helden et al., 2009). Despite its prevalence to cause infection in the Arabian Oryx, *Oryx bacillus* has not been reported to cause infection to the South African Gemsbok (*Oryx gazella*). However, it may be hypothesised that it would also be pathogenic to Gemsbok or that *Oryx bacillus*, like *Dassie bacillus* is less virulent than *M. tuberculosis* or *M. bovis*. Recently, *Oryx bacillus* strain was isolated causing disease in human patients in Australia (Fyfe, 2011). This observation suggests that *Oryx bacillus* can be transmissible within and between a range of hosts including humans and animals.

1.6 Next generation sequencing

The advent of new sequencing technologies, such as Roche 454, SOLEXA and SOLiD has seen an explosion of new data with the advantages of high throughput, reduced time and cost (Margulies et al., 2005; Bentley et al., 2008; Valouev et al., 2008). These methods are used to decode the exact order in which DNA occurs in the chromosome. The Maxam-Gilbert sequencing approach, also known as chemical sequencing, was the first sequencing platform capable of reading a maximum of 600 bps from either end of a DNA fragment. Since this was the only sequencing platform, genome assembly softwares were optimised to use fragments of this size.

In comparison with the traditional Sanger sequencing platform, the next generation sequencing (NGS) technologies exploit massive parallelisation at the DNA polymerization step, generating millions of DNA reads. The ABI SOLiD sequencer is currently capable of sequencing between 35-70 bp. Therefore, existing assembly software have been adapted to better suit the NGS read lengths. NGS technologies are being improved and used for the characterisation of evolutionary relationships in ancient genomes giving better understanding of genome evolution.

The ABI SOLiD system is based on ligating flourescently labelled dinucleotide probes to the DNA template under investigation (Tomkinson et al., 2006). Following emulsion PCR, beads are covalently bound to a glass slide along with the universal sequencing primers while ligases and labelled dinucleotide probes are added to the glass slide. The final DNA sequence is determined through recording the color code

which represents the first two bases of the dinucleotide (Metzker, 2005, 2010). SOLiD uses a version of a dinucleotide probe where the first and second nucleotides are analysed. Several cycles of DNA ligation and primer resetting are repeated. Based on different library preparation protocols one can differentiate single-end (SE), paired-end (PE) and mate-paired (MP) libraries available for sequencing.

A single raw read sequence corresponds to a color space sequence. Following sequencing, the NGS raw reads have to be assembled into contigs either by alignment to a known reference genome or by *de novo* assembly. Alignment can be described as a process of determining the most likely source within the genome sequence for the observed DNA sequence read, provided the genome sequence is known. Traditional alignment programs for the Sanger sequencing are for example BLAST and BLAT but they may not scale well with the NGS reads in terms of mapping accuracy.

The observation that NGS reads are much shorter has led researchers to design bioinformatics tools and algorithms specific for these reads, which include alignment, and *de novo* assembly tools (Zerbino et al., 2008). Most of these tools encompass a heuristic technique to help identify small subsets of the reference genome where reads are most likely able to align. Besides aligning the raw reads to a reference genome, the sequenced reads may be assembled into contigs *de novo*. The de Bruijn approach has been extensively used in the *de novo* assembly of genomes and transcriptomes (Zerbino et al., 2008).

The de Bruijn graph consists of a set of nodes, and edges, where each node represents a sequence of length *k*. While an edge with direction may exist between two nodes, it may only exist if k-1 length suffix of the one node is equal to k-1 length prefix of the other node (Zerbino et al., 2008). Therefore, the length parameter k is of great importance in an assembly. For example, a large k may give less convoluted and more linear graphs that is easy to traverse while a smaller k could result in a higher connected graph, that may require some post-processing to isolate the correct path (Alekseyev and Pevzner, 2007)**.** Any sequencing errors are propagated to the de Bruijn graph, and all sequencing errors that may occur toward the end of a read take a form of a ‘tip’, while those in the centre of a read, may manifest as a ‘bubble’ in the graph (Chaisson et al., 2009). The ‘tip’ node, which lacks an outgoing edge, hence it may not connect to any other read in an overlap. The ‘bubble’ represents alternative paths from one node to the other, whereby one path will result in the correct sequence while the other path may result in an erroneous sequence (Figure 1.5).

These next-generation DNA sequencing platforms promise to revolutionise biological and biomedical research, by enabling comprehensive analysis of genomes and transcriptomes. However, sequence reads produced by these NGS technologies are much shorter than the traditional Sanger reads and error rates are high. Furthermore, the volume of data produced creates a challenge for computational analysis.

Genomic sequence data of the reference genomes considered in this study was generated using the automated Sanger sequencing method as opposed to the ABI SOLiD sequencing of *Oryx bacillus*. The automated Sanger method is today considered as a 'first-generation' technology, and newer methods such as ABI SOLiD, RFLX 454 and Illumina are referred to as next-generation sequencing (NGS) technologies (Sanger et al., 1977; Metzker, 2010). These NGS technologies have drastically changed the acquisition of genomic data.

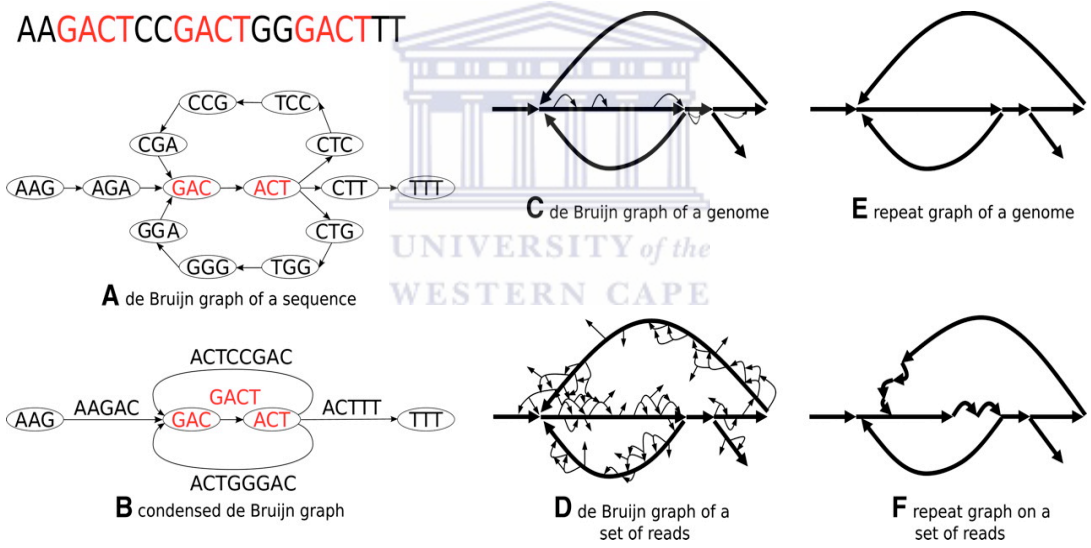


Figure 1.5. The deBruijn graph algorithm for construction of contigs.

A) The deBruijn graph of a sequence contains an intersection for every k-mer in the sequence together with an edge for every overlapping k-mers in the sequence. (B) All paths with nonbranching intersections are replaced from the condensed deBruijn graph to form a single sequence that generates the path. (C) The deBruijn graph of a genome may contain "tips" and "bubbles" representing repeats in the condensed deBruijn graph. (D) Sequencing errors may result in additional tips and bubbles. (E) In the repeat graph of a genome, tips and bubbles are removed. (F) The repeat graph of a set of reads cosntructed as a Eulerian assembly (Zerbino et al., 2008).

1.7 ABI SOLiD sequencing

The ABI SOLiD sequencing is a next-generation sequencing technology developed by Life Technologies and commercialised by ABI. It is a sequencing-by-ligation technology based on the polony multiplex sequencing technique (Shendure et al., 2005; Tomkinson et al., 2006). Sequencing by ligation involves the following stages (i) template preparation; (ii) sequencing; (iii) imaging (iv) read alignment and post-processing. These four stages can be further summarised into: primer attachment and hybridization; sequence ligation and washing; excite and fluorescent 4-color imaging; fluorophore cleavage; repeat ligation cycles.

Briefly, SOLiD sequencing technology entails template preparation for sequencing whereby the DNA is fragmented followed by adpaptor ligation using sequence specific adaptors, and single fragments are then amplified on beads in a PCR emulsion mixture. Beads are then randomly deposited onto a glass slide.

This platform is characterised by a di-base encoding with a sole purpose of improving reliability of alignments (Mardis, 2008). The di-base sequencing is also known as "color space". The ABI SOLiD platform uses a set of four fluorescently labeled probes each being comprised of eight bases. The first and second bases (for example AT or GC) are derived from one of the 16 possible dinucleotide sequences that are responsible for the specificity of hybridization and because there are only four fluorephores available, each color can be the results of several different sequences, resulting from 16 possible di-base combinations (Figure 1.6). Following template preparation, all probes are added to the flowcell simultaneously for the hybridization reaction and hybridization occurs based on rules of complementary of DNA base-pairing. DNA ligases are incorporated in the PCR mix to catalyse ligation of the probe next to the sequencing primer. Through excitation of the fluorescent label, the incorporated probes are identified and the intensity of fluorescence is captured by a camera. Subsequently, the fluorescent labels and the last three nucleotides of the probes are then cleaved in preparation of the next round of hybridization and ligation. On this platform, DNA bases are interrogated at positions five bases apart. Following several rounds of repeated hybridization and ligation, all probes and the first sequencing primer are removed, and a new sequencing cycle is initiated by annealing a sequencing primer one base closer to the beads. The process is repeated with five sequencing primers in total, until all bases of the template have been identified

(Mardis, 2008). Based on different library preparation protocols one can differentiate single-end (SE), paired-end (PE) and mate-paired (MP) libraries available for sequencing. For example, the fragment/single-end libraries are created by randomly shearing genomic DNA into fragments which are less than 1kb (kilobase) in size.

Sequencing errors may occur at any sequencing phase as a result of a misread base, but this approach is error prone. Therefore, subsequent analyses of SOLiD reads are rather generally performed in color space.

Following sequencing, the NGS raw reads have to be assembled into contigs either by alignment to a known reference genome or by *de novo* assembly. Alignment can be described as a process of determining the most likely source within the genome sequence for the observed DNA sequence read, provided the genome sequence is known. Traditional alignment programs for the Sanger sequencing are for example BLAST and BLAT but they may not scale well with the NGS reads in terms of the processing time and mapping accuracy. The observation that NGS reads are much shorter has led researchers to design bioinformatics tools and algorithms specific for these reads, which include alignment, and *de novo* assembly tools (Zerbino et al., 2008). Most of these tools encompass a heuristic technique to help identify small subsets of the reference genome where reads are most likely able to align. Besides aligning the raw reads to a reference genome, the sequenced reads may be assembled into contigs *de novo*. The de Bruijn approach has been extensively used in the de novo assembly of genomes and transcriptomes (Zerbino et al., 2008).

Polymorphisms/substitutions is one of the most common error types to be expected when using NGS sequencing technologies (Metzker, 2005, 2010). Furthermore, NGS sequencing platforms (Illumina, 454 as well as SOLiD) tend to under-represent AT-rich as well as GC-rich regions (Harismendy et al., 2009). With regards to SOLiD data, however, the problem has recently been shown to have resulted in the under-calling of true variants after an alignment (Shen et al., 2008). While determining the complete genome sequence of a species is still of great importance, great effort has been made to improve the efficiency of DNA sequencing. These next-generation DNA sequencing platforms promise to revolutionize biological and biomedical research, by enabling comprehensive analysis of genomes and transcriptomes (Margulies et al., 2005; Bentley et al., 2008; Valouev et al., 2008). However, the

volume of data produced creates a challenge for computational analysis.

We report the implementation of a SOLiD assembly pipeline capable of handling the NGS data produced by the ABI SOLiD platform. The pipeline utilises basic scripting tools and publicly available software packages for the *de novo* and reference assembly of ABI SOLiD short reads. The pipeline processes fragment or single-end to paired-end color space data to generate measurements for input quality evaluation and sequence alignment. It also includes a component for automatic detection and annotation of SNPs and Indels as well as visualization through IGV.

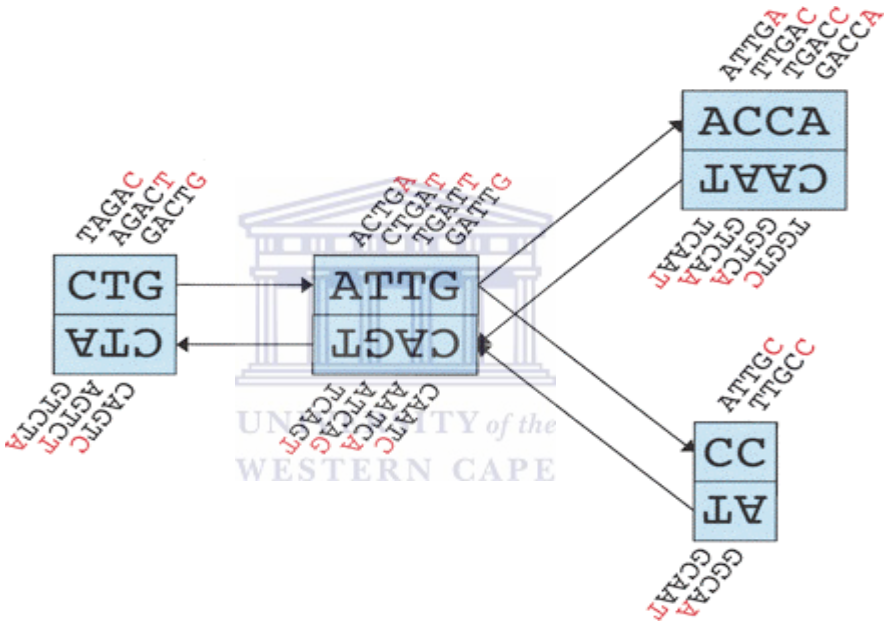


Figure 1.6. Velvet assembler deBruijn approach.

The diagram depicts graph traversal using the deBruijn approach. The nodes are represented by a single rectangle and series of overlapping k-mers listed directly above or below. In this diagram, k=5 and the last nucleotide of each k-mer is colored in red. The most likely sequence assembly at each node is represented in block letters inside the light blue rectangular box and it is the set of nucleotides derived as a result of the selected k-mer. Arcs are represented as arrows between nodes. The two nodes on the left could be merged into a single node without loss of information because they form a chain (Zerbino and Birney, 2008).

1.8 Genome alignment and assembly

Following sequencing, the short reads are assembled *de novo* into contiguous sequences or aligned to a known reference genome (Metzker, 2010). These high-throughput sequenced short reads may be mapped to a reference genome of a closely related species to identify variations (Brockman et al., 2008). As much as errors may affect variation calling, biases may result from the use of a reference genome (Degner et al., 2009; Harismendy et al., 2009). Placing reads within repetitive regions in the reference genome and the fact that there may be rearrangements in the assembled reads, where additional mutations could have occurred in regions not present in the reference genome, are major limitations to the alignment approaches (Metzker, 2010). But, with the development of accurate computational tools, these problems (data acquisition, characterization and analysis) can be addressed (Pop and Salzberg, 2008). Following the standard overlap layout consensus approaches conducted by previously used assemblers, the Velvet assembler uses a deBruijn approach to decompose a set of short reads into a set of shorter DNA sequences (Zerbino and Birney, 2008; Figure 1.6). The constructed graph is then merged into nodes in which two segments can get connected if they are adjacent in one of the original reads that traverses into an Euler path (Pop and Salzberg, 2008). The deBruijn approach shows an improvement and greater performance in the *de novo* assembly (Zerbino and Birney, 2008).

Although, next-generation sequencing platforms such as ABI SOLiD have revolutionised biological and biomedical research, they pose new challenges to the Bioinformatics community. For example, besides the enormous data generated, the read lengths are much shorter than the conventional Sanger sequencing method. Mapping the short reads to a reference genome poses different challenges in identifying the nature of mutations because mismatches can occur due to sequencing errors or due to differences between the query sequence and the reference sequence, or due to the existence of repetitive regions. Although de novo assembly of short reads without a reference sequence has been previously criticised because of their sensitivity to sequencing errors and chimeric sequences (Schuster 2008), the strategy is applicable to genomes without a known reference sequence or those lacking a high-quality finished genome. Compared to reference-genome assembly, de novo assembly can recover tags that are missing from the reference genome assembly, or tags that are from an unknown source.

1.9 Comparative genomics analysis of the MTC

Several techniques have been employed to address genomic sequence comparisons between the members of the MTC (Baess, 1979; Bradley, 1973). These include the use of hybridization-based methods employed for whole-genome sequence comparisons (Gordon et al., 1998; Brosch et al., 2002). These studies have demonstrated the extensive degree of relatedness between the genomes isolates of *M. tuberculosis* as well as those of *M. bovis* (Labidi and Thoen, 1989). For example, members of the MTC display little sequence variation when compared with each other (Cole, 2002). However, despite their relatedness at the genomic level, they differ remarkably with respect to their host range and pathogenicity, hence the *Oryx bacillus* is infectious to *Oryx leucoryx*, antelope (Greth et al., 1994; Behr et al., 1999). Characterization of the *Oryx bacillus* genome assembly relative to the seven reference genomes considered in this study, is of great importance and illustrates how comparative genomics may help understand phenotypic variations between *Oryx bacillus* and strains of the MTC (Table 1.1). Comparison of genomic sequences between *Oryx bacillus* and members of the MTC can identify the key differences and similarities among them, can thus provide insights into the genetic factors that contribute to *M. tuberculosis* virulence and drug resistance. For example, to shed light on the genetic changes that mediate drug resistance, Ioerger and co-workers (2009) mapped polymorphisms among the KZN strain family of *Mycobacterium tuberculosis* to drug-susceptibility profiles. The results were consistent with the drug-susceptibility profiles, that is, there was correlation between well-known mutations and resistance to isoniazid, rifampicin, kanamycin, ofloxacin, ethambutol, and pyrazinamide.

Although highly conserved, each MTC genome has several polymorphic regions and mobile insertion sequences (Zhu et al., 2009). These mutations are postulated to confer selective advantage to members of the MTC in changing environments (Foster, 2004).

Cell walls of pathogenic bacteria are known to demonstrate variation in protein sequences and macromolecular composition, reflecting selective pressures on these structures (Garnier et al., 2003; Zheng et al., 2008). For example, members of the MTC exhibit an abundance of lipids in the cell wall (up to 60% of its dry weight) much of which is attached to polysaccharides, which include glucan, mannan, arabinogalactan, and arabinomannan, marking the foundation for an external

permeability barrier (Brennan and Nikaido, 1995). Other cell wall component such as trehalose dimycolate (cord factor) are potent initiators of the host immune response to TB infection (Höner zu Bentrup and Russell, 2001). The high concentration of lipids in the cell wall has been associated with antibiotic resistance and resistance to lethal oxidations (Brennan and Nikaido, 1995). Hence, variation of cell wall components and gene expression are key to the evolution of newly studied emerging MTC strains (Garnier et al., 2003).

In the host macrophage, mycobacteria encounter an environment containing low oxygen and altered carbon source. Although the strict oxygen requirement of *M. tuberculosis* has been demonstrated *in vitro*, members of the MTC are able to adapt to environments with low oxygen availability (Höner zu Bentrup and Russell, 2001). For example, *in vitro* models using oxygen depletion in nutrient-rich medium or oxygen-rich nutrient starvation medium have demonstrated the survival of *M. tuberculosis* for extended periods in non-replicating and drug-tolerant state (Gengenbacher et al., 2010). Under these metabolic downshifts, the mycobacteria shift their metabolism to anaerobiosis by changing to nitrate respiration and reductive amination of glyoxylate. However, *M. bovis* is not known to have nitrate reductase in the absence of oxygen (Wayne and Lin, 1982). Thus, *M. bovis* demonstrates reduced virulence with disease progression because of the low oxygen availability in the mature granuloma (Wayne and Hayes, 1998; Höner zu Bentrup and Russell, 2001).

1.10 Gene duplication and multigene families

Gene duplication is a driving force for the introduction of novel gene functions. The analysis of the impact of gene expansion on the introduction of novel gene functions is made possible through the availability of sequenced genomes for a range of species. Gene duplication is a consequence of unequal crossing over, retroposition, or chromosomal (or genome) duplication. Unequal crossing over usually generates tandem gene duplication whereby duplicated genes are linked in a chromosome. Chromosomal or genome duplication occurs probably by a lack of disjunction among daughter chromosomes after DNA replication (Zhang, 2003). Duplicated genes are often referred to as paralogous genes, which form gene families, which can either be fixed or lost in a population. Previous studies outlined protocols for gene duplication analysis (Christoffels et al., 2004; Mulder et al., 2009). *Mycobacterium tuberculosis*

contains about 250 genes involved in fatty acid biosynthesis compared to 50 in *Escherichia coli*. This is partly attributed to gene duplication (paralogy). Kinsella and colleagues (2003) documented large gene duplications of mycobacterial genes involved in the fatty acid biosynthesis in *M. tuberculosi*s. Brosch and colleagues (2000) identified two major rearrangements in the genome of *M. bovis* that correspond to two tandem duplications, DU1 and DU2, of 29, 668 bp and 36, 161 bp respectively. About half of the proteins present in the tubercle bacillus have arisen from ancient gene duplication and adaptation events.

The PE (protein family characterized by Proline-Glutamic acid motif) and PPE (protein family characterised by Proline-Proline-Glutamic acid motif) multigene families of *M. tuberculosis* comprise about 10% of the coding potential of the genome (Karboul et al., 2006). The PE and PPE families are associated with the ESAT-6 (esx) gene cluster regions, which encode a secretion system of the potent T-cell antigen ESAT-6 family. These gene families, totalling more than 170 proteins, are characterized by a conserved amino-terminal segment with either a proline–glutamic acid (PE) or a proline–proline–glutamic acid (PPE) motif combined with a carboxy-terminal domain comprising varying numbers of short repetitive motifs.

The resistance-nodulation-division (RND) family consists of 12 membrane proteins proposed to have undergone expansion. RND proteins are a family of multi-drug resistance pumps that recognise and mediate the transport of a great diversity of cationic, anionic, or neutral compounds. The genome of *M. tuberculosis* contains 13 genes that encode RND proteins designated MmpL (mycobacterial membrane protein large). Four out of the 12 MmpL proteins (MmpL4, MmpL7, MmpL8, and MmpL1) are necessary for virulence (Domenech et al., 2005).

1.11 Genetic variation and annotation

Previous studies have used next-generation sequencing technologies for in-depth investigation of genetic variation, including Indels (insertion/deletions) and SNPs (Medvedev et al., 2009).

1.11.1 Single nucleotide polymorphism

There are three groups of SNPs, namely, the non-synonymous SNPs (nsSNP), synonymous SNPs (sSNP) as well as the intergenic SNPs. While the intergenic SNPs

may be subjected to selection pressure due to the fact that they may affect gene expression, the non-synonymous SNPs may be subjected to various selection pressures as they are often associated with amino acid changes. However, sSNPs are not associated with amino acid sequence changes and are hence regarded neutral to selective pressure. Synonymous SNPs are generally used as a powerful tool in molecular epidemiology. One disadvantage in studies using SNPs is the ascertainment bias in them, meaning that such analysis involve skewed datasets by the non-random nature of selection and discriminating features. However, identifying large SNP numbers from a wide diversity of strain sequences may overcome this problem with careful observations that the information generated would be sufficient to differentiate closely related strains for epidemiological statistic analysis.

1.11.2 Insertion sequences and prophages

Deletions, insertions and transposon elements: Transposons also known as "jumping genes" or transposable elements (TEs), are mobile genetic elements that can integrate in the genome, thereby causing an insertional mutation. Transposons can be active in both DNA and RNA (retrotransposons) forms, however, they are inert under normal conditions. It would be too harmful to the genome if a transposon would keep jumping around, because every new integration causes a novel mutation. This would rapidly lead to deleterious mutations that would damage the genome such that the organism would no longer reproduce. TEs might play regulatory roles, determining which genes are turned on and when this activation takes place (McClintock, 1965). It has been shown that certain transposons have regulatory mechanisms to restrict their copy number to a certain threshold (Schmidt et al., 2010). In prokaryotes, transposons often carry genes conferring antibiotic resistance or heavy metal resistance and they are often located on plasmids. Such plasmid-borne transposons are notorious for their "horizontal transfer", i.e. transfer from one species to the other.

Brosch and colleagues (2001) observed that nucleotide substitutions are not a significant source of genetic diversity either between strains of *M. tuberculosis* or between members of the *M. tuberculosis* complex. An *in silico* genome comparison of the virulent laboratory strain H37Rv and the epidemic strain CDC1551 revealed a polymorphism rate of approximately 1 in 3000 bp.

The major research question to address remains whether there are other genetic polymorphisms in *M. tuberculosis* that could account for the perceived phenotypic

differences between strains. The identification of a series of deletions in *M. bovis* relative to *M. tuberculosis* suggested that Indels could be one such mechanism (Gordon et al., 1998; Behr et al., 1999). In a recent microarray study, only one of 16 clinical isolates analyzed had the full complement of H37Rv genes, the other 15 having lost between three and 38 ORFs (Kato-Maeda et al., 2001). These deletions were associated with IS6110 insertion element present in variable numbers in the majority of *M. tuberculosis* strains. Previous studies of several mycobacterium species proved that recombination between adjacent insertion elements can lead to the deletion of the intervening genomic segment, suggesting this is an important mechanism of generating diversity (Fang et al., 1999; Brosch et al., 2000; Eckstein et al., 2000). Recombination between an identical 11 bp repeat has been shown to result in loss of the *katG* gene and resistance to the anti-mycobacterial drug isoniazid (Pym et al., 2001). Further support for the role of Indels comes from an *in silico* full-genome comparison of *M. tuberculosis* H37Rv and CDC1551 (Betts et al., 2000).

