Asian J. Adv. Basic Sci.: **2015**, *3(2),* 117-141
ISSN (Print): 2454 – 7492
ISSN (Online): 2347 – 4114
www.ajabs.org

Evaluation of ESX Sequence Variations within *Mycobacterium tuberculosis* Clinical and Laboratory Isolates

Melisha Sukkhu[1,2*] and Yeshnee Naidoo[1,2]

[1] *Department of Infection, Prevention and Control, Nelson R. Mandela School of Medicine, University of KwaZulu-Natal, Durban, SOUTH AFRICA*
[2] *South African Medical Research Council (SAMRC) TB Research Unit: Clinical and Biomedical, Durban, SOUTH AFRICA*
[*] *Correspondence: E-mail: sukkhu@gmail.com*

(Received 24 June, 2015; Accepted 30 June, 2015; Published 02 July, 2015)

ABSTRACT: The ESX family of genes (*esxA-W*) in *Mycobacterium tuberculosis (Mtb*) encodes 23 effector molecules influencing immunogenicity and pathogenicity. This study was aimed at identifying and evaluating variations in ESX sequence profiles in clinical and laboratory isolates, using amplicon sequencing, and examining how diversity might influence immune responses. Using spoligotyping and a strain specific test, the genotypes for all isolates were confirmed. Following amplicon sequencing, all 23 ESX genes were evaluated for variations between the Beijing, KwaZulu-Natal (KZN) and other genetic isolate groupings as well as within isolates. 23 ESX genes from 55 clinical isolates (20 Beijing, 25 KZN and 10 other) and 3 Laboratory strains (H37Rv, H37Ra and BCG) were sequenced. 482 single nucleotide polymorphisms (SNPs) were identified in 12 ESX genes relative to H37Rv. Majority of the identified 363 nsSNPs occured in Beijing isolates. No mutations were observed in *esxA, B, C, E, G, H, J, R, S* and *T*. Six unique nsSNPs were identified in the Beijing isolates: *esxI* (Q20L), *esxO* (E52G), 2 in *esxP* (T3S; N83D), *esxU* (P63S) and *esxW* (T2A). Three unique nsSNPs were identified in the KZN isolates: *esxK* (A58T), *esxL* (R33S) with the *esxL* polymorphism resulting from a dinucleotide change. These results identify new mutations in the ESX family, some of which were not detected by genome sequencing.

Keywords: Amplicon Sequencing; CFP-10; ESX; ESAT-6; Mycobacterium tuberculosis; Sequence Variation.

INTRODUCTION: To date, the tuberculosis disease, caused by *Mycobacterium tuberculosis* (*Mtb*), remains a major global health problem. According to the 2012 and 2013 WHO Global Tuberculosis report, majority of the estimated TB cases occurred in Asia (59%) and Africa (29%). Of the 22 high burden countries, South Africa is the 3rd country with the largest number of reported incident cases (0.4 million-0.6 million). The chronic nature of this disease has also been influenced by selective pressures, namely, the host immune response, changes in human demography, HIV co-infection and the usage of anti-TB drugs (Malik and Godfrey-Fausset, 2005; Gagneux *et. al.*, 2006; Nicol and Wilkinson, 2008; Kato-Maeda *et. al.*, 2001; Comas and Gagneux, 2011; Brites and Gagneux, 2012; WHO, 2012 and 2013).

The *Mtb* genome harbors the ESX gene family, comprising of five gene clusters (*esx-1* to *esx-5*), that are part of the Type VII protein secretion system (T7SS) in *Mtb*. This family of genes (*esxA* to *esxW*) encodes 23 proteins of approximately 100 amino acids (aa). Twenty-two of these genes occur in couples, with only one gene (*esxQ)* occuring in isolation. *esxA* and *esxB* encodes two proteins ESAT-6, and CFP-10, that were the first ESX family members to be identified. Five of the gene couples occur in larger homologous gene clusters or regions named ESX-1 (Rv3866-Rv3883c), ESX-2 (Rv3884c-Rv3895c), ESX-3 (Rv0282-Rv0292), ESX-4 (Rv3444c-Rv3450c) and ESX-5 (Rv1782-1798). Due to the high sequence identity between members, the ESX family can be classified into subfamilies. The TB10.4 subfamily includes *esxH* (from the ESX-3 cluster), while the Mtb9.9 and QILSS subfamilies contain 5 members together with *esxM* and *esxN* (from the ESX-5 cluster) (Tekaia *et. al.*, 1999; Louise *et. al.*, 2001; Skjøt *et. al.*, 2002).

ESX-1 is crucial for the survival and spread of the bacteria *in vivo* (Chapman *et al*., 2002; Novikov *et. al.*, 2011; Manzanillo *et. al.*, 2012; Watson, Manzanillo and Cox, 2012). As a consequence of being located directly adjacent to each other, *esxA* and *esxB* are co-transcribed and secreted from the cell as a heterodimer despite the absence of known secretion signals (Sørensen *et. al.*, 1995; van Pinxteren *et. al*., 2000; Pym *et. al*., 2003, Champion *et. al.*, 2012). A genomic deletion resulted in loss of the ESX-1 system from *M. bovis* BCG (Harboe *et. al.*, 1996; Behr *et. al.*, 1999; Gordon *et. al.*, 1999; Colangeli *et. al*., 2000;

Wards *et. al.,* 2000). However, the secretion of the ESAT-6 and CFP-10 antigens in the context of a recombinant BCG vaccine in the background of a reconstituted RD1 has been shown to restore attenuation and have a beneficial effect on protection against challenge with *Mtb* (Pym *et. al.,* 2003). As of 2012, a novel proteomics approach to evaluate and monitor the ESX-1 protein secretion system was successfully developed by Champion and colleagues (Champion *et. al.,* 2012). They were able to directly detect on agar, ESX-1 protein secretion and the presence and absence of ESAT-6 and CFP-10 from the surface of wild-type and mutant *M. marinum* colonies using whole colony matrix-assisted laser desorption/ionisation-time of flight (MALDI-TOF) mass spectrometry. This was the first report of a proteomics confirmation demonstrating that these two genes are an essential requirement for an intact ESX-1 system in pathogenic bacteria.

Subsequent *in silico* investigations and genetic screens have indicated that the genes surrounding the *esxAB* operon are vital in the secretion of ESAT-6 and CFP-10 (Gey van Pittius *et. al.,* 2001; Pallen *et. al.,* 2002; Hsu *et. al.,* 2003; Sassetti *et. al.,* 2003a; Sassetti *et. al.,* 2003b; Stanely *et. al.,* 2003; Guinn *et. al.,* 2004). Additional studies have shown that several ESX pairs form tight complexes including *esxG* and *esxH* (Rv0287 and Rv0288), *esxR* and *esxS* (Rv3019c and Rv3020c) and *esxO* and *esxP* (Rv2346c and Rv2347c) (Renshaw *et. al.,* 2002; Meher *et. al.,* 2006; Lightbody *et. al.,* 2008; Arbing *et. al.,* 2010). Thus *esxA* and *esxB* are very important proteins of *Mtb* involved in host-pathogen interaction, induction of strong T-cell mediated immune responses (Berthet *et. al.,* 1998) and phagosomal escape in both cell models (Stanely *et. al.,* 2007; de Jonge *et. al.,* 2007; Simeone *et. al.,* 2012) and animal models (Hsu *et. al.,* 2003; Junqueira-Kipnis *et. al.,* 2006; Carlsson *et. al.,* 2010). Studies have found that the substrates secreted by this system play a role in granuloma formation (Ramakrishnan *et. al.,* 1997; Talaat *et. al.,* 1998; Davis *et. al.,* 2002; Prouty *et. al.,* 2003; van der Sar *et. al.,* 2003; Broussard and Ennis, 2007; Smith *et. al.,* 2008; Stoop *et. al.,* 2011), probably through intracellular bacterial spread between macrophages and interference in the phagosomal membrane integrity, leading to "leaky membranes" and bacterial entrance into the cytosol (Stanely *et. al.,* 2003; Lewis *et. al.,* 2003; Gao *et. al.,* 2004; Guinn *et. al.,* 2004; Swaim *et. al.,* 2006; Volkman *et. al.,* 2004; Stanely *et. al.,* 2007; Brodin *et. al.,* 2010; Smith *et. al.,* 2008; Manzanillo *et. al.,* 2012; Simeone *et. al.,* 2012; Watson, Manzanillo and Cox, 2012).

Despite being the most characterized system, seven additional proteins that are part of this system, have been also reported to be secreted by ESX-1. This includes EspA (Rv3616c), EspB (Rv3881c), EspC (Rv3615c), EspE (Rv3864), EspF (Rv3865), PE35 (Rv3872), and EspR (Rv3849) (Fortune *et. al.,* 2005; Abdallah *et. al.,* 2007; McLaughlin *et. al.,* 2007; Xu *et. al.,* 2007; Raghavan *et. al.,* 2008; Bitter *et. al.,* 2009). The ESX-1 system is controlled by multiple regulators including the DNA binding transcription factor EspR (Rv3849), the PhoP two component regulator system and the serine protease MycP1 (Frigui *et. al.,* 2008; Gonzalo-Asensio *et. al.,* 2008; Raghavan *et. al.,* 2008; Ohol *et. al.,* 2010).

Likewise, similar secretion systems have been identified in other *Mtb* species and in a few Gram positive bacteria (Tekaia *et. al.,* 1999; Gey van Pittius *et. al.,* 2001; Pallen, 2002; Finn *et. al.,* 2006).

The ESX-2 system is located adjacent to the ESX-1 system bearing structural similarities to ESX-1 (Gey van Pittius *et. al.,* 2001). The *esxC* (Rv3890c) and *esxD* (Rv3891c) genes are found in this system, however, their functional role is still unknown (Sassetti *et. al.,* 2003a; Sassetti *et. al.,* 2003b). Due to a partial deletion of the membrane-bound component, MMAR_5460, it is believed that the ESX-2 system may be defective in *M. marinum* (Gey van Pittius *et. al.,* 2001; Abdallah *et. al.,* 2006). To date, this secretion system has not been investigated and therefore its function remains unknown.

The ESX-3 system is structurally similiar to the ESX-1 system, but encodes *esxG* and *esxH*. ESX-3 expression is controlled by the zinc uptake regulator Zur/FurB (Maciag *et. al.,* 2007) and regulated by diavalent cation levels (Serafini *et. al.,* 2009). Orthologous systems of ESX-3 are found in all mycobacterial species, whose genomes were analysed so far and ESX-3 is the most conserved, which is compatible with the essential character of the system that appears to play an elementary role in the mycobacterial lifecycle. In preliminary work it was recently suggested that ESX-3 prevents innate immune killing of mycobacteria (Sassetti *et. al.,* 2003a; Sassetti *et. al.,* 2003b). Hence, ESX-3 also represents an interesting potential new drug target that needs to be explored more comprehensively.

Being the smallest gene cluster, phylogenetic investigations have implied that the ESX-4 system is the progenitor system, based on its presence in other actinobacteria. The remaining systems arose with gene duplications, proceeded by divergence events as well as the recruitment and accumulation of other genes (Tekaia *et. al.,* 1999; Gey van Pittius *et. al.,* 2001; Gey

van Pittius *et. al.,* 2006). Whole genome mutagenesis studies have not predicted a requirement of the ESX-4 system for virulence or *in vitro* growth (Sassetti *et. al.,* 2003a; Sassetti *et. al.,* 2003b).

ESX-5 is a poorly understood group of proteins due to the lack of reported studies in *Mtb.* The ESX-5 locus encodes the *esxM* and *esxN* proteins that induce strong CD4+ T-cell responses in human and animal models (Alderson *et al.*, 2000; Jones *et al.,* 2010). Extensive studies have been conducted with *Mycobacterium marinum (Mma)* where the ESX-5 system was shown to be involved in the secretion of proteins from the PE/PPE family and the PPE-MPTR and PE-PGRS subgroups (Abdallah *et. al.,* 2006; 2009; 2011; Cascioferro *et. al.,* 2011; Bottai *et. al.,* 2012). This system is the more recent duplication and the region occurs only in slow growing mycobacteria (Gey van Pittius *et. al.,* 2001; Abdallah *et. al.,* 2006; Gey van Pittius *et. al.*, 2006). There are an additional gene pairs (*esxKL, esxIJ, esxOP* and *esxVW*) that belong to the ESX-5 system and encode a variant QILSS and Mtb9.9 motif, which are required for normal microbial growth, and likely to have a role in bacterial multiplication during active infection (Bukka *et. al.,* 2011). Proteins of the ESX-5 system may play a role in maintaining a fully functional cell envelope and the virulence of *Mtb* (Bottai *et. al.,* 2012).

The availability of the sequenced mycobacterial genome has led to a better understanding into the biology and genomics of the organism (Cole *et. al.,* 1998; Das, Ghosh and Mande, 2011). Several research groups implemented phylogenetic approaches in the analysis of clinical *Mtb* isolates from diverse geographical locations. The main aim was to associate unique Single Nucleotide Polymorphisms (SNPs) as genetic markers in different lineages. Those studies revealed definitive evidence for a clonal population structure for this complex as well as a lack of on going horizontal gene transfer (Filliol *et. al.,* 2006; Gagneux *et. al.,* 2006; Gutacker *et. al.,* 2006; Hershberg *et. al.,* 2008; Comas *et. al.,* 2010; Uplekar *et. al.,* 2011).

Since then, public databases of several mycobacterial sequences have been established making it possible to compare genes and establish whether the genes are conserved to specific species. They have also been successfully implemented in identifying *Mtb* outbreaks, where polymorphic genetic markers have been utilized in the discrimination and subtyping of *Mtb* strains (Brosch, *et. al.,* 2001). Three independent studies were published recently, assessing the genetic diversity of the ESX family members (Uplekar *et. al.,* 2011), their potential for antigenic variation (Comas *et. al.,* 2010) and the evolution and distribution of the ESX family proteins on a genomic and proteomic level (Deng *et. al.,* 2014). Uplekar and colleagues showed that some of the identified mutations did affect known ESX epitopes. All three studies also noted a high genetic variability in the Mtb9.9 and QILSS subfamilies. Whereas, Deng and colleagues showed a change of stability, gain or loss of globular domains and phosphorylation of serine/threonine may be responsible for the difference between the pathogenesis and virulence of the ESX proteins in attenuated and non-pathogenic mycobacteria. (Deng *et. al.,* 2014). It therefore became imperative and vital that further studies are executed in identifying and characterizing these proteins and their secretion systems to identify variability from a wider selection of strains.

Therefore, our study was aimed at investigating the ESX system of *Mtb* to (i) evaluate the sequence diversity by reference to the completed genome sequences and comparison with other publicly available *Mtb* sequences and (ii) analyze and identify potential genetic variations within and between the ESX systems in clinical isolates using amplicon sequencing.

MATERIAL AND METHODS:

Bacterial Culture and Growth Conditions: For this study, 55 Mycobacterial clinical isolates and laboratory strains (Tables 1a and b) were kindly provided by Dr. M. Pillay (Department of Medical Microbiology, UKZN). These isolates were obtained from archived collections in the Department of Medical Microbiology within the last 15 years that were isolated as part of routine Department of Health diagnostics. The strains were used from 3 studies that have ethical approval from BREC to store and molecularly analyze. These studies were: Rapid Detection Study (Ethic Number: E157/04); Department of Health (DOH) surveillance; *M. vaccae* clinical trial and Westville Prison Molecular Epidemiology Study (Ethic Number: H084/00).

Mycobacteria were maintained in either liquid Middlebrook 7H9 medium (Difco) or solid Middlebrook 7H11 agar enriched with Albumin Dextrose Catalase (ADC) (Difco) and 0.05% Tween 80. Cultures were grown at 37˚C, with agitation in the biosafety level three (BSL3) facilities within the Department of Medical Microbiology. Glycerol stocks of the cultures were maintained at -70˚C in 30% (v/v) glycerol.

