# Population Genetic Frameworks and Functional Genomics of *Mycobacterium bovis*

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Deepti J Joshi, M.V.Sc

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Dr. Srinand Sreevatsan

September 2012

# ACKNOWLEDGEMENTS

The author wishes to thank several people who were instrumental in completion of the research projects undertaken in this dissertation.

First and foremost, I would like to thank my advisor, Dr. Srinand Sreevatsan, for providing his support and guidance throughout my doctoral degree program. I thank him for his immense patience and for the facilities he provided me with, in his laboratory and outside, to enable the generation of this data and successful execution of the projects.

Thank you to committee members: Dr. Doug Foster, Dr. Sagar Goyal and Dr. Tim Johnson, for their continued support and communication throughout.

I am appreciative of the support and guidance of Dr. Beth Harris, Dr. Ray Waters, Dr. Tyler Thacker and Dr. Allen Jensen at the United States Department of Agriculture Veterinary Labs in Ames, Iowa.

Thank you, to the bioinformaticians at the Minnesota Supercomputing Institute, University of Minnesota, Dr. Zheng Jin Tu, Dr. Ying Zhang and Dr. Wayne Zhu for their invaluable support with the genome sequencing and transcriptomic profiling projects.

I am thankful to the VMED graduate program, College of Veterinary Medicine, for providing me with the opportunity to pursue my PhD, the entire faculty and staff there, especially to Lisa Hubinger.

I also thank the members of Dr. Sreevatsan's lab, who have been great colleagues to work with.

I received funding from the Rapid Agricultural Response Fund and USDA Specials Grant on Bovine TB awarded to my advisor Dr. Sreevatsan.

This grad school journey was made lighter by the constant support and encouragement of my family and friends, particularly my husband Dr. Abhishake, my son Vedant and my parents Dr. Vijaylaxmi and Jagadeesh.

# DEDICATION

This dissertation is dedicated to my lovely family, my husband – chotu, my baby – minku, my mother and my father, my brother and my grandma. Their unconditional love, support and encouragement have made me the independent and confident person I am today.

# RESEARCH ABSTRACT

Bovine tuberculosis is a zoonotic infection of cattle caused by *Mycobacterium bovis.* Approximately one-third of the world's population is infected with *M. tuberculosis* or *M. bovis,* "the world's most successful pathogen", the majority in developing countries. The global spread and increasing severity of tuberculosis are due in part to the high number of individuals infected with HIV, and in part to the increasing intensity of human-animal interactions as land use patterns around the globe rapidly change. The resistance of *M. bovis* to two frontline drugs used to treat tuberculosis—isoniazid and pyrazinamide—threatens to return tuberculosis-associated mortality rates to those of the pre-antibiotic era.

The severe and growing threat of *M. bovis* necessitates rapid, thorough national and international surveillance of strain distribution dynamics in the population. To date, piecemeal analysis of *Mycobacterium bovis* genomes and conventional genotyping methods have not themselves lent to a comprehensive resolution of its genetic diversity to explain the wide range of disease phenotypes caused by this zoonotic pathogen. Conventional genotyping methods target small hypervariable regions on the genome of *M. bovis* and provide anonymous allelic information insufficient to develop *M. bovis* phylogeny. Genome-wide single nucleotide polymorphisms (SNPs) studies in *M. tuberculosis* have shown sufficient resolution to develop trait-allele associations. We hypothesized that genetic and phenotypic diversity in *M. bovis* is enciphered in their genomes. To study genetic variations we first interrogated the *M. bovis* genome for 350 loci including geneic (n =306) and intergeneic (n =44) regions for SNPs. A collection of 75 *M. bovis* isolates associated with bovine bovine tuberculosis outbreaks in the US between 1990-2009 and isolated from a variety of mammalian hosts – cattle (*n*=25), deer (*n*=6), elk (*n*=10), elephant (*n*=2), swine (*n*=7), and humans (*n*=24) were used for the study. Sixty-one *M. tuberculosis* isolates from human, primates, birds, and elephants were also included in the analysis. Based on 206 variant SNPs among the *M. bovis* strains, five major clusters consistent with epidemiologic and other strain-typing information were identified. Forty-nine of the 51 human *M. tuberculosis* isolates were identical at the 350 loci. This SNP based phylogeny provides new insights into the evolution of *M. bovis* and a gateway to study strain genotype-disease phenotype correlations that we next undertook in an *in vitro* infection model of the disease with 4 virulent *M. bovis* strains isolated from human (*n*=1), cattle (*n*=2) and deer (*n*=1). We investigated their virulence based on entry and survival in macrophages and relative gene expression profile of previously identified virulence genes.  The results revealed that the 4 strains had differential survival patterns in the macrophage mode coupled with a variation in relative gene expression profile for 6 six virulence-associated genes *mce4C*, *PE6*, *speE*, *mmpL12*. These studies led me to conclude that *M. bovis* isolates from diverse geographic origins and host species represent an array of genetic profiles that may potentially relate to their phenotypic variation.

Next, to improve resolution of genomic variability among *M. bovis* strains circulating in the United States, we undertook genome sequencing of 2 strains based on phylogeny developed in the SNP study. The genome of *M. bovis* Corsentino comprises a circular chromosome of 4307383 bp with average G+C content of 65.4% and with 4008 predicted protein-coding regions. The genome of *M. bovis* NE elk comprises a circular chromosome of 4302584 bp with an average G+C content of 65.4%  and with 4009 predicted protein coding sequences. Genome comparisons against the UK origin reference strain AF2122/97 did not reveal any unique genes or large sequence polymorphisms. A total of 1139 and 1184 SNPs were identified in Corsentino and NE elk genomes when compared to AF2122/97 genome, respectively. Comparison of *M. bovis* Corsentino and *M. bovis* NE elk genomes identified ~900 SNPs between them. Comparative genomics with other members of the *Mycobacterium Tuberculosis Complex* revealed a high percentage of sequence similarity between the strains. Thus, this study provides new evidence in favor of low genetic variability in this organism, suggesting variations in gene expression and post-transcriptional or post-translational regulation events as the likely sources of host specificity and phenotypic variation. Alternately, we reasoned that host genetics may contribute significantly to the range of pathology and transmission cycles seen in bovine tuberculosis. The restricted allelic variation among *M. bovis* strains also supports the contention that long-term host-pathogen co-evolution has likely selected a few successful organisms.

We next set out to explore the biology of granuloma by transcriptional profiling of *M. bovis* during its infection cycle within the host. This study aimed to decipher mechanisms of pathogenecity and to identify virulence markers of *M. bovis* and to associate host responses within a granuloma. Mediastinal lymph nodes from two experimentally infected cattle and two age matched control cattle were obtained for the study. The infected animals displayed characteristic granulomatous pathology consistent with bovine tuberculosis. Total RNA was extracted and enriched for bacterial mRNA. The enriched samples were submitted for next-gen sequencing employing the Illumina RNA-Seq Platform for transcriptomics profiling. The contigs obtained from the sequencing were assembled against the bacterial reference genome of *M. bovis* strain AF2122/97 and the bovine genome (*Bos taurus*) to build the gene expression profiles of the bacteria as well as the host. However the enrichment protocol used failed, leading to poor quality of bacterial sequences and no significant gene expression profile could be obtained for the host sequences. We would recommend a re-evaluation and standardization of RNA extraction techniques for future studies.

In conclusion, our studies identified that SNP based genotyping was successful in building a phylogeny among isolates of *M. bovis* from a variety of hosts and geographic locations. We further demonstrated that SNP genotypic variations correlated with intra-macrophage survival. Future studies should use these genotypically well-characterized strains to evaluate pathogenesis of bovine tuberculosis at the cellular and molecular levels. We demonstrated by complete genome sequencing of 2 isolates that this organism has undergone severe evolutionary bottleneck resulting in host specialization to the bovine host as indexed by the restricted allelic variation. Future analyses should study genomewide SNPs, their location on genomes, and whether they result in amino acid changes or not, to decipher the extent of selective evolution *M. bovis* has undergone in the bovine host. Finally, while our transcriptional analysis of the granuloma failed to provide information, these studies should be repeated with further refinements in techniques to enable the elucidation of host-pathogen interaction as it occurs inside a granuloma.

# TABLE OF CONTENTS

# LIST OF TABLES

## LIST OF FIGURES

# CHAPTER 1: Literature Review

Bovine tuberculosis, caused by *Mycobacterium bovis* (*M. bovis)*, is a well-known worldwide zoonotic disease of cattle. *M. bovis* has a broad host range including wildlife, domestic livestock, non-human primates and humans. The public health risk has been mitigated around the world by the introduction of pasteurization, but the disease continues to cause production losses when poorly controlled (50). Bovine tuberculosis is a disease of major economic importance in the developed world affecting animal productivity and trade of animal products (93). In many developed parts of the world like the UK, the USA, Australia and New Zealand, bovine tuberculosis been detected in wildlife that serve as potential reservoirs of infection. The existence of feral reservoirs of this insidious infection can have severe consequences for livestock due to spillover epidemics at wildlife-livestock interface (158).