Approximately 52% of *M. tuberculosis* proteome is part of the big families derived from duplication events (Mulder et al., 2009). The concentration of Indels in the PE and PPE genes and the observation that these are highly polymorphic, suggests that these genes are the principal source of genetic diversity in *M. tuberculosis*. Although their function is currently unknown, various members have been implicated in the pathogenesis of both *M. tuberculosis* (Camacho et al., 1999) and *M. marinum* (Ramakrishnan et al., 2000).

Other polymorphic elements have been detected in *M. tuberculosis* such as the direct repeat (DR) region (Kremer et al., 1999) which forms the basis of spoligotyping, and the more recently identified mycobacterial interspersed repetitive units (MIRUs) (Magdalena et al., 1998; Supply et al., 2000). MIRUs are 40–100 bp elements often found as tandem repeats in intergenic regions of the *M. tuberculosis* complex. At least 12 of these are polymorphic, with variable numbers of the tandem repeats occurring in different strains of *Mycobacterium tuberculosis* (Magdalena et al., 1998; Supply et al., 2000; Mazars et al., 2001). Their function is unknown but they are similar to short sequence repeats (SSRs), which in other bacterial pathogens can modulate gene expression (Van Belkum et al., 1998).

1.12 Thesis Rationale

Tuberculosis has been declared as a global health emergency and it is a major cause of infectious disease deaths despite the availability of chemotherapy and vaccines (WHO, 2006). The causative agent of this world-wide disease is *Mycobacterium tuberculosis*, a member of the MTC. Recently, *Oryx bacillus* was isolated from a healthy buffalo on a farm where a buffalo was imported from a zoo in Portugal (Van Helden et al., 2009). *Oryx bacillus* is an MTC member and a causative agent of tuberculosis in antelopes. It was first isolated from the Arabian Oryx and camelids in 1986 (Greth et al., 1994). Despite its prevalence to cause infection in the Arabian Oryx, *Oryx bacillus* has not been reported to cause infection to the South African Gemsbok (*Oryx gazella*). However, it may be hypothesised that it would also be pathogenic to Gemsbok or that *Oryx bacillus*, like *Dassie bacillus* is less virulent than *M. tuberculosis* or *M. bovis*. In a recent study, this strain was isolated causing disease in human patients in Australia (Fyfe, 2011). This observation suggests that *Oryx bacillus* can be transmited within and between a range of hosts including humans. The sequencing and analysis of the *Oryx bacillus* genome will provide an understanding of the potential pathogenicity of this strain and ultimately disease control.

1.13 Aims of the study

i) To implement a SOLiD assembly pipeline for analysing NGS data.

ii) Assemble the *Oryx bacillus* genome.

iii) Detect genetic variations between *Oryx bacillus* and other members of the MTC.

1.14 Thesis outline

The rest of this thesis is composed of the following chapters:

Chapter 1 briefly reviews the history of tuberculosis and the bottlenecks of next generation sequence analysis. Additionally, this chapter provides an overview of SOLiD sequencing, genome assembly using the Velvet assembler and Novocraft computational pipeline for short read alignments using a reference genome.

Chapter 2 provides a summary of the materials and methods used in the analysis of *Oryx bacillus* SOLiD data. We further report the implementation of a SOLiD

assembly pipeline capable of handling the NGS data produced by the ABI SOLiD platform. The pipeline utilises basic scripting tools and publicly available software packages for the *de novo* and reference assembly of ABI SOLiD short reads.

In this chapter, we investigated the utility of ABI SOLiD raw reads for the reference-based and *de novo* assembly of the *Oryx bacillus* genome. The major aim of this chapter was to identify genetic variation between *Oryx bacillus* and seven members of the *M. tuberculosis* complex. A catalog of genes not covered by the *Oryx bacillus* reads was generated by mapping *Oryx bacillus* short reads to known coding sequences from the reference genomes (Table 1.1) using novoalignCS as described in section 2.1.3. Additionally, we attempted to identify deletions specific to *Oryx bacillus* and map the known regions of difference (RD) onto the current assembly. Results for a variety of analyses to assemble and annotate *Oryx bacillus* short reads are discussed.

This chapter further reports about allele variants predicted using a computational pipeline that was developed to analyse SOLiD data. Through successful alignments using novoalignCS, described in section 2.1.3, we focused on refining functional SNPs and short Indels observed in SOLiD data and attempted to elucidate their biological relevance. The observed results suggest that the virulence and host specificity of *Oryx bacillus* may be due to changes in genes encoding respiratory enzymes, stress-related products and metabolic enzymes as well as proteins involved in fatty acid catabolism.

Chapter 3 provides a summary of the results of the studies described in this thesis. Briefly, the chapter presents the predicted genetic variation between *Oryx bacillus* and seven members of the *M. tuberculosis* complex. We provide a catalogue of genes not covered by the *Oryx bacillus* reads (Table 1.1). Results for a variety of analyses to assemble and annotate *Oryx bacillus* short reads are also presented.

Chapter 4 is the discussion of results that were described in chapters two and three with a review of the similarities and differences between *Oryx bacillus* and MTC strains. The chapter also discusses the status of *Oryx bacillus* genome, the lessons learnt and the challenges to be addressed in the future.

Chapter 5 is the conclusions. This chapter outlines the status of the *Oryx bacillus* genome with a review of the similarities and differences between *Oryx bacillus* and MTC strains, the lessons learnt and the challenges to be addressed in the future.

CHAPTER 2

MATERIALS AND METHODS

Mycobacterium tuberculosis complex organisms are characterized by 99.9% similarity at the nucleotide level however; they differ widely in terms of their host ranges, phenotypes and pathogenicity. A fundamental requirement in their survival is the ability to adapt to changes in the host environment. We investigated the utility of ABI SOLiD raw reads for the *de novo* and reference-based assembly of the *Oryx bacillus* genome in order to elucidate distinct features among members of the MTC and *Oryx bacillus.*

The effective usage of NGS data in modern genetics strongly depends on bioinformatics tools with capabilities to handle downstream analyses of frequently generated sequence data. NovoalignCS was used to align *Oryx bacillus* ABI SOLiD color space reads to seven reference genomes considered in this study. *De novo* assembly Pipeline for SOLiD v.2.0 and Velvet software suite were used to assemble the mapped and unmapped reads into contigs. Contigs from unmapped reads were screened against GenBank non-redundant and Swissprot databases while contigs composed of reads that aligned to the reference genomes were assigned GO terms and metabolic pathway information.

The genome assembly for *Oryx bacillus* was carried out in two parts; alignment of the reads to a reference genome, and second, through *de novo* assembly of unmapped reads (Figure 2.2).

The *Oryx bacillus* ABI SOLiD short reads were mapped to five members of the *M. tuberculosis* to identify the key differences and similarities between them including genetic variations (Table 2.1). An overview of the methods used in this chapter is outlined in Figure 2.3, a continuation to the methods outlined in Figure 2.2.

2.1 Data Sources

Samples of *Oryx bacillus* were isolated from a healthy African buffalo captured from a game farm in Kwazulu-Natal in 2007. The buffalo tested positive to an intradermal tuberculin test. A series of experiments were carried out, by the Tygerberg Hospital group, to confirm the *Oryx bacillus* isolate. These include the 16S rRNA polymerase chain reaction (PCR) and an MTC-specific multiplex-PCR.

Sequencing of *Oryx bacillus* was done by the Co-factor sequencing company using ABI SOLiD sequencing platform to generate a fragment library of color space short reads. The obtained read length was 50 bpr (read per base) and a coverage of 350X (number of reads x read length / genome size). H37Rv reference genome was used to derive the coverage. A total of 31,271,059 color space fragment library short reads were supplied by Professor Nicolaas Claudius Gey Van Pittius, from the Tygerberg medical school, University of Stellenbosch.

Seven reference genomes (*Mycobacterium tuberculosis* H37Rv, H37Ra, CDC1551, F11, KZN_1435, *Mycobacterium bovis* AF2122/97 and *Mycobacterium bovis* BCG str. Pasteur 1173P2) (Table 1.1) that were used for mapping (see section 2.1.3) were downloaded from ftp://ftp.ncbi.nih.gov/genomes/Bacteria/ and formatted using a PERL script to convert from GenBank to FASTA format and generic feature format (GFF) and indexing the seven genome files.

Furthermore, protein coding sequences (CDS) of five reference genomes were used to compute functional SNP and SNPs located within genes (Table 2.1). The CDS's of reference genomes were downloaded from the tuberculosis database (http://genome.tbdb.org) and the Broad Institute (http://www.broadinstitute.org).

Table 2.1. The *Mycobacterium tuberculosis* reference strains used in this study.

Reference strain	Genome size	Total number of genes	Source
H37Rv	4.14Mb	3,999	http://genome.tbdb.org
H37Ra	4.42Mb	4,034	http://genome.tbdb.org
CDC1551	4.4Mb	4,189	http://genome.tbdb.org
F11	4.42Mb	3,959	http://genome.tbdb.org
KZN 1435	4.38Mb	4,060	http://www.broadinstitute.org

2.1.1 ABI SOLiD file formats

The raw input data files consists of the two-base encoded (color space) SOLiD reads in ".csfasta" format and a ".qual" file with values corresponding to the reads in the ".csfasta" file. The order of the reads in the .csfasta file should correspond to the order of the reads in the .qual file and missing color-calls in the .csfasta file should be encoded as a dot (.) (Figure 2.1).

```
>469_29_17_F3
T203303103012313303232311310133211223331321213103 20
>469_29_17_F3
T132113 . 21231311212222311021312211121122 20 . . 2123221
```

a) The 2-base encoded color space (.csfasta) format of ABI SOLiD reads.

```
>469_29_17_F3
30 31 24 22 25 17 20 21 17 29 22 30 15 2 31 15 21 4 3 28 10 24 26 18 22 17
24 4 8 12 10 14 5 21 15 5 23 12 13 7 6 15 14 17 6 18 21 12 11 13
>469_29_17_F3
31 24 29 31 20 24 2 31 30 27 31 27 22 30 28 29 32 21 31 31 23 22 31 30 23 31
16 17 22 13 8 21 31 17 7 31 8 29 23 13 8 22 2 1 14 8 27 20 10 17
```

b) The quality score (.qual) file format of ABI SOLiD reads.

Figure 2.1 (a-b). ABI SOLiD input file format.

a) Represents short read colorspace file format. b) Represents the quality score file format with values corresponding to the two-base encoded colorspace reads.

2.1.2 Quality assessment and quality control

Biases in NGS data occurs due to inconsistencies in the quality of reads such as length, quality scores and base distribution. Hence, the raw sequence reads were quality checked using the FastQC program in order to assess the quality of the data and to filter low quality reads (Barbraham-Bioinformatics, 2009). FastQC is a java program that aims to provide simple ways of doing quality control checks to validate the raw sequenced data and ensures the raw data carries no biases which may affect its usefulness. The FastQC analysis was run in an off-line mode as part of a methodology workflow, outlined in Figure 2.2**,** and the output was converted to HTML format.

As a first step, the reads were converted from csfasta/qual to color space FastQ, a

format acceptable in FastQC. FastQC analyses raw sequenced data through steps of ten modules. The first being an output of the general statistics, explaining the type of input file, a count of sequences being processed as well as the read length. Secondly, the program calculates a per sequence quality score using the mean scores of all sequences. It also calculates a per base sequence content which presented an even distribution of the four bases throughout the sequence. The program also computes the per sequence G+C content which shows a distribution of the G+C content throughout the sequences with a theoretical normal distribution and a mean. This module also signifies the presence of any contamination that may be present in the sequenced data. The number of uncalled bases throughout the library sequences was also calculated. This step was considered as the pre-processing step into the alignment process. Filtering poor-quality reads, chimeric sequence contaminants, and trimming adaptor sequences are examples of quality control procedures that were applied to raw reads for subsequent analysis.

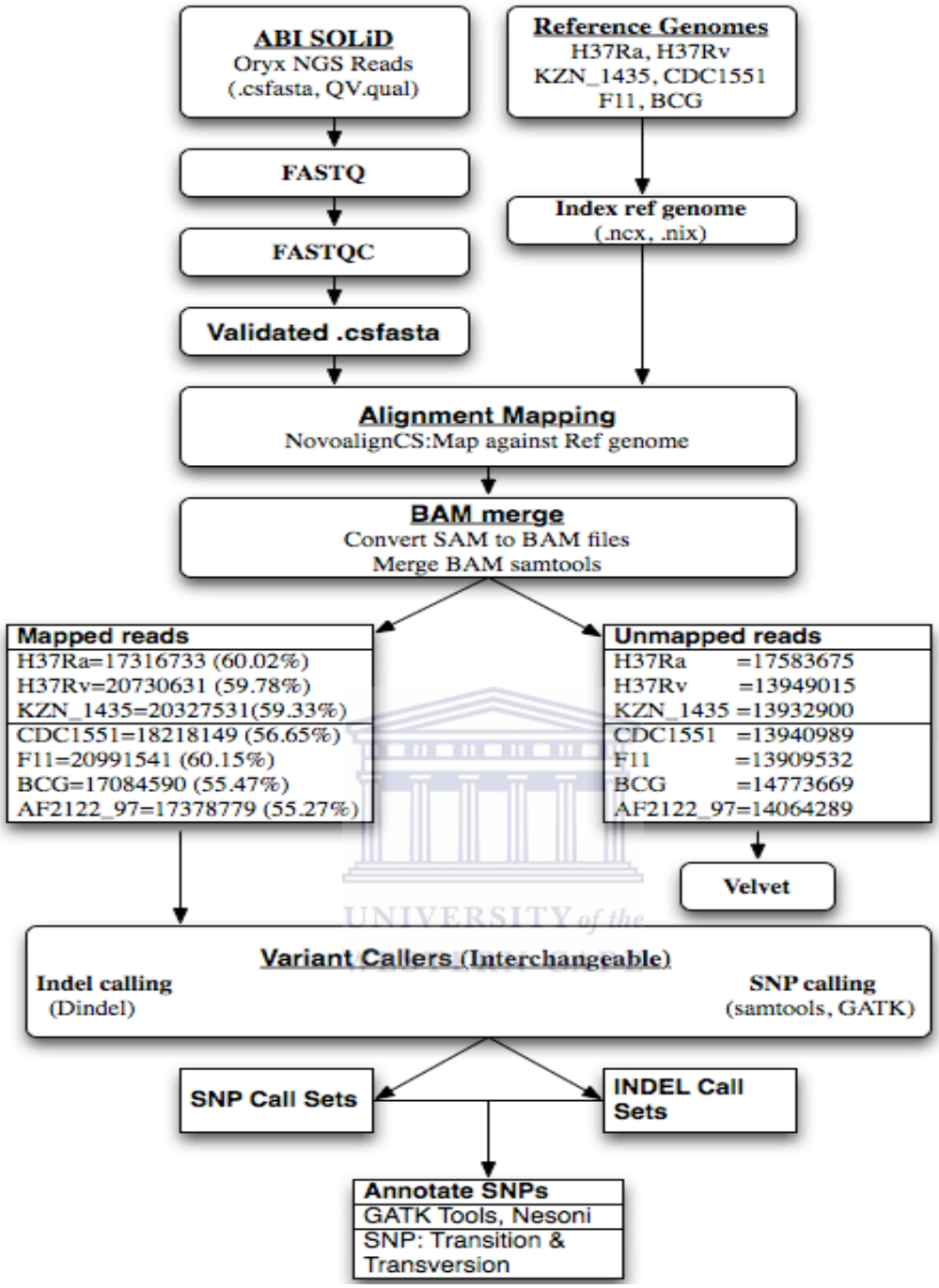


Figure 2.2. A computational workflow for the analysis of *Oryx bacillus* short reads. The *Oryx bacillus* short reads ('.csfasta' and '.qual') were converted to fastq format and pre-processed for quality control checks using FastQC. The raw reads were aligned to MTC reference genomes using NovoalignCS. The mapped reads were further analysed to derive genome coverage statistics using BEDTools. The unmapped reads were assembled *de novo* using Velvet.

2.1.3 Read alignment and mapping

The reads were aligned to the reference genome sequences using NovoalignCS (novocraft version 2.05.02; Hercus, 2009). Mapping was done in a two-step procedure. In the first step, reads were mapped to the whole genome fasta sequences to identify similarities and differences between *Oryx bacillus* and members of the MTC. In the second step, reads were mapped to CDS's for each reference genome to identify functional SNPs located within gene. Briefly, the coding sequences for each reference strain were first indexed using the novoindex option, to create a database of the indexed reference genome and used as input to novoalignCS. The program takes as input, the short read file in colorspace (.csfasta) format and the corresponding quality score (.qual) file and performs adapter trimming, base quality calibration, Bi-Seq alignment, and has options to report multiple alignments per read. Local gapped alignments were carried out with an allowance of eight mismatches per alignment. NovoalignCS uses the Needleman Wunsch algorithm to compute the alignment. While other aligners such as BWA and Bowtie have demonstrated optimal performance for the Illumina platform, NovoalignCS showed the best overall performance for SOLiD data (Wang, 2011).

The NovoalignCS algorithm includes an indexing step ('novoindex') which creates a database of the indexed reference genome to be fed into NovoalignCS as color space. The coding sequences for each reference strain were first indexed using the novoindex option, to create a database of the indexed reference genome to be used as input to novoalignCS. The purpose is to convert the reference genomes into color space instead of changing the SOLiD color space data into letterspace data. Novoindex creates a k-mer index that can easily be loaded into shared memory for access by multiple search processors during the NovoalignCS runs.

Following novoindex, novoalignCS was executed with the option '-r All' chosen to report multiple alignment locations. However, at default settings the alignment include searches for mismatch penalty, gap opening penalty and gap extension penalty. NovoalignCS, aligns reads to a reference genome using ambiguous nucleotide codes and uses an iterative search algorithm to find the best alignment and any other alignment with similar scores. The quality score was modified for use with the shorter 50 bp; the gap open and gap extend penalties both set to 10, while the

match and mismatch score were both set to one and eight, respectively. Two reads were called "aligned" if they had an overlapping segment of at least 100 bp with an identity of 96% or more. NovoalignCS takes as input, the base quality file (.qual) as well as the color space reads (.csfasta) and gives as output, the unsorted alignment in Sequence Alignment/Map (SAM) format. The unmapped reads flag ("readUnmappedFlag") was checked to determine whether or not a read is mapped. Finally, the output alignment files were changed from SAM format to a compressed format called Binary Alignment/Map (BAM). The alignment process was repeated across all reference genomes used.

2.1.4 Alignment statistics

Statistics about the alignment were collected using samtools "flagstat" command. This analysis included calculating the total number of reads aligned and not aligned to the reference genome using SAMtools. SAMtools is a program which describes the sum of software packages designed for parsing and manipulating alignments in the SAM/BAM format and it is available in C, Java and Perl languages. The package also provides several utilities for format conversions, alignment sorting, alignment merging, PCR duplication removal, generating alignments in a per-position format (Li et al., 2009). Read coverage with regards to the genome size was calculated using BEDTools, described below.

2.1.5 SAM file validation

The alignment output SAM file was validated using Picard, a pipeline for processing and delivering NGS data. Picard comprises Java-based command-line utilities that manipulates SAM files. Picard validates a SAM file to have retained its format, describing the alignment of the reads to a reference sequence. Picard checks the SAM file header for sequence dictionary, labels for the read groups ("@RG" header) for each sam record and also reports on any error detected by the SAMRecord.isValid class. This method deliberately returns null if there are no validation errors, because callers tend to assume that if a non-null list is returned, it is modifiable. Therefore, "SAMRecord" statistic output returns null when the file is valid but if the file is invalid, the program returns a list of error messages. Picard also validates if NM (nucleotide differences) exists.

2.1.6 Converting SAM to BAM

The SAM format is a tab-delimited text format with two sections: i) header – meta data about the alignments, and ii) the alignments themselves . Each alignment is on one line, has 11 tab-delimited fields and other optional fields. The SAM files were sorted to ease searching and extraction of reads. The SAM alignment output files were converted into indexed BAM format using the SAMtools "view" and "index" options (Li *et al*., 2009).

2.1.7 Post alignment processing of BAM files

The Novoalign algorithm can detect SNPs and short Indels in short reads from next-generation sequencing platforms (Hercus, 2009). Sequence aligners are unable to perfectly map reads containing insertions or deletions, alignment artefacts, and false positives SNPs, hence post alignment processing was undertaken as described below.

I. Local realignment around Indels

The local alignments were processed in three steps implemented in a workflow outlined in Figure 2.3. i) determining small misaligned intervals for subsequent realignment, ii) running the realigner over those intervals and iii) fixing the realigned reads. The Indel realigner of the BAM file was run using the full Smith-Waterman alignment algorithm in the Genome Analysis ToolKit (GATK), a java-based program. For this analysis, the RealignerTargetCreator option was run with a –T flag, which specifies the target to be realigned. Initially, all regions around potential Indels and clusters of mismatches were collected in step (i) of the analysis and merged into intervals. Realignment of reads overlapping these specified regions was done using GATK.

II. Co-ordinate sorting and indexing of the realigned BAM file using SAMtools

The BAM file was sorted and indexed using SAMtools (Li et al., 2009). The tool performs the sorting, indexing and merging of data allowing the option of extracting reads at any region of the alignment swiftly. SAMtools takes as input a SAM/BAM file and sorts it by chromosomal coordinates. Sorting by coordinate is used to avoid loading extra alignments into memory. The BAM file was sorted in the chromosome order of the analysed genome (referred to as "the reference order"). Realigned BAM file was sorted by coordinates for efficient memory allocation while loading the

alignments. Sorting by coordinate minimises loading extra alignments into memory. We used the "calmd" command to compute the edit distance (NM), that is, the number of nucleotide differences in the reference sequence that need to be changed to perfectly match the query and the number of reported alignments that contain the query in the current record (MD). The MD field refers to the number of times a particular read aligned to the reference sequence. The MD field aims to achieve SNP/indel calling without looking at the reference. Sorting is important in ensuring a quick and stable parallel processing of data. Therefore, the position-sorted BAM file was indexed, which is important for fast random access of the sorted alignment file.

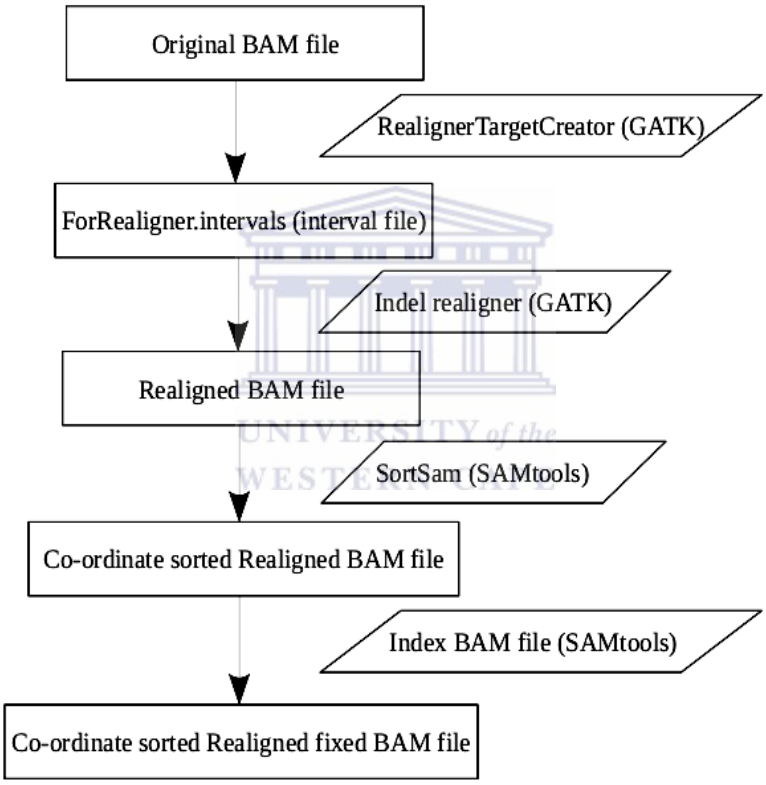


Figure 2.3. Workflow for post-processing analysis of the alignment BAM file.

2.1.8 Per base coverage calculations

The depth and breath of sequence coverage across each base of the reference genome was calculated using BEDTools (Quinlan and Hall, 2010). A module within BEDTools called "genomeCoverageBed" was used to conduct the calculations and takes as input the alignment BAM file as well as the reference fasta files. The output files are in a BED format. The output per base coverage BED file was further

analysed to identify regions with zero coverage. The resulting file was then separated into intervals using the zero read coverage regions as the start and end and the sum of all reads covering that stretch as an additional column. A plot was then generated using this information (Figure 3.7).

2.1.9 Per gene coverage calculations

By loading the alignment BAM file into memory, BEDTools "coverageBed" utility, analyses its features, line by line against the reference genome. "CoverageBed" summarises the depth and breadth of coverage of all genes by the reads in the alignment BAM file.

We generated a reference gene list from the reference genome GFF file which together with the alignment BAM file was taken as input files for BEDTools. After each interval encountered in the gene list file, coverageBed will report the number of features overlapping with the alignment file by at least one base. The output was a summary report of the length of each gene and the number of reads covering each gene.

2.1.10 Zero coverage genes

Oryx bacillus reads were aligned to the MTC reference genomes using NovoalignCS (see section 2.1.3). Reference genes without any read alignment support were extracted from the alignment file and a list of their orthologs in the MTC strains was compiled to confirm genes that have been deleted in the *Oryx bacillus* sequence assembly.

2.2 *De novo* assembly

Velvet is currently the most used genome assembler due to its simplistic approach, speedy execution and relatively accurate results (Zerbino and Birney, 2008). Velvet uses the "Tour bus" algorithm to detect and correct sequencing errors that occur in the middle of a read 'bubble'.