Genotypic Confirmation of Clinical Isolates: The Research Laboratory team in the Department of Medical Microbiology at the University of KwaZulu-Natal, Nelson Mandela Medical School under the leadership of Dr. M. Pillay performed confirmations of the isolates genotypes. Both the *IS6110* RFLP

method (adapted from van Soolingen *et. al.*, 1994) and the spoligotyping method (Isogen BioSciences, Maarssen, Netherlands) were used.

A strain specific molecular test, developed and performed by Miss N. Pillay in the Department of Medical Microbiology at the University of KwaZulu-Natal as part of her Masters dissertation, was additionally performed to confirm the isolates genotypes. This test was developed to assess the prevalence and distribution of the F15/LAM4/KZN within the province of KZN and used as a screening tool for a province wide drug susceptibility survey (Pillay, 2010).

Table 1(a): Bacterial strains utilized in this study.

Isolate Number	Isolates	RFLP / Strains	Isolated From	Isolation Date
1	R 26	KZN variant 1	Sputum	1995-96
2	R 62	Neither	Sputum	1995-96
3	R 224	Neither	Sputum	1995-96
4	R 226	Neither	Sputum	1995-96
5	R 252	Neither	Sputum	1995-96
6	R 253	KZN variant 1	Sputum	1995-96
7	R 257	KZN variant 1	Sputum	1995-96
8	R 295	KZN variant 1	Sputum	1995-96
9	R 299	Neither	Sputum	1995-96
10	R 300	Neither	Sputum	1995-96
11	R 339	KZN variant 1	Sputum	1995-96
12	R 351	KZN variant 1	Sputum	1995-96
13	R 375	KZN variant 1	Sputum	1995-96
14	R 389	KZN variant 1	Sputum	1995-96
15	R 402	Neither	Sputum	1995-96
16	R 413	KZN variant 1	Sputum	1995-96
17	R 426	KZN Archetype	Sputum	1995-96
18	R 434	KZN variant 1	Sputum	1995-96
19	R 443	F28	Sputum	1995-96
20	R 467	KZN variant 1	Sputum	1995-96
21	R 492	F28	Sputum	1995-96
22	R 502	KZN variant 1	Sputum	1995-96
23	R 503	KZN variant 1	Sputum	1995-96
24	R 504	KZN variant 1	Sputum	1995-96
25	R 506	KZN variant 1	Sputum	1995-96
26	R 525	KZN variant 1	Sputum	1995-96
27	R 576	Neither	Sputum	1995-96
28	R 623	KZN variant 1	Sputum	1995-96
29	H37Rv	Reference Strain	-	-

Table 1(b): Bacterial strains utilized in this study.

Isolate Number	Isolates	RFLP/Strains	Isolated From	Isolation Date
30	H37Ra	Avirulent Strain		
31	910 P4	Beijing	Sputum	1996-01
32	1784 P5	Beijing	Sputum	1996-01
33	1528 P8	Beijing	Sputum	1996-01
34	Vac 1435	KZN variant 1	Sputum	1996-01
35	Vac 4207	KZN variant 1	Sputum	1996-01
36	Vac 4258	KZN variant 1	Sputum	1996-01
37	Vac 2475	KZN variant 1	Sputum	1996-01
38	KZN 605	KZN variant 1	Sputum	1996-01
39	Vac 666	KZN variant 1	Sputum	1996-01
40	Vac 8426	KZN variant 1	Sputum	1996-01
41	43178	Beijing	Sputum	2002
42	WPO6 25210	Beijing	Sputum	2002

43	WPO6 10263	Beijing	Sputum	2002
44	WPO6 2705	Beijing	Sputum	2002
45	39321	Beijing	Sputum	2002
46	43117	Beijing	Sputum	2002
47	48246	Beijing	Sputum	2002
48	WPO6 21386	Beijing	Sputum	2002
49	WPO77021344	Beijing	Sputum	2002
50	WPO77027955	Beijing	Sputum	2002
51	WPO77037345	Beijing	Sputum	2002
52	WPO77075078	Beijing	Sputum	2002
53	WPO77084874	Beijing	Sputum	2002
54	WPO77023338	Beijing	Sputum	2002
55	WPO77041516	Beijing	Sputum	2002
56	WPO62731	Beijing	Sputum	2002
57	WPO77115992	Beijing	Sputum	2002
58	BCG	Reference Strain	-	-

***In Silico* Genome Comparisons of Publically Available Databases:** *In silico* comparisons of the ESX sequences was conducted using the DNAStar MegAlign software (LaserGene 7.2, Madison, Wisconsin, USA). Genome comparisons were employed using partial and completed genome sequences that were downloaded from public websites and databases (TBDB, TubercuList and GenoList) (Table 2). Comparisons were done with the corresponding ESX sequences from the *Mtb* H37Rv reference strain. This was done at the initiation of the study and does not reflect the rapid expansion of genome sequences that have become available in the last year.

Table 2: List of Public Available Websites and Databases for ESX sequences.

Website/Database	Website Adress	Reference
TBDB	http://www.tbdb.org	Reddy, *et. al.*, 2009
TubercuList	http://genolist.pasteur.fr/TubercuList	Lew, *et. al.*, 2011
GenoList	http://genodb.pasteur.fr/cgibin/WebObjects/GenoList	Lechat, *et. al.*, 2008

Genomic DNA Extraction using CTAB/NaCl Method: The genomic DNA extraction was achieved using the CTAB/NaCl method as previously described, (van Soolingen, *et. al.*, 1994). Scraped bacterial colonies were heat inactivated for 30 minutes at 80˚C in 500µl TE buffer (100mM Tris/HCL; 10mM EDTA, pH 8.0), and in turn treated with 50µl lysozyme (10mg/ml). This mixture was incubated for 1 hour at 37˚C, and subsequently treated with 75µl of the Proteinase K-SDS mixture (10mg/ml; 10% v/v). Following brief vortexing and a further 10 minutes incubation at 65˚C, 100µl of pre-warmed CTAB/NaCl (10% *N*-cetyl-*N,N,N,*-trimethyl-ammonium bromide, 0.73M NaCl) was added. The liquid contents were vortexed until the mixture appeared milky and was incubated for 10 minutes at 65˚C. This was followed by an addition of 750µl of chloroform/isoamyl alcohol (24:1, v/v), vortexing for 10 seconds following a 20 minute centrifugation at maximum speed. The nucleic acid was precipitated with the addition of 500µl isopropanol to the supernatant. Upon DNA thread formation, centrifugation was performed at maximum speed for 30 minutes. The DNA pellet was washed twice with 70% cold ethanol to remove residual CTAB/NaCl, and air-dried for 5 minutes at room temperature. The resultant pellet was re-dissolved in the appropriate amount of TE Buffer (100mM Tris/HCL; 10mM EDTA, pH 8.0) and stored at 4˚C. DNA concentration and purity was measured by optical density at 260nm.

PCR of ESX genes: Primers were designed for the 11 ESX gene pairs of interest using the Primer3 program (http://frodo.wi.mit.edu/cgibin/primer3/primer3_www.cgi) (Table 3(a)). In some instances it was necessary to amplify single genes in order to retrieve sufficient single PCR products. Individual ESX gene primers were also designed for *esxA, esxB, esxR, esxS, esxT and esxU* (Table 3b). The PCR reactions for the ESX gene pairs were performed using EconoTaq Buffer and 5U/µl of Econo Taq polymerase (Invitrogen, Saint Aubin, France), 25mM nucleotide mix, 2pM of each primer, 1-10ng of template DNA and nuclease-free water to a final volume of 25µl. For the PCR of the individual ESX genes, the GoTaq Colorless Mastermix (Promega Corp., USA) was applied as stipulated by protocol dictated by the manufacturer.

The thermal cycling was performed in a Bio-Rad T100 Thermal Cycler machine and the GeneAmp PCR system 9700 PCR machine (Applied Biosystems, Foster City, CA, USA) respectively. Primer specificity was confirmed through gel electrophoresis of the amplicons with Sybr safe (Invitrogen, Saint Aubin, France).

Table 3(a): Primers for Sequencing ESX Gene Pairs.

Genes	Primer Sequence (5' – 3')
esxCD	F: ATCGACAGGTCCGCAGAG R: TAGCAGCAAGCAGAAGGTG
esxEF	F: GTGCTGTGTGCTGGTGA R: GAGATCACCGCACCCAAAC
esxGH	F: GACCGCAACCAAAGAAC R: CCAGCACCCACGGAAAG
esxIJ	F: AGTCATAACCTGTCCGCCAC R: TCCCAGTTCAGCACCATCC
esxKL	F: GGCGCAGACTGTCGTTATTT R: AACACCCCAGCACTGACCAC
esxMN	F: AAGGAGAGGGGGAACATCC R: ATCCATCGCTACCTCA
esxOP	F: GGGCGCAGACTGTCATTATT R: CTTAGCGGAGGCACCAGAG
esxQ	F: TTCGATCGAAAGAGTGTCTA R: ACGAACATCGCCGCCAAC
esxVW	F: TTTAACAACTTCGCTGC R: AGTGTTCCCAACGACGAC

Table 3(b): Primers for Sequencing Individual ESX genes.

Genes	Primer Sequence (5' – 3')
esxA	F: AGGCCGGCGTCCAATACT R: TCAGAGTGCGCTCAAACGTA
esxB	F: GGTGAGCTCCCGTAATGACA R: GTGACATTTCCCTGGATTGC
esxR	F: CAAGCCAATTTGGGTGAGG R: CTACCGGATCCACCAACAG
esxS	F: GGGGCCGGATTTGGTCG R: GCACGCTGCAGAGCTTG
esxT	F: GCTCTACCACGTCCTGCAC R: AGCTTACCGACGAGCATCC
esxU	F: GTGGACGATTCTGCTCGAC R: CGGTGGTGTGGATCTCCT

Sequencing of ESX Genes: Upon amplification of PCR products, primers were eliminated by incubating 10µl of the PCR product with 1U of Shrimp Alkaline Phosphatase (SAP) (Fermentas Life Sciences, USA) and 10U of Exonuclease I (ExoI) (Fermentas Life Sciences, USA) for 25 minutes at 37°C followed by 15 minutes at 80°C. To 1µl of this treated reaction mixture, 0.4µl of Big Dye sequencing mix (Applied Biosystems, Foster City, CA, USA), 1.6µl of the 2pM primer and 2µl of the 5x Buffer (5mM $MgCl_2$/200 mM Tris-HCl, pH 8.8) and nuclease-free water were added to a final volume of 10µl. Thermal cycling was performed on this mixture with an initial denaturation step of 1 minute at 96□, followed by 35 cycles of 10sec at 96□, 5sec at 50□ and 4 minutes at 60□. The DNA precipitation reactions were performed in 96-well plates to a final volume of 10µl. To each well, 1µl of EDTA (125mM, pH 8.0) was added, followed by 26µl of a combined mixture of NaOAc (3M, pH 5.2) and 100% ethanol. Following centrifugation at 3000xg for 10 minutes at 18ºC, pre-chilled 35µl of 70% ethanol (v/v) was added to each well and centrifuged at 3000 x g for 5 minutes. The plate was dried in a thermal cycler at 50°C for 5 minutes. The reactions were dissolved in 10µl of Hi-Di Formamide (Applied Biosystems, Foster City, CA, USA) denatured in a thermal cycler (95°C for 3 minutes, 4°C for 3 minutes) and subjected to automated sequencing on an ABI Prism 310 genetic sequence analyzer (Applied Biosystems, Foster City, CA, USA). Resultant gene sequences were subjected to comparison and alignment with the DNAStar SeqMan sequence assembler and MegaAlign software (LaserGene 7.2, Madison, Wisconsin, USA).

SNP Detection: The 23 ESX sequences for the H37Rv reference strain were downloaded from the TubercuList database (Lew *et. al.*, 2011) (http://tuberculist.epfl.ch). Using the DNAStar MegAlign software (LaserGene 7.2, Madison, Wisconsin, USA), these sequences were aligned and a phylogenetic tree was complied, using the Geneious software (Drummond *et. al.*, 2012), to survey the clustering patterns of the sequences.

The ESX sequences from the 58 clinical isolates were compared to the corresponding sequences of *Mtb* H37Rv reference strain. Using the BLAST function available on the Tuberculist website (http://tuberculist.epfl.ch), positions of the variant nucleotides were documented as SNPs. The SNPs were further characterized by comparing the amino acid resulting from the substitution with the reference amino acid from H37Rv, into the sSNPs, no change in amino acid, and non-synonymous nsSNPs, resulting in a change in the amino acid mutations.

Detection of Selection: The dN/dS ratio allows for the measurement of the type of selection occurring on codon alignments. However, the number of SNPs occurring in the individual ESX genes was too low for inclusion in this ratio. Therefore, for each isolate, the 23 ESX genes were concatenated to generate single sequences, and used in the subsequent analyses. dS

and dN are the numbers of synonymous and nonsynonymous substitutions per site, respectively. The variance of the difference was computed using the bootstrap method (500 replicates). Analyses were conducted using the Nei-Gojobori method (Nei and Gojobori, 1986). The analysis involved the 58 nucleotide concatenated sequences. All positions containing gaps and missing data were eliminated. There were a total of 880 positions in the final dataset. Evolutionary codon-based analyses were conducted in MEGA5 (Tamura, *et. al.,* 2011). The program estimates the number of synonymous mutations per synonymous site (dS) and the number of nonsynonymous mutations per nonsynonymous site (dN) as well as the variances of the estimates. This estimate was then used in testing the null hypothesis that the genes are undergoing neutral (H_o: dN=dS), purifying (H_o: dN<dS) or positive (H_o: dN>dS) selection.

RESULTS AND DISCUSSION:

Survey of ESX Diversity of publically available genomes: An initial review of available sequenced genomes on public databases revealed the ESX sequences are highly conserved amongst the genomes surveyed (Table 4). This exercise was conducted at a time when a sizeable number of uploaded genome sequences were incomplete, due to the lack of closure at the time of data retrieval. As a result, this is not a complete assessment of ESX gene diversity in the public available databases relative to the H37Rv ESX sequences. However, this preliminary survey indicated that not all annotated members of the ESX gene family were potentially transcriptionally active due to deletion, frameshift or truncation. This is consistent with an earlier study cataloguing indels (Marmiesse *et. al.,* 2004) and confirms that some ESX members are not essential for human infection.

Genotypic confirmation of Clinical Isolates: In this study, a total of 55 clinical isolates and 3 laboratory strains were included to investigate the ESX genetic diversity. These strains were carefully selected to represent two important groups within KwaZulu-Natal. The first was the Beijing lineage that represents approximately 20 percent of all circulating strains in KwaZulu-Natal and the second a group of closely related strains known as the F15/LAM4/KZN lineage that was associated with a large outbreak of drug resistant tuberculosis (Pillay and Sturm, 2007). We reasoned that successful strains, implied by their ecological abundance, would be involved in ongoing cycles of transmission in the local population and would therefore be most likely to have been under the influence of immune selection.