## i. The Organism

*M. bovis* is a slow growing, aerobic, acid fast bacillus (236) belonging to the genus *Mycobacterium.* The organism is a part of *Mycobacterium tuberculosis* complex (MTBC) that includes several species of host specialized mycobacteria including, *M. tuberculosis, M. cannettii, M. africanum, M. bovis, M. pinnipedii, M. caprae and M. microti* that cause similar pathology in various mammalian hosts (133). The name *Mycobacterium* is derived from its waxy cell wall of which mycolic acids that form integral components engaged in the remarkable survival ability of these bacteria within infected hosts, virulence and evasion of immunity (134). Despite the different host tropisms, the *M. tuberculosis* complex is characterized by 99.9% or greater similarity at the nucleotide sequence level, and by virtually identical 16S rRNA sequences (28, 99, 224)

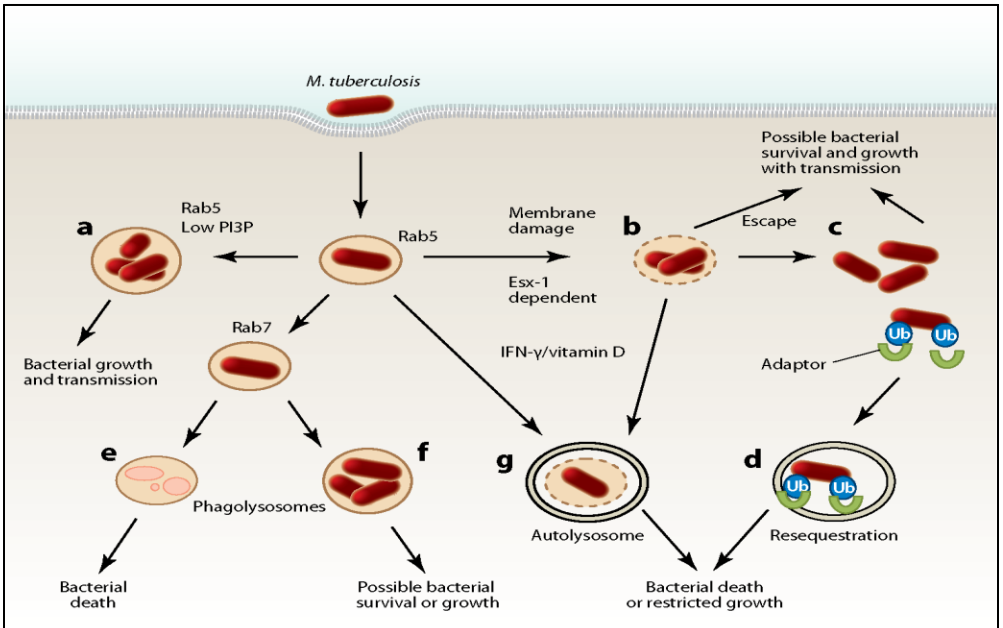
## ii. Nosology

Tuberculosis is usually a chronic debilitating disease in cattle, characterized by a period of latency before the disease develops but it can occasionally be acute and rapidly progressive, although early infections are often asymptomatic. Infections generally remain dormant for years and reactivate during periods of stress or in old age. Similarly, severe disease can develop in some deer within a few months of infection, while other deer do not become symptomatic for years (World Organization for Animal Health, OIE). In the late stages, common symptoms include progressive emaciation, a low–grade fluctuating fever, weakness and inappetence. Animals with pulmonary involvement usually have a moist cough that is worse in the morning, during cold weather or exercise, and may have dyspnea or tachypnea. In the terminal stages, animals may become extremely emaciated and develop acute respiratory distress. If the digestive tract is involved, intermittent diarrhea and constipation may be seen. In cervids, bovine tuberculosis may be a subacute or chronic disease, and the rate of progression is variable. In some animals, the only symptom may be abscesses of unknown origin in isolated lymph nodes, and symptoms may not develop for several years. In other cases, the disease may be disseminated, with a rapid, fulminating course (Adapted from Bovine Tuberculosis: Center for food security and public health, CFSPH, IA).

### iii. Pathogenesis of bovine tuberculosis

The main route of infection is via inhalation of contaminated aerosols and lungs are the primary organs affected, however, infection may also be acquired through the gastro-intestinal route and become systemic affecting other organs (O'Reilly, 1995 #1). In humans the main route of infection is via consumption of unpasteurized milk/milk products and close contact to infected animals (54).

Upon inhalation, respiratory droplets are deposited in the distal alveoli and the organism is presumed to first encounter and be ingested by alveolar macrophages (144). However other phagocytic cells recruited to the infected lung, including neutrophils, monocyte-derived macrophages, and dendritic cells can also ingest bacteria and probably play an important role in the outcome of the infection (203). The fate of the tubercle bacilli within the host phagocytic cells is an intriguing aspect of disease process and decides the outcomes of latency, active disease or clearance of infection. Pathology of *M. bovis* infection is characterized by typical granulomatous lesions with varying degrees of necrosis, calcification, and encapsulation (215). The possible fates of intracellular *Mycobacterium bovis* is presumed to be similar to that of *Mycobacterium tuberculosis* and is depicted in Figure 1.1 [adapted and re-produced with permission from Philips JA and Ernst JD (203)] . These include, (*a*) *M. tuberculosis* which can prevent phagosome maturation and grow in an early endosome–like compartment by inhibiting phosphatidylinositol 3-phosphate (PI3P) generation on the phagosome and impairing the recruitment of active, GTP-bound Rab7 while retaining Rab5; (*b*) the Esx-1 system which permeabilizes the phagosomal membrane, allowing direct cytosolic access; (*c*) in some cases, this process may result in the escape of the bacteria into the cytosol. The extent of cytosolic growth likely depends on the cell type; (*d*) in the case of *M. marinum*, the cytosolic bacteria are recognized by the host ubiquitin system and are resequestered in a membrane-bound compartment; (*e*) some ingested bacteria fail to prevent phagosome maturation, and they are delivered to the lysosome, where their replication is curtailed; (*f*) in certain contexts, they may be able to grow in lysosomes (*g*) Interferon (IFN)-γ and vitamin D can overcome the early endosome–like arrest of *M. tuberculosis*, thereby promoting delivery of bacteria to autolysosomes, where growth is curtailed.



Philips JA, Ernst JD. 2012.
Annu. Rev. Pathol. Mech. Dis. 7:353–84

Innate immunity plays an important role in the host defense against mycobacteria, and the first step in this process is recognition of the bacterium by cells of the innate immune system. Several classes of pattern recognition receptors (PPRs) are involved in the recognition of *M. tuberculosis*, including Toll-like receptors (TLRs), C-type lectin receptors (CLRs), and Nod-like receptors (NLRs) (114). Among the TLR family, TLR2, TLR4, and TLR9 and their adaptor molecule MyD88 play the most prominent roles in the initiation of the immune response against tuberculosis. In addition to TLRs, other PRRs such as NOD2, Dectin-1, Mannose receptor, and DC-SIGN are also involved in the recognition of *M. tuberculosis* (114). The phagocytosis and the subsequent secretion of IL-12 are processes initiated in the absence of prior exposure to the antigen and hence form a component of innate immunity. The other components of innate immunity are natural resistance associated macrophage protein (Nramp), neutrophils, natural killer cells (NK) and plasma lysozymes (207). The host acquired immune response can broadly be categorized into two groups: cytokines and cellular response. Post-phagocytosis alveolar macrophages are stimulated primarily by ligation of TLR2 and TLR4 with the major mycobacterial cell wall component lipoarabinomannan (LAM) to secrete proinflammatory cytokines like TNF-α, IL-1β, IL-6 and IL-12 leading to a pyroptosis response (152). IFN-γ, a T-cell cytokine that activates macrophages to produce reactive oxygen and nitrogen species, is probably the most important cytokine in immune response to mycobacteria (215). Upon the onset of acquired immunity, macrophages are activated by IFN-γ, mainly from T lymphocytes, which, among other effects, activate macrophage antimicrobial mechanisms. CD4 T cells, which recognize peptide antigens bound to MHC/HLA class II molecules, are essential for protective immunity to tuberculosis. This is supported by the fact, that HIV infection increases the rate of progression from latent to active tuberculosis by 5- to 10-fold, and only a modest reduction in CD4 T cell counts are sufficient to increase the incidence of tuberculosis (120). Several subsets of CD4 T cells have emerged as important contributors of the host immunity. These include Th1 cells that secrete IL-12 and IFN-α, Th17 that secrete IL-17 and Tregs (Regulatory T-cells) (110, 152). Studies with *M. tuberculosis* have demonstrated antigen-specific CD8 T cells produce cytokines such as IFN-γ in response to stimulation and can lyse *M. tuberculosis* infected antigen-presenting cells (125).

### iv. Zoonotic impacts of *M. bovis*

The World Health Organization (WHO) in conjunction with Food and Agricultural Organization of the United Nations (FAO) and World Organization for Animal Health (OIE) recently classified bovine tuberculosis as neglected zoonosis with special reference to developing countries. Bovine tuberculosis is an endemic disease in livestock in many African countries (168). In most developing countries *M. bovis* infection remains an uninvestigated problem and an accurate epidemiological and zoonotic impact is lacking. *Mycobacterium bovis* as a source of human infection is likely under reported due to specific nutrient requirements of this organism that are not used under routine laboratory conditions to isolate *M. tuberculosis* from sputum or other primary clinical sources. *Mycobacterium bovis* is unable to use glycerol as carbon source that is commonly used in culture media for *Mycobacterium tuberculosis* growth and needs supplementation with pyruvate. In the developed world bovine tuberculosis is a disease of significant economic importance affecting animal productivity and trade of animal products (158). *Mycobacterium bovis* as a zoonosis is a concern in the pastoral settings of the developing world where animal-human interface is close, and HIV prevalence is high (179). A recent study of all human tuberculosis cases in the USA from 1995 through 2005 estimated that only 1.4% of cases were being caused by *Mycobacterium bovis (96)* . In San Diego, California, over 45% of all culture-confirmed tuberculosis cases in children and 8% of all tuberculosis cases were found to be due to *Mycobacterium bovis* (212).