Oryx bacillus contigs were assembled from both aligned and unaligned reads using the Velvet short read assembler version 0.7.55 as implemented in the *de novo* Assembly Pipeline for SOLiD v.2.0. First, SAMtools was used to extract reads that did not align to the reference genome from the alignment BAM file. Second, Velvet assembler was run with the following parameters: k-mer hash length 31, coverage =

1, read category = fragment and refLength= expected reference genome length in base pairs (bps). To obtain a meaningful assembly, the coverage should be larger than 30x. Optimal assembly is achieved for coverage between 200x – 500x. However, coverage greater than 500x does not alter the assembly results. For this analysis, a default 300x coverage was selected to yield the results listed in Table 3.6.

The input sequence was a fragment library of 50 base per read. The *de novo* assembly process was implemented in three stages:

(i) Pre-assembly stage: The unaligned reads were assembled using the *de novo* Assembly Pipeline for SOLiD v.2.0 which includes, in its analysis, the Velvet assembler. The preprocessing step, also known as pre-assembly performs error correction of reads using the SOLiD Acuracy Enhancement Tool (SAET) embeded within the pipeline.

The unmapped reads were corrected for errors and prepared for intake by the Velvet assembler. The input sequences, in color space had the first base and first color removed and converted into double encoded reads (.de) where, 0->A, 1->C, 2->G, 3->T. The double encoded file formats were used in the assembly as input. This step also creates compatible file formats for assembly assistant engines.

(ii) Sequence assembly stage: Following read preprocessing stage, the output files were used as input into Velvet program which utilises the de bruijn algorithm to construct contigs. As one of the outputs, Velvet constructs a de bruijn graph of the *Oryx bacillus* aligned and unaligned reads. In the de bruijn graph, vertices correspond to k-mers present in the unmapped reads (k < read length) while an edge connects two vertices if corresponding k-mers are consecutive k-mers from a single read. The graph is simplified by the removal of erroneous or low frequency vertices and edges creating an Eulerian path (repeat graph) for the final assembly.

(iii) Post-processing stage: This step includes the assembly Assistant for SOLiD (AsiD) tool which is involved in the conversion of color-space assembly into base-space.

2.2.1 Sequence similarity search for unmapped contigs

The contigs generated from the unaligned reads were further screened against the GenBank non-redundant (NR) and Swiss-Prot databases using TBLASTX search algorithm with a criterion imposed at a cut-off expectation value (e-value) of 1e-5. The aim of a BLAST search was to to further characterise contigs as either unique to *Oryx bacillus* or as sequencing error. Contigs that did not have a match to any database sequence were classified as *Oryx bacillus* specific.

2.2.2 Functional annotation

Functional annotation of 7,953 contigs derived from unmapped reads relative to H37Rv was performed using GO terms of cellular component, biological process and molecular function. Gene Ontology terms were assigned to 7,953 assembled contigs based on sequence similarity with known UniProt-TrEMBL database proteins annotated with GO terms using BLASTX. A functional enrichment was done using Fisher's exact test whereby the SNP list was compared against the whole genome as a background.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al., 2002) provides a list of UniProt protein sequences, which have a KEGG Orthology (KO) number. This list was parsed to create a BLAST searchable database against which similarity searches were performed using BLASTX search algorithm. The cut-off e-value was set at 1e-5 and the number of hits per query sequence was set at 10. For each *Oryx bacillus* contig, annotation entries that are supported by one or more hits and match the defined cut-offs were collected after filtering out self-hits. The annotation of biochemical pathways was based on the KO identifier, which is a unique identifier assigned to the corresponding biochemical pathway. For each contig, the following information was captured: contig name, the KO identifier, the corresponding biochemical pathway to the KO identifier and a short description of the pathway.

2.3 SNP calling

SNPs were detected as sequence variations that occur when one nucleotide in the aligned sequence has been altered. The key to this analysis is to be able to distinguish true SNPs from sequencing errors. Hence, polymerase chain reaction (PCR) duplicates were filtered from the alignment before proceeding with the SNP and Indel calls. Three tools were considered in this analysis to validate functional SNPs, namely, GATK, SAMtools and Nesoni. SNP results were compared across all three SNP callers and a union between all the tools was further filtered and annotated into synonymous and non-synonymous SNPs. Furthermore, a consensus list of SNPs predicted by all three tools was manually curated using SAMtools alignment viewer (samtools tview). Below we provide an overview of the SNP calling tools used in this analysis.

The Genome Analysis Toolkit (GATK)

In its SNP calling analysis, GATK starts by examining the sequence reads overlapping each base of the reference genome. The base calls and quality scores overlapping each position are then examined and used to calculate the probability of observed bases and quality scores for each sequence given a potential underlying genotype (McKenna et al., 2010; Depristo et al., 2011). These quality genotype likelihoods are defined as:

$$GL_{ij}(g) = P(\mathbf{B}_{ij}, \mathbf{Q}_{ij} | G_{ij} = g) \quad \text{Equation (1)}$$

Whereby GL_{ij} is the genotype likelihood associated with a specific genotype *g* and position *j* in sequence *i.* It is calculated as the probability of observing the vectors of bases $\mathbf{B}_{ij}$ and $\mathbf{Q}_{ij}$ overlapping position *j* in the mapped reads for sequence *i*. To assign genotype, GATK uses a Bayes rule approach:

$$P(G_{ij} = g | B_{ij}, Q_{ij}) = P(\mathbf{B}_{ij}, \mathbf{Q}_{ij} | G_{ij} = g)\, P(G_{ij} = g) / K_{ij} \quad \text{Equation (2)}$$

Where K is a normalizing constant. Following the initial set of variations called, there are a number of post-processing steps that take place to remove any false positives. The output SNP files are stored in Variant Call Format (VCF), which represents the SNP position and the nature of the base change, also supports an identifier, a quality score scaled in PHRED ($-10\log_{10}$) units analogous to a base quality, and a possible filter field.

SAMtools pileup

The SAMtools pileup SNP caller, implemented in the SAMtools package, was used to call SNPs directly from the alignment BAM file (Li et al., 2009). Reads with a minimum mapping quality of 20 were considered in this analysis. The variation quality score above 20 ensures a base call accuracy of 99%. For each reference genome position, the pileup output file of high quality variations captured the following information: chromosome, SNP position, reference Base, consensus base, consensus quality, SNP quality, number of reads mapping to the given SNP position. The consensus quality was calculated as a Phred-scaled probability that the consensus is identical to the reference sequence.

The called SNPs were filtered by setting the maximum read depth (number of reads of the same nucleotide position) per SNP locus to the average read depth. Filtering stage was carried out to remove false SNPs that account for structural variations or alignment artefacts.

Nesoni

Nesoni (Harrison and Seemann, 2009), is an open source Python/Cython (C-extensions for Python) program used to analyse high-throughput sequencing data based on the alignment to a reference genome. It is largely used to analyse prokaryotic genomes because of their small size. To avoid ambiguity between read alignment and consensus calls, Nesoni tallies each base counts at each aligned position to each reference base and then compares them using the Fisher's exact test. Furthermore, the tool carries out protein level consequences tests, where the genome annotation file in Genbank format was used to produce a list of protein level variation between the reference and the sequenced strain. This information was used to classify SNPs into synonymous and non-synonymous groups based on the amino acid change. Only variations where a consensus base was called are reported by Nesoni. Nesoni outputs two report files in GFF and text formats. It also outputs a “consequence” text file which contains the protein level variations.

2.4 Short Insertion/Deletion (Indel) Calling

The process of indel identification and calculation of the likelihood of the observed indel through local realignment was carried out by SAMtools and Nesoni using BAM alignment file as input (see section 2.3). The SAMtools pileup computes a consensus letter for each alignment column. A dot or a comma in the read base column stands for a match to the reference, either on the forward (dot) or reverse (comma) strand. Any other nucleotide that appears in the read base column represents a mismatch. Upper case letters are mismatches on the forward strand whereas lower case letters are mismatches on the reverse strand. Additionally, there are meta-characters returned in the output file. For example, a '+' indicates an insertion, a minus a deletion. A '^' symbol marks the start of a read segment and is followed by the mapping quality of that base. A dollar '$' symbol marks the end of a read segment. Nesoni employs the Fisher's exact test to compare each position in the reference as to the bases deleted or inserted from the consensus sequence. It also includes the protein level consequence analyses by determining the amino acid changes due to Indels, whether frame-shifting or frame-preserving (see section 2.3).

2.5 Indels and SNP annotation/filtering

Selection of SNPs and Indels common among Nesoni, SAMtools and GATK was employed as one of the criteria for variant filtering. Genetic variations (Indels and SNPs) shared between the three tools were filtered using a custom PERL script with a criteria imposed based on gene ID, substitution position and variation nucleotide identity. Consensus SNPs and Indels present in all three tools were considered true and were further manually curated using SAMtools tview. For example, identical SNPs between two tools were extracted and compared to SNPs predicted by the third tool in a pair-wise manner. The consensus genetic variation list was further annotated using Nesoni consequence, which took as input the reference genome sequences stored in a Genbank format.

2.6 Validation of genetic variants

The mapped reads were assembled *de novo* into contigs using Velvet and the resultant contigs were used as the reference sequence in a second round of assembly whereby raw short reads were mapped back onto the contigs. Subsequently, SNP calling was performed using a three-tools approach (see section 2.3). This procedure was carried out to identify true SNPs and to exclude the possibility that the observed SNPs and Indels are artefacts arsing from sequencing errors or due to rearrangements in the MTC reference genomes.

CHAPTER 3

RESULTS

3.1 Quality assessment and quality control: basic Statistics

The quality control procedure was undertaken to assess properties of the reads such as length, quality scores and base distribution in order to retain high quality reads for downstream assembly or mapping. The quality analysis was conducted in a series of 10 modules implemented in FastQC program and the first was a basic statistics table that included the number of sequenced reads, the read length, base call comparisons depending on the position within the read, the percentage G+C content within the sequences and per base sequence quality (Table 3.1). Part of the report was the characteristics of the sequences concerning uncalled/unidentified base call (N) and the level of duplicated sequences. Following this analysis, no reads were trimmed or filtered as all read quality observed was good.

Table 3.1. FastQC generated statistics for the *Oryx bacillus* reads

Measure	Value
Filename	Oryx.solidcfg02_20100127_bcSample1_.single.fastq
File type	Conventional base calls
Total Sequences	31271059
Sequence length	49
%GC	49

(i) Per base sequence quality

FastQC presented the per base sequence quality results in a graph (Figure 3.1). Along the x-axis on the plot, are all the individual bases for the reads and for every base called in this analysis, there is a distribution quality plotted on the y-axis. The yellow block shows the interquartile range, from the 25th to the 75th percentile. The red line shows the median value and the black viscus which goes between the 10th and the 90th percentile as well as the blue line which is the median quality score. It is of importance that the quality scores remain high with the increase in read length. The good quality is mostly in the ranges of 20 and above. With increasing sequence length there was a simultaneous decrease in read quality (Figure 3.1).

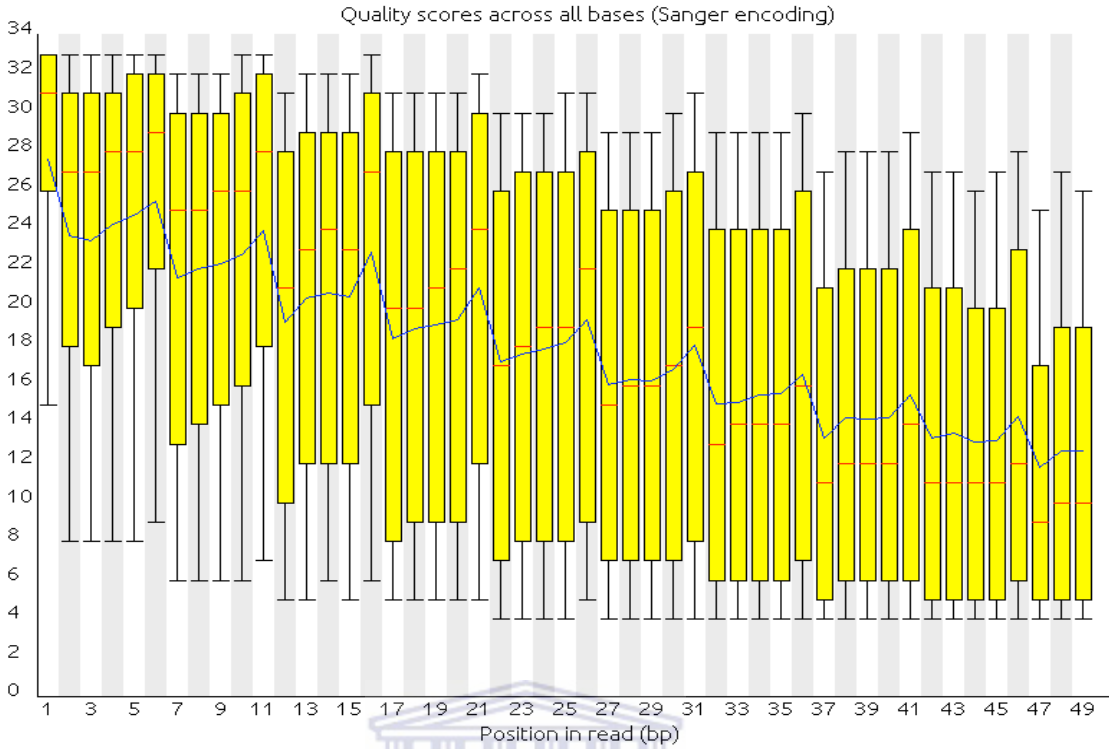


Figure 3.1. A plot showing quality scores across all bases.

(ii) Per sequence quality scores

Following quality distribution per base, FastQC generated the quality per sequence plot (Figure 3.2). For each sequence, FastQC computed the mean quality (Phred) score across all the bases of the particular sequence library, plotting out the distribution of the mean. In Figure 3.2, the sequences display a very tight distribution with universally high quality.

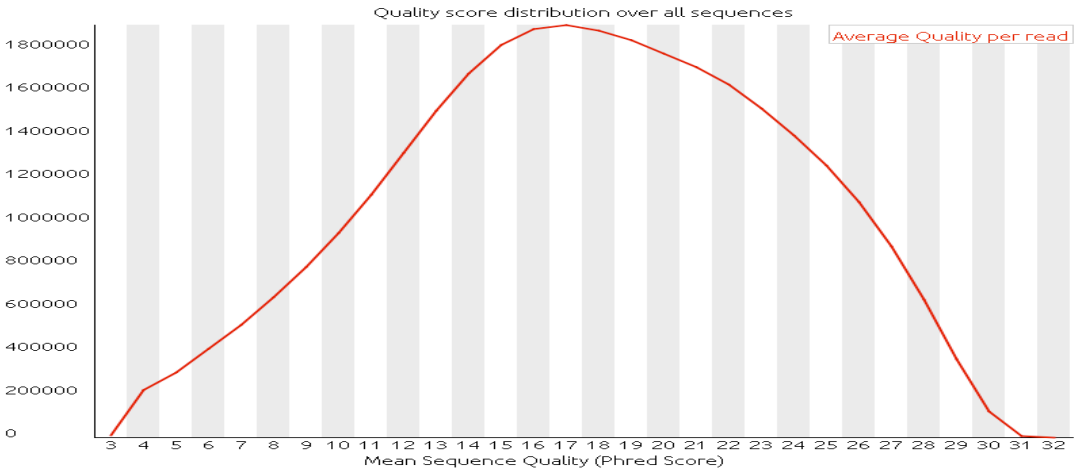


Figure 3.2. Distribution of quality scores across all sequences in the library.

(iii) The per base sequence content

The per base sequence content summarises the distribution of all the four bases (A, C, G and T) in a library of sequences because all of the bases should be equally distributed in a diverse library. Figure 3.3 displays an even distribution of the four bases which are not subject to change at any base position. The parallel lines across the plot display the distribution of the bases implying that the position looked at would not affect any base calls. Although there are spikes of over-representation of the T base in the start and end of the sequences, this is normal. A very biased position will have spikes reaching 50 to 80.

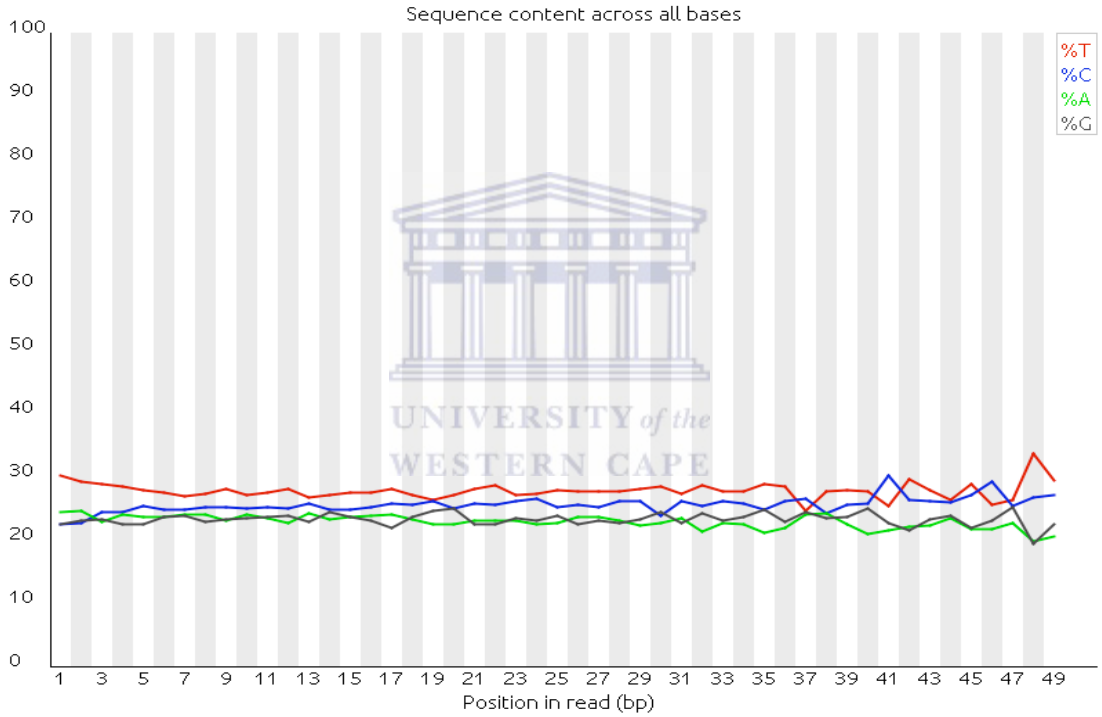


Figure 3.3. The sequence content across all bases in the library.

(iv) The per base G+C content

The mean G+C content for *Oryx bacillus* reads was 50%. *Mycobacterium tuberculosis* represents a high GC Gram-positive actinobacterium (Brosch et al., 2001). This analysis processes each base to check the G+C content across a read. A consistent G+C content across all bases of the reads were observed (Figure 3.4).

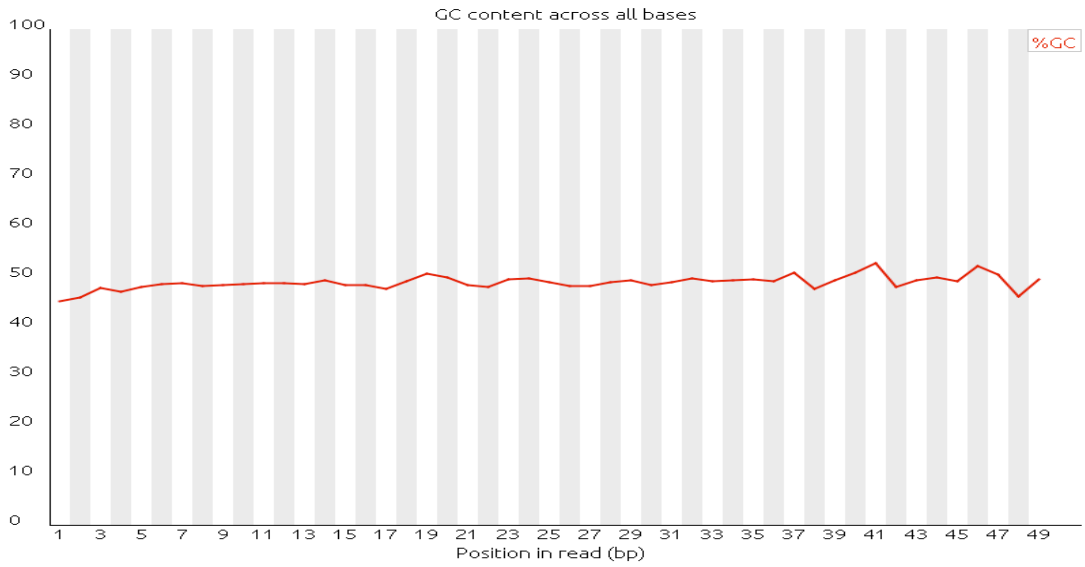


Figure 3.4. The mean % G+C content across all bases.

(v) Per sequence G+C content

Instead of only plotting the G+C content across each base in the library, FastQC also plots a distribution of the G+C content across a sequence (Figure 3.5). The red line is a distribution observed from the sample sequence while the blue line represents a theoretical normal distribution with the same mean and standard deviation as the sample real library contains. In Figure 3.5, the plot indicates a good overlay between the two distribution. The plots are nearly completely symmetrical around 50% mean GC content. The plot are an indication that there are no contaminants in the sample sequenced library.

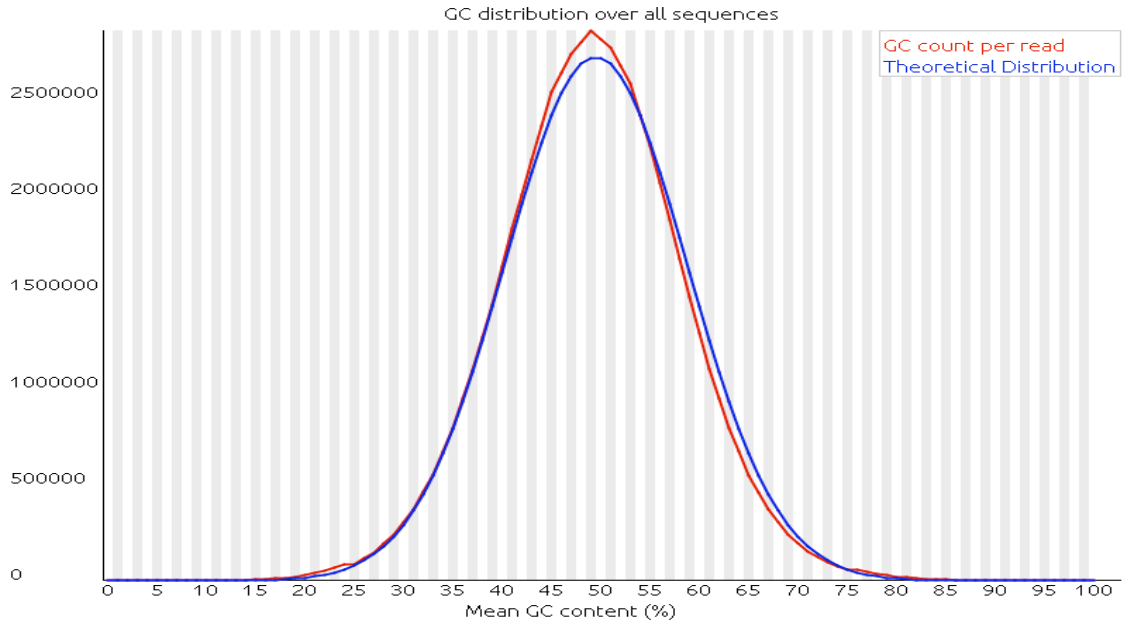


Figure 3.5. Distribution of the mean % G+C content over all sequences.

3.1.1 Sequence duplication levels

One way of identifying the uniqueness of each library is to carry out a sequence duplication level test (Figure 3.6). The raw sequence data was checked for the level of duplication. Each sequence in the fragment library should only occur once to avoid duplication and redundancy. In Figure 3.6, the plot immediately drops to two and everything else is zero. There was an overall sequence duplication level of 0.88% for this fragment library. This observation indicates that only 0.88% of sequences in the fragment library are non-unique and therefore duplicated. Following the sequence quality check, no reads were trimmed or filtered, as the library had good quality sequences.

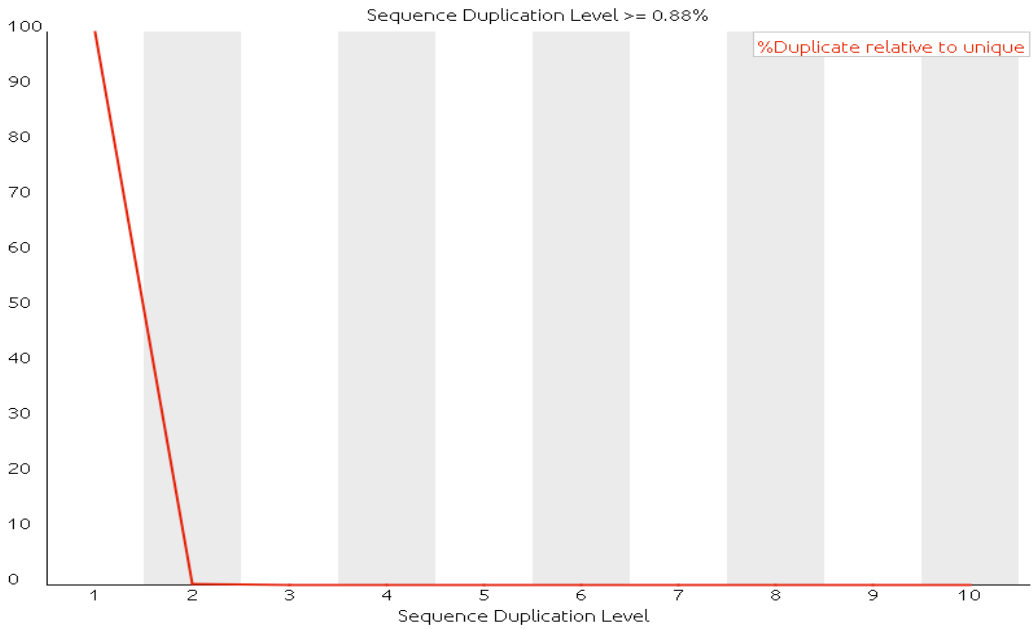


Figure 3.6. Sequence duplication levels.