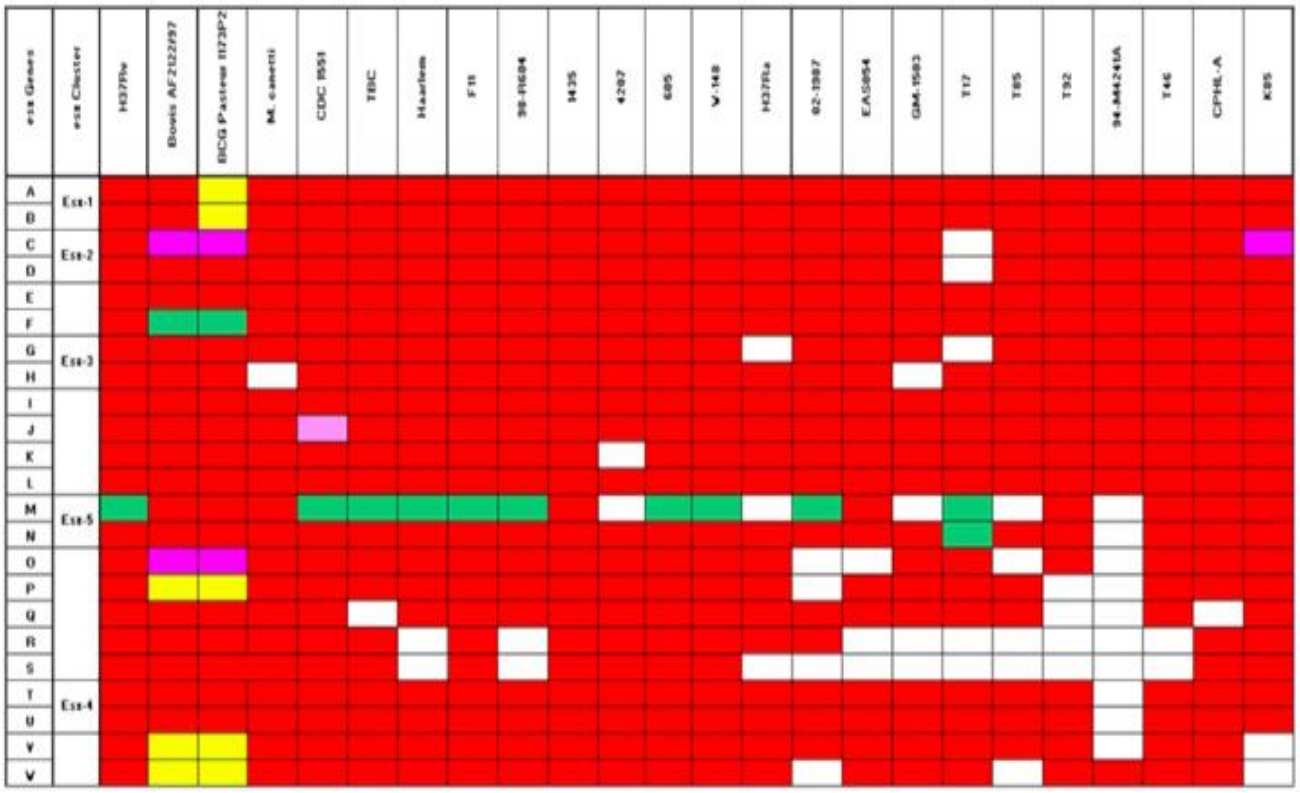
The clinical isolates originated from Tugela Ferry; Rapid Detection Study *M. vaccae* clinical trial and Westville Prison Molecular Epidemiology Study, conducted at the department of Infection, Prevention and Control (UKZN) to follow the transmission of *Mtb* strains in host populations. The IS6110 RFLP typing method, the gold standard of typing methods for *Mtb* (van Embden *et. al*., 1993), was used to confirm the genotypes of the clinical isolates and laboratory strains and dendograms were generated on the band placement for the isolates (Figures 1-3). Hawkey and colleagues reported on the difficulty of this method in sizing up the DNA fragments, since one cannot distinguish different sized fragments appropriately. In addition, software used in the analysis may not necessarily be accurate in the placement of the bands at the correct positions for the strain comparisons (Hawkey *et. al.,* 2003). With this typing method, all the Bejing isolates displayed the expected Beijing RFLP pattern (Figure 1). Gagneux and colleagues screened unique large sequence polymorphisms (LSPs) in the Beijing strains and reported those strains to be a monophyletic clade based on the RD105 deletion (Gagneux *et. al*., 2006), and these strains were confirmed to have RD105 deletions.

The isolates grouped as KZN (Figure 2) displayed the ST60 spoligotype signature listed on the online spoligotype database (Pillay and Sturm, 2007), except for BCG that had two unique bands that was not in agreement with the KZN patterns. In addition, the remaining 8 clinical isolates (Figure 3) lacked any signatures similar to a Beijing or KZN pattern, and were denoted as the "Other" grouping. The pattern observed in Figure 2 is a typical pattern unique to the F15 family that forms part of the Latino-American and Mediterranean family (LAM) and corresponds to the LAM4 subgroup, thus being named the F15/LAM4/KZN strain (Pillay and Sturm, 2007).

An additional confirmation of the isolates genotypes was performed. In Figure 4a, a representation of the isolates were selected and tested positive for the *helZ* and *fadE22* deletions, indicated by a smaller amplicon size for a KZN strain relative to the non-KZN H37Rv strain. These two deletions are unique to the KZN strains (Pillay, 2010). The isolates were then subjected to a multiplex spoligotype PCR assay to verify whether those positive results were in fact true positives for the deletions detected. With this test (Figure 4b), the presence of two bands confirms the positive test for KZN strains, whereas the Beijing isolates displayed a single band that was larger than 1031 bp. However, isolate 8426 displayed two bands in Figure 4a, and was not concordant to the spoligotype pattern (Figure 2) for a KZN strain. However, after being

subjected to the multiplex spoligotype PCR assay (Figure 4b), it was confirmed to be a KZN genotype. The characteristic pattern for the KZN strain shows that the strain has the spacer 20 but is missing the spacers 21-24 and spacer 40. Any deviation from this pattern indicated that the isolate was not of the KZN strain (Pillay, 2010). The remaining isolates were tested using both assays and their genotypes were successfully confirmed and correlated to the spoligotype patterns (results not shown).

Table 4: Survey of ESX genes in sequenced genomes available on public databases representing presence (red), deletion (yellow), stop codon (green), frame shift or truncation (purple) and unavailable sequence (white).



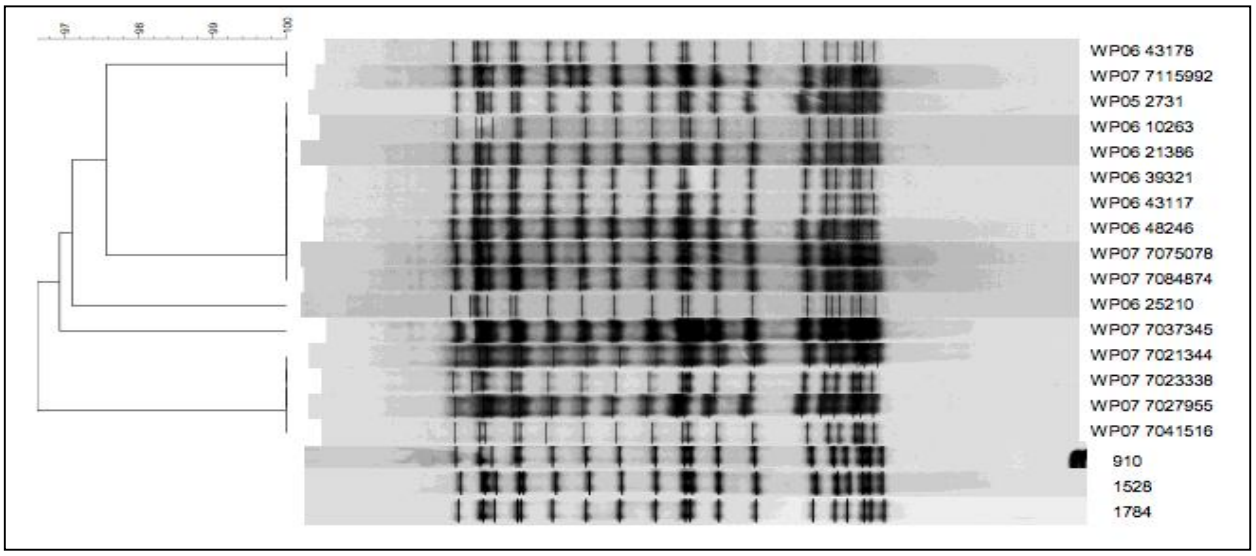


Figure 1: IS6110 RFLP pattern generated for clinical isolates corresponding to the Beijing genotype.

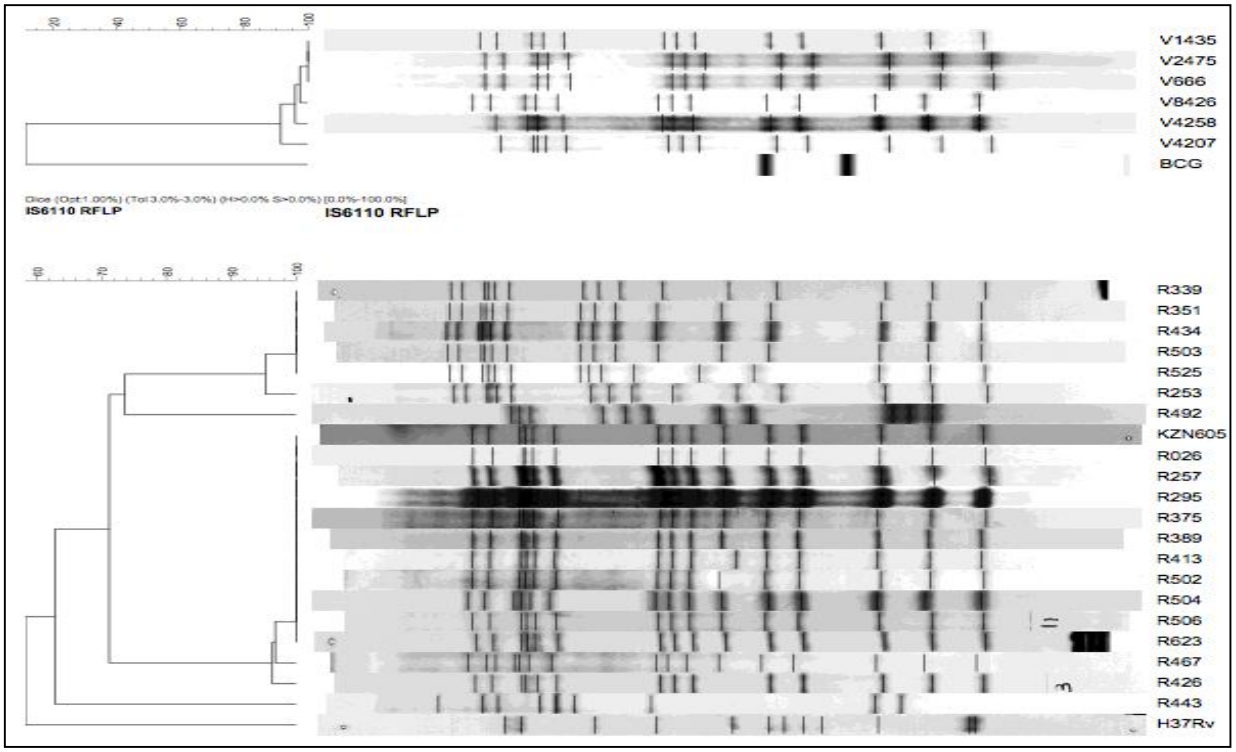


Figure 2: IS6110 RFLP pattern generated for clinical isolates corresponding to the F15/LAM4/KZN strain genotype.

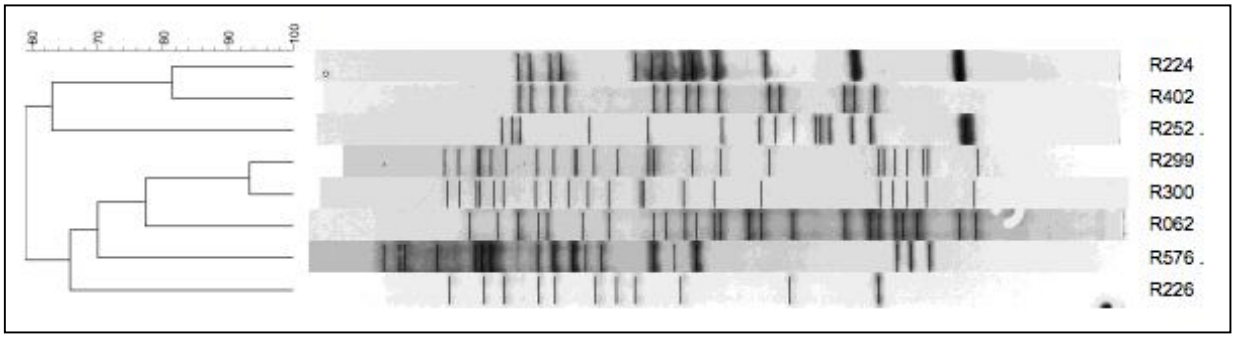
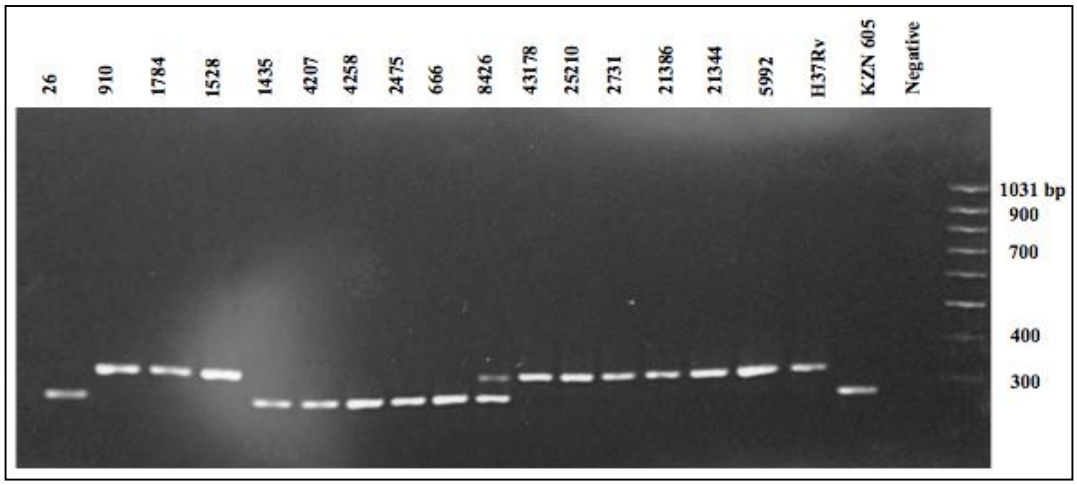
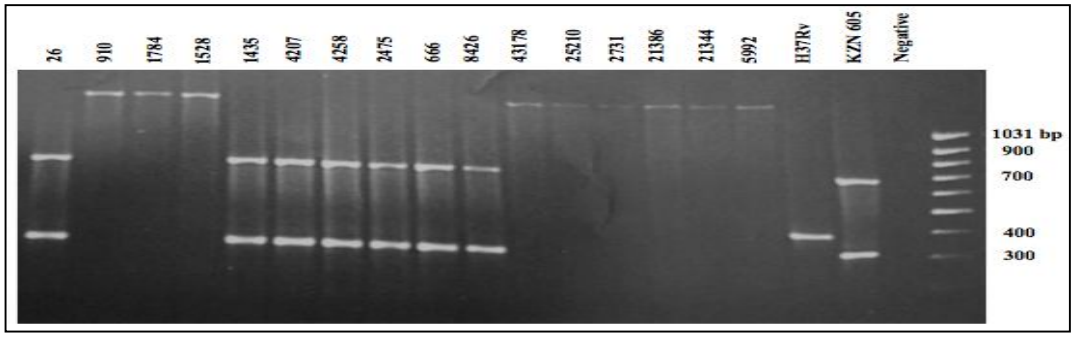


Figure 3: IS6110 RFLP pattern generated for clinical isolates corresponding to neither Beijing or F15/LAM4/KZN strain genotypes.