Tuberculosis in wildlife can pose serious difficulties for control and eradication of bovine tuberculosis. Among the developed countries, deer in United States, European wild boar in Spain, badgers in the UK and brush-tailed possums in New Zealand have been identified as some of the potential reservoirs and vectors of *M. bovis* infection although cases of *M. bovis* have been reported in more than 40-free ranging wild animal species worldwide (158). This appears to be the underlying theme for the maintenance and periodic spillover of the infection into domestic animals (158, 184, 187). *M. tuberculosis,* the cause of human tuberculosis is responsible for 2 million deaths annually worldwide and it is currently estimated that one third of the world's population is infected with *M. tuberculosis* with 9 million people each year becoming sick with TB (CDC, WHO). Human tuberculosis caused by *M. bovis* is clinically and pathologically indistinguishable from *M. tuberculosis*. Although the exact estimate of human *M. bovis* infections is hard to predict, the WHO reported in 1998 that 3.1% of tuberculosis cases in humans worldwide are attributable to *M. bovis* and that 0.4-10% of sputum isolates from patients in African countries could be *M. bovis*. The epidemiological pattern of bovine tuberculosis can be quite complex involving interaction between domestic animals, humans and wildlife. Recently Evans *et al* 2007 (64), reported a cluster of human TB caused by *M. bovis* in the UK suggesting human-human transmission, however human-to-human transmission of bovine tuberculosis is an exceptional event in the absence of immunosuppression (83). The post pasteurization era has seen a significant drop in cattle-to-human transmission of *M. bovis* infections with the risk estimated to be almost negligible in developed countries, and also aided by the mandatory test and slaughter of infected livestock under various tuberculosis eradication programs. However in developing countries, poverty, malnutrition and immunosuppression due to HIV/AIDS is a known complication in humans affected by *M. tuberculosis* and has recently emerged as an aggravating factor in *M. bovis* infections in humans at the livestock-human interface. The information available on the global animal health information database of the OIE states that 128 out of 155 countries reported the presence of *M. bovis* infection and/or clinical disease in their cattle population during the period between 2005 and 2008. Thus the continued maintenance and transmission of *M. bovis*, which is resistant to two frontline drugs used to treat tuberculosis – isoniazid and pyrazinamide, threatens to return tuberculosis-associated mortality rates to those of the pre-antibiotic era, and thereby necessitates rapid, thorough national and international surveillance of strain distribution dynamics in the population.

### v. Bovine Tuberculosis in USA

In the US, the bovine tuberculosis eradication program was initiated in 1917 and mandated pasteurization of milk and slaughtering of infected herds. By 1941, every county in US was officially accredited free of bovine tuberculosis and the infection rate was reduced to about 0.5%. Since then the first epidemic of bovine tuberculosis was reported in 1995 in the state of Michigan. In the last decade about 61 infected herds have been identified in different states of the US with Michigan and Minnesota topping the list at 32 and 12 infected herds, respectively. Minnesota acquired its tuberculosis free status in 1971. Its first reported outbreak since then came in the year 2005, reported in the Roseau and Beltrami counties in the northwest. In 2006, USDA increased the regulatory testing of cattle and declared the modified accredited advanced (MAA) status for MN. Identification of 4 infected herds in 2008 led to a downgrade of MN’s MAA to modified advanced (MA). This brought in the requirement for annual whole herd tuberculosis testing maintaining up-to-date contact information with the board of animal health for cattle producers. However, in October 2008 MN was approved for the split state status, where a large part of the state was upgraded to MAA and a smaller section in the high prevalence northwestern Minnesota remained Modified Accredited. As of May 2012, the state has obtained a bovine TB free status; however the identification of *M. bovis* infection in free ranging deer in MN adds to the complexity of the transmission and maintenance of this disease in animal populations and poses a serious threat to the eradication of bovine tuberculosis. According to the latest USDA TB eradication status (May 1, 2012) of the 50 states in the country, only Michigan holds the MA/MAA status and California holds MAA status for bovine TB. As for cervid TB, all the 50 states remain with the MA status.

The biological basis of bovine tuberculosis spread among cattle and deer populations is unclear and warrants further investigation. There is a desperate need to better understand the genetic attributes of *M. bovis* that contribute to their virulence, increased velocity of spread, and enhanced environmental survival, all of which would aid in developing better control strategies.

### vi. Evolutionary predictions for the *Mycobacterium tuberculosis* Complex organisms

*M. bovis* belongs to the *M. tuberculosis* complex (MTC) group of organisms - a family of 'ecotypes' of genetically very closely related mycobacteria. Each ecotype is adapted to a specific host species or group, although inter-species transmission can occur (158, 223). These ecotypes or "sibling" species of MTC include *M. tuberculosis sensu stricto* (s.s.), *M. africanum*, *M. bovis*, *M. caprae*, *M. microti*, *M. pinnipedii*, and *M. canettii* that cause tuberculosis in humans and animals (8, 67, 170, 224). However, *M. bovis* is unique among the MTC group as having wide host range including most mammalian species, humans, livestock and wild animals (46, 62, 169). Recent studies suggest that the common ancestor of *M. tuberculosis* complex emerged from its progenitor perhaps 40,000 years ago in East Africa. Some 10,000-20,000 years later, two independent clades evolved, one resulting in *M. tuberculosis* lineages in humans, while the other entered animal hosts and co-evolved in these new hosts over millennia resulting in the diversification of its host spectrum and formation of other *M. tuberculosis* complex member species including *M. bovis* (87, 259). This adaptation to animal hosts and co-evolution probably coincided with the domestication of livestock some 13,000 years ago.

In modern history, however, cattle have served as principal reservoir species for *M. bovis.* (158). Animal movement and trade have facilitated the spread of *M. bovis* infection across borders. In 1997 Sreevatsan *et al* (224) proposed an evolutionary pathway for *M. tuberculosis* complex characterized by polymorphisms in the KatG codon 463(194) and GyrA codon 95 (Thr) in the MTC bacterial genomes, in which they hypothesized that *M. bovis* is ancestral to the modern *M. tuberculosis*. A similar hypothesis was put forward by Diamond *et al* 2002 (55) and Brosch *et al* 2002 (28) that states that tuberculosis has evolved from an originally animal disease to human disease. On the other hand findings of Wirth *et al* 2008 (259) indicate that tuberculosis first emerged in humans and was subsequently transmitted to animals. Hence the evolutionary origin of *M. bovis* has so far remained ambiguous. Regardless, the zoonotic nature of this organism is real and requires further investigation.

### vii. *M. bovis* – BCG, the attenuated vaccine strain

Bacillus Calmette-Guérin (BCG) is the only available live attenuated vaccine used for the prevention of tuberculosis, derived from virulent *Mycobacterium bovis* to which it is closely related. It was derived by the repeated subculture of a strain of *Mycobacterium bovis* on potato slices soaked in glycerol and ox bile (31) leading to the in vitro accumulation of mutations and ultimately attenuation. Although the BCG vaccine has been one of the most widely used vaccines in the world for over 40 years, the genetic basis of BCG's attenuation has not been completely elucidated. The first study by Mahairas et al 1996, showed that the BCG strain has a regulatory mutation in a distinct genomic region described as RD1 (region of difference 1) (137). The RD1 locus was shown to be deleted from all BCG strains but present in all virulent strains of *M. bovis* and *Mycobacterium tuberculosis* studied. Subsequent work has shown that this deletion played a major role in the attenuation of BCG (126, 206). A major step toward defining the molecular basis of attenuation in BCG was the completion of the genome sequences of *M. bovis* BCG Pasteur and virulent M. bovis AF2122/97 (27, 77) . Genomic comparison of BCG Pasteur with *M. bovis* AF2122/97 identified a range of mutational differences, including deletions, duplications, and single nucleotide polymorphisms.

### viii. Molecular Sub-typing and Phylogenetic Analysis of MTC

The MTC group of organisms is highly clonal with little to no exchange of chromosomal DNA between them (17, 42, 86, 223) thus it is assumed that any mutation within the ancestral strain will be retained within the descendants and can be used to identify clonal complexes (17). A series of deletions known as regions of difference (RD) have been used to identify phylogenetic relationships between members of MTC group (28) and for *M. tuberculosis* different lineages and sublineages have also been characterized by specific deletions (244). Molecular sub typing of *Mycobacterium tuberculosis Complex* using spoligotyping (49, 90, 106, 107), mycobacterial interspersed repetitive units (MIRUs) (3, 49, 90, 167), and/or Variable number tandem repeat (VNTR) (3, 90, 95), has provided robust strain differentiation of the MTC group of organisms.

The current gold standard for MTC typing is the *Insertion Sequence 6110* Restriction Fragment Length Polymorphism (IS*6110* RFLP). IS*6110* is an insertion element found exclusively within the MTC; the assumption is that this restriction site is a result of the lack of genetic exchange with other mycobacterial species and that this element is randomly distributed in the genome (48, 146). The PCR amplified IS*6110* DNA is identified by a restriction endonuclease and electrophoresis based assay. The main drawbacks of this method are that it requires extensive strain cultivation, is technically and time demanding and expensive. It has low discriminatory power especially for *M. bovis* strains, which have a low copy number of IS*6110* (188). PCR based spoligotyping, is a method based on amplification of a Direct Repeat (DR) region and hybridization to oligonucleotides complimentary to the variable spacer regions between these DRs (106). The method is simple and robust, but the discriminatory power remains unsatisfactory and is unable to provide sufficient information to accurately establish genotypic relationships between clinical isolates (251). Methods based on minisatellites that contain Variable Number of Tandem Repeats (VNTRs) have been demonstrated to be effective and portable methods for MTC strain typing. However this method remains more suitable for global epidemiological surveillance and is technically challenging to apply to local epidemic investigations. Thus most of these techniques also fail to qualify as robust tools for the resolution of phylogenetic relationships.