3.2 Genome assembly of *Oryx bacillus*

3.2.1 Read alignment mapping

NovoalignCS was used for alignment of the SOLiD color space short reads to the reference genomes. This program is capable of handling both gapped and quality scored alignment. *Oryx bacillus* raw reads (31,271,059) were mapped to the reference genomes and the alignment results were formatted into SAM output format. The SAM files were converted into binary BAM format for storage (see Section 2.1.6).

A total of 20,730,631 (59.78% of the total reads) reads aligned to the reference genome *M. tuberculosis* H37Rv (Table 3.2). Of the total aligned reads, 3,408,587 (10.9%) were aligned to multiple chromosomal positions of the H37Rv reference genome. The uniquely-mapped reads, 17,322,044 (55.39%) were used for further downstream analysis (Figure 2.2). There were 13,949,015 (44.61% of the total reads) *Oryx bacillus* reads that were not aligned to any chromosomal position of the H37Rv reference genome. The unmapped reads were assembled into contigs using velvet (see section 2.2). Alignment results were summarised in Table 3.2 for all the reference genomes considered in this study. NovoalignCS was run in multiple alignment mode hence allowing redundancy of read mapping. There were significantly more reads mapping to H37Ra , F11 and H37Rv, with an overall 60% average mapping of *Oryx bacillus* short reads. An average of 19,382,475 (58 %) *Oryx bacillus* reads mapped to

the reference genome. In this thesis we used H37Rv mapping results as a starting point for our analysis even though *Oryx bacillus* and *M. tuberculosis* H37Rv have different host range.

Table 3.2 The sequence alignment statistics between *Oryx bacillus* and MTC reference genomes.

Reference Genome	Mapped (%)	Unmapped	Total reads	Uniquely-mapped reads	Multi-mapped reads
H37Rv	20,730,631 (59.78%)	13,949,015	31,271,059	17,322,044	3,408,587
H37Ra	20,946,104 (60.02%)	10,324,955	31,271,059	17,316,750	3,629,354
CDC1551	18,218,149 (56.65%)	13,940,989	31,271,059	17,330,070	888,079
F11	20,991,541 (60.15%)	13,909,532	31,271,059	17,361,527	3,630,014
KZN1435	20,327,531 (59.33%)	13,932,900	31,271,059	17,338,159	2,989,372
AF2122/97	17,378,779 (55.27%)	14,064,289	31,271,059	17,206,770	172,009
BCG	17,084,590 (55.47%)	14,773,669	31,271,059	16,497,390	587,200

3.2.2 Per base coverage

The alignment BAM files were used as input to BEDTools program to calculate coverage at each base of the reference genomes, that is, the number of reads that map to a given base on the reference genome. The coverage demonstrated in Figure 3.7 is the calculated average coverage per base (nucleotide) that is, read depth per nucleotide in the reference genome (see section 2.1.8). Furthermore, we identified regions of zero coverage by clustering *Oryx bacillus* reads according to the position that they map to on the reference genome, and hence, generated intervals of positions on the reference genome that were not supported by the alignment. The results were summarised as plots demonstrating coverage across the reference genome. Figure 3.7 shows genomic region on the x-axis and read coverage on the y-axis. The plots

display an even distribution of the *Oryx bacillus* reads over each of the reference genomes with some positions of the reference chromosome covered by more reads than average. There were also regions without any alignment (Table 3.3). The zero covered regions were calculated as per base not covered.

The *M. tuberculosis* H37Rv reference genome had 1,242 regions of zero coverage. These regions, when zoomed into vary from 1 to 1,882 nucleotides of zero coverage. With respect to the reference genome length, the total number of nucleotides with zero coverage averaged to 40,396 (0.92% of the whole genome), meaning that a 99.08% of the reference genome is covered by the *Oryx bacillus* reads.

Table 3.3. Per base zero coverage regions on the reference genome considered in this study.

Reference Genome	Reference Genome length	Zero Covered bases (%)	Min Coverage (nt)	Max Coverage (nt)	Total zero covered regions
H37Rv	4,411,532	40,396 (0.92%)	1	1,882	1,242
H37Ra	4,419,977	47,265 (1.07%)	1	6,552	1,240
CDC1551	4,403,837	47,960 (1.09%)	1	7,280	1,247
F11	4,424,435	41,277 (0.93%)	1	7,301	1,241
KZN_1435	4,398,250	32,572 (0.74%)	1	414	1,213
AF2122/97	4,345,492	46,028 (1.06%)	1	6,220	1,243
BCG	4,374,522	39,274 (0.9%)	1	6,220	1,210

3.2.3 Significance of zero-covered regions

Orxy bacillus regions of zero coverage could possibly represent i) deletions specific to *Oryx bacillus*; ii) species specific regions of difference (RD) commonly used in spoligotyping of *M. tuberculosis*; iii) these regions could represent sequencing errors. We attempted to elucidate the biological relevance of the zero-covered regions by mapping well annotated genes to these regions. A catalog of genes not covered by the *Oryx bacillus* reads was generated by mapping *Oryx bacillus* reads to known coding sequences from the reference genomes (Table 1.1) using novoalignCS (section 2.1.3). Reference genes with no read alignment support were cataloged (Table 3.4), however,

the *M. tuberculosis* KZN_1435 reference genome had all its coding sequences covered by the *Oryx bacillus* reads. The KZN_1435 strain is therefore not listed among the zero covered genes in table 3.4. Not all coding sequences of each of the analysed *M. tuberculosis* and the *M. bovis* reference genomes had genes with an alignment support of *Oryx bacillus* reads.

Oryx bacillus sequences did not align to the phiRv1 phage elements present in *M. tuberculosis* H37Rv, H37Ra and CDC1551. The F11 and BGC strains do not contain any of the phiRv1 phage elements. Furthermore, *Oryx bacillus* reads did not align to the 14th member of the MmpL gene family, MmpL14 (Table 3.5).

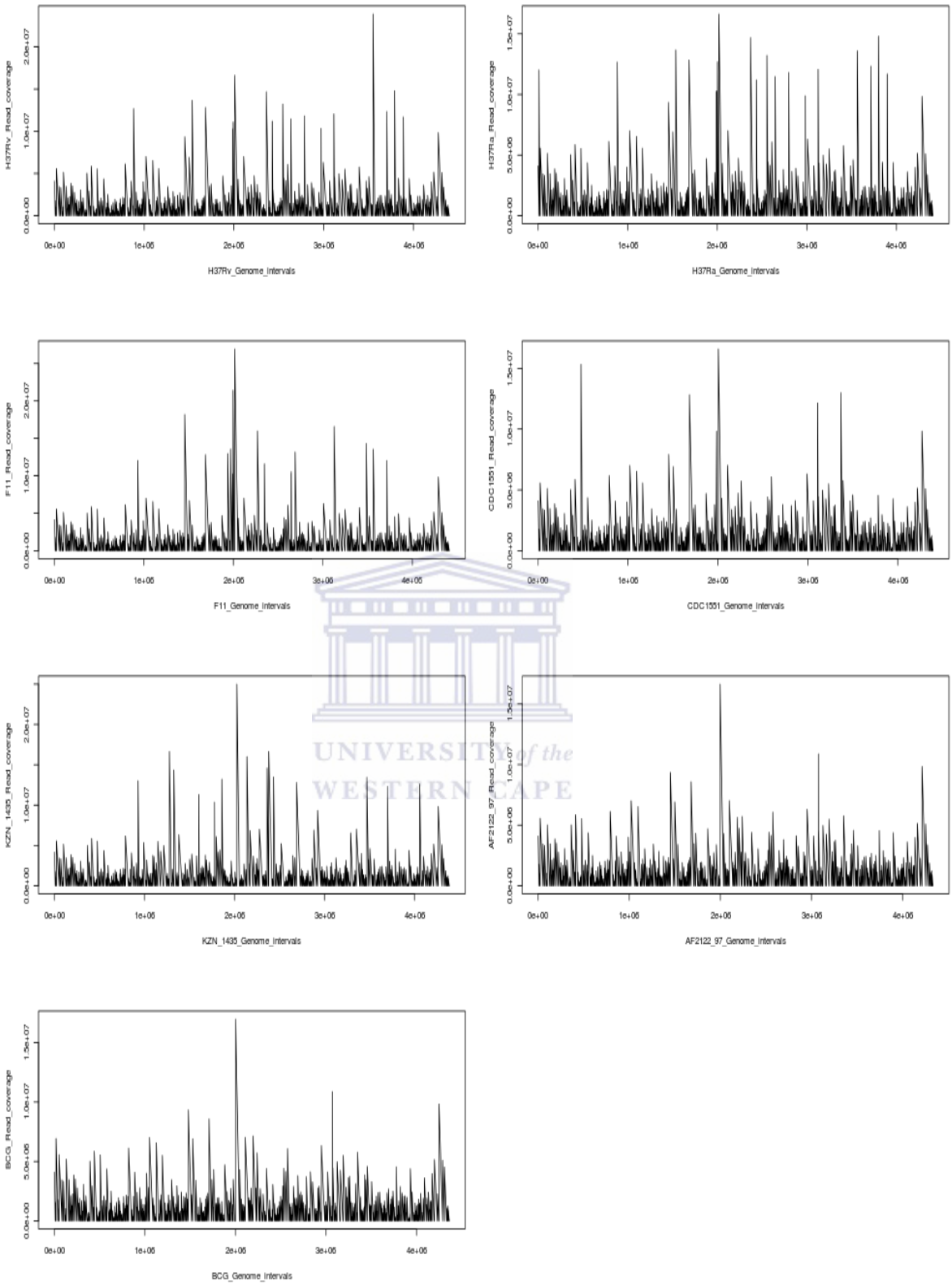


Figure 3.7. Genome coverage per base.

Each base of the reference genome (x-axis) was plotted against the read depth (y-axis) per nucleotide in the reference genome.

Table 3.4. Reference genes that overlap the zero coverage regions.

H37Rv	H37Ra	CDC1551	F11	AF2122/97	BCG	Gene Name
Rv1573	MRA1584	MT3573.15		Mb1599		PhiRv1 phage
Rv1575	MRA1586	MT3573.13				PhiRv1 phage
Rv1580c	MRA1591	MT3573.5		Mb1606c		PhiRv1 phage
Rv1583c	MRA1594	MT3573.2				PhiRv1 phage
Rv1584c	MRA1595	MT3573.10		Mb1610c		PhiRv1 phage
Rv1585c	MRA1596	MT3573.1		Mb1611c		PhiRv1 phage
	MRA1768D	MT1802	TBFG11778	Mb1787	BCG1797	MmpL 14
	MRA1768B	MT1800	TBFG11776	Mb1785c	BCG1795c	Glycosyl transferase
	MRA1768C		TBFG11777	Mb1786	BCG1796	sulfite oxidase
		MT1801				Molybdopterin oxidoreductase

Putative functions, encoded by *Oryx bacillus* zero-covered regions, were inferred based on the reference genes that overlapped regions without read alignment. In cases where the gene is not found in a strain, the colomn is left empty.

3.2.4 Deletions specific to *Oryx bacillus*

Mycobacterium tuberculosis complex deletion denoted as "RD" ("Region of Difference") forms the basis of Spoligotyping and they have previously been used to differentiate *M. tuberculosis* strains into ancestral and "modern" strains (Brosch et al., 2002). We attempted to identify deletions specific to *Oryx bacillus* and mapped the known RDs onto the current assembly.

Previous studies identified RvD2, RvD3 and RvD4 deletions to be *Oryx bacillus* specific (Brosch et al., 2002; Marmiesse, 2004; Mostowy et al., 2005). However, the presence of sequence reads within RvD3 and RvD4 regions contradicts this observation. Table 3.5 lists putative regions deleted from the *Oryx bacillus* genome. The RD3 contains the phiRv1 phage-like element with 14 ORFs. *Oryx bacillus* read alignment support for these 14 genes ranged from zero to nine reads. RD5 contains three genes (MT1799-MT1801) with the number of reads covering each gene. MT1799 is a phospholipase C and had 23% read coverage. The RD13 region contains three genes (Rv1255c-1257c) which have 100% read coverage.

Table 3.5. Reference genes that correspond to regions of difference specific to *Oryx bacillus*.

Region	Location	Reference genes	Read coverage
RD3	1780199-1780699	Rv1575	0 reads (0.0000000)
	1788162-1789163	Rv1587c	1 read (0.0429142)
	1780643-1782064	Rv1576c	4 reads (0.1392405)
	1782072-1782584	Rv1577c	1 read (0.0916179)
	1782758-1783228	Rv1578c	0 reads (0.0000000)
	1784497-1785912	Rv1582c	9 reads (0.2245763)
	1786307-1786528	Rv1584c	0 reads (0.0000000)
	1787096-1788505	Rv1586c	1 read (0.0340426)
	1786584-1787099	Rv1585c	0 reads (0.0000000)
	1783309-1783623	Rv1579c	3 reads (0.3079365)
	1779314-1779724	Rv1573	0 reads (0.0000000)
	1783620-1783892	Rv1580c	0 reads (0.0000000)
	1785912-1786310	Rv1583c	0 reads (0.0000000)
	1783906-1784301	Rv1581c	1 read (0.1212121)
RD5	1977868-1979412	MT1799	356 reads (0.2304207)
	1979541-1980686	MT1800	0 reads (0.0000000)
	1980876-1982015	MT1801	0 reads (0.0000000)
RD13	1402778-1403386	Rv1255c	609 reads (1.0000000)
	1403386-1404603	Rv1256c	1218 reads (1.0000000)
	1404717-1406084	Rv1257c	1368 reads (1.0000000)
RvD2	1973919-1975103	Mb1785c	0 reads (0.0000000)
	1975253-1976392	Mb1786	0 reads (0.0000000)
	1976593-1979430	Mb1787	0 reads (0.0000000)
RvD3	1993803-1994270	MT1812	536 reads (1.0000000)
	1994368-1994520	MT1813	155 reads (1.0000000)
RvD4		PPE (62)	111504 reads (0.9888627)

3.2.5 Visualization

The alignment BAM file was uploaded in to the integrated genome viewer (IGV) for visualisation. IGV allows zooming into a specific region of the alignment by entering the reference genome coordinates. In Figure 3.8 regions of zero read alignment were visualised. The region presented in Figure 3.8(a) overlaps the *MmpL14* gene (MT180, MRA1768D and TBFG11778). The read depth (also referred to as coverage) was presented as peaks on the graph while reads were presented as a stack of blocks shown below the peaks. Regions of zero coverage or where there is no sequence alignment are indicated by the absence of observable peaks. Figure 3.8(b) displays the RD3 region containing 14 ORFs from the H37Rv strain with zero *Oryx bacillus* read coverage.

Figure 3.8(a) Zero coverage region overlaps *MmpL14* gene encoded by CDC1551 reference genome.

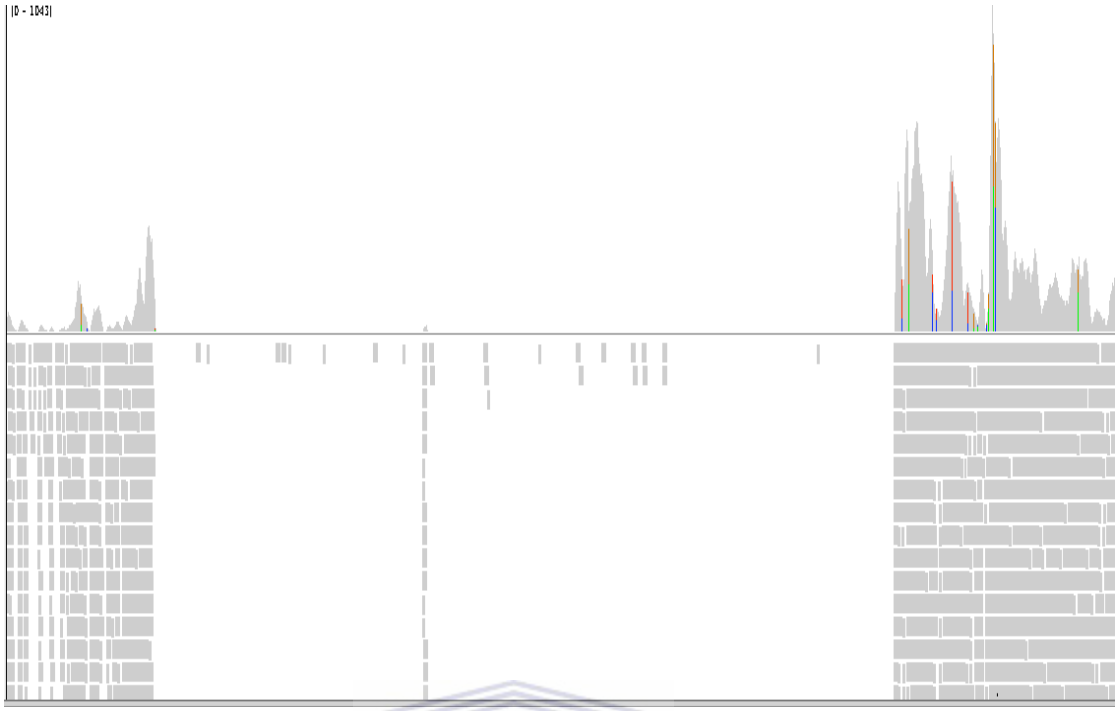


Figure 3.8(b) Zero coverage region overlaps phage-like element (phiRv1; RD3) in H37Rv reference genome.

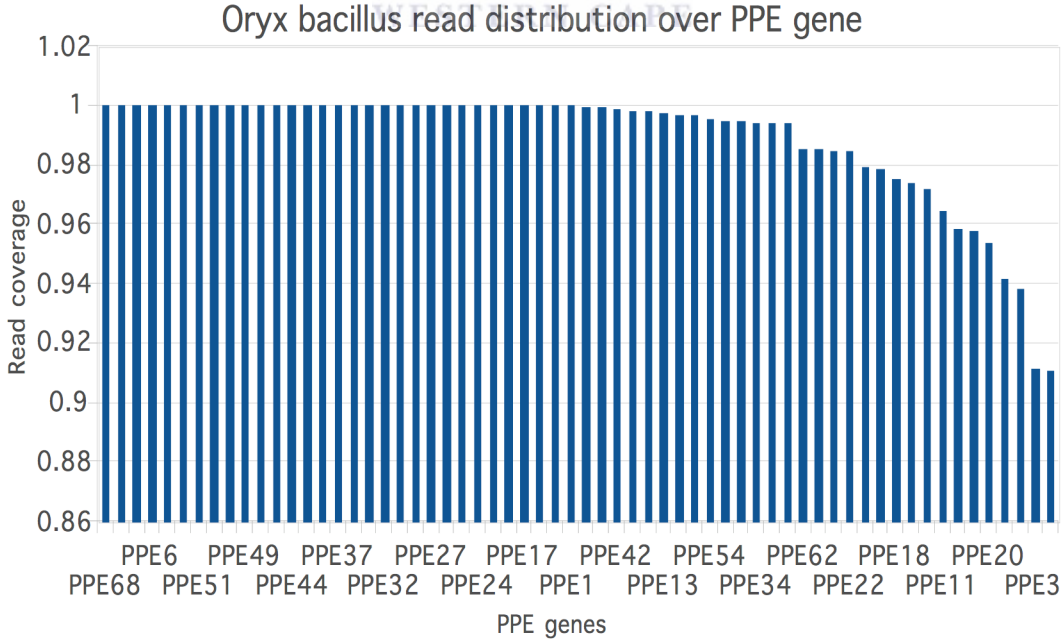


Figure 3.9. The *PPE* gene coverage by *Oryx bacillus* reads.

3.2.6 Unmapped read assembly

Oryx bacillus contigs were assembled from both aligned and unaligned reads using the Velvet short read assembler version 0.7.55 as implemented in the *de Novo* assembly pipeline for SOLiD v.2.0 (see section 2.2).

De novo assembly of 13,949,015 unmapped *Oryx bacillus* reads resulting from alignment to *M. tuberculosis* H37Rv reference genome generated 7,953 contigs and subsequently, the N50 was computed to estimate the quality of the assembly (Table 3.6). Velvet computes N50 by first ordering all contigs by size and then summing up their lengths until the summed length exceeds 50% of the total length of all contigs. For the assembly of reads that did not align to H37Rv reference genome, the computed N50 was 600 with the longest contig being 4,706 bps (and the shortest being a 100 bps) and a sum contig length of 4,078,065. The length distribution of these contigs was presented as graphs for all reference genomes used in this study with the length plotted on the x-axis and frequency of each contig plotted on the y-axis (Figure 3.10).

Overall the length of the contigs which contained 50 percent of the genome (N50) varied between 4,074 kb and 4,808 kb. The shortest contigs (minimum 100 bps) were the most frequent.

Table 3.6. A summary of the assembly statistics of the unmapped reads.

Coverage	Read type	Reference genome	Total contigs	Contigs N50	Contigs max length	Sum contigs length
300	Fragment 50	H37Rv	7,953	600	4,706	4,078,065
		H37Ra	9,521	583	3,942	4,808,059
		CDC1551	7,951	600	4,015	4,076,931
		F11	7,950	602	4,015	4,074,491
		KZN_1435	7,950	603	4,015	4,076,917
		AF2122/97	8,055	595	4,002	4,109,877
		BCG	8,091	592	4,021	4,117,677

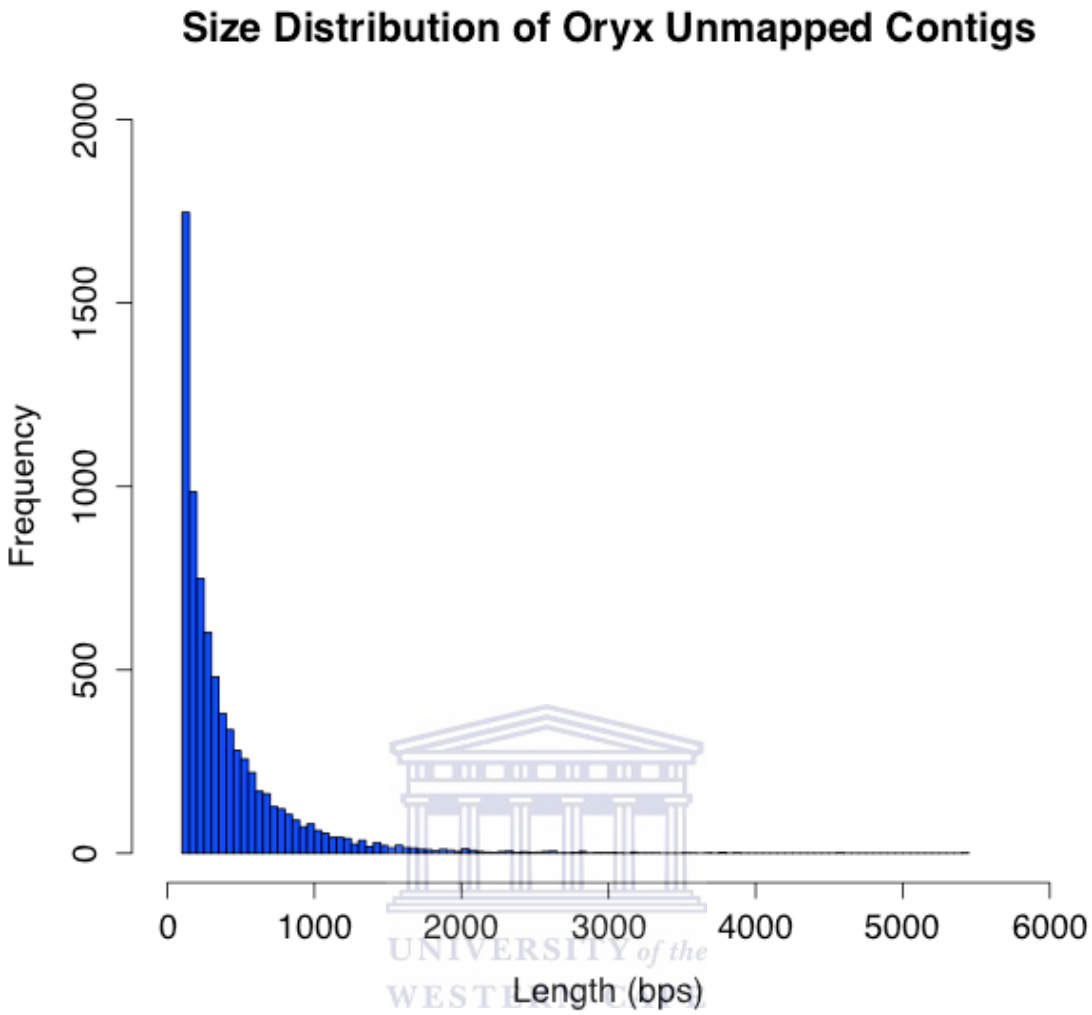


Figure 3.10. A cumulative plot for the length distribution of the generated contigs.