(a)



(b)

Figure 4: Agarose gel of amplified DNA for clinical isolates using a strain specific PCR that selectively amplifies specific DR region spacer sequences. Agarose gel (a) shows the amplified DNA for KZN and non-KZN strains. KZN strains have a band close to 266 bp that is similar to the KZN 605 positive control. Agarose gel (b) is the verification assay for false positive strains. KZN strains have a similar amplification pattern to the KZN 605 positive control.

PCR Amplification of ESX genes: The PCR primers for the ESX genes were initially optimised with the H37Rv and BCG genomic DNA (Fig 5). PCR amplification was successful as each gene pair was amplified at the expected band size (Table 5). The lack of a PCR product for *esxA*, *esxB*, *esxOP* and *esxVW* in BCG was expected, as it is well documented that these genes are absent in the organism (Maheiras *et. al.*, 1996). This result is indicative of the specificity in the primer design for the amplification of the genes of interest. As a result, these primer sets were applied to amplify the ESX genes in genomic DNA extracted from the 55 clinical isolates (results not shown) and used in subsequent sequencing reactions.

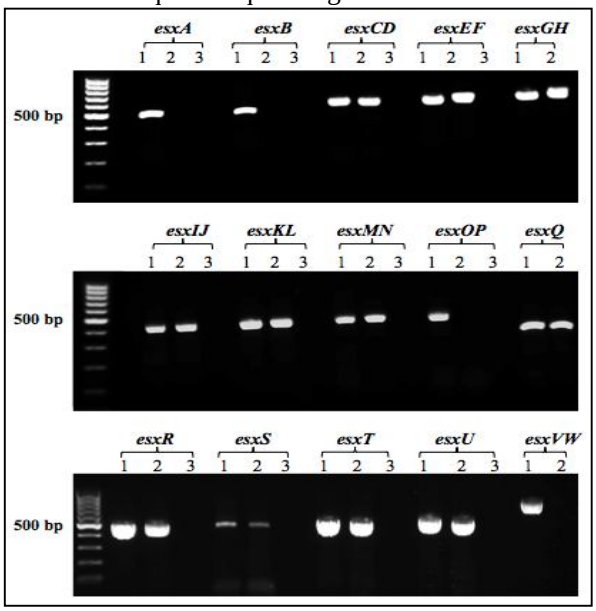


Figure 5: PCR amplification of ESX genes and gene pairs on a 1.5% (w/v) agarose gel with H37Rv (1) and BCG (2) genomic DNA and no DNA control (3). A 100bp molecular weight marker was run on the gel to verify amplified products.

Table 5: Expected PCR product sizes for ESX gene pairs.

ESX Genes	Expected Molecular Weight in bp (base pairs)
esxA	510 bp
esxB	525 bp
esxCD	610 bp
esxEF	776 bp
esxGH	850 bp
esxIJ	644 bp
esxKL	670 bp
esxMN	725 bp
esxOP	764 bp
esxQ	551 bp
esxR	500 bp
esxS	584 bp
esxT	510 bp
esxU	520 bp
esxVW	625 bp

Clustering of the 23 ESX Sequences from H37Rv Reference Strain: A phylogenetic tree was constructed using the Geneious software (Drummond *et. al.*, 2012) from the H37Rv ESX sequences (Figure 6). The three ESX subfamilies are represented within the tree and cluster into distinctive clades as a consequence of their high sequence identity between members. However, within the Mtb9.9 subfamily, the *esxI* and *esxV* sequences lack any SNPs thus rendering those sequences to be identical. This was further confirmed upon alignment of both the nucleotide and protein sequences (Figure 7). However, these genes could be amplified independently because of unique flanking sequences.

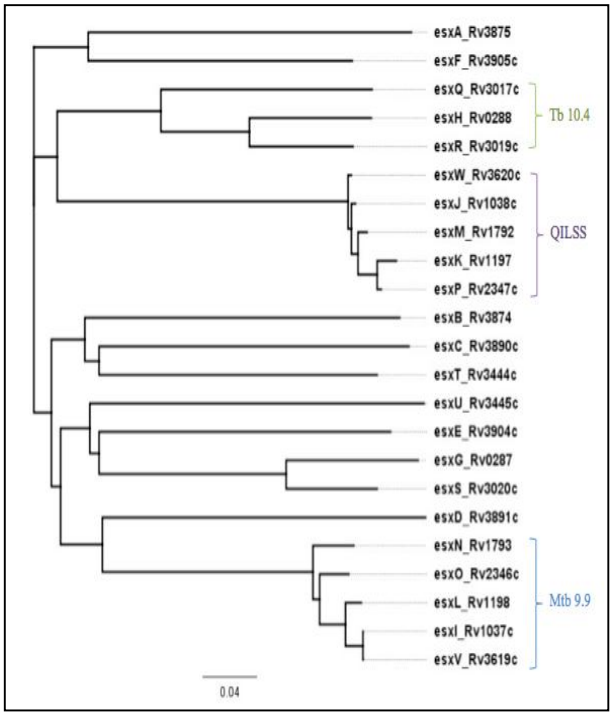


Figure 6: Phylogenetic tree representing the clustering of all 23 ESX members and subfamilies within H7Rv using the Geneious software (Drummond *et. al.*, 2012).

ESX Sequence Diversity within Clinical Isolates: Comparative sequence analysis was conducted by screening all the ESX sequences for variation using the MegAlign software, with the H37Rv ESX sequences as a reference. All 23 ESX genes were successfully sequenced in the 55 clinical isolates as well as the 3 laboratory strains (H37Rv, H37Ra and *M. bovis BCG*). Polymorphisms relative to H37Rv were located in 12 of the 23 ESX genes (Figure 8, Tables 6 and 7) with the majority of the polymorphisms occurring in the Beijing isolates (Figure 9). The number of nsSNPs was three and a half times more than the number of sSNPs for the Beijing isolates (Figure 10). Mestre and colleagues analyzed polymorphisms in the DNA repair, replication and recombination (3R) genes from a collection of 305 Beijing isolates. Interestingly, they reported that the number of nsSNPs was twice the number of the sSNPs from their Beijing dataset (Mestre *et. al.*, 2011).

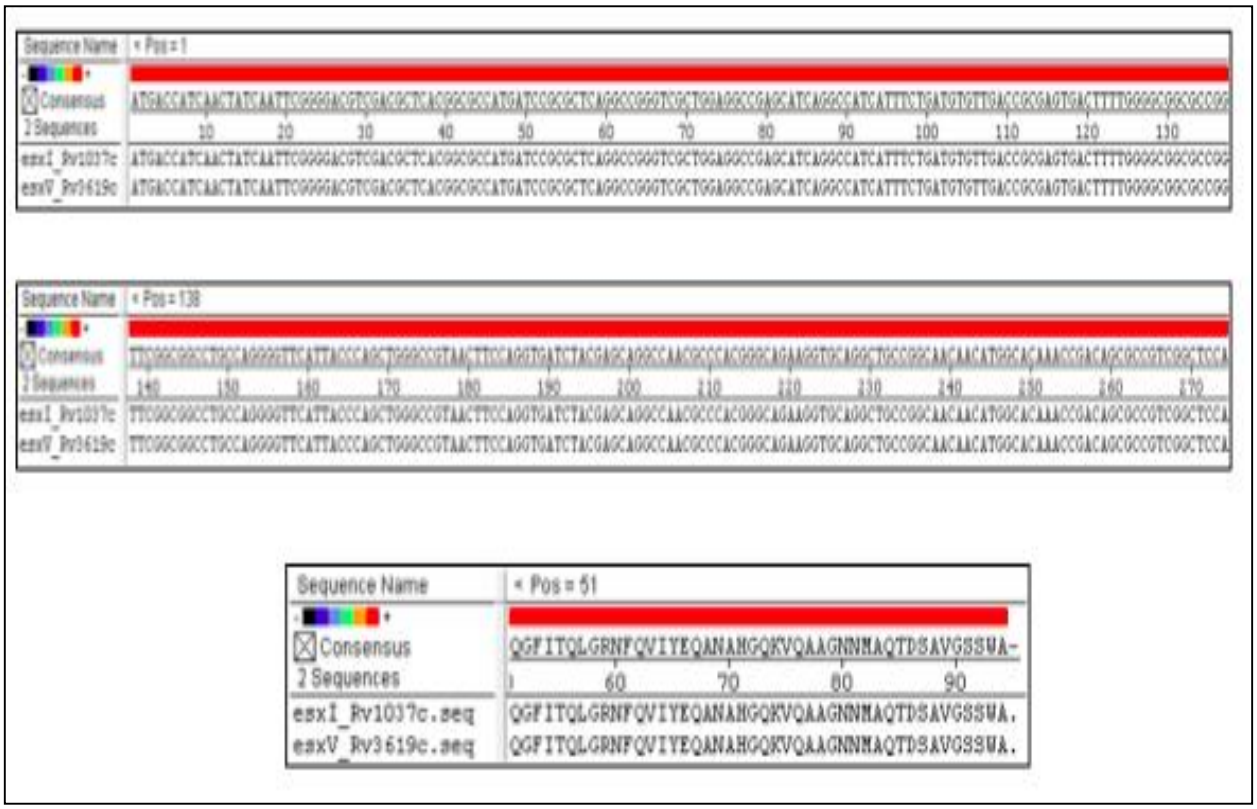


Figure 7: Nucleotide and protein sequence alignment of *esxI* and *esxV* sequences from H37Rv using the DNAStar MegAlign software.

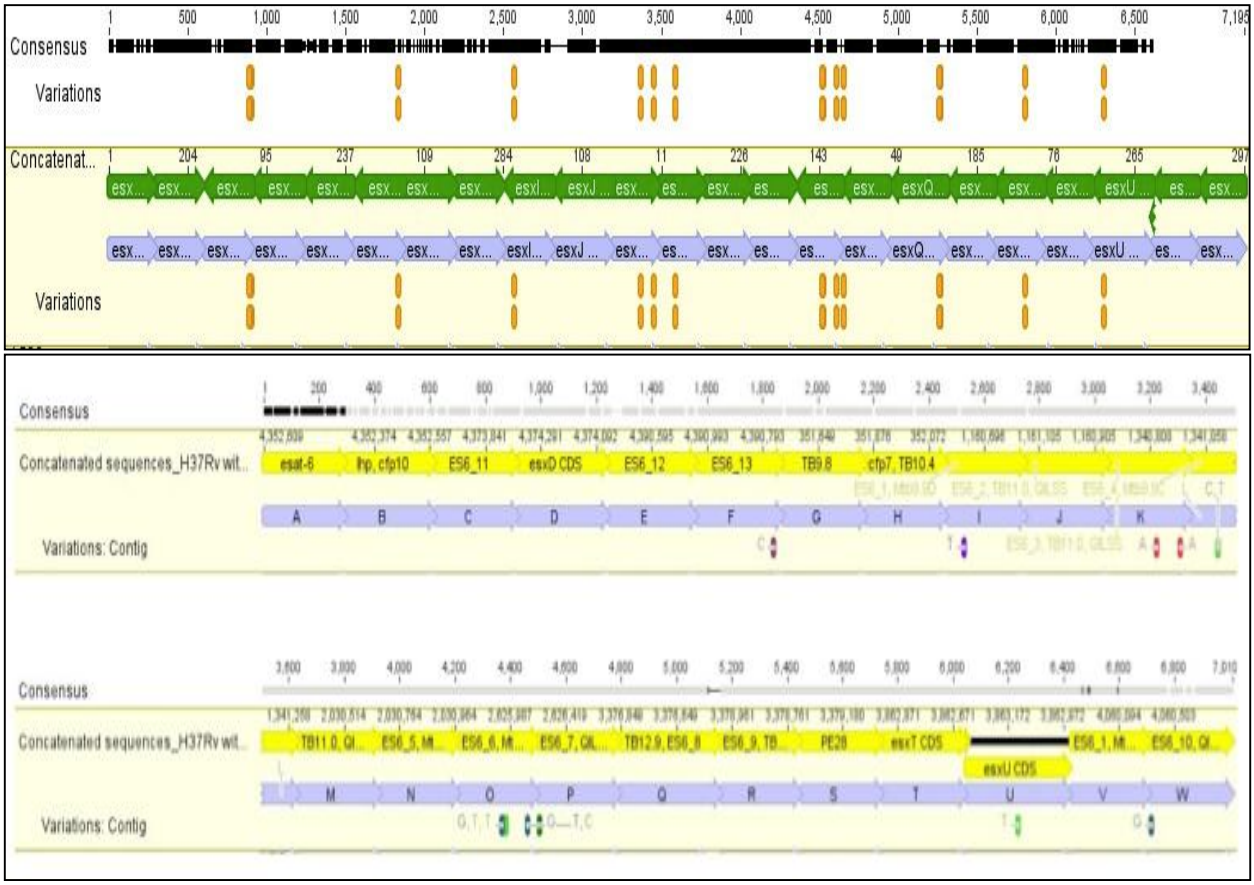


Figure 8: Snapshot of clustal alignment of concatenated sequences using Geneious software of all the SNPs identified in the 23 ESX genes sequences for the 55 clinical isolates and 3 laboratory strains (blue arrows). The sequences were aligned to the concatenated H37Rv reference sequence (yellow and green arrows with coding sequence annotations) and aligned using the ClustalW software available on the Geneious package (Drummond *et. al.*, 2012).

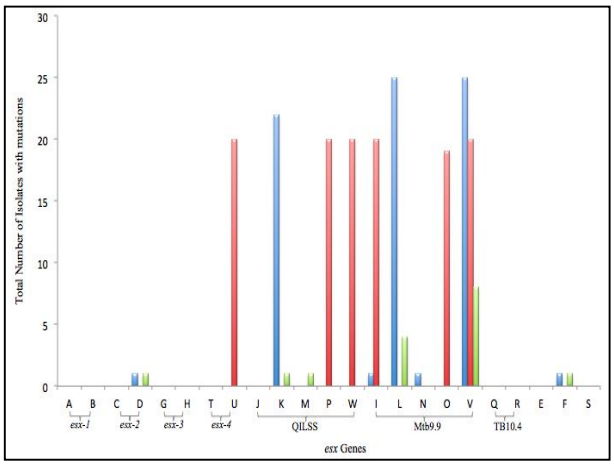


Figure 9: Distribution of SNPs relative to H37Rv for the 23 ESX genes and subfamilies seen across the three isolate groupings. Blue bars: KZN Isolates, Red bars: Beijing Isolates and Green bars: Other Isolate groupings.

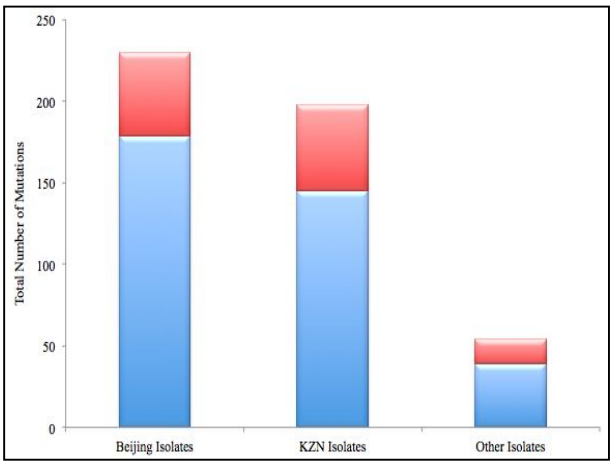


Figure 10: Total numbers of Non-Synonymous (blue) and Synonymous (red) SNPs across the three isolate groupings: 20 Beijing Isolates; 25 KZN Isolates and 13 Other Isolates.