Identifying genetic variants is expected to aid in unraveling *M. bovis* evolution, epidemiologically define its zoonotic importance, and gain insight into *M. bovis* allele-pathogenicity trait (108). Global lineages of *M. bovis* have been described as African 1, African 2, European 1 and European 2 that are established in different geographic regions of the world (17, 168, 211, 221, 222). However, population genetics of *M. bovis* related to strain variation based on genome wide sequence variations remains largely undefined. In the recent past, polymorphisms like Large Sequence Polymorphisms (LSPs) (2, 28, 109, 164, 178, 244) and single nucleotide polymorphisms (SNPs) (68, 76, 85, 86, 224) in the MTC genomes have been recognized to serve as good evolutionary genetic markers and used to reveal phylogenetic relationships between isolates, however these have mainly focused on *M. tuberculosis* and *M. bovis* -BCG strains. By extension, such genetic information specific to *M. bovis* isolates would guide in assessing their prevalence through space and time in a variety of host species and provide markers for molecular epidemiologic and virulence assessment (163). *M. bovis* differs from *M. tuberculosis* in key biological properties, such as transmissibility, host range and antigenic variability, where factors unique to *M. bovis* are expected to play a role (22).

Despite extensive knowledge of many aspects of tuberculosis, the diagnosis and identification of infecting mycobacterial species is a challenging task. Assuming that they are all derived from a common ancestor, it is intriguing that some are exclusively human tuberculosis whereas others have a wide host range e.g. *M. bovis* (28). Typing of strains is very important for disease surveillance and control. Molecular tools for studying strains are currently being widely characterized and are emerging to provide answers to our gaps in knowledge regarding this pathogen. The identification of clonal complexes of *M. bovis* dominant in larger geographic locations indicates there could be other groups localized to other regions of the world. However the genetic forces that shape the population structure of this organism in a given geographic location need to be described. The practical implications of such knowledge is fundamental for understanding the spread of infection and designing location specific disease distribution, farming practices and specific control measures.

### ix. Single nucleotide polymorphisms (SNPs)

A single nucleotide polymorphism or SNP is a single nucleotide variation at a specific location in the genome that is by definition found in more than 1% of the population (26). Recent studies involving genome-wide analysis of SNPs are attempting to overcome the limitations of previously described techniques and made possible by the availability of whole genome sequence data for MTC strains. SNPs can provide rich information on genetic variation and are important tools in evolutionary studies; they are relatively easy to assay and provide for large-scale population genetics studies (85, 86). SNPs have also been hypothesized to play a role in the molecular attenuation of the BCG vaccine and hence may also provide insights for vaccine development (76). High-throughput SNP analysis can be particularly beneficial for confirming associations between specific SNPs and a phenotype of interest such as drug resistance. SNP analysis is becoming increasingly important for studies of drug resistance, evolution and molecular epidemiology of the MTC group. Most causes of drug resistance in *M. tuberculosis* appear to be the result of SNPs in particular target genes (92). Gutacker *et al* 2002 (85, 86) reported that synonymous SNP genotyping rapidly describes relationships among closely related strains of pathogenic microbes and allows construction of genetic frameworks for examining the distribution of biomedically relevant traits such as virulence, transmissibility and host range. In their study on *M. tuberculosis*, 432 MTC strains from global sources were genotyped on the basis of 230 synonymous SNPs identified by genome comparison and the clustering pattern relative to their epidemiologic data. In another study, genetic relationships between 5069 *M. tuberculosis* strains recovered from patients enrolled in 4 population-based studies in the US and Europe, was observed by analysis of 36 sSNPs. This SNP-based phylogenetic framework provided new insight into worldwide evolution of *M. tuberculosis* and a gateway for investigating genotype-disease phenotype relationships in large number of samples. Filiol *et al* 2006 (68) analyzed a global collection of *M. tuberculosis* strains using 212 SNP markers. Eventually they designed an algorithm to identify two minimal sets of either 45 or 6 SNPs that could be used in future investigations to enable global collaborations for studies on evolution, strain differentiation and biological differences of *M. tuberculosis*. Pelayo *et al* 2009 (76) used comparative genomics to identify over 700 SNPs that differed between MBO and BCG strains. SNPs showed phylogenetic clustering that was consistent with geographical origin of the strains refining previous BCG strain genealogy and discriminated between virulent *M. bovis* strains isolated in UK and France. SNPs were analyzed to further unravel mechanisms responsible for attenuation of tuberculosis vaccine BCG. Monot *et al* 2009, (160) showed the presence of 78 informative SNPs surveyed in about 400 isolates enabling classification of *M. leprae* into 16 SNP subtypes of limited geographic distribution that correlated with the patterns of human migrations and trade routes. In a study by Gicquel *et al* 2008, SNPs analysis of 3R genes in *M. tuberculosis* strains from across the world made it possible to distinguish between 80% of clinical isolates. SNPs in 3R genes may play an important role in evolution of highly clonal bacteria and also further facilitate epidemiologic studies of these bacteria through development of high-resolution tools (57). Insertion sequence *IS900* is used as a target for the identification of *Mycobacterium avium* subspecies *paratuberculosis.* Many studies have reported SNPs within *IS900*; a recent study analyzed the *IS900* sequence in a panel of isolates representing all three strains of MAP (I, II, III) and revealed conserved type-specific polymorphisms that could be utilized as a tool for diagnostic and epidemiological purposes (33). Kaser *et al* 2009 (108), analyzed the *M. ulcerans* (cause of Buruli Ulcer) genome for large sequence polymorphism (LSP) haplotype-specific insertion sequence elements among 83 *M. ulcerans* strains and identified SNPs that could differentiate between regional strains, in this highly clonal organism. SNPs have also recently been described for the malaria parasite *Plasmodium falciparum*, which like MTC exhibits restricted genetic diversity and this is considered key to its success as a pathogen. SNPs have been used as a tool for studying demographic history of the parasite, its population structure and linkage equilibrium within its genome (176). One of the first studies of SNPs in *Bordetella pertusis*, a genotypically homogenous pathogen, identified over 1500

SNPs distributed throughout the genome including 5 synonymous SNPs in virulence genes. This study laid the foundation for studying SNPs in this species of bacteria to further understand its evolution and diversity (Maharjan, 2008 #103).

Information derived from this comprehensive comparative genomic analysis to identify single nucleotide polymorphisms (SNPs) in the *M. bovis* genome to develop phylogeny will help investigate hypotheses fundamental to understanding the transmission and survival traits of *M. bovis*. The estimates of overall genetic relationships for all strains provided by SNP genotyping will make possible the mapping of traits onto the *M. bovis* phylogenetic tree, which can be done for multiple virulent isolates in a an effective manner. Better understanding of the underlying principles that determine pathogenesis and transmission of bovine TB will directly aid in better diagnostics, vaccine candidates, provide biologically valid data for risk analysis of transmission via feed or water and risk modeling for disease transmission, and therefore lead to better science-based control strategies.

### x. Genotype-Phenotype Associations

Evidence from genotype-phenotype studies suggests that genetic diversity in pathogens have clinically relevant manifestations that can impact the outcome of infection and epidemiologic success (142). The understanding of the molecular basis of this pathogen's success in causing disease is necessary to implement effective control strategies and therapeutic options. One reason for *M. bovis* to be considered a dangerous microbe lies in its ability to survive latently within infected hosts amid a robust host immune response. Infection of a host with *M. bovis* is initiated following the inhalation of droplets (aerosols) containing a small number of bacilli (110). Once in the lung, bacilli are internalized through phagocytosis by the resident macrophages of the lung, the alveolar macrophages. Alveolar macrophages activated by the appropriate stimuli can effectively transfer the phagocytosed *M. bovis* to the destructive environment of lysosomes, but some bacilli are able to escape lysosomal delivery and survive within the macrophage (5, 110, 213). Infected macrophages can then either remain in the lung or are disseminated to other organs in the body. During the pathogenic process *M. bovis* is thought to be exposed to a number of different stress conditions like acidic pH, reactive nitrogen and oxygen species and nutrient starvation etc (117). Many of the proteins induced in response to stress are thought to be involved in survival of the pathogen inside the host. Studies have also looked at expression of genes encoding a range of functional activities and are known to vary between *M. bovis* and *M. tuberculosis* (79) with a possible biological impact of this variation on strain phenotype. Functional genomic techniques such as proteomics and transcriptomics allow the biology of MTC to be explored on a global scale, giving information as to which genes and proteins are expressed under which conditions (20, 175). Understanding the transcriptome helps to interpret the functional elements of the genome and revealing molecular constituents of cells and understanding disease. Previous study (79) comparing the trancriptomes of *M. tuberculosis* and *M. bovis* revealed differential expression of genes encoding a range of functions with biological implications like host tropism, cell wall and secreted proteins etc.

Accumulating evidence from genotype-phenotype studies suggests that genetic diversity among *M. tuberculosis* isolates may have clinically relevant manifestations that could impact on outcome of infection (34, 56, 139, 142, 190). Strain-dependent variations in replication rates, immunogenicity, pathogenesis, survival, and transmission potential have been described elsewhere (56, 243). Recent reports have associated strain-specific microbial factors that alter the host immune response with enhanced virulence (177, 210). An in vitro study  (205) in human macrophages, comparing phylogenetically defined ‘‘ancient’’ and ‘‘modern’’ lineages, noted reduced cytokine responses in the latter group and genetic diversity appeared to have functional consequences during intracellular infection of bone marrow–derived macrophages (98), where transcriptomic profiles were lineage specific (142). A biologically active lipid species- a polyketide synthase-derived phenolic glycolipid (PGL) produced by a subset of *M. tuberculosis* isolates belonging to the W-Beijing family is known to be ‘hyperlethal’ in murine disease models with the disruption of PGL leading to loss of hypervirulence phenotype without affecting bacterial load during disease (208). Thus, it appears that genetically diverse *M. tuberculosis* clinical isolates can differ in their phenotypic characteristics. Although phylo-geographical lineages of *M. tuberculosis* and *M. bovis* have been described, it is still unclear whether there is a microbial basis to explain why some variants cause widespread disease and other closely related strains remain limited in spread. Also research so far has not been able to delineate the correlation between the genotype of a particular strain to its ability to successfully exploit the host macrophage environment.