3.2.7 Annotation of unmapped contigs

Unmapped *Oryx bacillus* contigs (7,953) resulting from assembly using different reference genomes were screened against NR and Swiss-Prot databases (see section 2.2.1). This procedure successfully assigned putative functions to 7,415 (93.23%) contigs that were assembled from reads without sequence alignment to H37Rv reference genome. However, 538 contigs from this assembly had no matches. Of the 7,415 contigs with matches, 2,448 unique top hits were identified, 737 of these hits were matched by multiple queries with overlap while 520 of these hits were matched by multiple queries without overlap. The 1,612 queries matching these 520 hits were concatenated into 520 unique transcripts. Hence, the original 7,953 assembled sequences were collapsed down to 6,861 clustered sequences through this process. Additionally, database searches using unmapped contigs derived from the other reference genomes were summarised (Table 3.7).

Table 3.7. A summary of sequence similarity search results for unmapped contigs

Reference genome	Total contigs	Number of matches	Number of unique hits	Number of hits with overlap	Number of hits without overlap
H37Rv	7,953	7,415	2,448	737	520
H37Ra	9,521	8	7	1	0
CDC1551	7,951	16	15	1	0
F11	7,950	10	8	2	0
KZN_1435	7,950	8	7	1	0
AF2122/97	8,055	12	10	1	0
BCG	8,091	18	16	1	0

3.2.8 Functional annotation: GO analysis

Functional annotation of 7,953 contigs of *Oryx bacillus* sequences resulting from assembly using H37Rv reference genome was performed as described in section 2.2.2 using GO terms of cellular, biological and molecular function.

Gene Ontology terms were assigned to 4,951 (62.25%) assembled contigs while 2,714 (34.14 %) were mapped to known biochemical pathways. GO terms were assigned to each contig based on sequence similarity with known proteins annotated with GO terms (see section section 2.2.2). Empirical distribution of GO terms was estimated by calculating the frequency of occurrence of each GO term in the fragment library with a criterion imposed to eliminate duplicates (Figure 3.11 - 3.12).

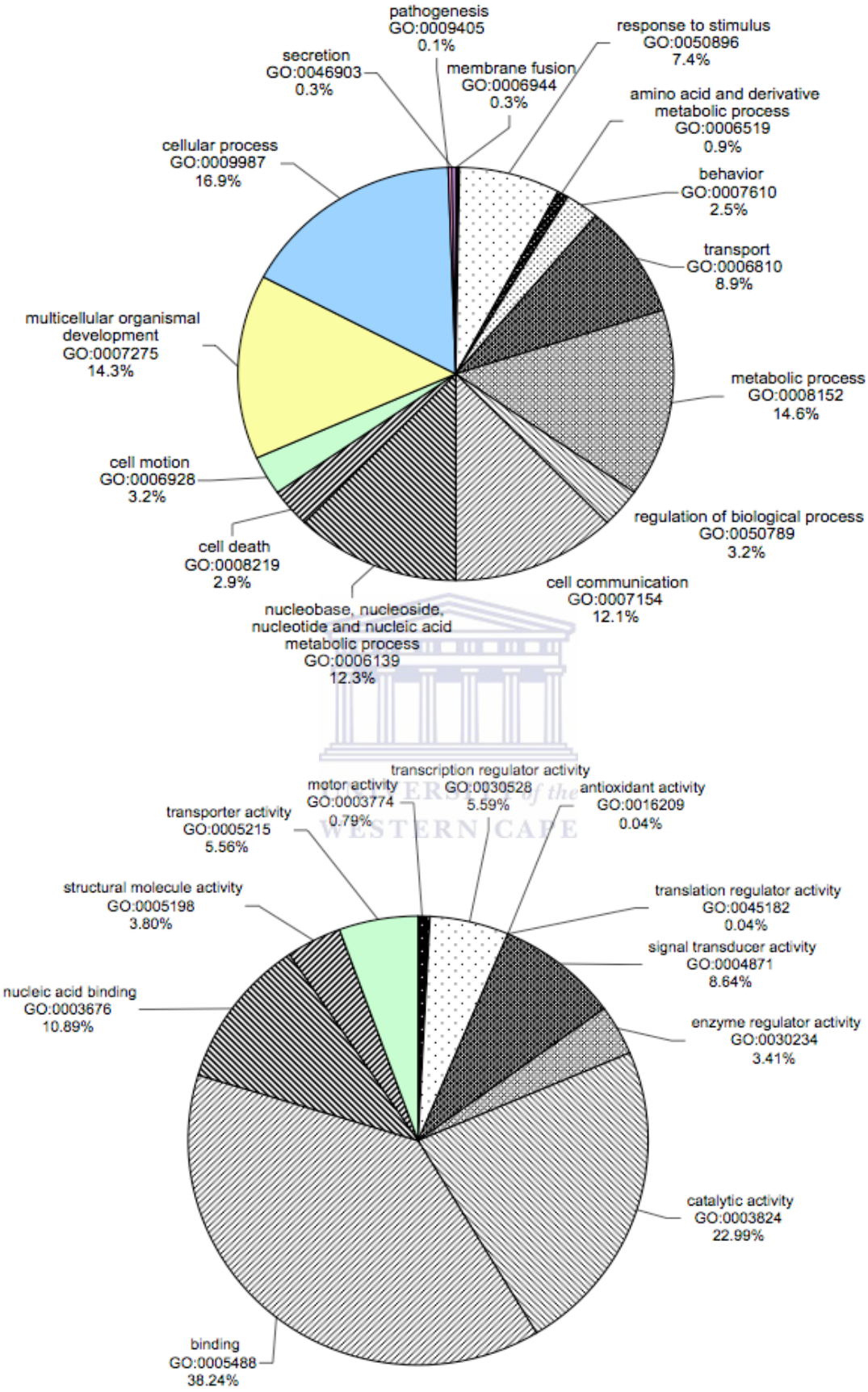


Figure 3.11. Functional annotation of *Oryx bacillus* unmapped reads by GO terms: (a) Biological process b) Molecular function.

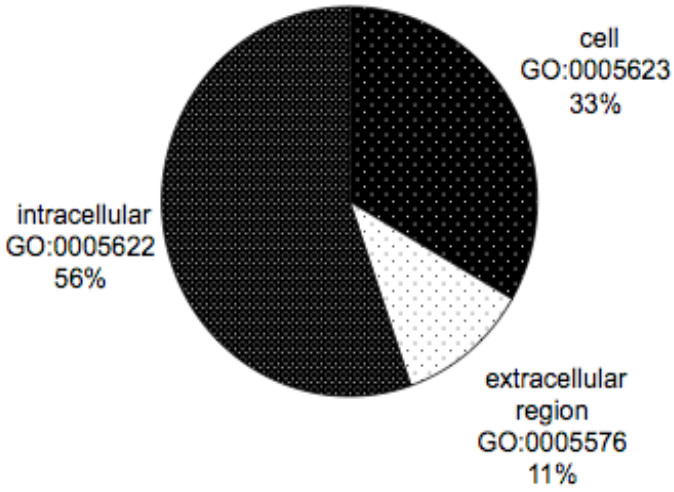


Figure 3.12. Functional annotation of *Oryx bacillus* unmapped reads by GO Cellular component terms.

3.3 SNP and Indel calling

We attempted to elucidate the biological relevance of the observed mutations based on the functionality of the genes (in the reference genome) in which the genetic variant leads (or is predicted to lead) to alteration in protein function and hence differences in virulence.

3.3.1 Computational prediction of genetic variations in *Oryx bacillus*

We identified a total of 2,680 genetic variations which were categorised into 845 synonymous and 1,724 non-synonymous SNPs together with 44 insertions and 67 deletions (Table 3.8). A highly configurable sequence analysis pipeline employing pre-existing tools was used to predict these variations based on the number of reads per allele as well as the quality score which weighted the variant calls using the error profile of the reads. Alignment between *Oryx bacillus* and *M. tuberculosis* H37Rv identified a higher number of variant alleles compared to the alignment between *Oryx bacillus* and other members of the MTC. For example, there were 672 (including 363 non-synonymous SNPs, 278 synonymous SNPs, 12 insertions and 19 deletions) out of the 1,340 variant-alleles that were predicted between *Oryx bacillus* and *M. tuberculosis* H37Rv reference genome (Figure 3.13).

Predicted variant-alleles were manually curated and catalogued (Table 3.9). A higher number of non-synonymous (a single nucleotide change leads to substitution of a

different amino acid) compared to synonymous SNPs (do not alter protein sequence) were observed and a majority of the nucleotide substitutions were transitions. These (transitions) are point mutations involving interchanges of purines (A<->G) or of pyrimidines (C<->T) while transversions are interchanges of purine for pyrimidine bases. However, transitions are less likely to result in amino acid substitutions (due to "wobble" concept), and therefore they persist as "silent substitutions" in populations. Non-synonymous mutations often render the encoded protein non-functional, hence they are often associated with disease (Bauer-Mehren et al., 2009).

Certain categories of protein families were over-represented in SNPs, for example, trans-membrane proteins, integral membrane proteins, transcriptional activator proteins, replication associated proteins, PPE family, PE-PGRS family, MCE-family protein, cell wall associated lipoproteins, ABC transporters,penicillin-binding proteins, enzymes (such as cytochrome P450, enoyl-CoA, glycine dehydrogenase, polyketide synthase, oxidoreductase, kinases), proteins associated with the tri-carboxylic acid (TCA) cycle. However, majority of the SNPs overlapped hypothetical proteins.

3.3.2 Sequence Alignment and mapping

On average, 50,5% *Oryx bacillus* reads were aligned to reference *M. tuberculosis* strains. The *M. tuberculosis* strain CDC1551, had the least number of read alignment (Table 3.8).

Table 3.8. Summary of *Oryx bacillus* reads mapped to the CDS's of five reference genomes

Vs Oryx	Mapped (%)	Unmapped	Total reads	Uniquely-mapped reads	Multi-mapped reads
H37Rv	17490362 (51.05)	13780697	31271059	14501568	2988794
H37Ra	17651722 (51.36)	13619337	31271059	14555436	3096286
CDC1551	15318708 (47.83)	15952351	31271059	14561316	757392
F11	17692364 (51.48)	13909532	31271059	14599091	3093273
KZN_1435	17265306 (51.02)	14005753	31271059	14695268	2570038

3.3.3 Genetic variations between *Oryx bacillus* and *M. tuberculosis* strains

We adopted a three-tools approach to analyse genetic variations of *Oryx bacillus*. Genetic variations predicted by SAMtools, Nesoni and GATK using the *M. tuberculosis* reference strains were summarised to generate a list of the overlapping variations (Figure 3.13 - 3.17) . Filtering this list and linking it to candidate genes in the reference genomes yielded 2,680 manually curated SNPs and Indels (Table 3.9).

SNP prediction and annotation

We used three SNP calling tools to predict SNPs and identify overlapping variant alleles between them. The accuracy and consistency of SNP calling tools varies depending on the in-built algorithms, hence a three-tools approach was adopted as a measure to reduce false positives. Figure 3.13 – 3.17 shows the relationship between the number of SNPs identified by the three variant callers. SNPs at the intersection between the three SNP callers were considered as true SNPs. The high number of SNPs generated by SAMtools might be caused by misalignments of sequenced reads to the reference genome and less stringent parameters for filtering SNPs. Nesoni tallies the raw base counts at each aligned position and does a comparison using Fisher's exact test to find variable nucleotide positions in *Oryx bacillus* relative to a reference sequence. Thus, making Nesoni a more strict variant caller.

Additional filtering criteria was imposed on SNPs and Indels called by SAMTools using the "samtool.pl varFilter". For the alignment, reads with a minimum mapping quality score of 20 were accepted while the minimum mapped read depth to the reference was greater than three and the maximum mapped read depth was set to 60. Further filtering was based on SNP base quality score and Phred-scaled probability. The Phred-scaled probability is the consensus quality score between the "reference base" and the "read bases" and it refers to the likelihood that the consensus is wrong. SNP quality score is the Phred-scaled probability that the consensus is identical to the reference. The minimum mapping quality for SNPs was set to 25.

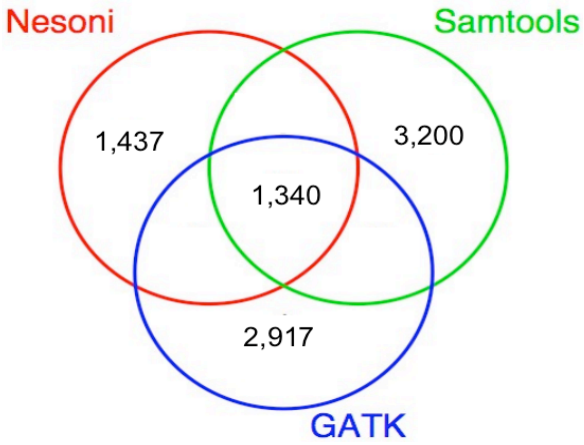


Figure 3.13. Venn diagram of the number of SNPs identified by three NGS SNP callers between *Oryx bacillus* and H37Rv.

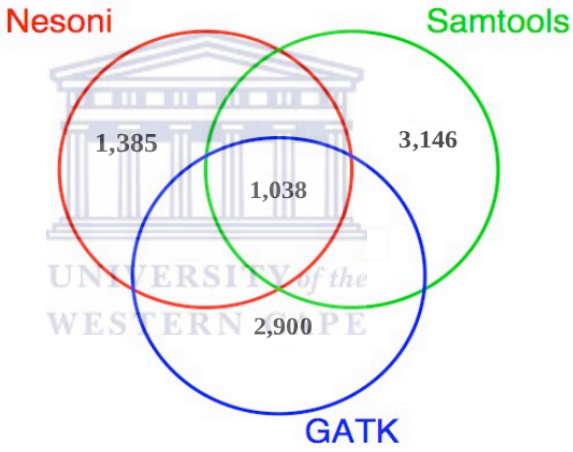


Figure 3.14. Venn diagram of the number of SNPs identified by three NGS SNP callers between *Oryx bacillus* and H37Ra.

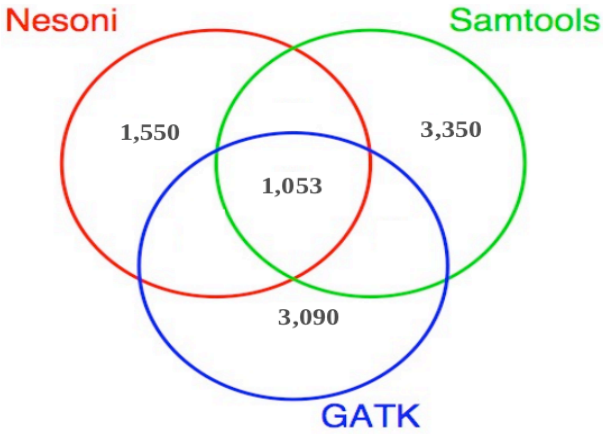


Figure 3.15. Venn diagram of the number of SNPs identified by three NGS SNP callers between *Oryx bacillus* and CDC1551.

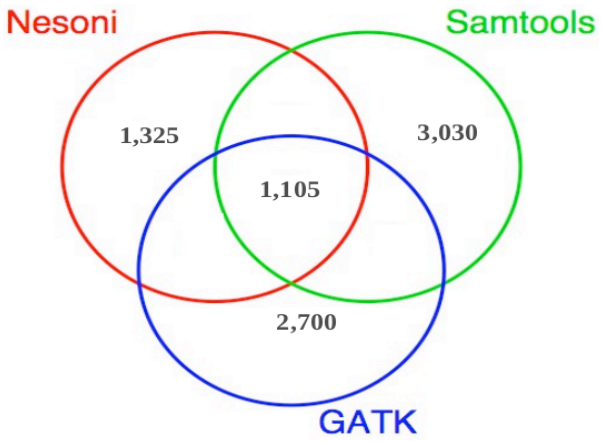


Figure 3.16. Venn diagram of the number of SNPs identified by three NGS SNP callers between *Oryx bacillus* and F11.

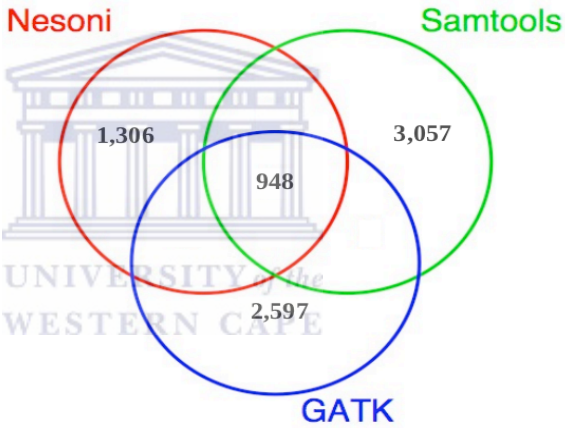


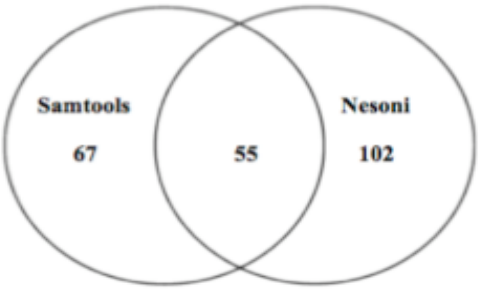
Figure 3.17. Venn diagram of the number of SNPs identified by three NGS SNP callers between *Oryx bacillus* and KZN_1435.

Indel prediction and annotation

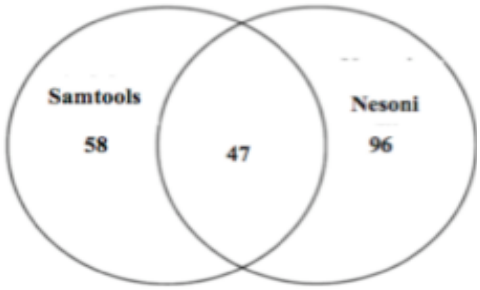
Indels called by the two variant callers were compared to identify variant alleles at the intersection between the two tools. This approach was undertaken, to identify indels (true indels) at the intersection between the two tools, due to lack of consistency between indel calling tools. Using SAMtools "tview", we manually filtered indels based on the position of the indel on the reference genome, the reference gene name and the number of reads supporting the indel (read depth). False indels not present at the recorded position were excluded from the analysis. Figure 3.18 shows the relationship between the numbers of the same Indels identified by two variant callers with overlapping number of reads generated between the two Indel callers.

Table 3.9. Curated genetic variations between *Oryx bacillus* and *M. tuberculosis* strains.

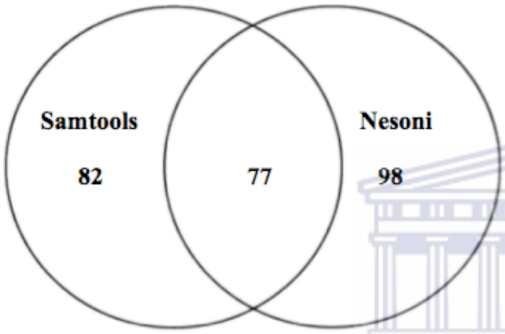
Genetic variations compared to *Oryx bacillus*	Number of variations	H37Rv	H37Ra	CDC1551	F11	KZN_1435
Synonymous	845	278	158	136	137	136
Non-synonymous	1724	363	345	344	341	331
Transitions (s/ns)	1708	420	340	324	326	298
Transversions (s/ns)	860	221	163	155	152	169
Insertions (≥1bp)	44	12	5	9	10	8
Deletions (≥1bp)	67	19	10	15	19	4
Total number	2680	672	518	504	507	479



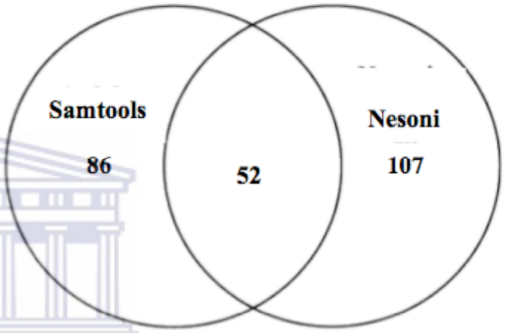
a) *Oryx bacillus* and MTC H37Rv.



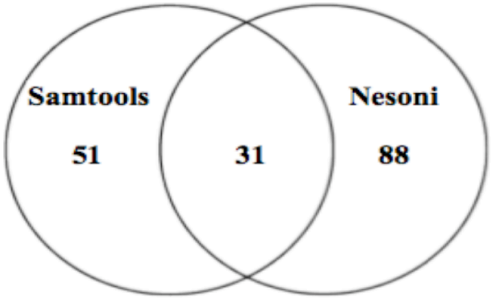
b) *Oryx bacillus* and MTC H37Ra.



c) *Oryx bacillus* and MTC CDC1551 .



d) *Oryx bacillus* and MTC F11.



e) *Oryx bacillus* and MTC KZN_1435.

Figure 3.18 (a-e). Genetic variation (Indels) between *Oryx bacillus* and MTC strains. Number of Indels predicted using Nesoni and SAMtools is shown. Indels that overlap between the two tools were considered as true indels.

3.3.4 Functional analysis of non-synonymous SNPs

Functional categories of these nsSNP are outlined in Table 3.10 with the majority of the nsSNPs occurring in genes coding for cellular metabolism and cell wall and cell processes. These mutations were further categorised into three groups to reflect SNPs associated with genes expressed during the MTC persistence and adaptation to changing environments. Three categories were captured (respiratory enzymes, stress-related products and metabolic enzymes as well as proteins involved in fatty acid catabolism) to reflect the potential of *Oryx bacillus* adaptation to changing environments (Table 3.11).

3.3.5 Mutations associated with transcription factors and global regulations

A number of nsSNP in *Oryx bacillus* mapped to MTC reference genes encoding regulatory proteins. Mutations in this category were limited because most nsSNPs occur in coding sequences resulting in alteration of the amino acid and hence the encoded protein. For example, we identified a G -> A *Oryx bacillus* nucleotide substitution at position 27 of an *lrp/asnC-family* transcriptional regulator encoded by *M. tuberculosis* KZN_1435 strain (TBMG_01195) with 414 short reads alignment support, hence confirming the nucleotide substitution.

The two component sensor kinase, a gene family known to contribute to the hypoxic response of *M. tuberculosis*, carried a mutation in *Oryx bacillus* compared to the reference genomes H37Rv (Rv0845), H37Ra (MRA_0853) and F11 (TBFG_10862). We observed two substitutions on this gene, with the first being a transversion (C –> A) at position 177, with 365 *Oryx bacillus* reads confirming the nucleotide substitution. The second variation (G –> A) at position 1133 of the gene with 80 reads confirming the nucleotide substitution (Figure 3.19).

Additionally, a deletion (T -> -) in *Oryx bacillus* was observed on the *pstB2* gene at position 186 with 295 read confirmation.

3.3.6 Mutations affecting genes associated with metabolism and respiration

We mapped *Oryx bacillus* reads to members of the MTC in order to identify and quantify polymorphisms affecting genes associated with metabolism and respiration. The observed disruptions of genes associated with oxidative stress and nitrate metabolism in *Oryx bacillus* are consistent with previous studies (Wayne and Lin, 1982; Wayne, 1994; Fritz et al., 2002; Sohaskey et al., 2003; Homolka et al., 2010). Mutations in these genes can potentially impair the capability of the bacilli to respond and survive stressful environmental changes. However, the disruptions in *Oryx bacillus* were mainly conservative and were classified as transitions. For example, the *narX* gene, a putative fused nitrate reductase , has been implicated in the persistence of *M. tuberculosis* in the host and it enables the mycobacterium to utilize nitrate as an energy source (Stermann et al., 2004). *Oryx bacillus* carried a transition (A –> G) mutation in the *narX* (TBMG_02259) gene when compared to *M. tuberculosis* KZN_1435 reference genome and the SNP occurred at position 230 with 180 read alignment support for the nucleotide substitution (Table 3.11).

The activity of *narG*, a gene that codes for a membrane-bound nitrate reductase in both *M. bovis* and *M. tuberculosis*, has been reported under hypoxic conditions. Mapping results identified a variant allele in the *NarG* gene when *Oryx bacillus* was compared to *M. tuberculosis* H37Rv, H37Ra and F11 reference genomes. We identified a transition (A –> G) SNP at position 2074 of the reference sequences (Rv1161/MRA_1172/TBFG_11186) with 329 *Oryx bacillus* reads mapping to the SNP locus.

Changes in gene expression due to metabolic downshift from an oxygen-rich environment switching to nitrate respiration are associated with MTC adaptation and persistence to oxidative stress in the host. For example, the *NarK2* gene, a putative nitrite-extrusion protein, was mutated in *Oryx bacillus* compared to *M. tuberculosis* KZN_1435 (TBMG_02258). This gene (*NarK2*) is a member of the dosR-controlled NRP regulon that regulates the expression of 48 genes in response to either hypoxia or nitric oxide (Sohaskey et al., 2003). We identified a transition (A –> G) at position 104, with 96 *Oryx bacillus* reads mapping to this locus.

Table 3.10. Functional categories of non-synonymous SNPs, identified in *Oryx bacillus* compared to five reference genomes.

Functional category	H37Rv	H37Ra	CDC1551	F11	KZN_1435
Metabolism and respiration	94	74	93	91	66
Cell wall and cell processes	78	60	63	49	42
Virulence, detoxification, adaptation	14	8	5	5	10
Regulatory proteins	10	12	8	11	13
Lipid metabolism	33	39	45	39	35
insertion seqs and phages	3	3	4	4	5
information pathways	20	21	14	18	19
PE/PPE multigene family	9	8	8	8	8
Hypothetical proteins	102	105	98	115	133
Total	636	345	344	341	331

3.3.7 Stress-response and general metabolic proteins

A number of mutations in genes involved in glycerol metabolism were identified and correlated with the metabolic potential of *Oryx bacillus*. *Oryx bacillus* encodes a lesion in the *pykA* gene that encodes pyruvate kinase (PK), an enzyme that catalyses the final committed step in glycolysis: conversion of phosphoenol pyruvate (PEP) and adenosine diphosphate (ADP) to pyruvate and adenosine triphosphate (ATP).