A total of 482 SNPs were identified, of which 363 were nsSNPs and 119 were sSNPS. A closer inspection of the nsSNPs distribution amongst the isolate groupings revealed 179 occurred in the Beijing isolates with 145 in the KZN isolates and 39 in the Other isolate grouping (Figure 10). Similarly, the distribution of the sSNPs amongst the isolate groupings revealed 51 occurred in the Beijing isolates with 53 in the KZN isolates and 15 in the Other isolate grouping (Figure 3.10). No mutations occurred within 10 ESX genes (*esxA, B, C, E, G, H, J, R, S* and *T*). Similarly a study conducted by Musser and colleagues involving the sequencing of 24 *Mtb* antigens from 16 clinical isolates resulted in no variation for *esxA* and *esxB* (Musser *et. al*., 2000). These findings are also in agreement with the 2010 study conducted by Davila and colleagues where no sequence variation was observed for the *esxA* and *esxH* genes sequenced from 88 clinical isolates (Davila *et. al*., 2010). Despite the report of SNPs in *esxH* (Comas *et. al.*, 2010 and Uplekar *et. al.*, 2011, Deng *et. al*., 2014) and *esxR, G* and *S* (Deng *et. al.*, 2014), this was not the case in our data. Based on our data, the lack of SNPs in *esxA, B, C, E, G, H, J, R, S* and *T* genes suggests that these gene regions are highly conserved among different genetic groups of *Mtb* analyzed in this study.

Overall, the clinical isolates harboured more than one non-synonymous mutation in *esxI, K, L, O, P, V* and *W* relative to H37Rv (Table 8). Four SNPs associated with the KZN strain grouping, with 3 being nsSNPs and the other a sSNP, respectively. Of the 3 nsSNPs (Table 8), one occurred in *esxK* (A58T) and 2 in *esxL* (R33S) in 22 KZN isolates, whereas the sSNP occurred in *esxK* (E86) for 17 KZN isolates (Table 9). Similarly, 9 unique SNPs were identified only in the Beijing isolates, of which 6 were nsSNPs (Table 8) and 3 were sSNPs (Table 9). The nsSNPs included *esxI* (Q20L), *esxO* (E52G), 2 in *esxP* (T3S; N83D), *exU* (P63S) and *esxW* (T2A). Of the 3 remaining sSNPs, 2 occurred in *esxO* (I54; L57) and 1 in *esxP* (A2), respectively.

5 SNPs occurred in isolates belonging to the 3 different lineages of the data set. Four SNPs occurred in *esxV,* which included 3 nsSNPs (Q20L, S23L and A57V) and 1 sSNP (H13), while the remaining SNP, being synonymous, occurred in *esxF* (S93) (Tables 8 and 9). Interestingly, the SNPs identified for *esxV,* occurred in almost all the isolates for the dataset. The Q20L, S23L nsSNPs identified in our dataset also confirmed a similar finding in the dataset of Uplekar and colleagues (Uplekar *et. al*., 2011).

Interestingly, two nsSNPs, one in *esxF* (W58stop) and one in *esxD* (T49A), occurred only in a *M. bovis BCG* laboratory strain and the KZN 8426 isolate. Upon comparison to the genome sequences of *M. bovis* BCG Pasteur and *M. bovis* AF2122/97, these particular mutations were confirmed to be *M. bovis* specific. Nonetheless, isolate 8426 was confirmed to be a KZN strain by spoligotype (Figure 2) and the genotypic test (Figure 4b) and no further mutations within the ESX family for this isolate was recognised to be similar to *M. bovis* BCG. Although the PCR and sequencing was replicated and repeated independently, the mutation for this isolate remained unchanged.

Additionally, 5 other SNPs in the BCG laboratory strain were validated as *M. bovis* specific. These included 2 nsSNPs in *esxM* (M48Q) and 3 sSNPs, of which 2 occurred in *esxK* (S3 and N39) and 1 in *esxM* (G47), respectively. The *esxI* gene had 2 nsSNPs (L55F and A88G) for isolate 426, which was not observed in the other isolates. Similarly, isolate 389 showed a single sSNP in *esxN* (V90) that was not present in the other isolates. Both isolates 389 and 429 were confirmed to represent the KZN grouping in our dataset.

SNPs encoding stop codons have substantial impacts on the structure and functionality of proteins. *esxM* is the only other gene that possesses a stop codon (stop59Q), which occurred in the *BCG* laboratory strain and was not present in any of the other strains. This result was contrary to the findings by Uplekar and colleagues in which this mutation was accounted for in 9 clinical isolates (Uplekar *et. al*., 2011).

Overall the 55 clinical isolates used in our study were approximately half the 108 isolates used in the previous study conducted by Uplekar and colleagues. Taking into account this difference the absolute number of SNPs identified in this study is compatible with the numbers identified in the study of Uplekar and colleagues (Uplekar *et. al*., 2011).

SNP Diversity across ESX gene sub-families: In Figure 9, isolates harbouring SNPs were grouped according to the ESX gene subfamilies to establish the SNP distribution and diversity across them. Overall, genes encoded within ESX-1 to ESX-4 presented with low levels of SNP variation except for *esxD* and *esxU*. Majority of the SNPs occur in genes constituting the ESX-5, Mtb9.9 and QILSS subfamilies (Figure 9). Intriguingly, the co-occurrence of 2 nsSNPs was noticed in the neighbouring positions 97 and 98 of *esxL* (Tables 6 and 8). This particular trait was also observed in *esxM* of the BCG laboratory strain containing a single sSNP in codon 47 followed by 2 nsSNPs in codon 48 (Tables 6 to 9). In our dataset, the SNPs reported for *esxV* are highly prevalent in over 90% of the clinical isolates. The remainder of nucleotide sequences for the ESX-5, Mtb9.9 and QILSS subfami-

lies were identical upon comparison to the *Mtb* H37Rv sequences.

ESX Phylogenetic and Evolutionary Analysis of the Clinical Isolates: The information from the spoligotype patterns and the genotypic confirmation assay were inspected closely to correlate to lineage specificity based on the SNPs identified in the sequence dataset (Tables 6 to 9). As expected from the topology of the phylogenetic tree, two distinct clades were observed in Figure 2.11, based on the concatenated sequences containing all the SNPs identified in this study.

Most mutations can decrease the function of the protein and can eventually be eliminated from the population through a negative or purifying selection. Nevertheless, in other instances, strains with a higher mutation rate may have a selective advantage under certain conditions (Mestre *et. al.,* 2011). In extremely unusual circumstances, a mutation can be beneficial and is as a consequence fixed into the population by positive or diversifying selection (Filliol *et. al.,* 2006; Gutacker *et. al*., 2006). To infer the type of evolution occurring in this data set, the dN/dS ratio was used to measure the type of selection in operation on the ESX genes. In order to test the hypothesis that positive selection was in operation as these sequences diverged, the codon-based Z test was used on all 58 concatenated codon alignments. This yielded a value of 0.328. It was noted that our dataset harboured majority of the SNPs in the Mtb9.9 and QILSS genes (Figure 9), with sSNPs occurring in a larger proportion in the clinical isolates compared to the nsSNP (Figure 10), thus influencing the low dN/dS value. The analysis was repeated, but excluded the Mtb9.9 and QILSS subfamily of genes resulting in a value of 1.307 (dN/dS >1) indicating positive selection for these genes. Our findings replicate the results reported by Uplekar and colleagues. (Uplekar *et. al.,* 2011). A recent publication by Comas and colleagues investigated the antigenic variation and diversity of human T-cell epitopes amongst 21 strains representative of the 6 major lineages of the MTBC. Using the Illumina sequencer, whole genome sequences were generated from the 21 strains from which more than 9000 unique SNPs were identified. Their analysis of the sequences revealed that the T-cell epitopes showed little sequence variation and estimated low dN/dS ratio values for changes observed in essential genes to that of the non-essential genes (Comas *et. al.,* 2010).

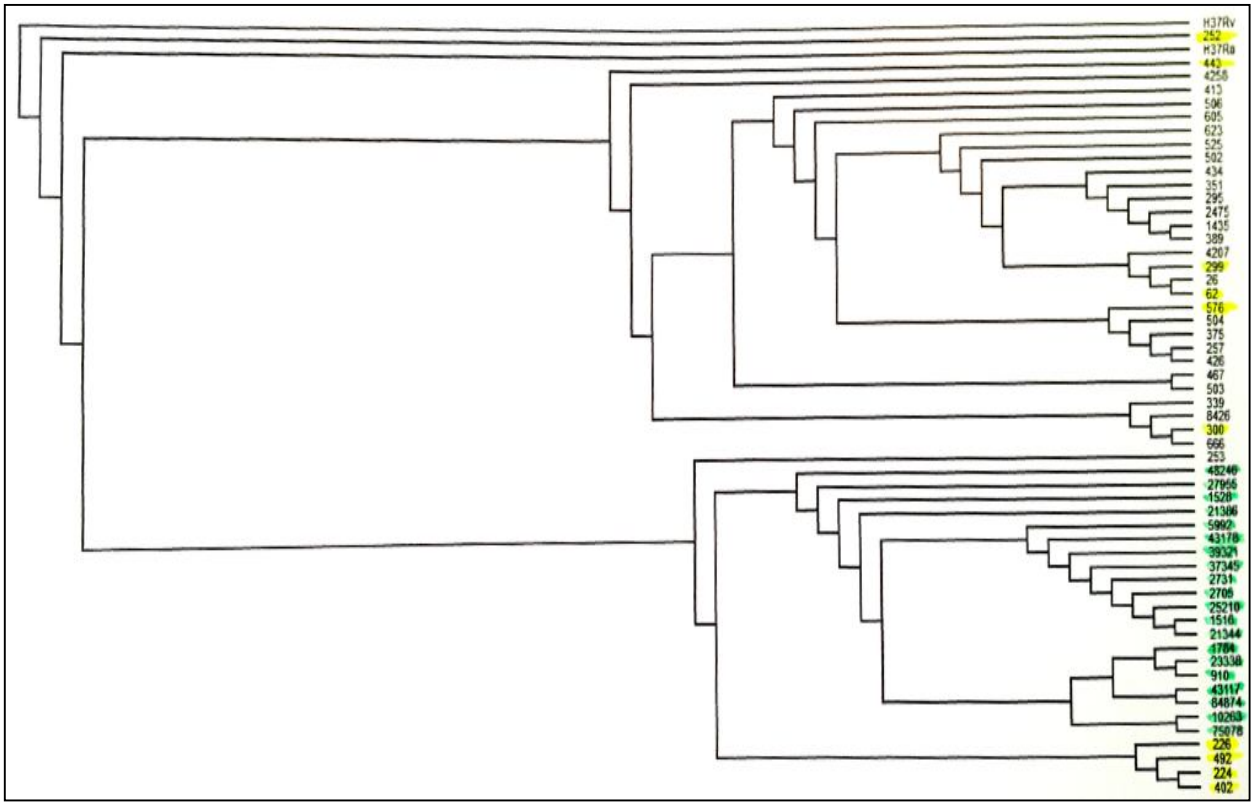


Figure 11: Phylogenetic tree of concatenated sequences using the DNAStar TreeView software rooted to H37Rv. Highlighted in Green: the Beijing isolates, Yellow: the Other Isolate grouping, and unhighlighted sequences: the KZN grouping of isolates. The concatenated BCG sequences was not included as part of the phylogenetic tree.

Table 6: Table of Non-Synonymous Mutations Identified from the 58 sequenced Clinical Isolates.

			Number of Non-Synonymous Mutations		Non-Synonymous Mutations		
esx Cluster:	esx Genes:	Total Clinical Isolates Sequenced	Ref. Strain (H37Rv)	Clinical Isolates	Isolates:	Mutation:	Number of Isolates Mutation Occurs in:
esx-1	A	58	-	-	-	-	-
	B	58	-	-	-	-	-
esx-2	C	58	-	-	-	-	-
	D	58	-	1	40; 58	T49A (a145g)	2
	E	58	-	-	-	-	-
	F	58	-	1	40; 58	W58Stop (g173a)	2
esx-3	G	58	-	-	-	-	-
	H	58	-	-	-	-	-
	I	58	-	1	31-33; 41-58	Q20L (a59t)	21
	J	58	-	-	-	-	-
	K	58	-	1	32	A58P (g172c)	1
				1	2; 4; 8-14; 16; 20; 22-28; 34; 36; 38-40	A58T (g172a)	25
	L	58	-	1	4	L23S (t68c)	1
				2	2; 8-14; 16; 18-20; 22-28; 30; 34; 38-40	R33S (c97t; g98c)	24
esx-5	M	58	-	-	-	-	-
	N	58	-	-	-	-	-
	O	58	-	1	31-33; 41-57	E52G (a155g)	20
	P	58	-	1	31-33; 41-57	T3S (a7t)	20
	Q	58	-	-	-	-	-
	R	58	-	-	-	-	-
	S	58		-	-	-	-
esx-4	T	58	-	-	-	-	-
	U	58	-	1	31-33; 41-57	P63S (c187t)	20
	V	58	-	1	1-28; 31-57	Q20L (a59t)	56
				1	1-28; 31-57	S23L (c68t)	56
				1	1-28; 31-57	A57V(c170t)	56
	W	58	-	1	31-33; 41-57	T2A (a4g)	20

Table 7: Table of Synonymous Mutations Identified from the 58 sequenced Clinical Isolates.

		Number of Synonymous Mutations			Synonymous Mutations		
esx Cluster:	*esx* Genes	Total Clinical Isolates Sequenced	Ref. Strain (H37Rv)	Clinical Isolates	Isolates:	Mutation:	Number of Isolates Mutation Occurs in:
esx-1	A	58	-	-	-	-	-
	B	58	-	-	-	-	-
esx-2	C	58	-	-	-	-	-
	D	58	-	-	-	-	-
	E	58	-	-	-	-	-
	F	58	-	1	1-2; 8-14; 16-19; 22; 26; 34-39	S93S (g279c)	21
esx-3	G	58	-	-	-	-	-
	H	58	-	-	-	-	-
	I	58	-	-	-	-	-
	J	58	-	-	-	-	-
	K	58	-	1	3; 21; 58	S3S (a9g)	3
				1	21	H10H (c30t)	1
				1	58	N39N (c117t)	1
				3	4; 21	I40I (t120c); S41S (c123g); G42G (t126c)	2
	L	58	-	-	-	-	-
esx-5	M	58	-	-	-	-	-
	N	58	-	-	-	-	-
	O	58	-	3	31-33; 41-57	I54I (c162t); L57L (c169t); A83A (a249g)	20
	P	58	-	1	31- 33; 41-57	A2A (a6c)	20
	Q	58	-	-	-	-	-
	R	58	-	-	-	-	-
	S	58	-	-	-	-	-
esx-4	T	58	-	-	-	-	-
	U	58	-	-	-	-	-
	V	58	-	1	1-28; 31-57	H13H (c39t)	56
	W	58	-	-	-	-	-

Table 8: Binary representation for Non-synonymous SNPs identified in the 23 ESX genes across 55 clinical isolates and the 3 laboratory strains. Presence of the mutation is represented by 1 and no mutation represented by 0. The dataset is a color coded representation of isolate groupings. Green: Beijing Isolates, Blue: KZN Isolates and Yellow: Other Isolate grouping.