### xi. Lessons learned from the complete genome sequence of *M. bovis*

The research advances in genome sequencing of pathogenic bacteria have helped reveal their genetic blueprints offering unparalleled insights into their virulence factors. The PATRIC database (http://patricbrc.vbi.vt.edu) currently hosts 131 genomes of the genus Mycobacteria of which 42 are completed genomes. Of these 42 completed genomes, 24 belong to the *M. tuberculosis* complex including 18 strains of *M. tuberculosis* s.s, 3 *M. bovis BCG* and 1 each of *M. canetii, M. africanum* and *M. bovis*. Strain AF2122/97 is the only available genome sequence of a virulent *M. bovis* isolate, obtained from a infected cow in the UK (77). The genome size is 4,345,492 bp and contains around 4000 genes accounting for > 91% coding capacity of the genome, revealing potential virulence factors and antigens. Strikingly, the genome sequence of *M. bovis* is >99.95% identical to that of *M. tuberculosis*, but deletion of genetic information has led to a reduced genome size (77). Furthermore, there are no genes unique to *M. bovis*, implying that differences at the transcriptional level or post transcriptional modifications may be the key to the host tropisms of human and bovine bacilli (77). The genome sequence therefore offers major insight on the evolution, host preference, and pathobiology of *M. bovis*.

Whole-genome sequencing provides detailed information on genetic differences between bacteria. The study of genetic variability within natural populations of pathogens may provide insight into their evolution and pathogenesis. The availability of multiple *M. tuberculosis* genomes (Cole, 1998 #77; Gordon, 1999 #208; Fleischmann, 2002 #212) and their comparative analyses has revealed novel information about associations between strains, their host populations, and environment. Various studies have utilized a variety of low and high resolution comparative genome techniques to identify differences in the genomes of *M. bovis* BCG vaccine strains and *M. tuberculosis* laboratory and clinical strains identifying a number of sequence differences between the different mycobacterial species and strains (82, 97, 137, 202).

An analysis (109) of genomic deletions among a population of pathogenic *M. tuberculosis* provided a novel perspective on genomic organization and evolution when compared to the reference strain H37Rv. The study implied that deletions are likely to contain ancestral genes whose functions are no longer essential for the organism's survival, whereas genes that are never deleted constitute the minimal mycobacterial core genome. Their overall finding was that as the amount of genomic deletion increased, the likelihood that the bacteria could cause pulmonary cavitation decreased, suggesting that the accumulation of mutations tends to diminish their pathogenicity. However, the deletions here represent only a subset of the total genetic variability; as they do not include sequence present in the clinical isolates but absent from the reference strain H37Rv. Current evidence based on sequence-based analysis suggests that as a species *Mycobacterium tuberculosis* exhibits very little genomic sequence diversity with remarkably few single nucleotide polymorphisms within coding regions, including genes coding for targets of the host immune system (Brosch, 2000 #211; Musser, 2000 #113; Sreevatsan, 1997 #127). Several studies have also described large sequence polymorphisms (LSPs) among the *M. bovis* BCG vaccine strains and virulent *M. bovis* as well as among other tubercle bacilli (13-15, 82, 109, 137) that have provided useful genetic markers in phylogenetic analyses. Most genetic variability that has been detected is associated with transposable elements and drug resistance phenotypes (Beck-Sague, 1992 #213; Fleischmann, 2002 #212; Jereb, 1993 #214; Almeida Da Silva, #216). This would lead to an assumption that *M. tuberculosis* should exhibit very little phenotypic variation in immunologic and virulence factors. However, evidence of phenotypic diversity among clinical isolates conflicts with this hypothesis (142). In the *M. tuberculosis* strain H37Rv, the G+C content plotted across the genome is found to be relatively uniform (42). The only areas with exceptionally high G+C content (>80%)

corresponded to the PGRS (polymorphic G+C rich sequences) gene family (Brosch, 2000 #211) which were unique to mycobacteria are largely implicated to play a role in antigenic variation of this bacteria (237). Regions with higher A+T content were found mostly in the housekeeping genes. This suggests that the acquisition of virulence genes in the form of pathogenecity islands by horizontal transfer (89) may not have occurred in MTC. Prophages have also been described in *M. tuberculosis* genome (21, 225) and it is speculated that one of the prophages was lost during the attenuation process of *M. bovis* to *M. bovis*- BCG (137). The contribution of existing prophages in the *M. tuberculosis* genome to the disease pathogenesis remains questionable as they are absent in some of the clinical isolates and such comparative studies are lacking. Database comparisons (i.e similarity of *M. tuberculosis* genes to other genes of known function) have led to the tentative attribution of function to about 40% of the total genes and these are predominantly involved in core metabolism. Another 44% genes have some functional information or similarity to other gene functions; although they belong to the class of conserved hypothetical proteins (29). The remaining genes are considered to be unique to mycobacteria and these constitute about 20% of the chromosome and are devoted to genes encoding two different classes of proteins: enzymes involved in fatty acid metabolism and acidic, glycine-rich polypeptides of unknown function the PE and PPE proteins (42, 145). These two large gene families of PE and PPE proteins show no significant similarity to proteins of known function and contain conserved Pro-Glu and Pro-Pro-Glu motifs at the amino terminus respectively (40, 42).  There are approximately 99 members of the PE family and 61 of these belong to the PGRS subfamily, containing multiple tandem repeats of the tripeptide Gly-Gly-Ala. The PPE family has 68 members that fall into at least 3 subfamilies of which the most intriguing is the MPTR (major polymorphic tandem repeat) class and several of these proteins are predicted to consist of >3000 amino acids with repeat motifs. The PGRS members are also large proteins and can contain upto 1400 amino acid residues. These sequences of these genes are responsible for the extensive polymorphisms in the *M. tuberculosis* complex and are hypothesized to play a role in antigenic variation and evasion of host immune response

(41) . About 51% of the total genes are said to have arisen from gene duplication events, which is similar to other eubacteria like *Escherichia coli* and *Bacillus subtilis* (119, 143), however the degree of sequence conservation is much higher suggestive of extensive functional redundancy or that *M. tuberculosis* is of recent evolutionary descent (40, 170, 224). The basis for this remarkable genetic homogeneity is interesting but not clearly known and reflects either a very efficient DNA repair system or replication machinery of very high fidelity (40).

To date research has primarily focused on comparative genome analysis of *M. tuberculosis* sequenced strains and clinical isolates of *M. tuberculosis* or studies focused on *M. bovis*-BCG strains to elucidate mechanisms of their molecular attenuation. *M. bovis*, primarily a pathogen of veterinary importance has received relatively less attention, however given the public health challenge and reemergence of bovine tuberculosis in many developed countries, it warrants focusing on the genome dynamics of this pathogen. With the availability of the genome sequence of *M. bovis*, one can address the genetic basis of key phenotypic traits of the bovine tubercle bacillus (93). Comparative analyses have shown that deletion of genetic information has been the dominant force in shaping the genome, with *M. bovis* not presenting any unique genes *per se* compared with other members of the *M. tuberculosis* complex (77). However substantial studies in this direction are hindered by the lack of availability of more than one *M. bovis* genome sequence for comparative studies. A reference *M. bovis* strain from the USA needs to be sequenced to direct studies relating to population genetics and strain dynamics of this pathogen. With the combination of molecular epidemiological data and recent advances in mycobacterial genomics and ultra deep massively parallel next generation sequencing technologies, can provide insights into genetic and phenotypic diversity and phylogeny of *M. bovis* strains circulating globally. The goal of bacterial population genetic research is to understand the relationships between genetic diversity, clonal lineages and bio-medically relevant phenotypes such as virulence, transmissibility, host specialization and evolutionary success.
Comparative genome analysis thus can provide new insights for better understanding the

evolutionary events of this species and improving drugs, vaccines, diagnostics and most importantly tools for controlling bovine tuberculosis.

### xii. Transcriptomics of *M. bovis* and the host

The aims of trancriptomics are to catalogue all species of transcripts, including mRNAs, non-coding RNAs and small RNAs to determine the transcriptional structure of genes and to quantify the changing expression levels of each transcript during the disease process. Members of the *Mycobacterium tuberculosis* complex show distinct host preferences, yet the molecular basis for this tropism is unknown. Comparison of the *M. tuberculosis* and *M. bovis* genome sequences revealed no unique genes in the bovine pathogen per se, indicating that differences in gene expression may play a significant role in host predilection. Comparative analyses of the *M. tuberculosis* and *M. bovis* genomes have revealed the basis for distinguishing phenotypes such as the pyruvate requirement of *M. bovis* in glycerol-based media, or the reason for eugenic / dysgonic colony morphology (111). However, comparative genomics in itself does not reveal the basis for the complexity of phenotype between *M. tuberculosis* and *M. bovis*. Extra information needs to be layered onto the genome data, such as gene expression profiling, metabolic network analyses, signaling pathways, etc., to fully explore the biology of these pathogens (79).

In a transcriptomic based study (209) comparing the bovine and human tubercle bacilli, differential expression was detected in 258 genes, representing a 6% of the total genome. The main variations were found in genes encoding proteins involved in intermediary metabolism and respiration, cell wall processes, and hypothetical proteins. Interestingly, compared to *M. tuberculosis*, the expression of a higher number of transcriptional regulators was detected in *M. bovis*. Studies have also revealed that the human and bovine pathogens show differential expression of genes encoding a range of functions, including cell wall and secreted proteins, transcriptional regulators, PE/PPE proteins, lipid metabolism and toxin-antitoxin pairs (79). A study comparing pathogen transcriptomes of a virulent *Mycobacterium bovis* isolate to that of the attenuated vaccine strain BCG showed 133 genes with a minimum 2 fold difference of expression (22).

mRNA expression among clinical isolates of *M. tuberculosis* demonstrates that genes with important functions like the  T-cell antigens, those involved in lipid metabolism, bacterial stress response and PE/PPE genes can vary in their expression levels between strains grown under identical conditions (75, 115, 166) providing evidence for intra-species genetic diversity and strain-to-strain variation.