Under microaerophilic conditions and extended nitrosative stress, *M. tuberculosis* expresses the *Hmp* (Rv3571/TBMG_03610/TBFG_13604) protein, a putative flavohemoglobin which is mutated in *Oryx bacillus* when compared to *M. tuberculosis* H37Rv, KZN_1435 and F11. A nucleotide change (A –> G) at position 18 in the *hmp* gene was observed in *Oryx bacillus* with 35 read alignment support (Table 3.11). The hmp gene of *M. tuberculosis* is implicated in the protection of the organism from NO killing under hypoxic conditions (Hu et al., 1999).

3.3.8 Mutations affecting cell envelope and virulence

Alignment of *Oryx bacillus* reads to members of the MTC identifed mutations in several genes that are known to be under immune surveillance, for example genes encoding cell wall and cell processes proteins such as polyketide synthase (*pks7*). We identified polymorphisms in reads that overlapped the lipoprotein (*lppO*), a gene that encodes a protein that promotes the resuscitation of dormant or non-growing bacilli. A nsSNP was identified in *Oryx bacillus* reads that mapped to *lppO* (TBFG_12313) gene of *M. tuberculosis* strain F11, with a transitional (T –> C) nucleotide substitution at position 242 and a total of 241 reads overlapping this locus.

We identified a deletion (-C/*) in *Oryx bacillus* reads that overlapped the polyketide synthase (*pks7*) (Rv1661/MRA_1672/MT1701/TBFG_11679) gene, at position 5,426 and a total of 357 *Oryx bacillus* reads mapped to this locus (Figure 3.20). The *pks7* gene is involved in the formation of fatty acid components of the cell wall and has been implicated in virulence. Disruption of this gene causes defective production of phthiocerol dimycocerosates (Rousseau et al., 2003).

The mammalian cell entry (MCE) family of proteins are involved in the invasion and persistence in the host macrophages. Nucleotide substitutions in the MCE-family were observed when *Oryx bacillus* reads were mapped to *M. tuberculosis* (Table 3.11). For example, *mce1F* (Rv0170/ MRA_0178/ TBFG_10171/ MT0179/ TBMG_00171) gene contains a transition (T –> C) substitution at position 1109 of Rv0170 gene with a total of 231 short reads alignment support. *Oryx bacillus* carried mutations inthe reads that mapped to the *mce2* operon. Three nucleotide substitutions (C->T; A->C; T->G) and a deletion (C->-) at position 1283 of *M. tuberculosis* strain CDC1551 occurred inthe *mce2C* gene (Table 3.11).

Table 3.11. Mutations in genes implicated in invasion and persistence in the host macrophages.

Reference Locus Tag	Gene	Variation	Position	Oryx reads	Function Category
TBMG_02259	narX	A->G	230	183	Respiration
TBFG_11186	narG	A->G	2074	329	Respiration
TBMG_02258	nark2	A->G	104	96	Respiration
Rv1832	gcvB	A->G	1342	141	Central imetabolism
Rv3571	hmp	A->G	18	35	Flavohemoglobin-like
MRA_2956	fadD26	T->G	184	577	Lipid metabolism
MRA_2957	ppsA	A->C	16	115	Lipid metabolism
MRA_2957	ppsA	G->A	2135	99	Lipid metabolism
MRA_2957	ppsA	C->T	2409	39	Lipid metabolism
MRA_2957	ppsA	T->C	2433	87	Lipid metabolism
MRA_2957	ppsA	G->A	4318	281	Lipid metabolism
Rv2932	ppsB	T->C	3294	225	Lipid metabolism
Rv2932	ppsB	G->T	3809	328	Lipid metabolism
Rv2934	ppsD	G->A	553	342	Lipid metabolism
Rv2934	ppsD	G->A	816	96	Lipid metabolism
MRA_2961	ppsE	C->T	585	286	Lipid metabolism
MRA_2961	ppsE	A->C	2956	117	Lipid metabolism
MRA_2961	ppsE	A->G	3301	129	Lipid metabolism
TBFG_12313	lppO	T->C	242	241	cell wall and cell processes
Rv0170	mce1B	T->C	536	381	virulence,detoxification
Rv0174	mce1F	T->C	1109	231	virulence,detoxification
Rv0589	mce2A	T->C	152	428	virulence,detoxification
Rv0591	mce2C	C->T	146	454	virulence,detoxification
Rv0591	mce2C	A->C	1392	76	virulence,detoxification
Rv0591	mce2C	T->G	1407	112	virulence,detoxification
MT0621	mce2C	-C/-C	1283	18	virulence,detoxification, adaptation
Rv1967	mce3B	T->C	44	14	virulence,detoxification, adaptation
Rv1971	mce3F	C->A	1187	41	virulence,detoxification, adaptation

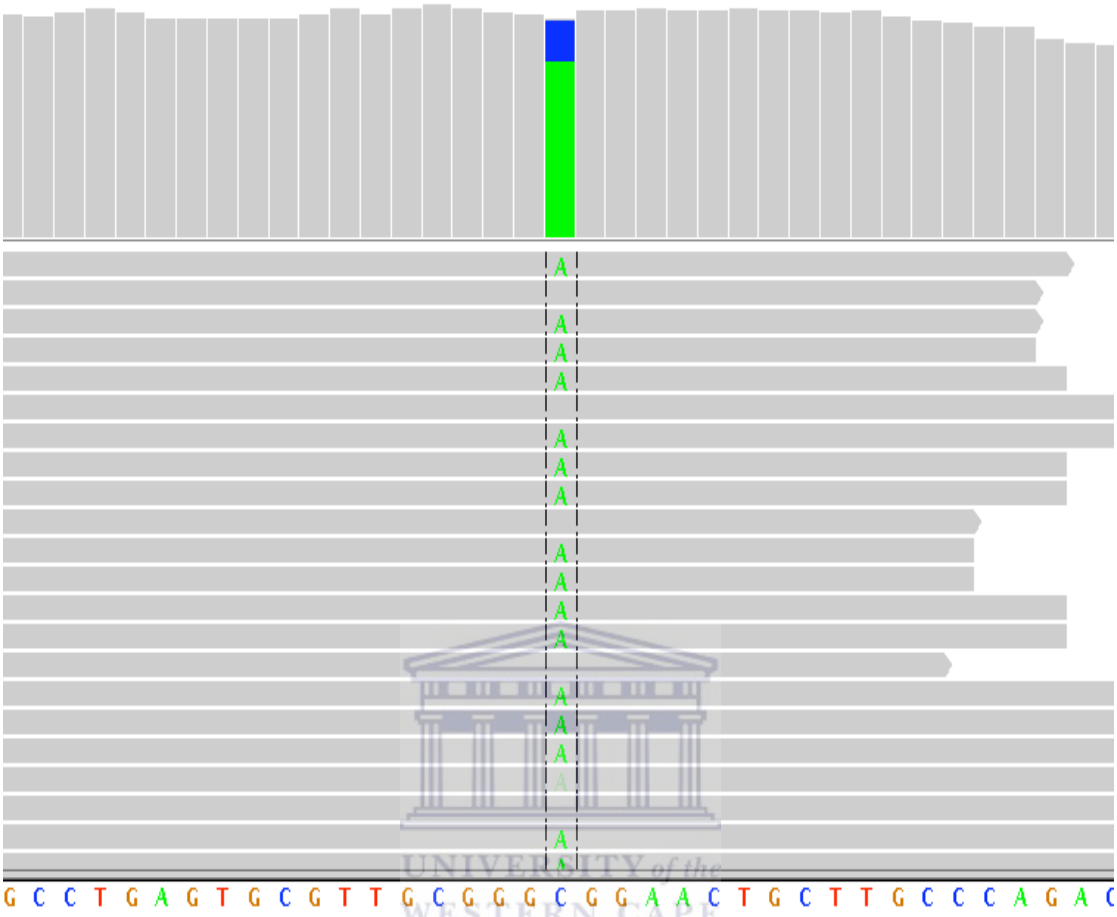


Figure 3.19. Visualisation of nsSNP using IGV.

A nsSNP observed in two component sensor kinase (Rv0845) gene in *Oryx bacillus* reads compared to *M. tuberculosis.*

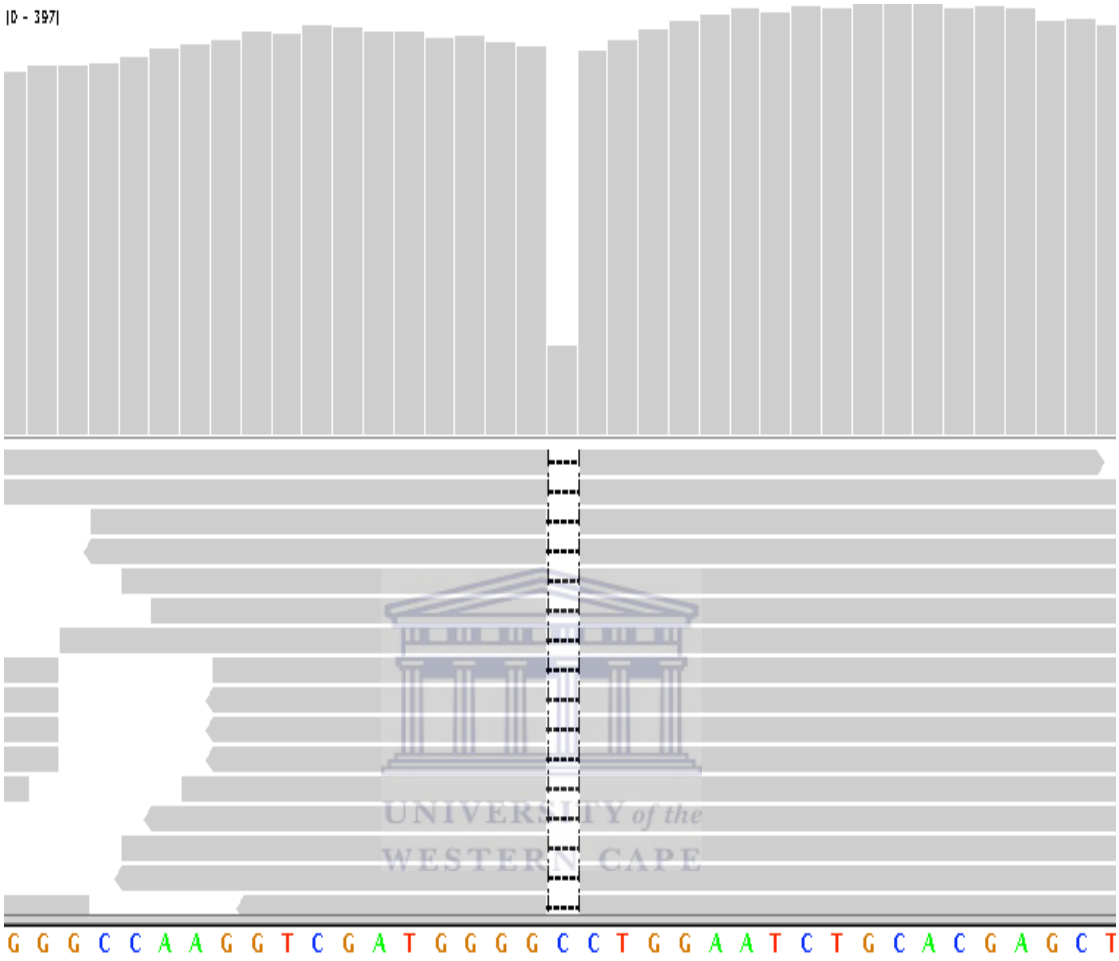


Figure 3.20. Visualisation of Indels using IGV.
A base deletion observed in the *pks7* (Rv1661) gene by *Oryx bacillus* when compared to *M. tuberculosis* H37Rv, H37Ra, CDC1551 and F11 reference strains.

3.3.9 Mutations related to drug resistance

Genetic variations that correlate with resistance to isoniazid, fluoroquinolones and ethambutol were observed in *Oryx bacillus* reads that mapped to the multi-drug resistant KZN_1435, H37Rv and H37Ra reference strains (Aubry et al., 2006). *Oryx bacillus* overlapped three mutations in the *gyrA* (Rv0006/MRA_0006) gene, responsible for resistance to fluoroquinolones, at positions 1324, 61 and 284, respectively (Table 3.12).

Mutations in *katG* (TBMG_02084) gene, the catalase/peroxidase that activates the pro-drug isoniazid, have most frequently been associated with drug resistance (Zhang et al., 1992). Previous studies have demonstrated that mutations in *katG* are associated with isoniazid resistance (Hazbón et al., 2006). We observed two polymorphisms (C->T and G->C) in *katG* (TBMG_02084) gene at positions 609 and 1405, respectively when *Oryx bacillus* was compared to the KZN_1435 strain (Table 3.12).

Additionally, *Oryx bacillus* contained a nsSNP in the reads that overlapped the membrane protein, arabinosyltransferase (*embB*), which is associated with resistance to ethambutol (EMB). A transition (C->T) mutation in *Oryx bacillus* compared to KZN_1435 *embB* (TBMG_03842) gene was observed at position 351, hence *Oryx bacillus* putatively prevent ethambutol from interfering with biosynthesis of the arabinogalactan layer in the cell wall (Ioerger et al., 2009).

In a recent study, Motiwala and coworkers, (2010) identified a nucleotide substitution (A –> G) at position 1046 in the *atsD* gene (Rv0663/TBFG_10676/TBMG_00674), an aryl-sulfate sulphohydrolase, implicated in drug resistance in *M. tuberculosis* (Motiwala et al., 2010). The same mutation (A –> G) was observed in *Oryx bacillus* when compared to *M. tuberculosis* H37Rv, F11 and KZN_1435 at position 1046 of the *atsD* gene.

Table 3.12. Non-synonymous SNP containing genes relative to drug resistance in *Oryx bacillus*

Gene	Mutation	Function	Known resistance effect
Rv0006 (gyrA)	G->C	DNA gyrase	fluoroquinolones
MRA_0006 (gyrA)	C->T	DNA gyrase	fluoroquinolones
MRA_0006 (gyrA)	G->A	DNA gyrase	fluoroquinolones
TBMG_02084 (katG)	C->T	Catalase/peroxidase	isoniazid
TBMG_02084 (katG)	G->C	Catalase/peroxidase	isoniazid
TBMG_03842 (embB)	C->T	Membrane protein, arabinosyltransferase	ethambutol
Rv3926 (drrA)	C->G	Membrane transporter	

CHAPTER 4

DISCUSSION

Oryx bacillus is a newly emerging *M. tuberculosis* strain in South Africa and its sequencing is giving insight into its evolution and disease causing mechanisms. With a fragment library of 50 bases per read, this strain was aligned to seven fully sequenced and annotated genome sequences of the MTC to identify conserved genomic features. As a member of the MTC, *Oryx bacillus* is said to have genetic patterns identical to *M. africanum* (Huard et al., 2006). However, it is believed to have emerged from *M. bovis* (Vasconcellos et al., 2010).

4.1 Genome assembly

Next generation sequencing technologies are rapidly becoming a cost-effective option for the analysis of genomes to discover genetic variations which might have implications in health and disease. The increasing availability of complete genome sequences of various members of the MTC, along with the evolving bioinformatics analysis tools, permits detailed inspection into their evolutionary scenario, pathogenicity, immunogenicity and the possibility of finding potential drug targets at a much greater pace. The effective analysis of NGS technologies data in modern genetics highly depends on efficient bioinformatics algorithms which are capable of handling any downstream analyses of NGS data (McPherson, 2009). As the bottleneck shifted from sequence generation to analysis, new and innovative analysis tools are in high demand to conquer the computational challenges in future (Flicek, 2009). This study successfully mapped 31,271,059 *Oryx bacillus* raw reads to the reference genomes. The methods implemented in this study attempted to resolve challenges associated with the analysis of ABI SOLiD sequence data.

In comparing our results with previous studies that used ABI SOLiD data, 100 % of *Oryx bacillus* raw reads passed quality filters (see section 2.1.2). Overall, our assembly incorporated an average of 60% good quality reads that mapped to reference genomes. For example, a total of 20,730,631 *Oryx bacillus* reads were evenly distributed and aligned to H37Rv reference genome while 13,949,015 were not aligned, leaving 1,242 zero covered regions. Based on the H37Rv genome length,

40,396 (0.92%) nucleotides in the genome of H37Rv had no sequence alignment support from *Oryx bacillus*, that is, zero read alignment (Table 3.2). This observation is consistent with the previous results that *Oryx bacillus* sequences are >99.08% identical at the nucleotide level to H37Rv reference genome (Brosch et al., 2002; Marmiesse, 2004; Mostowy et al., 2005). Furthermore, members of the MTC are nearly completely identical (>99.9%) at the nucleotide level (Cole, 2002). On the contrary, H37Ra reference genome had a total alignment support of 20,946,104 (60 %) Oryx reads and 1,240 regions with no coverage across the genome. There were 47,265 (1.07% of the whole genome) nucleotides with zero read alignment support. *M. tuberculosis* H37Ra is the avirulent counterpart of the virulent strain H37Rv and both strains were derived from their virulent parent, H37, but when compared to each other, H37Ra is 8,445 nucleotides (bp) larger than that of H37Rv (Zheng et al., 2008). *M. tuberculosis* strain F11 (family 11) was first isolated in South Africa and is said to have contributing to the high TB prevalence in the Western Cape province. This strain is as virulent and highly infectious as the H37Rv strain (Victor et al., 2004). F11 had 60.15% (20,991,541) of the *Oryx bacillus* reads successfully aligned to it with 1,241 regions of zero alignment. However, a total of 41,277 (0.93%) nucleotides of F11 genome were not covered by the *Oryx bacillus* reads.

The KZN_1435 strain is highly virulent and endemic to the Kwazulu-Natal region of South Africa (Ioerger et al., 2010). It is a multi-drug resistant strain and a member of the KZN strain family of *M. tuberculosis*. We successfully mapped 59.33% (20,327,531) of *Oryx bacillus* reads to KZN_1435 although 1,213 regions had zero alignment. Approximately 32,572 (0.74%) nucleotides of the reference genome had zero alignment support. An average of 55% of the *Oryx bacillus* reads aligned to each BCG and AF2122/97 strains of *M. bovis*. The coverage peak pattern demonstrated similarities (Figure 3.7). Although, BCG may have a slightly higher number of nucleotides in its genome, 39,274 nucleotides had zero alignment support while AF2122/97 had 46,028 nucleotides with zero coverage.

We attempted to elucidate the biological relevance of the zero-covered regions by mapping well annotated genes to these regions (see section 2.1.9-2.1.10). A number of genes were disrupted by these deletions and they have defined biological roles (Table 2.5). The prophage-like element phiRv1 resides in the region of difference, RD3, which is deleted in *M. bovis* strain BCG. The RD3, phiRv1 is situated in between the *bioB* and *bioD* genes (Lari et al., 2001). This region encodes at least 14

ORFs in H37Rv. The small number of ORFs do not have the capability to generate infectious particles (Bibb and Hatfull, 2002). This region is deleted in *Oryx bacillus*, an observation illustrated by poor alignment suppport of the 14 genes reported to reside in the RD3 (Figure 3.8(b)). Furthermore, *Oryx bacillus*, like *M. tuberculosis* H37Rv and CDC1551 reference genomes, contains a second prophage-like element phiRv2. This shared region, may contribute to the hypothesis that *Oryx bacillus* might possess characteristics capable of infecting humans. However, the missing RD3 in *Oryx bacillus* is also deleted in the two South African isolated *M. tuberculosis* strains, F11 and KZN_1435.

The RvD2 region of deletion is specific to *M. tuberculosis* and it includes a sugar transferase (glycosyl transferase) an oxidoreductase besides the *MmpL14* gene, which are both missing from the *Oryx bacillus* sequence alignment (Lari et al., 2001). The deletion of the RvD2 region has been proven to possess no consequences for the virulence of *M. tuberculosis*.

Previous studies identified RvD2, RvD3 and RvD4 deletions to be *Oryx bacillus* specific (Brosch et al., 2002; Marmiesse, 2004; Mostowy et al., 2005). In contrast, most RDs and all the RvD regions are highly conserved in the *M bovis* strains (AF2122/97 and BCG) including the RD3 region, present in most clinical *M. tuberculosis* strains (CDC1551 and H37Ra) (Cole et al., 1998).

Although *Oryx bacillus* is believed to contain deletions of all the six regions listed in Table 3.5, the alignment results only shows that the RD13, RvD3 and RvD4 are maintained. The RD5 (MT1799-1801) has poor read alignment with the MT1799 gene partially covered (23% of the gene length) by *Oryx bacillus* reads, and the MT1800-1801 genes had zero alignment. The MT1799 is defined as a phospholipase C coding gene with partial identity to the genes *plcA, plcB* and *plcC (*plcABC cluster*)* from the *M. tuberculosis* H37Rv reference genome (Raynaud et al., 2002). This gene is defined as a phospholipase C (PLC), a well known bacterial virulence factor (Lari et al., 2001). However, from the alignment results, *Oryx bacillus* contains the plcA-C genes with a 100% read coverage. Therefore the absence of MT1799 from *Oryx bacillus* might not necessarily conclude to the absence of PLC entirely, hence, the complementation with the entire plcABC cluster has the ability to restore full PLC activity in this strain (Raynaud, 2002).

The *M. tuberculosis* specific RvD2 deletion contains three genes (Mb1785-87)

derived from *M. bovis* strain AF2122/97 but this region is deleted in *Oryx bacillus* with zero read coverage. The RvD3 and RvD4 regions seem to be maintained with an alignment of over 98% of each gene length contained in these regions. The *Oryx bacillus* deletion RvD4 contains the PPE gene family, which is composed of 62 ORFs from the reference genome *M. bovis* strain AF2122/97. The PPE gene family was highly supported by a read depth of 111,504 and 99% coverage (Figure 3.8). *Orxy bacillus* regions of zero coverage could possibly represent i) deletions specific to Oryx; ii) species specific regions of difference (RD) commonly used in spoligotyping of *M. tuberculosis*; iii) these regions could represent sequencing errors.

4.2 SNP and Indel calling

In this thesis, we determined genetic variations in *Oryx bacillus*, a subtype of *M. bovis*, with reference to *M. tuberculosis*. These analyses provide accurate genomic information regarding the genetic differences between *Oryx bacillus* and *M. tuberculosis*, which are useful for a better understanding of the potential pathogenicity of the *Oryx bacillus* strain.

Based on our analyses, it is notable that the degree of genetic variation between members of the MTC and *Oryx bacillus* is confined to genes encoding cell wall and cell processes along with those encoding cellular metabolism and respiration. However, genes coding for regulatory proteins play vital roles in metabolic processes, including transcription, cell development and interactions with host cells. Across the aligned coding sequences from the five *M. tuberculosis* strains used in this study, *Oryx bacillus* showed 2,680 variations, including 845 synonymous and 1,724 non-synonymous SNPs together with 44 insertions and 67 deletions. The high number of nsSNPs may be a product of the close evolutionary relationship between *Oryx bacillus* and *M. tuberculosis* or a possible occurrence of pseudo-genes.

The majority of genetic variations observed in this study were transitions. Although transitions (1,708) occur at a higher frequency than transversions (860), they are less likely to result in amino acid substitutions (due to “wobble” concept), and therefore they persist as “silent substitutions” in populations (Table 3.9).

One of the differences between *M. bovis* BCG and *M. tuberculosis* is the requirement for pyruvate under conditions when glycerol is the sole carbon source (Garnier et al., 2003). *M. bovis* carries a SNP in the *pykA* gene that encodes pyruvate kinase (PK) enzyme. The lack of a functional PK implies that *M. bovis* is uanble to metabolise

glycerol hence it uses lipids as a sole carbon source (Keating et al., 2005). Like *M. bovis*, *Oryx bacillus* appears to have lost the functionality of *pykA* gene.

Variations observed in genes encoding respiratory proteins suggest that *Oryx bacillus,* like *M. bovis*, lost the functionality of nitrate reductase due to a mutation in the *narX* gene, a putative fused nitrate reductase. *NarX* gene encodes a respiratory nitrate reductase and it is expressed under hypoxic conditions, hence allowing survival with minimal metabolic activity (Höner zu Bentrup and Russell, 2001). In contrast to *M. tuberculosis*, the disrupted *narG* gene in *Oryx bacillus* further implies that *Oryx bacillus* probably lost the ability to exploit nitrate as the sole energy source. Thus, because there is low oxygen availability in the mature granuloma, *Oryx bacillus* might show reduced virulence as the infection progresses. Although both *M. tuberculosis* and *M. bovis* BCG express the NarG virulence factor, an anaerobic nitrate reductase, its role in virulence is unclear and it does not support anaerobic growth. The up-regulation of nitrate reductase activity under hypoxic conditions is attributed to nitrate and nitrite transport gene *narK2* (Sohaskey et al., 2003).

Activated macrophages are known to produce nitric oxide (NO) and related radicals that help in the macobactericidal response (Höner zu Bentrup and Russell, 2001). The *hmp* gene (flavohemoglobin) plays a role in protecting *M. tuberculosis* from nitrosative stress and the conversion of nitric oxide (NO) to nitrate (NO_3-) in oxygen-rich or nitrous (N_2O) in oxygen-deprived environment (Poole and Hughes, 2000; Höner zu Bentrup and Russell, 2001). However, it is not clear whether the nucleotide substitution observed in this gene when *Oryx bacillus* was compared to *M. tuberculosis* might impair *hmp* gene function.

The genome of *M. tuberculosis* contains a significant number of genes devoted to polyketide synthesis (Zheng et al., 2008). One of them has a nucleotide deleted in *Oryx bacillus* compared to *M. tuberculosis* H37Rv, H37Ra, CDC1551 and F11. The polyketide synthase (*pks7*) (Rv1661/MRA_1672/MT1701/TBFG_11679) gene is involved in the formation of fatty acid components of the cell wall and has been implicated in virulence. Disruption of this gene causes defective production of phthiocerol dimycocerosates (Rousseau et al., 2003). A nucleotide (-C/*) deletion was observed at position 5,426 of the *pks7* gene and 357 *Oryx bacillus* reads mapped to this locus (Figure 3.20).