Nucleotide						a145g		g373a				a59t	a163c	c261g		g172a	c97t	g98t	a142c	t143a	t175c		a155g	a7t	a247g					c187t	a59t	c68t	c170t	a6g
Protein						T49A		W58Stop				Q20L	L55F	A88G		A58T	R33S	R33S	M48Q	M48Q	Stop59Q		I52G	T3S	N83D					P63S	Q20L	S23L	A57V	T2A
RFLP / Strains:	Isolates:	Isolate Number:	esxA	esxB	esxC	esxD	esxE	esxF	esxG	esxH	esxI	esxJ	esxK	esxL	esxM	esxN	esxO	esxP	esxP	esxQ	esxR	esxS	esxT	esxU	esxV	esxV	esxW						
Beijing	910	31	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1
Beijing	1784	32	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1
Beijing	1528	33	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1
Beijing	43178	41	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1
Beijing	25210	42	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1
Beijing	10263	43	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1
Beijing	2705	44	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1
Beijing	39321	45	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1
Beijing	43117	46	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1
Beijing	48246	47	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1
Beijing	21386	48	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1
Beijing	7021344	49	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1
Beijing	7027955	50	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	1	1	1	1	1
Beijing	7037345	51	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1
Beijing	7075078	52	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1
Beijing	7084878	53	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1
Beijing	7023338	54	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1
Beijing	7041516	55	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1
Beijing	2731	56	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1
Beijing	7115902	57	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1
KZN	26	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	253	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	257	7	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	295	8	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	339	11	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	351	12	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	375	13	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	389	14	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	413	16	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	426	17	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	434	18	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	467	20	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	502	22	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	503	23	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	504	24	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	506	25	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	525	26	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	623	28	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	1435	34	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	4207	35	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	4258	36	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	2475	37	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	605	38	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	666	39	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
KZN	8426	40	0	0	0	1	0	1	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
Other	62	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
Other	224	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
Other	226	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
Other	252	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Other	299	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
Other	300	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
Other	402	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	1	0
Other	643	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
Other	492	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	1	0
Other	576	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
Reference Strain	H37Rv	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Avirulent Strain	H37Ra	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Reference Strain	BCG	58	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 9: Binary representation for Synonymous SNPs identified in the 23 ESX genes across 55 clinical isolates and the 3 laboratory strains. Presence of the mutation is represented by 1 and no mutation represented by 0. The dataset is color coded according to isolate grouping. Green represents Beijing Isolates, blue representing KZN Isolates and yellow for the Other Isolate grouping.

Nucleotide								g279c						a9g	c117t	g258a		c141t	c267g	c162t	c169t	a6c	c30t	c168t					c39t	
Protein								S93						S3	N39	E86		G47	V90	I54	L57	A2	H10	A57					H13	
RFLP / Strains:	Isolates:	Isolate Number:	esxA	esxB	esxC	esxD	esxE	esxF	esxG	esxH	esxI	esxJ	esxK	esxK	esxK	esxK	esxL	esxM	esxN	esxO	esxO	esxP	esxP	esxQ	esxR	esxS	esxT	esxU	esxV	esxW
Beijing	910 P4	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
Beijing	1784 P5	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Beijing	1528 P8	33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	0
Beijing	43178	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
Beijing	WP06 25210	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	0
Beijing	WP06 10263	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Beijing	WP06 2705	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
Beijing	39321	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	0
Beijing	43117	46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Beijing	48246	47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Beijing	WP06 21386	48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	0
Beijing	WP07 7021344	49	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	0
Beijing	WP07 7027955	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
Beijing	WP07 7037345	51	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	0
Beijing	WP07 7075078	52	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Beijing	WP07 7084874	53	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Beijing	WP07 7023338	54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Beijing	WP07 7041516	55	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	0
Beijing	WP06 2731	56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
Beijing	WP07 7115992	57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
KZN variant 1	R 26	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KZN variant 1	R 253	6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
KZN variant 1	R 257	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KZN variant 1	R 295	8	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KZN variant 1	R 339	11	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
KZN variant 1	R 351	12	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
KZN variant 1	R 375	13	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
KZN variant 1	R 389	14	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
KZN variant 1	R 413	16	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
KZN Archetype	R 426	17	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
KZN variant 1	R 434	18	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KZN variant 1	R 467	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KZN variant 1	R 502	22	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
KZN variant 1	R 503	23	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
KZN variant 1	R 504	24	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
KZN variant 1	R 506	25	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
KZN variant 1	R 525	26	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
KZN variant 1	R 623	28	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
KZN variant 1	Vac 1435	34	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
KZN variant 1	Vac 4207	35	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KZN variant 1	Vac 4258	36	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
KZN variant 1	Vac 2475	37	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KZN variant 1	KZN 605	38	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
KZN variant 1	Vac 666	39	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KZN variant 1	Vac 8426	40	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Neither	R 62	2	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Neither	R 224	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Neither	R 226	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Neither	R 252	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Neither	R 299	9	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Neither	R 300	10	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Neither	R 402	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F28	R 443	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
F28	R 492	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
Neither	R 576	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Reference Strain	H37Rv	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Avirulent Strain	H37Ra	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Reference Strain	BCG	58	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

ESX Sequence Diversity of Clinical Isolates Using Full Genome Sequencing: With more recent advances in genome sequencing it was possible to augment our dataset with ESX sequence data from 130 full genome sequences of clinical isolates from KZN. These were sequenced as part of the ongoing collaboration between KwaZulu-Natal Research Institute for Tuberculosis and HIV/AIDS (K-RITH) and The Broad Institute. These isolates were from two sources. Ninety strains were collected from 20 sites across the province of KZN as part of a drug susceptibility study, and 80 of these isolates were at least resistant to one first-line anti-tuberculosis drug. The second study collected isolates from a single site in Durban, King Dinuzulu hospital, which is the centralized referral centre for all drug-resistant tuberculosis in KZN. The majority of these isolates were resistant to at least one drug. Full genome sequencing using the Illumina platform was carried out on two libraries, a short fragment library and a jumping library and all SNPs were identified using BROAD developed variant calling software. A total of 10,781 SNP positions, synonymous and non-synonymous, were identified and used to generate a phylogeny shown in Figure 12. nsSNP Mutations in the 23 ESX genes were identified and have been plotted against each strain in the peacock graph shown in Figure 12. The comparative analysis was performed relative to the H37Rv sequence.

The phylogeny was able to resolve the Beijing and KZN families. In Figure 12 the KZN strains are labeled as LAM4. In terms of the distribution of nsSNPs there were differences between the amplicon and genome sequenced clinical isolates. There was no overlap between these two collections of strains in that no strains were sequenced by both methods. In the am-

plicon sequenced group of strains no nSNPs were found in 12 genes (*esx A, B, C, E, G, H, J, N, Q, R, S, T*) (Table 2.6 and 2.8). In the genome sequenced group 5 *esx* genes (*esxA, G, H, J, Q, S* and *T*) had no nsSNPs (Figure 12). The differences can be accounted for by 7 nsSNPs in 5 genes (Figure 12) that occurred in strains other than the KZN or Beijing and reflect a broader sample size in the full genome sequenced group.

Similarly an excellent correlation was observed for the nsSNPs in the Beijing and KZN strains sequenced by both methods. In Figure 12, the largest number of polymorphisms occurred in the Beijing clade, probably in part due to the phylogenetic distance from the reference H37Rv used in this analysis. Unique nsSNPs, for the Beijing and KZN isolates that were identified earlier from our dataset, also occur in Figure 12.

The unique nsSNPs for the Beijing isolates that were found by both sequencing methods were: *esxI* (Q20L); e*sxO* (E52G); e*sxP* (T3S); *esxU* (P43S) and *esxW* (T2A). In the previous amplicon sequencing description above we annotated the *esxU* mutation at codon 63, but subsequent to this analysis there was a change in the annotated start codon (http://genolist.pasteur.fr/TubercuList) by -20 AA. However, in the amplicon sequencing an additional mutation was found in e*sxP* corresponding to N83D (Tables 6 and 8). Interestingly this portion of the gene is conserved amongst the QILSS family of genes and is likely that the stringency of the SNP calling software used by the BROAD bioinformatics systems would exclude a SNP in 1 out of the 5 QILSS genes because of its low occurrence. Similarly when we analysed the clade specific nsSNPs for the KZN isolates both methodologies revealed: *esxK* (A58T) and *esxL* (R33S). The *esxL* mutation corresponds to two nucleotide changes (g97t and g98t) (Table 2.6 and 2.8). In Figure 12 these have been annotated as individual codon changes (R33C and R33P) but taken together they result in an R33S mutation. Two mutations in *esxV* that occurred in nearly all the isolates (Q20L and S23L) were also found in both datasets, including in the dataset of Uplekar and colleagues (Uplekar *et. al.*, 2011), but an additional *esxV* mutation (A57V) was found across the same broad range of strains only by amplicon sequencing (Table 6 and 8). Similarly e*sxV* is an Mtb9.9 gene and the residue 57 is in a highly conserved portion of the protein.

CONCLUSION: The aim of this study was to evaluate heterogeneity in the QILSS and Mtb9.9 families and this is the first study assessing the ESX gene diversity and SNP variation in strains circulating within the South African population. Our amplicon data suggests that *esxA, B, C, E, G, H, J, R, S* and *T* are highly conserved. However with the inclusion of additional sequences from a collection of strains that were genome sequenced, the number of invariant genes declined as nSNPs were also identified in *esxB, C, E, R.* It is likely that as more clinical isolates are sequenced the number of invariant genes will decline further and for a complete picture a broader global collection of strains will need to be characterized. This will be important if immune selection is to be evaluated in the context of different host genetic backgrounds.

The distribution of nSNPs was not uniform across the *esx* genes and this possibly could result from a lack of immune selective pressure by the host immune system on the invariant members. The concentration of nSNPs in the QILSS and Mtb9.9 genes as appose to the ESX-1 and ESX-3 *esx* members can be explained by the functional roles of the latter two systems. In both these systems deletion of *esx* genes leads to attenuation or lack of viability and therefore mutations in these genes are likely to carry a high fitness cost. In contrast if ESX-5 has an immune-regulatory role, then mutations in the associated *esx* genes could carry a selective advantage. Therefore a further in depth investigation of the interaction with the host is required.

As previously stated a limiting factor in this study was the number of isolates used, hence the *esxA, G, H, J, S* and *T* genes cannot be inferred as being globally conserved. One distinct difference from the study by Comas and colleagues, is that we found *esxH* to be invariant, compared to their study that found to it to be hypervariable (Comas *et. al.*, 2010). They only looked at 26 strains and therefore their conclusions are not likely to be valid, so further studies at a later stage should include a larger genetically diverse number of isolates with a broader worldwide representation of the strain diversity.

One intriguing observation that has come out of these sequencing studies is the sites of sequence variation leading to nSNPs occur predominantly at sites that are already variable amongst the QILSS and particularly the Mtb9.9 family. An example of this is the A58T, and A58P mutations. The A58T mutation was observed in *esxK* and *esxP* and converted these proteins to have the same sequence as *esxW, J, M.* We also found 1 isolate with an A58P mutation at this residue. It is easy to imagine that this variation is an experimental artefact of sequencing highly similar genes. This is extremely unlikely as these types of mutations were seen both by amplicon and genome sequencing done in different laboratories. In addition similar gene conversion events were reported in the amplicon sequencing study carried out by Uplekar and colleagues (Uplekar *et. al.*, 2011). Of note the Broad institute

used both short and jumping libraries for sequencing. The jumping libraries were approximately 2000 to 3000bps in length and allow pair end sequences separated by approximately 2000bps to be more accurately positioned. This overcomes some of the difficulties of resolving diversity in repetitive regions that was encountered in earlier sequencing efforts.

Homologous recombination between the highly similar sequences of the QILSS and Mtb9.9 genes is the most likely mechanism generating this kind of diversity. It is not clear how altering the number of copies of a single allele from three to four out of five (for example in the case of an A58T mutation in *esxK*) would significantly alter immunogenicity. It seems unlikely that such a relatively small proportional effect on gene number could translate into a significant change in immune phenotype. Alternatively if the genes were transcribed differently between strains then it might represent an allelic switch.

Additional bioinformatics analysis on a larger number of sequences is required to more formally address this question of recombination events. Understanding how diversity in the immune-dominant ESX families is generated and maintained is clearly highly important for vaccine development. It is hoped this will ultimately contribute in the management of this disease burden in South Africa.

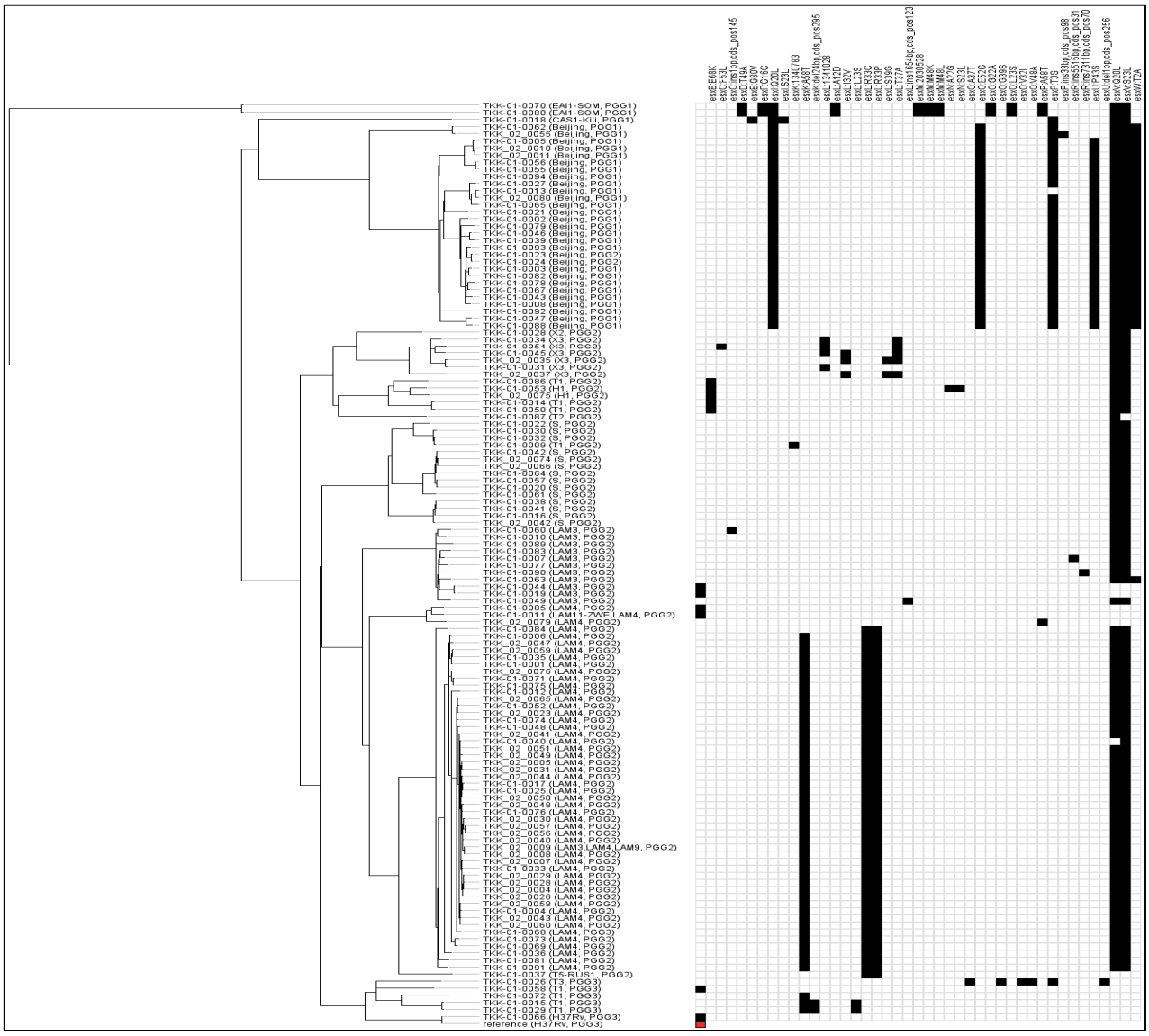


Figure 12: Phylogenetic tree based on all SNPs identified in 130 full genome sequenced isolates rooted to H37Rv. The grid to the right indicates the presence (bold) or absence (blank) of nSNPs identified in 23 ESX genes relative to H37Rv. The unlabelled column on the left signifies if there are no mutations present.