The elucidation of the bacterial transcriptome to identify virulence factors is essential to improve diagnostic and therapeutic tools. Similarly, functional genomics studies that highlight molecular mechanisms governing the host response to *M. bovis* infection are equally important.  These studies can enable the identification of novel transcriptional markers of bovine tuberculosis that can also augment current diagnostic tests and surveillance programs. Transcriptomic approaches have been used to identify gene expression profiles to define biomarkers of tuberculosis in mice, primates and humans in different infection conditions (247, 267). Likewise, studies in cattle, aiming to determine gene expression profiling, have been reviewed by Waters *et al*  focusing on ex vivo studies and macrophage infection (252).

Microarray of mRNA abundance was used to investigate the gene expression program of peripheral blood mononuclear cells (PBMC) from cattle infected with *M. bovis* (151). Analysis of total gene expression changes across a 24 hour time course infection revealed an immunosuppressive pattern of gene expression in response to stimulation with bovine purified protein derivative (PPD). Perturbation of the PBMC transcriptome was most apparent at time points 3 hours and 12 hours post-stimulation, with 81 and 84 genes differentially expressed respectively. This pattern of temporal gene expression is consistent with results reported for Johne's disease-positive cattle (51); in vitro stimulation with *M. avium sub* species *paratuberculosis* induced rapid changes in infected cattle PBMC gene expression within 2–4 hours after exposure to antigen. A more recent study by Magee et al (136) compared the gene expression profiles of *M. bovis* challenged monocyte-derived-macrophages (MDM) with that of non-challenged control MDM and identified 3,064 differentially expressed genes 2 hours post-challenge, with 4,451 and 5,267 differentially expressed genes detected at the 6 hour and 24 hour time points, respectively (adjusted *P*-value threshold ≤0.05). Notably, the number of downregulated genes exceeded the number of upregulated genes in the *M. bovis*-challenged MDM across all time points. These previous (136, 149-151) studies have shown that *M. bovis* infection is associated with the repression of host gene expression.

*Mycobacterium tuberculosis* survives in antigen-presenting cells (APCs) such as macrophages and dendritic cells. Macrophage recognition of mycobacteria occurs through the interaction of mycobacterial pathogen-associated molecular patterns (PAMPs) with host pathogen recognition receptors (PRRs), such as the Toll-like receptors (TLRs), expressed on the macrophage cell surface (91). PRR activation induces signaling pathways resulting in the production of endogenous NF-κB-inducible cytokines that promote an adaptive immune response characterized by the release of proinflammatory interferon-gamma (IFN-γ) from T cells and natural killer (NK) cells (44). In turn, IFN-γ induces microbicidal activity in infected macrophages and enhances the expression of major histocompatibility complex (MHC) class I and II molecules necessary for the presentation to T cells of mycobacterial antigens on the macrophage surface (70). These molecular mechanisms culminate in the formation of granulomas-organized complexes of immune cells comprised of lymphocytes, non-infected macrophages and neutrophils that contain mycobacterial-infected macrophages and prevent the dissemination of bacilli to other organs and tissues—however, in most cases the pathogen is not eliminated by the host (91, 136). The persistence of mycobacteria within granulomas is the hallmark of tuberculosis infection. This latent infection can progress to active tuberculosis whenever the host immunity is compromised. Survival within the granulomas- through the subversion of the host immune response is achieved through a diverse set of molecular mechanisms.

Research using cDNA microarrays has generally focused on gene expression profiles of *M. bovis* challenged macrophages i.e. cell-based in vitro assays (135, 149-151, 235, 249, 254, 256). The focus being on genes encoding proteins that are key players in the host immune response or gene candidates that can serve as disease biomarkers. Very few studies (73, 154, 155, 174, 246) have  looked at host or microbial gene expression signatures within these granulomas directly, especially none within cattle – the natural host. The recent availability of a complete *Bos taurus* genome (63), coupled with the continuing development of high-throughput genomic technologies should enable such transcriptional analysis. Application of functional genomics approaches to host–pathogen interactions will be the key to identifying and defining pathologically important genes, molecules, pathways and host–pathogen interactions (135).

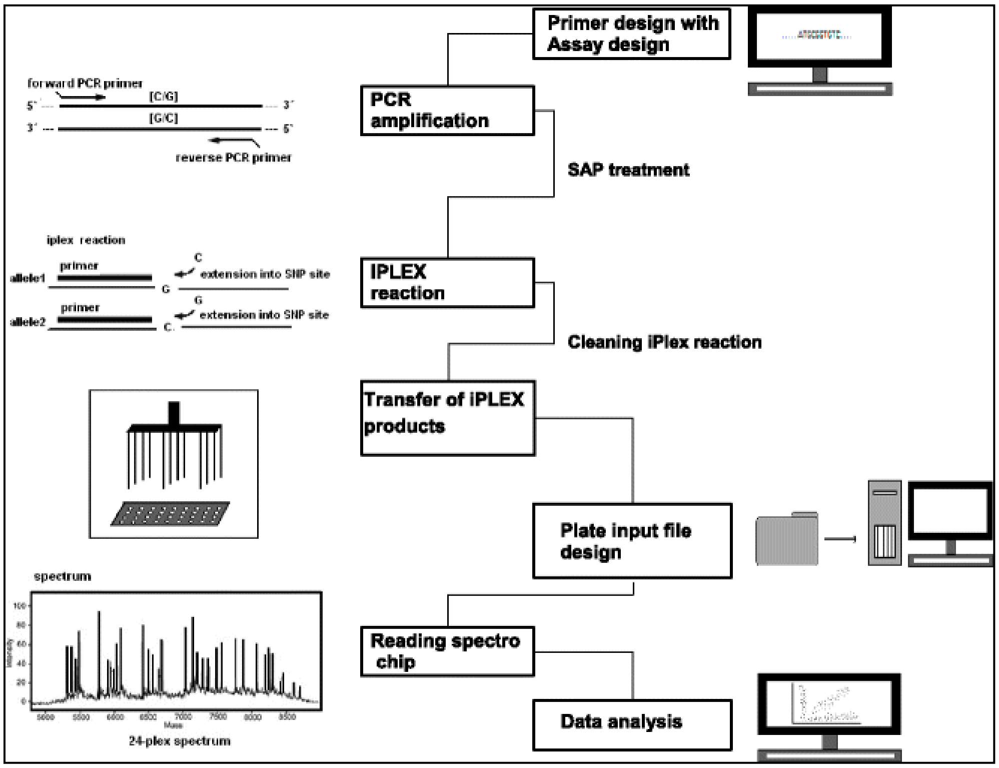
### xiii. SNP Genotyping - Sequenom Massarray$^{TM}$ Platform

SNPs are the most common source of genetic variation and important markers that link sequence variations to phenotypic changes. Thus, the scientific community has invested major resources to develop accurate, rapid, and cost-effective technologies for SNP analysis. Genotyping typically involves the generation of allele-specific products for SNPs of interest followed by their detection for subtype determination (113). In most technologies, PCR amplification of a desired SNP-containing region is performed initially to introduce specificity and increase the number of molecules for detection following allelic discrimination.  A number of SNP genotyping methods currently in use are described in the literature including,  TaqMan technology (218, 265), molecular beacons (220, 248, 265),  hairpin primer assay (92), single nucleotide extension microarray and fluorescence resonance energy transfer probes (18, 121, 123, 265) and many others are described in the literature (122, 231). Some of the commercially available high throughput platforms in use that can multiplex allowing for greater number of SNP discovery include Affymetrix$^{TM}$ (245), Illumina$^{TM}$ (35) and Sequenom$^{TM}$ (72) massarrays.

The Sequenom MassARRAY iPLEX$^{TM}$ assay consists of an initial locus-specific PCR, followed by single base extension using mass-modified dideoxynucleotide terminators of an oligonucleotide primer which anneals immediately upstream of the polymorphic site of interest. Using MALDI-TOF mass spectrometry, the distinct mass of the extended primer identifies the SNP allele (71, 72). The experimental procedure (24, 71, 72, 124) is subdivided into three different steps (Figure 1.2), with two intermediate cleaning reactions, before detection of the extension products. Target regions where the markers of interest are located are amplified first [polymerase chain reaction (PCR) amplification] to increase the quantity of specific template DNA. After inactivating unincorporated dNTPs with Shrimp alkaline phosphatase (SAP) (*PCR* reaction cleanup), a primer extension (iPLEX) reaction is done with mass-modified ddNTPs. The reaction incorporates different nucleotides according to the allele that is present immediately downstream of the 3′ end of the primer. After treating the extended primers with resin (iPLEX reaction cleanup) to optimize detection, and then spotting them into a chip that contains a specific matrix (transfer of the iPLEX reaction products), the resulting products are analyzed by MALDI-TOF MS (matrix-assisted laser desorption/ionization time-of-flight mass spectrometry). In each assay, the specific ddNTP that was incorporated can be identified by the increase in mass of the primer. MALDI-TOF MS with the current analysis software distinguishes molecules that differ by at least 10 Da. The MALDI-TOF MS mass detection range and precision currently set the limit of 40 assays per reaction. MALDI-TOF MS represents an emerging and powerful technique for DNA analysis because of its high speed, accuracy, no label requirement, and cost-effectiveness (180). So far, many MALDI-TOF MS approaches have been developed for rapid screening of SNPs, variable sequences repeat, epigenotype analysis, quantitative allele studies, and for the discovery of new genetic polymorphisms (47, 157, 180, 217, 238). The peak spectrum resulting from MALDI-TOF MS analysis can be analyzed with software that traces back primer masses to assayed alleles. Sequenom supplies software (SpectroTYPER, SpectroCALLER and SpectroACQUIRE) that automatically translates the mass of the observed primers into a genotype for each reaction.