Furthermore, we observed a deletion in the *atsA* (Rv0711/MRA_0719/TBMG_00725) gene, responsible for recycling sulfate whose loss of function may reflect lack of

sulfated glycolipids in *Oryx bacillus*. The high degree of variations observed in the MCE-family, involved in the invasion and prolonged existence in host macrophages, potentially contributes to virulence. Previous studies demonstrated that mutations in this gene family contribute not only to prolonged *in vitro* survival of the *M. tuberculosis* in the lungs, but also to *M. tuberculosis* virulence (Gioffré et al., 2005). Nucleotide substitutions in the MCE-family were observed when *Oryx bacillus* reads were mapped to *M. tuberculosis* (Table 3.11). Furthermore, *Oryx bacillus* carried mutations in the *mce1*, *mce2* and *mce3* operons that contain genes implicated in virulence and prolonged survival in macrophages.This observation implies that *Oryx bacillus* might be able to survive longer within macrophages.

During the first recorded outbreak of TB in the Arabian Oryx (*Oryx leucoryx*) in 1986, *in vitro* tests indicated that *Oryx bacillus* was sensitive to isoniazid (INH), ethambutol and rifampicin (Greth et al., 1994). However, based on the genetic variations identified in genes related to drug resistance in *M. tuberculosis*, our results contradict this observation (see section 3.3.9). *Oryx bacillus* contains mutations in *gyrA, katG* and *embB*, which might putatively confer resistance against first and second-line injectable drugs.

4.3 Limitations of this study

This study carried out a comprehensive genomic comparison of *Oryx bacillus* with all the currently fully sequenced members of the MTC. Assembly by reference genome may have its limitations, for example, without an existing reference genome that is about 99% identical to the *Oryx bacillus* isolate, it will be impossible to align more reads. Although NovoalignCS (novocraft version 2.05.02; Hercus, 2009) allows for eight mismatches per read, it is generally highly conservative. Hence, it is suitable while using a reference genome that is highly similar to the sequenced genome.

Sequence aligners are often unable to perfectly align reads containing insertions or deletions. Mismatches in a particular read can interfere with the gap, especially in low complexity regions. Strict parameters had to be employed for the differentiation between functional SNPs and sequencing errors.

The genome of *Oryx bacillus* was sequenced using ABI SOLiD generating a fragment library of short reads. Although a fragment library is suited for (re)sequencing of low complexity genomes, the disadvantages of this type of library include difficulty in mapping to repetitive regions, inability to detect structural rearrangements and is

often associated with less genome coverage.

4.4 Conclusions

This study has provided a comprehensive description of possible genetic variations, likely multiple mutations in *Oryx bacillus*, which may account for virulence and other phenotypic features that distinguish *Oryx bacillus* from members of the MTC. The results from this analyses show that high quality draft genomes can be obtained through *de novo* assembly of short read sequences. *Oryx bacillus* regions of zero coverage could possibly represent i) deletions specific to *Oryx bacillus*; ii) species specific regions of difference (RD) commonly used in spoligotyping of *M. tuberculosis*; iii) these regions could represent sequencing errors.

The observed results suggest that the virulence and host specificity of *Oryx bacillus* may be due to variations in genes encoding diverse functions such as cellular metabolism, respiration, cell surface antigenic variations, virulence factors, drug-resistance and proteins related to stress response.

Furthermore, our results suggest that the observed genetic variations in *Oryx bacillus* occur under selective pressures imposed by the host, leading to variation amongst members of the *Mycobacterium tuberculosis* complex. These findings have implications not only for improved understanding of pathogenesis of *Oryx bacillus* but also for development of new vaccines and new therapeutic agents. Furthermore, these results have implications for the development of a rapid diagnostic test for TB.

CHAPTER 5

CONCLUSIONS

Oryx bacillus is a newly emerging *M. tuberculosis* strain in South Africa and its sequencing is giving insight into its evolution and disease causing mechanisms. With a fragment library of 50 bases per read, this strain was aligned to seven fully sequenced and annotated genome sequences of the MTC. As a member of the MTC, *Oryx bacillus* is said to have genetic patterns identical to *M. africanum* (Huard et al., 2006). However, it is believed to have emerged from *M. bovis* (Vasconcellos et al., 2010).

The effective analysis of NGS data technologies highly depends on efficient bioinformatics tools which are capable of handling any downstream analyses (McPherson, 2009). As the bottleneck shifted from sequence generation to analysis, new and innovative analysis tools are in high demand to conquer the computational challenges in future (Flicek, 2009). A paradigm shift illustrating the convergence found in genetics and bioinformatics resulting from information accrued by NGS needs to be bridged by developing high-performance computing and intensive bioinformatics support (Flicek, 2009; Zhang et al., 2011).

In this thesis, a highly configurable sequence analysis pipeline employing pre-existsing tools, tailored specifically for ABI SOLiD (single end) sequence data, was developed. The pipeline covers the whole process of genome assembly, starting from raw sequence data pre-processing, sequence alignment and alignment statistics along with *de novo* assembly to variant detection. The pipeline incorporates read alignment using the full Needleman-Wunsch algorithm implemented in NovoalignCS with affine gap penalties and a limit of 8 mismatches per read, for single end reads, was used with external software incorporated for short read realignment, local realignment around indels, base quality score recalibration, and duplicate removal (Hercus C, 2009). Furthermore, it is capable of analysing sequences encoded in color space and handles single end reads. Other alignment programs such as MAQ, SOAP, Bowtie, were not chosen as they were either slower than NovoalignCS, did not support gapped alignment or allowed less mismatches per read alignment (Li et al., 2008; Langmead et al., 2009). Bowtie however, does not simply adopt any limit to mismatched read

alignments (Langmead et al., 2009). The preprocessing stage of this pipeline generated satisfactory results with no need for trimming or filtering any of the raw reads. Thus, all obtained reads were used for further downstream analysis, including assembly and variant calling.

5.1 Genome assembly of *Oryx bacillus*

Members of the *Mycobacterium tuberculosis* complex are characterized by 99.9% similarity at the nucleotide level; however, they differ widely in terms of their host ranges, phenotypes and pathogenicity (Brosch et al., 2002; Sreevatsan et al., 1997). To decipher distinct features among members of the MTC and *Oryx bacillus* raw reads, this study successfully mapped 31,271,059 reads to the reference genomes (Table 1.1) using a highly configurable sequence analysis pipeline (see section 2.1.3). Overall, the assembly yielded an average of 60% good quality reads that mapped to the reference genomes (Table 3.2). There were also regions without any alignment "zero covered regions" (Table 3.4). We attempted to elucidate the biological relevance of the zero-covered regions by mapping well annotated genes to these regions (see section 2.1.10).

Previous studies identified regions of difference RvD2, RvD3, RvD4, RD5, RD7, RD8, RD9, RD10 and RD13 specific to *Oryx bacillus* (Brosch et al., 2002; Marmiesse, 2004; Mostowy et al., 2005). Consistent with previous results, *Oryx bacillus* contains deletions of RD5, RD7, RD8, RD9, RD10 and RvD2. However, the presence of sequence reads with 100% coverage within RD13, RvD3 and RvD4 regions (Table 3.5) contradicts previous observations that these deletions are specific to *Oryx bacillus* (Brosch et al., 2002; Marmiesse, 2004; Mostowy et al., 2005).

We report for the first time a new deletion in *Oryx bacillus*, that corresponds to region of difference "RD3" in *M. tuberculosis* H37Rv. The RD3 region, that encodes a prophage-like element *phiRv1*, is deleted in *Oryx bacillus*, an observation illustrated by poor alignment support of the 14 genes reported to reside in this region (Table 3.5). *PhiRv1* includes a gene (Rv1586c) which is a member of the serine intergrase family with sequence similarity to transposon resolvases and DNA invertases (Bibb, Hancox, & Hatfull, 2005). This (RD3) deletion is present in the virulent *M. tuberculosis* H37Rv and *M. bovis* but it is absent in *M. bovis* BCG, *M. africanum* and *M. microti*. Furthermore, *Oryx bacillus*, like *M. tuberculosis* H37Rv and CDC1551 reference genomes, contains a second prophage-like element phiRv2. Both *phiRv* and

phiRv2 are about 10 kb in length and are similar to phages found in bacteriophages from *Streptomyces* (Cole et al., 1998). The presence of the RD3 deletion, which encodes bacteriophage-like elements (*phiRv1* and *phiRv2*) in *Oryx bacillus,* might be associated with phenotypic consequences such as reduced pathogenicity and host range when compared to *M. bovis*, a virulent strain with a wide host range.

The RD5 region contains three genes (*plcA, plcB, plcC*) encoding phospholipase C, an important virulence factor in *M. tuberculosis* H37Rv. This region also encodes the ESAT-6 family of genes encoding secreted T-cell antigens implicated in virulence (Tekaia et al., 1999). The RD7 deletion contains 14 ORFs that encode the *mce3* operon that codes for an invasin-like protein and an integral membrane protein while RD8 region encodes epoxide hydrolase genes (*EphA-F*) implicated in detoxification in addition to PPE and ESAT-6 family proteins (Gordon et al., 1999). The RD9 deletion encompasses four ORFs encoding an oxidoreductase (Rv2073c and Rv2074), an exported protein (Rv2075c) and a precorrin methyltransferase involved in the synthesis of cobalamin (*cobL*). The RD10 deletion overlaps two ORFs, *echA1* and *Rv0223*, which encode an enoyl CoA hydratase and an aldehyde dehydrogenase respectively.

Further differences between *Oryx bacillus* and members of the MTC involved deleted regions RvD1 and RvD2. RvD1 region overlapped three ORFs (RvD1-ORF1, RvD1-ORF2 and Rv2024c) encoding hypothetical proteins. The RvD2 region of deletion is specific to *M. tuberculosis* and it includes a sugar transferase (glycosyl transferase) an oxidoreductase besides the *MmpL14* gene, which are both missing from the *Oryx bacillus* sequence alignment (Lari et al., 2001). Consistent with this, *Oryx bacillus* contains a TbD1 locus consisting of the the *MmpL* gene family. The deletion of the RvD1 and RvD2 regions has been proven to possess no consequences for the virulence of *M. tuberculosis* (Gordon et al., 1999).

Although members of the *Mycobacterium tuberculosis* complex are over 99.9% similar at the nucleotide level, only 60% of *Oryx bacillus* raw short reads could be mapped to the MTC reference genomes. Unmapped *Oryx bacillus* reads were assembled into contigs (7,953) using different reference genomes and were screened against GenBank NR and Swiss-Prot databases (see section 2.2.1). This procedure successfully assigned putative functions to 7,415 (93.23%) contigs that mapped to members of the MTC and related prokaryotic bacteria. Several sequences coding for virulence factors were identified based on similarity to GenBank and Swissprot

entries. These were clustered into functional groups summarised below.

Enzymes: Short reads found in *Oryx bacillus* genome encode proteins with similarity to several enzymes important in fatty acid biosynthesis and the TCA cycle. Homologs of enzymes previously described in mycobacetria were identified. For example, homologs that putatively encode isocitrate lyase (Rv0467), an enzyme that converts isocitrate to succinate in the glyoxylate shunt hence allowing *Oryx bacillus* to grow on acetate or fatty acids as the sole carbon source. Other enzyme putatively encoded in this genome were alanine dehydrogenase, alcohol dehydrogenase, glutamine synthase, beta-lactamase, superoxide dismutase, catalase peroxidase (Rv1908c, *katG*) and phospholipase C. These enzymes probably have a potential role in the pathogenesis of tuberculosis.

Transcriptional Regulators: transcription regulators control the expression of many genes. For example, homologs of sigma factors were identified. Sigma A (Rv2703*, sigA*) regulates the expression of most mycobacterial housekeeping genes. We identified homologs of sigma E (Rv1221*, sigE*) essential for bacterial response to external stimuli, sigma F (Rv3286c*, sigF*) essential for sporulation, sigma H (Rv3223c*, sigH*) like sigma R is involved in response to oxidative stress.

Anaerobic respiration and oxidative stress proteins: Nitrate reductase is implicated in oxidative stress response. Several homologs of *narG* (Rv1161), *sodC* (Rv0342) and *katG* (Rv1908c) enzyme were identified in *Oryx bacillus* and they putatively allow the bacteria to adjust to oxidative stress encountered in the granulomatous tissues. Unlike *M. tuberculosis*, *Oryx bacillus* contains disruptions in *narG* and *narX* genes, hence, this implies that *Oryx bacillus* probably lost the ability to exploit nitrate as the sole energy source. Furthermore, the role of *narG* gene is unclear. Sohaskey and coworkers, (2003) attributed the up-regulation of nitrate reductase activity under hypoxic conditions to nitrate and the nitrite transport gene *narK2* (Sohaskey et al., 2003).

Cell surface components: mycobacterial cell wall and envelope is a glycolipid bilayer. We identified homologs of cell wall components such as Erp (Rv3810) an exported repetitive protein that is only found in pathogenic bacteria (Smith, 2003). *FadD26* (Rv2930) is a long-chain-fatty-acid AMP ligase involved in fatty acid degradation. Disruption of *FadD26* affects the expression of polyketide synthase required for phthiocerol biosynthesis (Azad et al., 1999).

Gene ontology terms were assigned to 4,951 (62.25%) assembled contigs while 2,714

(34.14%) were mapped to known biochemical pathways (see section 2.3.2). Terms associated with the contigs encoding genes that are highly enriched in Oryx included response to stimulus (GO:0050896), pathogenesis (GO:0009405), binding (GO:0005488), catalytic activity (GO:0003824), transcription (GO:0030528) and translation regulator activity (GO:0045182). Empirical distribution of GO terms was estimated by calculating the frequency of occurrence of each GO term in the fragment library with a criterion imposed to eliminate duplicates (Figure 3.11 - 3.12).

5.2 Genetic variations

Genetic polymorphisms in *Oryx bacillus* could account for the perceived phenotypic differences between MTC strains and *Oryx bacillus*. The identification of a series of deletions in *Oryx bacillus* relative to MTC strains suggested that insertion and deletion events (indels) could be one such mechanism (Gordon et al., 1998; Behr et al., 1999). Genetic variations were determined following alignment of *Oryx bacillus* raw reads to coding sequences of five *M. tuberculosis* strains. This analysis provided accurate genomic information regarding the genetic differences between *Oryx bacillus* and *M. tuberculosis*, which are useful for a better understanding of the pathogenesis of the *Oryx bacillus* strain.

We identified a total of 2,680 genetic variations which were categorised into 845 synonymous and 1,724 non-synonymous SNPs together with 44 insertions and 67 deletions. The majority of these variations were transitions. Although transitions occur at a higher frequency than transversions, they are less likely to result in amino acid substitutions due to the degeneracy of the genetic code at the Wobble or third codon position. Consistent with previous results, we observed an increased number of non-synonymous SNPs in coding regions when *Oryx bacillus* reads were mapped to MTC strains (Garnier et al., 2003; Zheng et al., 2008). A possible explanation for this observation is that nsSNPs potentially alter the structure or function of the encoded proteins, and are therefore, likely to account for disease susceptibility and altered drug response (Wang et al., 2009). Alternatively, the observed nsSNPs could be false positives created by sequencing errors and incomplete annotation information. In summary, comparison between *Oryx bacillus* and H37Rv identified a high number of genetic variations than comparisons that involved other MTC strains. One possible explanation for this observation might be that H37Rv has accumulated various mutations during repeated resequencing and *in vitro* passages (Zheng et al., 2008).

The existence of possible sequencing errors in the original H37Rv genome sequence cannot be ruled out (Zheng et al., 2008). However, high number of nsSNPs substitutions may be a product of the close evolutionary relationship between *Oryx bacillus* and *M. tuberculosis* or a possible occurrence of pseudo-genes.

A number of mutations in genes involved in glycerol metabolism were identified and correlated with the metabolic potential of *Oryx bacillus*. For example, a lesion in *pykA,* a gene that encodes pyruvate kinase (PK) enzyme that catalyses the final committed step in glycolysis: conversion of phosphoenol pyruvate (PEP) and adenosine diphosphate (ADP) to pyruvate and adenosine triphosphate (ATP). This observation implies that complex lipids are an essential carbon source because of the inability of *Oryx bacillus* to catabolise carbohydrates for energy. Sequence similarity searches using *Oryx bacillus* contigs identified homologous genes involved in beta-oxidation, the glyoxylate shunt and gluconeogenesis. These are essential pathways required for *in vivo* growth and persistence. For example, McKinney and coworkers, (2000) showed that the glyoxylate cycle enzyme, isocitrate lyase, is required for persistence in macrophages and mice. However, for the organism to enter the persistent state, the glyoxylate shunt within the tricarbonxylic acid (TCA) cycle has to be activated as a prerequisite to facilitate a metabolic shift to acetyl CoA as a primary carbon source.

Polymorphisms involving genes that are known to be under immune surveillance could account for the perceived virulence associated with *Oryx bacillus*, for example genes encoding cell wall and cell processes such as polyketide synthase (*pks7*) required for phthiocerol biosythesis. We identified polymorphisms in reads that overlapped the lipoprotein (*lppO*), a gene that encodes a protein that promotes the resuscitation of dormant or non-growing bacilli (see section 3.3.8). Disruption of genes encoding proteins involved with cell wall and cell processes could account for observed resistance to first-line drugs (see section 3.3.9).

Oryx bacillus reads overlapped known drug-resistance mutations (Aubry et al., 2006). For example, variations that correlate with resistance to isoniazid, fluoroquinolones and ethambutol were observed in *Oryx bacillus* reads that mapped to the multi-drug resistant KZN_1435, H37Rv and H37Ra reference strains. The antelope strain, *Oryx bacillus*, encodes drug-resistance mutations in *gyrA, katG* and *embB* genes, which explains resistance against first and second-line injectable drugs in comparison to *M.*

tuberculosis KZN 1435. MDR strains are associated with resistance to multiple drugs, such as the front-line drugs isoniazid and rifampicin (Ioerger et al., 2009). These results suggest that *Oryx bacillus* is a multi-drug resistant (MDR) strain and can evolve in to a virulent form without other XDR-specific mutations.

It is notable that most genetic variation between members of the MTC and *Oryx bacillus* were confined to genes encoding cell wall and cell processes along with those encoding cellular metabolism and respiration. Furthermore, a single base deletion was observed in the *atsA* gene, a system needed to recycle sulfate via the hydrolysis of sulfate esters, and may reflect lack of sulfate glycolipids in *Oryx bacillus* (Lamichhane et al., 2003; Sassetti et al., 2003; Ventura et al., 2007). The number of polymorphisms associated with transcription factors were limited because most nsSNPs occur in coding sequences resulting in alteration of the amino acid and hence the encoded protein (see section 3.3.5). However, genes coding for regulatory proteins play vital roles in metabolic processes, including transcription, cell development and interactions with host cells.

5.3 Limitations of this study

These next-generation DNA sequencing platforms have revolutionised biological and biomedical research, by enabling comprehensive analysis of genomes and transcriptomes. However, the error rates associated with NGS data are higher and the reads are shorter. Also, the volume of data produced created a challenge for computational analysis.

Genome coverage is important when dealing with NGS data. Numerous challenges are associated with NGS short reads. For example, reads are sampled from random locations of the genome and therefore, it is necessary to have an adequate number of reads to represent the whole genome. However, coverage does not have direct effect on the accuracy of the alignment because each read is independently mapped to the reference genome. A fragment library might not be an ideal sequencing library because it produces insufficient coverage and generates short reads randomly.

Additionally, sequence aligners are often unable to perfectly align reads containing insertions or deletions (indels) particularly with increasing Indel size. Strict parameters had to be employed for the differentiation between functional SNPs and sequencing errors.

5.4 Future perspective

Future research should be directed at prediction of ORFs and putative gene models. Furthermore, elucidation of protein-protein interaction maps (interactome) of *Oryx bacillus* may provide an insights of how pathogens exploit protein interactions to manipulate host immunity.

The method used to estimate SNPs is not entirely without errors. The high number of nsSNPs could contain artefactual contamination. The first potential artefact is quality control in the NGS data used. Thus the identified nsSNPs may not have any impact on gene expression. Experimental confirmation of the predicted SNPs and their influence on the gene will be required.

5.5 References

Alekseyev, M. A., & Pevzner, P. A. (2007). Colored de Bruijn graphs and the genome halving problem. IEEEACM Transactions on Computational Biology and Bioinformatics IEEE ACM, 4(1), 98-107.

Aubry, A., Veziris, N., Cambau, E., Truffot-Pernot, C., Jarlier, V., & Fisher, L. M. (2006). Novel Gyrase Mutations in Quinolone-Resistant and -Hypersusceptible Clinical Isolates of Mycobacterium tuberculosis : Functional Analysis of Mutant Enzymes. Society, 50(1), 104-112.

Azad, A. K., T. D. Sirakova, N. D. Fernandes, and P. E. Kolattukudy. 1997. Gene knockout reveals a novel gene cluster for the synthesis of a class of cell wall lipids unique to pathogenic mycobacteria. J. Biol. Chem. 272: 16741–16745.

Baess, I. (1979). Deoxyribonucleic acid relatedness among species of slowly-growing mycobacteria. Acta pathologica et microbiologica Scandinavica Section B Microbiology, 87(4), 221-226.

Barbraham Bioinformatics. (2009). FastQC: A quality control application for FastQ data.

Bauer-Mehren, A., Furlong, L. I., Rautschka, M., & Sanz, F. (2009). From SNPs to pathways: integration of functional effect of sequence variations on models of cell signalling pathways. BMC Bioinformatics, 10 Suppl 8, S6.

Behr, M A, Wilson, M. A., Gill, W. P., Salamon, H., Schoolnik, G. K., Rane, S., & Small, P. M. (1999). Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science, 284(5419), 1520-1523.

Bentley DR, S Balasubramanian, HP Swerdlow, GP Smith, J Milton, CG Brown, et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature, 456(7218), 53-59.

Betts, J. C., Dodson, P., Quan, S., Lewis, A. P., Thomas, P. J., Duncan, K., & McAdam, R. A. (2000). Comparison of the proteome of Mycobacterium tuberculosis strain H37Rv with clinical isolate CDC 1551. Microbiology, 146 Pt 12, 3205-3216.

Bibb, L. A., & Hatfull, G. F. (2002). Integration and excision of the Mycobacterium tuberculosis prophage-like element, phiRv1. Molecular Microbiology, 45(6), 1515-1526.

Bibb, L. A., Hancox, M. I., & Hatfull, G. F. (2005). Integration and excision by the large serine recombinase phiRv1 integrase. Molecular Microbiology, 55(6), 1896-1910.

Bishai, W. R., Dannenberg, A. M., Parrish, N., Ruiz, R., Chen, P., Zook, B. C., Johnson, W., et al. (1999). Virulence of Mycobacterium tuberculosis CDC1551 and H37Rv in rabbits evaluated by Lurie's pulmonary tubercle count method. (S. H. E. Kaufmann, Ed.)Infection and Immunity, 67(9), 4931-4934.

Bradley, S. G. (1973). Relationships among mycobacteria and nocardiae based upon deoxyribonucleic acid reassociation. Journal Of Bacteriology, 113(2), 645-51.

Brennan, P. J., & Nikaido, H. (1995). The envelope of mycobacteria. Annual Review of Biochemistry, 64(1), 29-63.

Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W. L., Russ, C., et al. (2008). Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Research, 763-770. doi:10.1101/gr.070227.107.

Brosch, R, Gordon, S V, Marmiesse, M, Brodin, P., Buchrieser, C., Eiglmeier, K., Garnier, T., et al. (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), 3684-3689.

Brosch, R, Gordon, S V, Pym, A., Eiglmeier, K, Garnier, T, & Cole, S T. (2000). Comparative genomics of the mycobacteria. International journal of medical microbiology IJMM, 290(2), 143-152.

Brosch, R, Pym, A. S., Gordon, S V, & Cole, S T. (2001). The evolution of mycobacterial pathogenicity: clues from comparative genomics. Trends in Microbiology, 9(9), 452-458.

Brosch, Roland, Philipp, W. J., Stavropoulos, E., Colston, M. J., Cole, Stewart T, & Gordon, Stephen V. (1999). Genomic Analysis Reveals Variation between Mycobacterium tuberculosis H37Rv and the Attenuated M. tuberculosis H37Ra Strain. (S. H. E. Kaufmann, Ed.)Infection and Immunity, 67(11), 5768-5774.

Calmette, A. (1928). Ann. Inst. Pasteur, 42(1).

Calmette, A., and C. G. (1924). Vaccination des bovides contre la tuberculose 'et methode nouvelle de prophylaxie de la tuberculose bovine. Ann. Inst. Pasteur, 38(371).

Camacho, L. R., Ensergueix, D., Perez, E., Gicquel, Brigitte, & Guilhot, C. (1999). Identification of a virulence gene cluster of Mycobacterium tuberculosis by signature-tagged transposon mutagenesis. Molecular Microbiology, 34(2), 257-267. doi:10.1046/j.1365-2958.1999.01593.x

Camus, J.-C., Pryor, M. J., Médigue, C., & Cole, Stewart T. (2002). Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. Microbiology, 148(Pt 10), 2967-2973. Mendeley Ltd.

Castets, M., N. Rist, H. B. (1969). La variété africaine du bacille tuberculeux humain. Medecine d'Afrique Noire, 1969, 16, 321-322, 16, 321-322.

Chaisson, M. J., Brinza, D., & Pevzner, P. A. (2009). De novo fragment assembly with short mate-paired reads: Does the read length matter? Genome Research, 19(2), 336-346.