ACKNOWLEDGEMENT: The authors would like to thank the following groups and individuals: The staff of Africa Centre for the sequencing services provided. Dr. Alexander Pym for discussions. Dr. Stephen Gordon and group of UCD, Ireland for advise and training provided. Pym Lab (K-RITH), both technical staff and students for support and assistance. The staff at the Department of Infection, Prevention and Control (UKZN) for the isolate confirmations. All the partners of the NOVSEC-TB consortium for support and assistance.

This study was supported by funding from the European Community's Seventh Framework [FP7/2007-2013]) under grant agreement no. 201762 and the DST/NRF grants. The College of Health Sciences (CHS), TESA and NRF/DST for the funding awarded during this study.

REFERENCES:

1. Abdallah, A. M., Verboom,T., Hannes,F., Safi, M., Strong, M., Eisenberg, D., Musters, R. J. P., Vandenbroucke-Grauls, C. M. J. E., Appelmelk, B. J., Luirink, L. and Bitter, W. (2006) A specific secretion system mediates PPE41 transport in pathogenic mycobacteria, *Mol. Microbiol.*, 62, 667-679.

2. Abdallah, A. M., Gey van Pittius, N. C., DiGiuseppe Champion, P. A., Cox, J., Luirink, J., Vandenbroucke-Grauls, C. M. J. E., Appelmelk, B. J. and Bitter, W. (2007). Type VII secretion-mycobacteria show the way, *Nat. Rev. Microbiol.*, 5, 883-891.

3. Abdallah, A. M., Verboom, T., Weerdenburg, E. M., Gey van Pittius, N. C., Mahasha, P. W., Jimenez, C., Parra, M., Cadieux, N., Brennan, M. J., Appelmelk, B. J. and Bitter, W. (2009). PPE and PE_PGRS proteins of *Mycobacterium marinum* are transported via the type VII secretion system ESX-5, *Mol. Microbiol.*, 73, 329-340.

4. Abdallah, A. M., Bestebroer, J., Savage, N. D., de Punder, K., van Zon, M., Wilson, L., Korbee, C. J., van der Sar, A. M., Ottenhoff, T. H., van der Wel, N. N., Bitter, W. and Peters, P. J. (2011) Mycobacterial secretion systems ESX-1 and ESX-5 play distinct roles in host cell death and inflammasome activation, *J. Immunol.*, 187, 4744-4753.

5. Alderson, M. R., Bement, T., Day, C. H., Zhu, L., Molesh, D., Skeiky, Y. A., Coler, R., Lewinsohn, D. M., Reed, S. G. and Dillon, D. C. (2000) Expression cloning of an immunodominant family of *Mycobacterium tuberculosis* antigens using human CD4(+) T cells, *J. Exp. Med.*, 191, 551-560.

6. Arbing, M. A., Kaufmann, M., Phan, T., Chan, S., Cascio, D. and Eisenberg, D**.** (2010) The crystal structure of the *Mycobacterium tuberculosis* Rv3019c-Rv3020c ESX complex reveals a domain-swapped heterotetramer, *Protein. Sci.*, 19, 1692-1703.

7. Ashiru, O. T., Pillay, M. and Sturm, A. W. (2010) Adhesion to and invasion of pulmonary epithelial cells by the F15/LAM4/KZN and Beijing strains of *Mycobacterium tuberculosis*, *J. Med. Microbiol.*, 59, 528-533.

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

9. Berthet, F., Rasmussen, P., Rosenkrands, I., Andersen, P. and Gicquel, B. (1998) A *Mycobacterium tuberculosis* operon encoding ESAT-6 and a novel low- molecular-mass culture filtrate protein (CFP-10), *Microbiol*., 144, 3195-3203.

10. Bitter, W., Houben, E. N. G., Bottai, D., Brodin, P., Brown, E. J., Cox, J. S., Derbyshire, K., Fortune, S. M., Gao, L.-Y., Liu, J., Gey van Pittius, N. C., Pym, A. S., Rubin, E. J., Sherman, D. R., Cole, S. T. and Brosch, R. **(**2009) Systematic Genetic Nomenclature for Type VII Secretion Systems, *PloS. Pathog*., 5, e1000507.

11. Bottai, D., Di Luca, M., Majlessi, L., Frigui, W., Simeone, R., Sayes, F., Bitter, W., Brennan, M. J., Leclerc, C., Batoni, G., Campa, M., Brosch, R. and Esin, S. (2012) Disruption of the ESX-5 system of *Mycobacterium tuberculosis* causes loss of PPE protein secretion, reduction of cell wall integrity and strong attenuation, *Mol. Microbiol*., 83, 1195-1209.

12. Brites, D. and Gagneux, S. **(**2012) Old and new selective pressures on *Mycobacterium tuberculosis*, *Infect. Genet. Evol.*, 12, 678-685.

13. Brodin, P., Poquet, Y., Levillain, F., Peguillet, I., Larrouy-Maumus, G., Gilleron, M., Ewann, F., Christophe, T., Fenistein, D., Jang, J., Jang, M., Park, S., Rauzier, J., Carralot, J., Shrimpton, R., Genovesio, A., Gonzalo-Asensio, J. A., Puzo, G., Martin, C., Brosch, R., Stewart, G. R., Gicquel, B. and Neysolles, O. (2010) High content phenotypic cell-based visual screen identifies *Mycobacterium tuberculosis* acyltrehalose-containing glycolipids involved in phagosome remodeling, *PLoS Pathog.*, 6, e1001100.

14. Brosch, R., Pym, A. S., Gordon, S. V. and Cole, S. T. (2001) The evolution of mycobacterial pathogenicity: clues from comparative genomics, *Trend. Microbiol.*, 9, 452-458.

15. Broussard, G. W. and Ennis, D. G. (2007) *Mycobacterium marinum* produces long- term chronic infections in medaka: a new animal model for studying human tuberculosis, *Comp. Biochem. Physiol.*, 145, 45-54.

16. Bukka, A., Price, C. T., Kernodle, D. S. and Graham, J. E. (2011) *Mycobacterium tuberculosis* RNA Expression Patterns in Sputum Bacteria Indicate Secreted Esx Factors Contributing to

Growth are Highly Expressed in Active Disease, *Front Microbiol*., 2, 266.

17. Carlsson, F., J. Kim, C. Dumitru, K. H. Barck, R. A. Carano, M. Sun, L. Diehl, and E. J. Brown. (2010) Host-detrimental role of Esx-1-mediated inflammasome activation in mycobacterial infection, *PLoS Pathog.*, 6, e1000895.

18. Cascioferro, A., M. H. Daleke, M. Ventura, V. Dona, G. Delogu, G. Palu, W. Bitter, and R. Manganelli. (2011) Functional dissection of the PE domain responsible for translocation of PE_PGRS33 across the mycobacterial cell wall, *PLoS One*, 6, e27713.

19. Champion, M., Williams, E. A., Kennedy, G. M. and Champion, P. A. D. (2012) Direct detection of bacterial protein secretion using whole colony proteomics, *Mol. Cell. Proteom*., 11, 596-604.

20. Chapman, A. L., Munkanta, M., Wilkinson, K. A., Pathan, A. A., Ewer, K., Ayles, H., Reece, W. H., Mwinga, A., Godfrey-Faussett, P. and Lalvani, A. (2002) Rapid detection of active and latent tuberculosis infection in HIV-positive individuals by enumeration of *Mycobacterium tuberculosis*-specific T cells, *AIDS*, 16, 2285-2293.

21. Colangeli, R., Spencer, J.S., Bifani, P., Williams, A., Lyaschenko, K., Keen, M. A., Hill, P. J., Belisle, J. and Gennaro, M. L. (2000) MTSA-10, the product of the Rv3874 gene of *Mycobacteria tuberculosis*, elicits tuberculosis-specific, delayed-type hypersensitivity in guniea pigs, *Infect. Immun.*, 68, 990-993.

22. Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M. A., Rajandream, M. A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J. E., Taylor, K., Whitehead, S. and Barrell, B. G. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence, *Nature*, 393, 537-544.

23. Comas, I., Chakravartti, J., Small, P. M., Galagan, J., Niemann, S., Kremer, K., Ernst, J. D. and Gagneux, S. (2010) Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved, *Nat. Genet*., 42, 498-503.

24. Comas, I. and Gagneux, S. (2011) A role for systems epidemiology in tuberculosis research, *Trends Microbiol*., 19, 492-500.

25. Das, C., Ghosh, T. S. and Mande, S. S. (2011) Computational analysis of the ESX-1 region of *Mycobacterium tuberculosis*: insights into the mechanism of type VII secretion system, *PLoS One*, 6, e27980.

26. Davila, J., Zhang, L., Marrs, C. F., Durmaz, R. and Yang, Z. (2010) Assessment of the genetic diversity of *Mycobacterium tuberculosis esxA*, *esxH*, and *fbpB* genes among clinical isolates and its implication for the future immunization by new tuberculosis subunit vaccines Ag85B-ESAT-6 and Ag85B-TB10.4, *J. Biomed. Biotechnol.*, 2010, 208371.

27. Davis, J. M., Clay, H., Lewis, J. L., Ghori, N., Herbomel, P. and Ramakrishnan, L. (2002) Real-time visualization of mycobacterium-macrophage interactions leading o initiation of granuloma formation in zebrafish embryos, *Immunity*, 17, 693-702.

28. Deng, W., Xiang, X. and Xie, J. (2014) Comparative genomic and proteomic anatomy of mycobacterium ubiquitous Esx family proteins: implications in pathogenicity and virulence, *Curr Microbiol*., 68, 558-567.

29. Drummond, A. J., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., Field, M., Heled, J., Kearse, M., Markowitz, S., Moir, R., Stones-Havas, S., Sturrock, S., Thierer, T. and Wilson, A. (2012) *Geneious*, v5.6, http://www.geneious.com.

30. van Embden, J. D., Cave, M. D., Crawford, J. T., Dale, J. W., Eisenach, K. D., Gicquel, B., Hermans, P., Martin, C., McAdam, R. and Shinnick, T. M. (1993) Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology, *J. Clin. Microbiol.*, 31, 406-409.

31. Filliol, I., Motiwala, A. S., Cavatore, M., Qi, W., Hazbón, M. H., Bobadilla del Valle, M., Fyfe, J., García-García, L., Rastogi, N., Sola, C., Zozio, T., Guerrero, M. I., León, C. I., Crabtree, J., Angiuoli, S., Eisenach, K. D., Durmaz, R., Joloba, M. L., Rendón, A., Sifuentes-Osornio, J., Ponce de León, A., Cave, M. D., Fleischmann, R., Whittam, T. S. and Alland, D. (2006) Global Phylogeny of *Mycobacterium tuberculosis* Based on Single Nucleotide Polymorphism (SNP) Analysis: Insights into Tuberculosis Evolution, Phylogenetic Accuracy of Other DNA Fingerprinting Systems, and Recommendations for a Minimal Standard SNP Set, *J. Bact*., 188, 759-772.

32. Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L. and Bateman, A. (2006) Pfam: clans, web tools and services, *Nucleic Acids Res*., 34, D247-D251.

33. Fortune, S. M., Jaeger, A., Sarracino, D. A., Chase, M. R., Sassetti, C. M., Sherman, D. R., Bloom, B. R. and Rubin, E. J. (2005) Mutually dependent secretion of proteins required for mycobacterial virulence, *PNAS.*, 102, 10676-10691.

34. Frigui, W., Bottai, D., Majlessi, L., Monot, M., Josselin, E., Brodin, P., Gar□nier, T., Gicquel, B., Martin, C., Leclerc, C., Cole, S. T. and Brosch, R. □(2008) Control of *M. tuberculosis* ESAT-6 Secretion and Specific T Cell Recognition by PhoP, *PLoS Pathog.*, 4, e33.

35. Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., de Jong, B. C., Narayanan, S., Nicol, M., Niemann, S., Kremer, K., Gutierrez, M. C., Hilty, M., Hopewell, P. C. and Small, P. M. (2006) Variable host–pathogen compatibility in *Mycobacterium tuberculosis*, *PNAS.*, 103, 2869-2873.

36. Gao, L. Y., Guo, S., McLaughlin, B., Morisaki, H., Engel, J. N. and Brown, E. J. (2004) A mycobacterial virulence gene cluster extending RD1 is required for cytolysis, bacterial spreading and ESAT-6 secretion, *Mol. Microbiol.*, 53, 1677-1693.

37. Gey van Pittius, N. C., Gamieldien, J., Hide, W., Brown, G., Siezen, R. and Beyers, A. (2001) The ESAT-6 gene cluster of *Mycobacterium tuberculosis* and other high G+C Gram-positive bacteria, *Genome Biol*., 2, research0044.1-research0044.18.

38. Gey van Pittius, N. C., Sampson, S. L., Lee, H., Kim, Y., van Helden, P. D. and Warren, R. M. (2006) Evolution and expansion of the *Mycobacterium tuberculosis* PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions, *BMC Evol. Biol*., 6, 95.

39. Gonzalo-Asensio, J., Mostowy, S., Harders-Westerveen, J., Huygen, K., □Hernandez-Pando, R., Thole, J., Behr, M., Gicquel, B. and Martín, C. (2008) PhoP: A Missing Piece in the Intricate Puzzle of *Mycobacterium tuberculosis* Virulence, *PLoS One*, 3, e3496.

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

41. Guinn, K. M., Hickey, M. J., Mathur, S. K., Zakel, K. L., Grotzke, J. E., Lewinsohn, D. M., Smith, S. and Sherman, D. R. (2004) Individual RD1-region genes are required for export of ESAT-6/CFP-10 and for virulence of *Mycobacterium tuberculosis*, *Mol. Microbiol*., 51, 359-370.

42. Gutacker, M. M., Mathema, B., Soini, H., Shashkina, E., Kreiswirth, B. N., Graviss, E. A. and Musser, J. M. (2006) Single-Nucleotide Polymorphism–Based Population Genetic Analysis of *Mycobacterium tuberculosis* Strains from 4 Geographic Sites, *J. Infect. Dis*., 193, 121-128.

43. Harboe, M., Oettinger, T., Wiker, H. G., Rosenkrands, I. and Andersen, P. (1996) Evidence for occurence of the ESAT-6 protein in *Mycobacteria tuberculosis* and virulent *Mycobacterium bovis* and for its absence in *Mycobacterium bovis* BCG, *Infect. Immun*., 64, 16-22.

44. Hawkey, P. M., Smith, E. G., Evans, J. T., Monk, P., Bryan, G., Mohamed, H. H., Bardhan, M. and Pugh, R. N. (2003) Mycobacterial Interspersed Repetitive Unit Typing of *Mycobacterium tuberculosis* compared to IS*6110*-Based Restriction Fragment Length Polymorphism Analysis for investigation of apparently clustered cases of tuberculosis, *J. Clin. Microbiol*., 41, 3514-3520.

45. Hershberg, R., Lipatov, M., Small, P. M., Sheffer, H., Niemann, S., Homolka, S., Roach, J. C., Kremer, K., Petrov, D. A., Feldman, M. W and Gagneux, S. (2008) High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography, *PLoS Biol*., 6, e311.

46. Hsu, T., Hingley-Wilson, S. M., Chen, B., Chen, M., Dai, A. Z., Morin, P. M., Marks, C. B., Padiyar, J., Goulding, C., Gingery, M., Eisenberg, D., Russell, R. G., Derrick, S. C., Collins, F. M., Morris, S. L., King, C. H. and Jacobs, W. R., Jr. (2003) The primary mechanism of attenuation of bacillus Calmette-Guerin is a loss of secreted lytic function required for invasion of lung interstitial tissue, *PNAS*., 100, 12420-12425.

47. Jones, G. J., Gordon, S. V., Hewinson, R. G. and Vordermeier, H. M. (2010) Screening of predicted secreted antigens from *Mycobacterium bovis* reveals the immunodominance of the ESAT-6 protein family, *Infect. Immun*., 78, 1326-1332.