A newly developed 16-plex iPLEX assay for MTC SNP genotyping (23) produced fully concordant results with the previously used technique (85, 86) SNaPshot primer extension method (Applied Biosystems, Foster City, CA) that allowed reliable differentiation of MTC species and recognition of lineages, thus demonstrating its potential value in diagnostic, epidemiological, and evolutionary applications.

The figure 1.2 below adapted from *Bradic et al* (24) and re-produced with permission summarizes the Sequenom assay-



### xiv. Whole Genome Sequencing - Illumina$^{TM}$ HiSeq 2000 Platform

The first whole genome sequence of a bacterium to be completed was that of *Haemophilus influenza* Rd in 1995 (69). Since then the total number of bacterial genome sequences has increased exponentially and is available through public domain databases like the NCBI- National Center for Biotechnology Information (www.ncbi.nlm.nih.gov). The data from these genome sequencing projects has been used to study evolutionary relationships among bacterial species (60, 260), for phylogenetic analysis (61, 255) and comparative genomics to discover virulence and pathogenecity markers (10, 131, 228, 266), strain phenotype characteristics (191, 198, 234) and others like drug resistance in bacteria (94, 264).

The first report on the sequence of 10 consecutive bases in a DNA strand of a bacteriophage was published in 1968 (262) and it was only in 1977 that sequencing methods like Sanger (216) and Maxam-Gilbert (148) came into being that could obtain longer nucleotide reads. For about 30 years after that, Sanger sequencing went through many improvements and automation that made whole genome sequencing possible (100). However, despite (or indeed because of) much progress in the area of genome sequencing, it became clear that even more information was to be gained not only from sequencing one genome per species but rather from sequencing and comparing the genomes of different individuals or strains / lines from the same species (181).

In the past decade, large-scale sequencing has been revolutionized by the development of next-generation sequencing (NGS) technologies. Sequencing technologies such as Illumina / Solexa, ABI/SOLiD, 454/Roche, and Helicos have provided unprecedented opportunities for ultra deep, massively parallel, high-throughput functional genomic research. These have significantly increased the number of bases and coverage per genomic region per sequencing reaction while decreasing the costs per base (181). NGS technologies generally yield shorter read lengths but provide sufficient coverage to enable genome sequencing and assembly, especially for prokaryotic genomes (181). NGS has also been applied for microbiome and metagenomics studies and for the detection of sequence variations within individual genomes, like SNPs, insertions / deletions (indels), or structural variants. Since NGS technologies are amenable multiplexing without compromising sequence quality or coverage, they have the ability to sequence multiple bacterial genomes simultaneously and deliver and interpret the resultant sequence information in near "real-time" (59). Further NGS technologies are associated with reduced sequencing cost by orders of magnitude making whole-genome sequencing a possible way for obtaining global genomic information from a given population (227). All commercially available NGS technologies differ from automated Sanger sequencing in that they do not require cloning of template DNA into bacterial vectors. Apart from being less labor-intensive, this has the distinct advantage that cloning biases. In most NGS approaches, template DNA is fragmented, bound to a substrate, and amplified by PCR to generate clonal representations of the original fragments that are spatially separated for subsequent sequencing (156, 219). A current exception to this is the Helicos system, which does not require template amplification but rather directly sequences single-template molecules (25). Sequencing itself is achieved by a number of methods that make use of different enzymes (polymerases or ligases) and chemistries to generate light signals that are recorded by highly sensitive detection methods (195). All NGS technologies allow for a high degree of parallelization, in which millions to billions of sequencing reactions occur simultaneously in small reaction volumes, thereby permitting higher throughput than automated Sanger sequencing (181, 195). Illumina / Solexa genome analyzer (Illumina, San Diego, CA) is a widely used NGS platform, also used in the human genome project (16). While it has a higher throughput than the Roche 454 genome sequencer, it yields shorter read lengths (25-100 bp). The short read lengths in particular present significant hurdles when it comes to assembling large sequence stretches especially in repeat rich sequences. However many bioinformatic algorithms are being developed to overcome these challenges. And with a reference genome available even relatively short reads can be mapped with high confidence to the reference sequence. Another important improvement is the ability to sequence both ends from a DNA fragment (paired-end sequencing) now implemented for most of the commercially available NGS platforms; paired-end data allow the scaffolding of contigs (contiguous sequences) in the absence of contiguous coverage of intervening sequences (181, 195).

Given the fast growth of NGS technologies, the main challenge is to cope with the analysis of vast production of sequencing database through advanced bioinformatics tools.  Despite these shortcomings, the advent of NGS is already a major breakthrough in molecular biology, genetics, and beyond, as well as a great leap forward for genomics and systems biology analyses (181, 195).

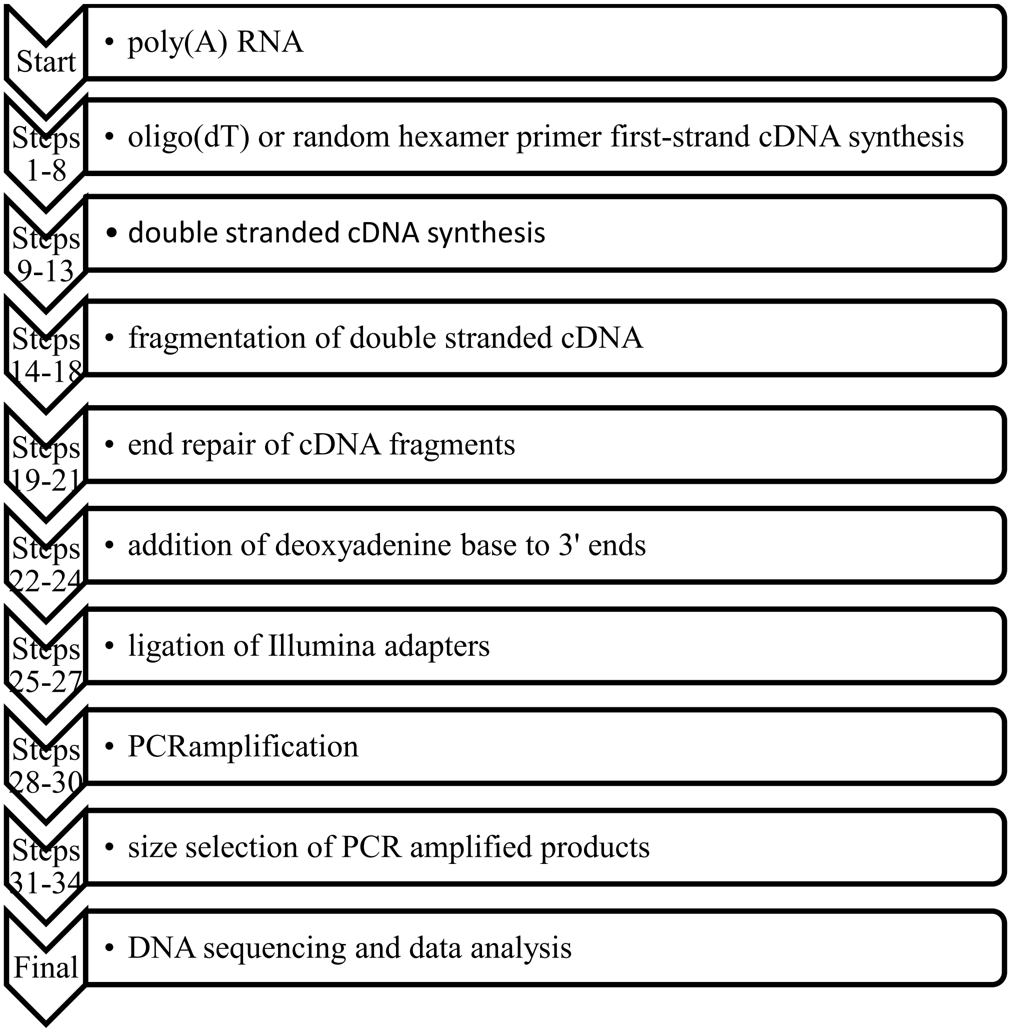
## Transcriptomic Profiling using RNA-Seq

The availability of a large number of microbial and mammalian genomes allows for in depth studies of these organisms and their interactions with their host species. The transcriptome is the complete set of transcripts in a cell, both in terms of type and quantity. Various technologies have been developed to characterize the transcriptome of a population of cells, including hybridization-based microarrays and Sanger sequencing–

based methods (19, 53, 263). In the last decade high-throughput NGS technologies (discussed previously) have revolutionized the field of trancriptomics. RNA sequencing (RNA-Seq) uses deep-sequencing technologies and involves direct sequencing of complementary DNAs (cDNAs) followed by mapping the reads to a reference genome (172). RNA-Seq captures almost all of the expressed transcripts for a snapshot of cells in theory, while microarrays rely on prior information that cannot detect novel splicing variants, novel genes, and novel transcripts (37). In addition, RNA-Seq has low background noise and high sensitivity, requires less RNA sample, and is becoming more cost-effective with the rapid advancements in the technology (141, 250).

RNA-Seq has been widely to infer alternative splicing (74, 229), quantify the expression of genes and transcripts (161, 241), detect gene  fusions (138, 201), reveal long noncoding RNAs (lncRNAs) (88), and identify single nucleotide variants (SNVs) in expressed exons  (38). RNA-Seq has also been applied in bacterial transcriptomics (11, 78, 81, 84, 116, 118, 127, 214, 258), to resolve interesting questions about biological processes in the bacterial cell, allowing better quality genome annotation (204) and especially to explore gene-trait associations.