Christoffels, A., Koh, E. G. L., Chia, J.-M., Brenner, S., Aparicio, S., & Venkatesh, B. (2004). Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Molecular Biology and Evolution, 21(6), 1146-1151

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., et al. (1998). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature, 393(6685), 537-544.

Cole, Stewart T. (2002). REVIEW ARTICLE Comparative and functional genomics of the Mycobacterium tuberculosis complex. Strategy, 2919-2928.

Comstock, G. W. (1994). Efficacy of BCG vaccine. Jama The Journal Of The American Medical Association.

Crevel, R. V., Ottenhoff, T. H. M., & Meer, J. W. M. V. D. (2002). Innate Immunity to Mycobacterium tuberculosis. Society, 15(2), 294-309. doi:10.1128/CMR.15.2.294

Cubillos-Ruiz, A., Morales, J., & Zambrano, M. M. (2008). Analysis of the genetic variation in Mycobacterium tuberculosis strains by multiple genome alignments. BMC research notes, 1(1), 110.

Daniel, T. M. (2006). The history of tuberculosis. British medical journal, 2(3229), 987-988.

Degner, J. F., Marioni, J. C., Pai, A. a, Pickrell, J. K., Nkadori, E., Gilad, Y., & Pritchard, J. K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics (Oxford, England), 25(24), 3207-12.

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics, 43(5), 491-8.

Dheda, K., Warren, R. M., Zumla, A., & Grobusch, M. P. (2010). Extensively drug-resistant tuberculosis: epidemiology and management challenges. Infectious Disease Clinics Of North America, 24(3), 705-25.

Domenech, P., Reed, M. B., & Barry, Clifton E. (2005). Contribution of the Mycobacterium tuberculosis MmpL Protein Family to Virulence and Drug Resistance. Infection and Immunity, 73(6), 3492-3501.

Eckstein, T. M., Inamine, J. M., Lambert, M. L., & Belisle, J. T. (2000). A genetic mechanism for deletion of the ser2 gene cluster and formation of rough morphological variants of Mycobacterium avium. Journal Of Bacteriology, 182(21), 6177-6182. American Society for Microbiology.

Fang, Z., Doig, C., Kenna, D. T., Smittipat, N., Palittapongarnpim, P., Watt, B., & Forbes, K. J. (1999). IS6110-mediated deletions of wild-type chromosomes of Mycobacterium tuberculosis. Journal Of Bacteriology, 181(3), 1014-1020. American Society for Microbiology.

Flicek, P. (2009). The need for speed. Genome Biology, 10(3), 212.

Forbes, M., Kuck, N. A., & Peets, E. A. (1962). Mode of action of Ethambutol. Journal Of Bacteriology, 84(5), 1099-1103.

Foster, P. L. (2004). Adaptive mutation in Escherichia coli. Journal Of Bacteriology, 186(15), 4846-4852.

Fritz, C., Maass, S., Kreft, A., & Bange, F. C. (2002). Dependence of Mycobacterium bovis BCG on anaerobic nitrate reductase for persistence is tissue specific. Infect Immun, 70(1), 286-291.

Furesz, S., and Timball, M. T. (1963). The antibacterial activity of rifamycins. Chemotherapia, 7(200).

Fyfe, J., & Globan, M. S. A. (2011). Identification and Molecular Characterisation of *Oryx bacillus* Isolates Causing Disease in Human Patients.

Gandhi, N. R., Moll, A., Sturm, A. W., Pawinski, R., Govender, T., Lalloo, U., Zeller, K., et al. (2006). Extensively drug-resistant tuberculosis as a cause of death in

patients co-infected with tuberculosis and HIV in a rural area of South Africa. Lance, 368(9547), 1575-1580.

Garnier, T., Eiglmeier, K., Camus, J.-C., Medina, N., Mansoor, H., Pryor, M., Duthoy, S., et al. (2003). The complete genome sequence of Mycobacterium bovis. Proceedings of the National Academy of Sciences of the United States of America, 100(13), 7877-82.

Gengenbacher, M., Rao, S. P. S., Pethe, K., & Dick, T. (2010). Nutrient-starved, non-replicating Mycobacterium tuberculosis requires respiration, ATP synthase and isocitrate lyase for maintenance of ATP homeostasis and viability. Microbiology, 156(Pt 1), 81-87.

Gioffré, A., Infante, E., Aguilar, D., Santangelo, M. D. L. P., Klepp, L., Amadio, A., Meikle, V., et al. (2005). Mutation in mce operons attenuates Mycobacterium tuberculosis virulence. Microbes and infection Institut Pasteur, 7(3), 325-334.

Gordon, D., Abajian, C., & Green, P. (1998). Consed: a graphical tool for sequence finishing. Genome Research, 8(3), 195-202. Cold Spring Harbor Laboratory Press.

Gordon, S. V., Brosch, R., Billault, A., Garnier, T., Eiglmeier, K., & Cole, S. T. (1999). Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. Mol Microbiol, 32(3), 643-655.

Greth, A., Flamand, J. R., & Delhomme, A. (1994). An outbreak of tuberculosis in a captive herd of Arabian oryx (Oryx leucoryx): management. Veterinary Record, 134(7), 165-167.

Gutierrez, M. C., Brisse, S., Brosch, Roland, Fabre, M., Omaïs, B., Marmiesse, Magali, Supply, P., et al. (2005). Ancient origin and gene mosaicism of the progenitor of Mycobacterium tuberculosis. PLoS pathogens, 1(1), e5.

Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., Schork, N. J., et al. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome biology, 10(3), R32.

Harrison, P., & Seemann, T. (2009). From high-throughput sequencing read alignments to confident , biologically relevant conclusions with Nesoni. Small, 2009-2009.

Hazbón, M. H., Brimacombe, M., Bobadilla Del Valle, M., Cavatore, M., Guerrero, M. I., Varma-Basil, M., Billman-Jacobe, H., et al. (2006). Population Genetics Study of Isoniazid Resistance Mutations and Evolution of Multidrug-Resistant Mycobacterium tuberculosis. Antimicrobial Agents and Chemotherapy, 50(8), 2640-2649.

Hendrix, R. W. (1999). Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. Proceedings of the National Academy of Sciences, 96(5), 2192-2197.

Hercus C. (2009). www.novocraft.com.

Homolka, S., Niemann, S., Russell, D. G., & Rohde, K. H. (2010). Functional genetic diversity among Mycobacterium tuberculosis complex clinical isolates:

delineation of conserved core and lineage-specific transcriptomes during intracellular survival. PLoS Pathog, 6(7), e1000988.

Höner zu Bentrup, K., & Russell, D. G. (2001). Mycobacterial persistence: adaptation to a changing environment. Trends in microbiology, 9(12), 597-605.

Hu, Y., Butcher, P. D., Mangan, J. A., Rajandream, M. A., & Coates, A. R. (1999). Regulation of hmp gene transcription in Mycobacterium tuberculosis: effects of oxygen limitation and nitrosative and oxidative stress. J Bacteriol, 181(11), 3486-3493.

Huard, R. C., Fabre, M., De Haas, P., Van Soolingen, Dick, Cousins, D., & Ho, J. L. (2006). Novel Genetic Polymorphisms That Further Delineate the Phylogeny of the Mycobacterium tuberculosis Complex. Journal Of Bacteriology, 188(12), 4271-4287.

Ioerger, T. R., Feng, Y., Chen, X., Dobos, K. M., Victor, T. C., Streicher, E. M., Warren, R. M., et al. (2010). The non-clonality of drug resistance in Beijing-genotype isolates of Mycobacterium tuberculosis from the Western Cape of South Africa. BMC Genomics, 11(1), 670.

Ioerger, T. R., Koo, S., No, E.-G., Chen, X., Larsen, M. H., Jacobs, W. R., Pillay, M., et al. (2009). Genome Analysis of Multi- and Extensively-Drug-Resistant Tuberculosis from KwaZulu-Natal, South Africa. (B. Marais, Ed.)PLoS ONE, 4(11), 9.

Jones, D., Metzger, H. J., Schatz, A., & Waksman, S. A. (1944). Control of gram-negative bacteria in experimental animals by streptomycin. Science, 100(2588), 103-105.

Kanehisa, M, S Goto, S Kawashima, and A Nakaya. 2002. The KEGG databases at GenomeNet. Nucleic Acids Res 30, no. 1: 42-46.

Karboul, A., Gey van Pittius, N. C., Namouchi, A., Vincent, Véronique, Sola, C., Rastogi, N., Suffys, P., et al. (2006). Insights into the evolutionary history of tubercle bacilli as disclosed by genetic rearrangements within a PE_PGRS duplicated gene pair. BMC evolutionary biology, 6, 107.

Karlson, A. G., & Lessel, E. F. (1970). Mycobacterium bovis nom. nov. International Journal of Systematic Bacteriology, 20(3), 273-282.

Kato-Maeda, M., Bifani, P. J., Kreiswirth, B. N., & Small, P. M. (2001). The nature and consequence of genetic variability within Mycobacterium tuberculosis. Journal of Clinical Investigation, 107(5), 533-537.

Kawahara-Miki, R., Tsuda, K., Shiwa, Y., Arai-Kichise, Y., Matsumoto, T., Kanesaki, Y., Oda, S.-I., et al. (2011). Whole-genome resequencing shows numerous genes with nonsynonymous SNPs in the Japanese native cattle Kuchinoshima-Ushi. BMC Genomics, 12(1), 103.

Keating, L. A., Wheeler, P. R., Mansoor, H., Inwald, J. K., Dale, J., Hewinson, R. G., et al. (2005). The pyruvate requirement of some members of the Mycobacterium tuberculosis complex is due to an inactive pyruvate kinase: implications for in vivo growth. Mol Microbiol, 56(1), 163-174.

Kinsella, R. J., Fitzpatrick, D. A., Creevey, C. J., & McInerney, J. O. (2003). Fatty acid biosynthesis in Mycobacterium tuberculosis: Lateral gene transfer, adaptive

evolution, and gene duplication. Proceedings of the National Academy of Sciences of the United States of America, 100(18), 10320-10325.

Koch, R. (1882). Die Aetiologie der Tuberculose. Berliner Klinische Wochenschrift, 15, 10 pp.

Koch, R. (1982). Classics in infectious diseases. The etiology of tuberculosis: Robert Koch. Berlin, Germany 1882. Reviews Of Infectious Diseases, 4(6), 1270-1274.

Kremer, K., Van Soolingen, D., Frothingham, R., Haas, W. H., Gaillot, O., Martín, C., Palittapongarnpim, P., et al. (1999). Comparison of Methods Based on Different Molecular Epidemiological Markers for Typing of Mycobacterium tuberculosis Complex Strains: Interlaboratory Study of Discriminatory Power and Reproducibility. Journal of Clinical Microbiology, 37(8), 2607-2618.

Labidi, A., & Thoen, C. O. (1989). Genetic relatedness among Mycobacterium tuberculosis and M. bovis. Acta Leprologica, 7 Suppl 1, 217-221.

Lamichhane, G., Zignol, M., Blades, N. J., Geiman, D. E., Dougherty, A., Grosset, J., Broman, K. W., et al. (2003). A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to Mycobacterium tuberculosis. Proceedings of the National Academy of Sciences of the United States of America, 100(12), 7213-8.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10(3), R25.

Lari, N., Rindi, L., Garzelli, C., Sperimentale, P., Mediche, B., & Epidemiologia, I. (2001). Identi ® cation of one insertion site of IS 6110 in Mycobacterium tuberculosis H37Ra and analysis of the RvD2 deletion in M . tuberculosis clinical isolates. Society, 50(April), 805-811.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., et al. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078-2079.

Long, R., Nobert, E., Chomyc, S., van Embden, J., McNamee, C., Duran, R. R., et al. (1999). Transcontinental spread of multidrug-resistant Mycobacterium bovis. Am J Respir Crit Care Med, 159(6), 2014-2017.

Magdalena, J., Vachée, A., Supply, Philip, & Locht, Camille. (1998). Identification of a new DNA region specific for members of Mycobacterium tuberculosis complex. Journal of Clinical Microbiology, 36(4), 937-943.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics, 9(1), 387-402.

Margulies, M., M Egholm, WE Altman, S Attiya, JS Bader, LA Bemben, et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature, 437(7057), 376-380.

Marmiesse, M. (2004). Macro-array and bioinformatic analyses reveal mycobacterial “core” genes, variation in the ESAT-6 gene family and new phylogenetic markers for the Mycobacterium tuberculosis complex. Microbiology, 150(2), 483-496.

Marrero, J., Rhee, K. Y., Schnappinger, D., Pethe, K., & Ehrt, S. (2010). Gluconeogenic carbon flow of tricarboxylic acid cycle intermediates is critical for Mycobacterium tuberculosis to establish and maintain infection. Proceedings of the National Academy of Sciences of the United States of America, 107(21), 9819-9824.

Mazars, Edith, Lesjean, Sarah, Banuls, A.-L., Gilbert, M., Vincent, Véronique, Gicquel, Brigitte, Tibayrenc, M., et al. (2001). High-resolution minisatellite-based typing as a portable approach to global analysis of Mycobacterium tuberculosis molecular epidemiology. Proceedings of the National Academy of Sciences of the United States of America, 98(4), 1901-1906.

McClintock, B. (1965). Components of action of the regular Spm and Ac. Carnegie Institution of Washington Year Book, 64, 527-536.

McDermott, W., & Tompsett, R. (1954). Activation of pyrazinamide and nicotinamide in acidic environments in vitro. American review of tuberculosis, 70(4), 748-754.

McKenna, A., Hanna, Matthew, Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20(9), 1297-1303.

McKinney, J.D., Honer zu Bentrup, K., Munoz-Elias, E.J., Miczak, A., Chen, B., Chan, W.T., et al. (2000) Persistence of Mycobacterium tuberculosis in macrophages and mice requires the glyoxylate shunt enzyme isocitrate lyase. Nature 406: 735–738.

McPherson, J. D. (2009). Next-generation gap. Nature Methods, 6(11), S2-S5.

Medvedev, P., Stanciu, M., & Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. Nature Methods, 6(11 Suppl), S13-S20.

Metzker, M. L. (2005). Emerging technologies in DNA sequencing. Genome research, 15(12), 1767-76.

Metzker, M. L. (2010). Sequencing technologies - the next generation. Nature reviews. Genetics, 11(1), 31-46.

Mostowy, S., Inwald, J., Gordon, S., Martin, C., Warren, R., Kremer, K., Cousins, D., et al. (2005). Revisiting the Evolution of Mycobacterium bovis. (T. Bäck, Ed.)Society, 187(18), 6386-6395.

Motiwala, A. S., Dai, Y., Jones-López, E. C., Hwang, S.-H., Lee, J. S., Cho, S. N., Via, L. E., et al. (2010). Mutations in extensively drug-resistant Mycobacterium tuberculosis that do not code for known drug-resistance mechanisms. The Journal of Infectious Diseases, 201(6), 881-888.

Mulder, N., Rabiu, H., Jamieson, G., & Vuppu, V. (2009). Comparative analysis of microbial genomes to study unique and expanded gene families in Mycobacterium tuberculosis. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases, 9(3), 314-21.

Mustafa, A. S., & Al-Attiyah, R. (2009). Identification of Mycobacterium tuberculosis-specific genomic regions encoding antigens inducing protective cellular immune responses. Indian Journal of Experimental Biology, 47(6), 498-504.

Niemann, S., Rüsch-Gerdes, S., Joloba, M. L., Whalen, C. C., Guwatudde, D., Ellner, J. J., Eisenach, K., et al. (2002). Mycobacterium africanum Subtype II Is Associated with Two Distinct Genotypes and Is a Major Cause of Human Tuberculosis in Kampala, Uganda. Journal of Clinical Microbiology, 40(9), 3398-3405.

Pablos-Mendez, A., Raviglione, M. C., Laszlo, A., Binkin, N., Rieder, H. L., Bustreo, F., Cohn, D. L., et al. (1998). Global surveillance for antituberculosis-drug resistance 1994-1997. New England Journal of Medicine, 338(23), 1641-1649.

Palomino, J., Leão, S., & Ritacco, V. (2007). Tuberculosis 2007-From Basic Science to patient care. (J. Palomino, S. Leão, & V. Ritacco, Eds.)Tuberculosis (pp. 30, 527).

Papavinasasundaram, K. G., Chan, B., Chung, J. H., Colston, M. J., Davis, E. O., & Av-Gay, Y. (2005). Deletion of the Mycobacterium tuberculosis pknH gene confers a higher bacillary load during the chronic phase of infection in BALB/c mice. J Bacteriol, 187(16), 5751-5760.

Peirs, P., Lefevre, P., Boarbi, S., Wang, X. M., Denis, O., Braibant, M., et al. (2005). Mycobacterium tuberculosis with disruption in genes encoding the phosphate binding proteins PstS1 and PstS2 is deficient in phosphate uptake and demonstrates reduced in vivo virulence. Infect Immun, 73(3), 1898-1902.

Pieters, J., & Gatfield, J. (2002). Hijacking the host: survival of pathogenic mycobacteria inside macrophages. Trends in microbiology, 10(3), 142-6.

Poole, R. K., & Hughes, M. N. (2000). New functions for the ancient globin family: bacterial responses to nitric oxide and nitrosative stress. Molecular Microbiology, 36(4), 775-783.

Pop, M., & Salzberg, S. L. (2008). Bioinformatics challenges of new sequencing technology. Trends in genetics : TIG, 24(3), 142-9.

Pym, A. S., Domenech, P, Honoré, N., Song, J., Deretic, V., & Cole, S T. (2001). Regulation of catalase-peroxidase (KatG) expression, isoniazid sensitivity and virulence by furA of Mycobacterium tuberculosis. Molecular Microbiology, 40(4), 879-889.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841-842. Oxford University Press.

Ramakrishnan, L., Federspiel, N. A., & Falkow, S. (2000). Granuloma-specific expression of Mycobacterium virulence proteins from the glycine-rich PE-PGRS family. Science, 288(5470), 1436-1439.

Raynaud, C., Guilhot, C., Rauzier, J., Bordat, Y., Pelicic, V., Manganelli, R., Smith, I., et al. (2002). Phospholipases C are involved in the virulence of Mycobacterium tuberculosis. Molecular Microbiology, 45(1), 203-217.

Rousseau, C., Sirakova, T. D., Dubey, V. S., Bordat, Y., Kolattukudy, P. E., Gicquel, B., & Jackson, M. (2003). Virulence attenuation of two Mas-like polyketide synthase mutants of Mycobacterium tuberculosis. Microbiology, 149(Pt 7), 1837-1847.

Runyon, E. H. (1959). Anonymous mycobacteria in pulmonary disease. The Medical clinics of North America, 43(1), 273-290.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain terminating inhibitors. PrProcNatlAcadSciUSAoceedings of the National Academy of Sciences USA, 74(12), 5463-5467.

Sassetti, C. M., Boyd, D. H., & Rubin, E. J. (2003). Genes required for mycobacterial growth defined by high density mutagenesis. Molecular microbiology, 48(1), 77-84.

Schatz, A., Bugie, E., & Waksman, S. (1944). Streptomycin, a substance exhibiting antibiotic activity against gram-positive and gram-negative bacteria. Proc Soc Exp Biol Med, 55, 66-69.

Schmidt, J. M., Good, R. T., Appleton, B., Sherrard, J., Raymant, G. C., Bogwitz, M. R., Martin, J., et al. (2010). Copy Number Variation and Transposable Elements Feature in Recent, Ongoing Adaptation at the Cyp6g1 Locus. (D. J. Begun, Ed.)PLoS Genetics, 6(6), 11.

Schuster, S.C. (2008). Next-generation sequencing transforms today's biology. Nat Methods 5: 16–18

Shen, Y., Sarin, S., Liu, Y., Hobert, O., & Pe'er, I. (2008). Comparing platforms for C. elegans mutant identification using high-throughput whole-genome sequencing. PloS one, 3(12), e4012.

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D., and Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309, 1728-1732.

Smith, I. (2003). Mycobacterium tuberculosis Pathogenesis and Molecular Determinants of Virulence. Society, 16(3), 463-496.

Sohaskey, C. D. & Wayne, L. G. (2003). Role of narK2X and narGHJI in hypoxic upregulation of nitrate reduction by Mycobacterium tuberculosis. J Bacteriol, 185(24), 7247-7256.

Sreevatsan, S., Stockbauer, K. E., Pan, X., Kreiswirth, B. N., Moghazeh, S. L., Jacobs, W R, Telenti, A., et al. (1997). Ethambutol resistance in Mycobacterium tuberculosis: critical role of embB mutations. Antimicrobial Agents and Chemotherapy, 41(8), 1677-1681.

Steenken, W. J., Wolinsky, E., Pratt, P. C., & Smith, M. M. (1952). Streptomycin in guinea pigs with discrete chronic tuberculous lesions. Am Rev Tuberc, 66(2), 194-212.

Stermann, M., Sedlacek, L., Maass, S., & Bange, F. C. (2004). A promoter mutation causes differential nitrate reductase activity of Mycobacterium tuberculosis and Mycobacterium bovis. J Bacteriol, 186(9), 2856-2861.

Supply, P, Mazars, E, Lesjean, S, Vincent, V, Gicquel, B, & Locht, C. (2000). Variable human minisatellite-like regions in the Mycobacterium tuberculosis genome. Molecular Microbiology, 36(3), 762-771.

Tekaia, F., Gordon, S. V., Garnier, T., Brosch, R., Barrell, B. G., & Cole, S. T. (1999). Analysis of the proteome of Mycobacterium tuberculosis in silico. Tuber Lung Dis, 79(6), 329-342.

Tomkinson, A. E., Vijayakumar, S., Pascal, J. M., & Ellenberger, T. (2006). DNA ligases: structure, reaction mechanism, and function. Chemical Reviews, 106(2), 687-699.

Valouev A., J Ichikawa, T Tonthat, J Stuart, S Ranade, H Peckham, et al. (2008). A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res, 18(7), 1051-1063.

Van Belkum, A., Scherer, S., Van Alphen, L., & Verbrugh, H. (1998). Short-Sequence DNA Repeats in Prokaryotic Genomes. Microbiology and Molecular Biology Reviews, 62(2), 275-293. American Society for Microbiology.

Van Helden, P D, Parsons, S. D. C., & Gey Van Pittius, N. C. (2009). "Emerging" mycobacteria in South Africa. Journal of the South African Veterinary Association, 80(4), 210-214.

Vasconcellos, S. E. G., Huard, R. C., Niemann, S., Kremer, Kristin, Santos, A. R., Suffys, P. N., & Ho, J. L. (2010). Distinct genotypic profiles of the two major clades of Mycobacterium africanum. BMC Infectious Diseases, 10(1), 80.

Ventura, M., Canchaya, C., Tauch, A., Chandra, G., Fitzgerald, G. F., Chater, K. F., & Van Sinderen, D. (2007). Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. Microbiology and molecular biology reviews MMBR, 71(3), 495-548.

Veyrier, F. J., Dufort, A., & Behr, Marcel A. (2011). The rise and fall of the Mycobacterium tuberculosis genome. Trends in Microbiology, 19(4), 156-161.

Victor, T. C., De Haas, P. E. W., Jordaan, A. M., Van Der Spuy, G. D., Richardson, M., Van Soolingen, D., Van Helden, Paul D, et al. (2004). Molecular Characteristics and Global Spread of Mycobacterium tuberculosis with a Western Cape F11 Genotype. Journal of Clinical Microbiology, 42(2), 769-772.

Wang, L. L., Li, Y., & Zhou, S. F. (2009). A bioinformatics approach for the phenotype prediction of nonsynonymous single nucleotide polymorphisms in human cytochromes P450. Drug Metab Dispos, 37(5), 977-991.

Wayne, L. G. (1994). Dormancy of Mycobacterium tuberculosis and latency of disease. European journal of clinical microbiology infectious diseases official publication of the European Society of Clinical Microbiology, 13(11), 908-914.

Wayne, L. G., & Hayes, L. G. (1998). Nitrate reduction as a marker for hypoxic shiftdown of Mycobacterium tuberculosis. Tubercle and lung disease: the official journal of the International Union against Tuberculosis and Lung Disease, 79(2), 127-32.

Wayne, L. G., & Lin, K. Y. (1982). Glyoxylate metabolism and adaptation of Mycobacterium tuberculosis to survival under anaerobic conditions. Infection and Immunity, 37(3), 1042-1049.

World Health Organization (2011). Global tuberculosis control. http://www.who.int/tb/publications/global_report/en/index.html.

World Health Organization (2009). Global tuberculosis control WHO report 2009 197. Tuberculosis, 197-216

World Health Organization (2006). WHO report 2006 Global Tuberculosis control. Geneza WHO.

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research, 18(5), 821-9.

Zhang, C.-C., Friry, A., & Peng, L. (1998). Molecular and genetic analysis of two closely linked genes that encode, respectively, a protein phosphatase 1/2A/2B homolog and a protein kinase homolog in the cyanobacterium Anabaena sp. strain PCC 7120. Journal Of Bacteriology, 180(10), 2616-2622.

Zhang, J. (2003). Evolution by gene duplication: an update. Trends in Ecology & Evolution, 18(6), 292-298.

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), 95-109.

Zhang, Y, Heym, B., Allen, B., Young, D., & Cole, S. (1992). The catalase-peroxidase gene and isoniazid resistance of Mycobacterium tuberculosis. Nature, 358(6387), 591-593.

Zheng, H., Lu, L., Wang, B., Pu, S., Zhang, X., Zhu, G., Shi, W., et al. (2008). Genetic Basis of Virulence Attenuation Revealed by Comparative Genomic Analysis of Mycobacterium tuberculosis Strain H37Ra versus H37Rv. (D. Davis, Ed.)PLoS ONE, 3(6), 12.

Zhu, L., Wang, Q., Tang, P., Araki, H., & Tian, D. (2009). Genome wide association between insertions/deletions and the nucleotide diversity in bacteria. Molecular Biology and Evolution, 26(10), 2353-2361.