48. de Jonge, M. I., Pehau-Arnaudet, G., Fretz, M. M., Romain, F., Bottai, D., Brodin, P., Honore, N., Marchal, G., Jiskoot, W., Endgland, P., Cole, S. T. and Brosch, R. (2007) ESAT-6 from *Mycobacterium tuberculosis* Dissociates from Its Putative Chaperone CFP-10 under Acidic Conditions and Exhibits Membrane-Lysing Activity, *J. Bacteriol*., 189, 6028-6034.

49. Junqueira-Kipnis, A. P., Basaraba, R. J., Gruppo, V., Palanisamy, G., Turner, O. C., Hsu, T., Jacobs, W. R. Jr., Fulton, S. A., Reba, S. M., Boom, W. H. and Orme, I. M. (2006) Mycobacteria lacking the RD1 region do not induce necrosis in the lungs of mice lacking interferon-gamma, *Immunol*., 119, 224-231.

50. Kato-Maeda, M., Bifani, P. J., Kreiswirth, B. N. and Small, P. M. (2001) The nature and consequence of genetic variability within *Mycobacterium tuberculosis*, *J. Clin. Invest*., 107, 553-537.

51. Lechat, P., Hummel, L., Rousseau, S. and Moszer, I. (2008) GenoList: an integrated environment for comparative analysis of microbial genomes, *Nucleic Acids Res*., 36, D469-D474. [http://genodb.pasteur.fr/cgibin/WebObjects/GenoList.woa/wa/goToMainPage] [Accessed Online: 02 February 2013].

52. Lew, J. M., Kapopoulou, A., Jones, L. M. and Cole, S. T. (2011) TubercuList-10 years after, p. 1-7, Tuberculosis, vol. 91, Edinb.

[http://genolist.pasteur.fr/TubercuList/] [Accessed Online: 15 January 2013].

53. Lewis, K. N., Liao, R., Guinn, K. M., Hickey, M. J., Smith, S., Behr, M. A. and Sherman, D. R. (2003) Deletion of RD1 from *Mycobacterium tuberculosis* mimics bacille Calmette-Guérin attenuation, *J. Infect. Dis*., 187, 117-123.

54. Lightbody, K. L., Ilghari, D., Waters, L. C., Carey, G., Bailey, M. A., Williamson, R. A., Renshaw, P. S. and Carr, M. D. (2008) Molecular features governing the stability and specificity of functional complex formation by *M. tuberculosis* CFP-10/ESAT-6 family proteins, *J. Biol. Chem.*, 283, 17681-17690.

55. Louise, R., Skjøt, V., Agger, E. M. and Andersen, P. (2001) Antigen discovery and tuberculosis vaccine development in the postgenomic era, *Scand. J. Infect. Dis*., 33 643–647.

56. Maciag, A., Dainese, E., Rodriguez, G. M., Milano, A., Provvedi, R., Pasca, M. R,, Smith, I., Palù, G., Riccardi, G. and Manganelli, R. (2007) Global analysis of the *Mycobacterium tuberculosis* Zur (FurB) regulon, *J. Bacteriol*., 189, 730-740.

57. Maheiras, G. G., Sabo, P. J., Hickey, M. J., Devinder, C. S. and C. K. Stover. (1996) Molecular analysis of genetic differences between *Mycobacterium bovis BCG* and virulent *M. bovis*, *J. Bacteriol*., 178, 1274-1282.

58. Malik, A. N. and Godfrey-Fausset, P. (2005) Effectes of genetic variability of *Mycobacterium tuberculosis* strains on the presentation of disease, *Lancet Infect. Dis*., 5, 174-183.

59. Manzanillo, P. S., Shiloh, M. U., Portnoy, D. A. and Cox, J. S. (2012) *Mycobacterium tuberculosis* activates the DNA-dependent cytosolic surveillance pathway within macrophages, *Cell Host Microbe*, 11, 469-480.

60. Mazurek, G. H., LoBue, P. A., Daley, C. L., Bernardo, J., Lardizabal, A. A., Bishai, W. R., Iademarco, M. F. and Rothel, J. S. (2001) Comparison of a whole-blood interferon gamma assay with tuberculin skin testing for detecting latent *Mycobacterium tuberculosis* infection, *JAMA*., 286, 1740-1747.

61. McLaughlin, B., Chon, J. S., MacGurn, J. A., Carlsson, F., Cheng, T. L., Cox, J. S. and Brown, E. J. (2007) A mycobacterium ESX-1-secreted virulence factor with unique requirements for export, *PLoS Pathog*., 3, e105

62. Meher, A. K., Bal, N. C., Chary, K. V. and Arora, A. (2006) *Mycobacterium tuberculosis* H37Rv ESAT-6-CFP-10 complex formation confers thermodynamic and biochemical stability. *FEBS J*., 273, 1445-1462.

63. Meier, T., Eulenbruch, H. P., Wrighton-Smith, P., Enders, G. and Regnath, T. (2005) Sensitivity of a new commercial enzyme-linked immunospot assay (T SPOT-TB) for diagnosis of tuberculosis in clinical practice, *Eur. J. Clin. Microbiol. Infect. Dis*., 24, 529-536.

64. Mestre, O., Luo, T., Dos Vultos, T., Kremer, K., Murray, A., Namouchi, A., Jackson, C., Rauzier, J., Bifani, P., Warren, R., Rasolofo, V., Mei, J., Gao, Q. and Gicquel, B. (2011) Phylogeny of *Mycobacterium tuberculosis* Beijing Strains Constructed from Polymorphisms in Genes Involved in DNA Replication, Recombination and Repair, *PLoS One*, 6, e16020.

65. Musser, J. M., Amin, A. and Ramaswamy, S. (2000) Negligible genetic diversity of *Mycobacterium tuberculosis* host immune system protein targets: evidence of limited selective pressure, *Genetics*, 155, 7-16.

66. Nei, M. and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions, *Mol. Biol. Evol*., 3, 418-426.

67. Nicol, M. P. and Wilkinson, R. J. (2008) The clinical consequences of strain diversity in *Mycobacterium tuberculosis*, *Trans. R. Soc. Trop. Med. Hyg*., 102, 955-565.

68. Novikov, A., Cardone, M., Thompson, R., Shenderov, K., Kirschman, K. D., Mayer-Barber, K. D., Myers, T. G., Rabin, R. L., Trinchieri, G., Sher, A. and Feng, C. G. (2011) *Mycobacterium tuberculosis* Triggers Host Type I IFN Signaling To Regulate IL-1β Production in Human Macrophages, *J. Immunol*., 187, 2540-2547.

69. Ohol, Y. M., Goetz, D. H., Chan, K., Shiloh, M. U., Craik, C. S. and Cox, J. S. (2010) *Mycobacterium tuberculosis* MycP1 protease plays a dual role in regulation of ESX-1 secretion and virulence, *Cell. Host Microbe*., 7, 210 –220.

70. Pallen, M. J. (2002) The ESAT-6/WXG100 superfamily - and a new Gram-positive secretion system?, *Trend Microbiol*., 10, 209-212.

71. Pillay, M. and Sturm, A. W. (2007) Evolution of the Extensively Drug-Resistant F15/LAM4/KZN Strain of *Mycobacterium tuberculosis* in Kwazulu-Natal, South Africa, *Clin. Infect. Dis*., 45, 1409-1414.

72. Pillay, N. (2010) Development of a genotypic test for the detection of KZN strain of *Mycobacterium tuberculosis*, University of KwaZulu-Natal, Durban, South Africa. (unpublished MSc *Thesis*).

73. van Pinxteren, L. A., Ravn, P., Agger, E. M., Pollock, J. and Andersen, P. (2000) Diagnosis of tuberculosis based on the two specific antigens ESAT-6 and CFP10, *Clin. Diagn. Lab. Immunol*., 7, 155-160.

74. Prouty, M. G., Correa, N. E., Barker, L. P., Jagadeeswaran, P. and Klose, K. E. (2003) Zebrafish *Mycobacterium marinum* model for mycobacterial pathogenesis, *FEMS Microbiol. Lett.*, 225, 177-182.

75. Pym, A. S., Brodin, P., Majlessi, L., Brosch, R., Demangel, C., Williams, A., Griffiths, K. E.,

Marchal, G., Leclerc, C. and Cole, S. T. (2003) Recombinant BCG exporting ESAT-6 confers enhanced protection against tuberculosis, *Nat. Med.*, 9, 533-539.

76. Raghavan, S., Manzanillo, P., Chan, K., Dovey, C. and Cox, J. S. (2008) Secreted transcription factor controls *Mycobacterium tuberculosis* virulence, *Nature*, 454, 717–721.

77. Ramakrishnan, L., Valdivia, R. H., McKerrow, J. H. and Falkow, S. (1997) *Mycobacterium marinum* causes both long-term subclinical infection and acute disease in the leopard frog (*Rana pipiens*), *Infect. Immun.*, 65, 767-773.

78. Reddy, T. B., Riley, R., Wymore, F., Montgomery, P., DeCaprio, D., Engels, R., Gellesch, M., Hubble, J., Jen, D., Jin, H., Koehrsen, M., Larson, L., Mao, M., Nitzberg, M., Sisk, P., Stolte, C., Weiner, B., White, J., Zachariah, Z. K., Sherlock, G., Galagan, J. E., Ball, C. A. and Schoolnik, G. K. (2009) TB database: an integrated platform for tuberculosis research, *Nucleic Acids Res.*, 37(Database issue), D499-508. [http://www.tbdb.org] [Accessed Online: 18 June 2013].

79. Renshaw, P. S., Panagiotidou, P., Whelan, A., Gordon, S. V., Hewinson, R. G., Williamson, R. A. and Carr, M. D. (2002) Conclusive Evidence That the Major T-cell Antigens of the *Mycobacterium tuberculosis* Complex ESAT-6 and CFP-10 Form a Tight, 1:1 Complex and Characterization of the Structural Properties of ESAT-6, CFP-10, and the ESAT-6 CFP-10 Complex. Implications For Pathogenesis and Virulence, *J. Biol. Chem.*, 277, 21598-21603.

80. van der Sar, A. M., Musters, R. J., van Eeden, F. J., Appelmelk, B. J., Vandenbroucke-Grauls, C. M. and Bitter, W. (2003) Zebrafish embryos as a model host for the real time analysis of *Salmonella typhimurium* infections, *Cell. Microbiol.*, 5, 601-611.

81. Sassetti, C. M. and Rubin, E. J. (2003a) Genetic requirements for mycobacterial survival during infection, *PNAS.*, 100, 12989-12994.

82. Sassetti, C. M., Boyd, D. H. and Rubin, E. J. (2003b) Genes required for mycobacterial growth defined by high density mutagenesis, *Mol. Microbiol.*, 48 77-84.

83. Serafini, A., Boldrin, F., Palu, G. and Manganelli, R. (2009) Characterization of a *Mycobacterium tuberculosis* ESX-3 conditional mutant: essentiality and rescue by iron and zinc, *J. Bacteriol.*, 191, 6340-6344.

84. Simeone, R., Bobard, A., Lippmann, J., Bitter, W., Majlessi, L., Brosch, R. and Enninga, J. (2012) Phagosomal Rupture by *Mycobacterium tuberculosis* Results in Toxicity and Host Cell Death, *PLoS Pathog.*, 8, e1002507.

85. Skjøt, R. L. V., Brock, I., Arend, S. M., Munk, M. E., Theisen, M., Ottenhoff, T. H. M. and Andersen, P. (2002) Epitope Mapping of the Immunodominant Antigen TB10.4 and the Two Homologous Proteins TB10.3 and TB12.9, Which Constitute a Subfamily of the esat-6 Gene Family, *Infect. Immun.*, 70, 5446-5453.

86. Smith, J., Manoranjan, J., Pan, M., Bohsali, A., Xu, J., Liu, J., McDonald, K. L., Szyk, A., LaRonde-LeBlanc, N. and Gao, L. Y. (2008) Evidence for pore formation in host cell membranes by ESX-1-secreted ESAT-6 and its role in *Mycobacterium marinum* escape from the vacuole, *Infect. Immun.*, 76, 5478-5487.

87. van Soolingen, D., Dehaas, P. E., Hermans, P. W. and Vanembden, J. D. (1994) DNA fingerprinting of *Mycobacterium tuberculosis*, *Meth. Enzymol.*, 235, 196-205.

88. Sørensen, A. L., Nagai, S., Houen, G., Andersen, P. and Andersen, A. B. (1995) Purification and characterization of a low-molecular-mass T-cell antigen secreted by *Mycobacterium tuberculosis*, *Infect. Immun.*, 63, 1710-1717.

89. Stanley, S. A., Raghavan, S., Hwang, W. W. and Cox, J. S. (2003) Acute infection and macrophage subversion by *Mycobacterium tuberculosis* require a specialized secretion system, *PNAS.*, 100, 13001-13006.

90. Stanely, S. A., Johndrow, J. E., Manzanillo, P. and Cox, J. S. (2007) The type I IFN response to infection with *Mycobacterium tuberculosis* requires ESX-1 mediated secretion and contributes to pathogenesis, *J. Immunol.*, 178, 3143-3152.

91. Stoop, E. J., Schipper, T., Huber, S. K., Nezhinsky, A. E., Verbeek, F. J., Gurcha, S. S., Besra, G. S., Vandenbroucke-Grauls, C. M., Bitter, W. and van der Sar, A. M. (2011) Zebrafish embryo screen for mycobacterial genes involved in the initiation of granuloma formation reveals a newly identified ESX-1 component, *Dis. Model Mech.*, 4, 526-536.

92. Swaim, L. E., Connolly, L. E., Volkman, H. E., Humbert, O., Born, D. E. and Ramakrishnan, L. (2006) *Mycobacterium marinum* infection of adult zebrafish causes caseating granulomatous tuberculosis and is moderated by adaptive immunity, *Infect. Immun.*, 74, 6108-6117.

93. Talaat, A. M., Reimschuessel, R., Wasserman, S. S. and Trucksis, M. (1998) Goldfish, *Carassius auratus*, a novel animal model for the study of Mycobacterium marinum pathogenesis, *Infect. Immun.*, 66, 2938-2942.

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

95. Uplekar, S., Heym, B., Friocourt, V., Rougemont, J. and Cole, S. T. (2011) Comparative genomics of *esx* genes from clinical isolates *Mycobacterium*

tuberculosis provides evidence for gene conversion and epitope variation, *Infect. Immun*., 79, 4042-4049.

96. Volkman, H. E., Clay, H., Beery, D., Chang, J. C., Sherman, D. R. and Ramakrishnan, L. (2004) Tuberculous granuloma formation is enhanced by a *Mycobacterium* virulence determinant, *PLoS Biol*., 2, e367.

97. Wards, B. J., de Lisle, G. W. and Collins, D. M. (2000) An Esat-6 knockout mutant of *Mycobacterium bovis* produced by homologuous recombination will contribute to the development of a live tuberculosis vaccine, *Tuber. Lung. Dis*., 80, 185-189.

98. Watson, R. O., Manzanillo, P. S. and Cox, J. S. (2012) Extracellular *M. tuberculosis* DNA tragets bacteria for autophagy by activating the host DNA-Sensing pathway, *Cell*, 150, 803-815.

99. WHO. (2012) Global Tuberculosis Report. [Accessed Online: 4 February 2013].

100. WHO. (2013) Immunization, Vaccines and Biologicals. [Accessed Online: 4 February 2013].

101. Xu, J., Laine, O., Masciocchi, M., Manoranjan, J., Smith, J., Du, S. J., Edwards, N., Zhu, X., Fenselau, C., and Gao, L. Y. (2007) A unique Mycobacterium ESX-1 protein co-secretes with CFP-10/ESAT-6 and is necessary for inhibiting phagosome maturation, *Mol. Microbiol.*, 66, 787-800.