In RNA sequencing, a population of RNA (total or fractionated such as poly A+) is converted to a library of cDNA fragments with adaptors attached to one or both ends. Each molecule, with or without amplification, is then sequenced in a high-throughput manner to obtain short sequences from one end or both ends (250). The read lengths are typically 30-400 bp which are then referenced to sequenced genome to construct a genome-wide transcript map. The basic protocol described by Nagalakshmi et al (172) involves the generation of a double-stranded cDNA library using random or oligo(dT) primers. The resulting library exhibits a bias towards the 5' and 3' ends of genes, which is useful for mapping the ends of genes and identifying transcribed regions. The cDNA is made from poly(A)$^{+}$ RNA, then fragmented by DNase I and ligated to adapters. These adapter-ligated cDNA fragments are then amplified and sequenced in a high-throughput manner to obtain short sequence reads. An alternate protocol describes the generation of a double-stranded cDNA library using random primers, but starting with poly(A)$^{+}$ RNA fragmented by partial hydrolysis. This provides a more uniform representation throughout the genes, which is helpful in quantifying exon levels, but is not as good for end mapping. Sequencing is done with an Illumina Genome Analyzer. Most of the reagents required are available in kits from commercial sources. The figure 1.3 below adapted from their manual summarizes the steps in RNA- seq method -


Start
• poly(A) RNA
Steps 1-8
• oligo(dT) or random hexamer primer first-strand cDNA synthesis
Steps 9-13
• double stranded cDNA synthesis
Steps 14-18
• fragmentation of double stranded cDNA
Steps 19-21
• end repair of cDNA fragments
Steps 22-24
• addition of deoxyadenine base to 3' ends
Steps 25-27
• ligation of Illumina adapters
Steps 28-30
• PCRamplification
Steps 31-34
• size selection of PCR amplified products
Final
• DNA sequencing and data analysis

**xiv. Bioinformatics/biocomputational needs to handle NGS data**

NGS technologies have demonstrated the capacity to sequence DNA/cDNA/RNA at unprecedented speed, thereby enabling previously unimaginable scientific achievements and novel biological applications. But, the massive data produced by NGS also presents a significant challenge for data storage, analyses, and management solutions (269). Advanced bioinformatic tools are necessary for the successful application of NGS technology. NGS technologies are characterized by a massive throughput for relatively short-sequences (30-100 bp), and they are currently the most reliable and accurate method for decoding genetic profiles. The first (and most crucial) step in sequence analysis is the conversion of millions of short sequences (reads) into valuable genetic information by their mapping to a known (reference) genome (45) or generating a de novo assembly. This has led to development of a wide variety of computational tools specifically designed to cope with the type and amount of sequencing data generated by NGS. PATRIC (Pathosystems Resource Integration Center) is one such database that offers a wide-array of tools including specialized searches, comparative analyses tools, visual browsers, and annotation pipelines. These tools help users harness the breadth and depth of PATRIC's data. It is a free resource database available at http://patric.vbi.vt.edu/.

Both Sanger and NGS techniques result in light signals that have to be decoded to determine the base sequence in the DNA. A widely used base-calling software for Sanger sequencing reads is phred, and the corresponding quality scores are called phred scores (65, 66). A number of file formats to represent sequence data and/or quality scores have already been developed for Sanger reads, and one format for the combined base sequence and phred scores that has also been adopted for NGS reads is the FASTQ format (39). There are also variants of the FASTQ format used by Illumina however there are many other input file formats available specific to different NGS applications and platforms. The first step post NGS is to convert the original files into an input format applicable to the tools used for downstream processing like mapping or assembly. There are many tools available that help converting the original files into the desired choice of input formats such as - http://bioinf.comav.upv.es/sff_extract/index.html or http://maq.sourceforge.net/fq_all2std.pl (181). However, to date there is no one standard format applicable to all. The downstream analysis like mapping to a reference genome also creates large files and there are tools available for mapping data in the Sequence/Alignment Map (SAM) format or its compressed equivalent BAM (128); that can now be used by a number of downstream applications, including several genome viewers.

*De novo assembly*: An assembly is a hierarchical data structure that maps the sequence data to a putative reconstruction of the target. It groups reads into contigs and contigs into scaffolds. Contigs provide a multiple sequence alignment of reads plus the consensus sequence. The scaffolds define the contig order and orientation and the sizes of the gaps between contigs. The most widely accepted data file format for an assembly is FASTA, wherein contig consensus sequence are represented by strings of the characters A, C, G, T, plus some other characters (http://droog.gs.washington.edu/parc/images/iupac.html) with special meaning. Dashes can represent extra bases omitted from the consensus but present in a minority of the underlying reads. Scaffold consensus sequence may have N's in the gaps between contigs. The number of consecutive N's may indicate the gap length estimate based on spanning paired ends (159). The contig N50 is the length of the smallest contig in the set that contains the fewest (largest) contigs whose combined length represents at least 50% of the assembly. Since NGS technologies give shorter reads, greater coverage of genomes is required for assembly as compared to Sanger sequencing because (i) the short read lengths require more reads within a region to confidently assemble contigs and (ii) the NGS reads have a higher error rate. However with enough coverage the higher error rate is not a problem (181). Since 2005, several assembly software packages have been created or revised specifically for de novo assembly of next-generation sequencing data, some of the published packages are SSAKE, SHARCGS, VCAKE, Newbler, Celera Assembler, Euler, Velvet, ABySS, AllPaths, and SOAPdenovo (159). The NGS assemblers are divided into three categories based on graphs. The Overlap/Layout/Consensus (19) methods rely on an overlap graph (171). The de Bruijn Graph (DBG) methods use some form of K-mer graph (101). The greedy graph algorithms may use OLC or DBG(159) . A graph is an abstraction used widely in computer science. It is a set of nodes plus a set of edges between the nodes. Many of these new assemblers take the approach known of the de Bruijn graph to performing assemblies. This approach is attractive as it does not require all reads to be aligned to all other reads and it can compress redundant sequence. Velvet is a set of algorithms that manipulate the de Bruijn graph for genome assembly. Velvet is capable of assembling bacterial genomes, with N50 contig lengths of up to 50 kb, and simulations on 5-Mb regions of large mammalian genomes, with contigs of ~3 kb (268). Recently the de novo assembly of a 40 Mb eukaryotic genome of a fungus *Sordaria macrospora* was done from short reads of Solexa using Velvet (182). The de Bruijn graph of Velvet algorithm is a compact representation based on short words (*k*-mers) that is ideal for high coverage and very short read (25–50 bp) data sets (197).

*Mapping of reads to a reference genome and detection of sequence variants*: With the exception of de novo assembly the mapping of reads to a reference genome is usually the first step for downstream analysis of sequencing. Reference guided assemblies enable comparative genomics leading to discovery of SNPs, indels, sequence polymorphisms, and structural variants. Similar to de novo assembly, mapping to reference genome requires computational algorithms and tools and there are many such available (199, 240). MOSAIK is one of such available resources at - http://bioinformatics.bc.edu/marthlab/Mosaik . Their program MosaikAligner pairwise aligns each read to a specified series of reference sequences. MosaikSort resolves paired-end reads and sorts the alignments by the reference sequence coordinates. Finally, MosaikAssembler parses the sorted alignment archive and produces a multiple sequence alignment which is then saved into an assembly file format. Unlike many current read aligners, MOSAIK produces gapped alignments using the Smith-Waterman algorithm. Additionally, the program goes beyond producing pairwise alignments and produces reference-guided assemblies with gapped alignments. These features make it ideal for SNP and short indel discovery. However when mapping RNA-seq reads there is a specific problem because reads that span splice junctions cannot be mapped to a genomic site (181). Several programs were developed to identify splice junctions either during the mapping or by first mapping all "mappable" reads and then identifying those reads that connect the transcribed regions that were identified in the previous mapping step (199).

TopHat is a read-mapping algorithm designed to align reads from a RNA-Seq experiment to a reference genome without relying on known splice sites (239). Similar to the mapping of genomic reads, the multiread problem also occurs during mapping of RNA-seq reads. This is due to repeat regions including paralogous genes but can also arise from, e.g., alternative splice forms and thereby cause not only mapping but also quantification problems (181). As with genomic mapping, there is no single solution to these concerns, but users can choose between a number of programs that offer different solutions to these problems (199, 241).

*Analysis of genetic variation*: While sequencing of one genome per species allows insights into the species' biology, genetic variation that leads to phenotypic variation between individuals cannot be deduced from one genome sequence alone. Individual genetic variations range from SNPs to small indels and rearrangements to large structural variants. Genomic comparisons can be used to address questions like mutation rates in evolving populations and their correlation to organismic adaptation, differences between a pathogenic and non-pathogenic strain and other such fundamental biological questions (181). Biocomutational tools are available for comparative genomic analysis and annotation of sequenced data that are freely available online (257). MAUVE (52) is one such program that allows for identification and alignment of conserved genomic DNA in the presence of rearrangements and horizontal transfer events. Similarly, Artemis Comparison Tool - ACT (32) is a mainstream tool for visualization, graphical presentation and annotation of sequence data. Bacterial genome annotation can be broadly categorized into structural and functional annotation. Structural annotation is dependent on algorithmic interrogation of experimental evidence to discover the physical characteristics of a gene. Functional annotation is dependent on sequence similarity to other known genes or proteins in an effort to assess the function of the gene (12). Sequence annotation pipelines are comprised of a variety of software modules and, in some cases, human experts. The reference databases, computational methods and knowledge that form the basis of these pipelines are constantly evolving, and thus there is a need to reprocess genome annotations on a regular basis (226). The RAST server is a fully automated service for annotating bacterial genomes. It identifies protein-encoding, rRNA and tRNA genes, assigns functions to the genes, predicts which subsystems are represented in the genome and uses this information to reconstruct the metabolic network (7).

Given the multitude free and open source bioinformatic resources available today, there is not one standard tool that can meet needs of a full genome analysis. Using a combination of any of these new technologies and software tools that best suit the experimental design; should enable analyses of sequences from NGS projects.

# CHAPTER 2

## Single nucleotide polymorphisms in *Mycobacterium bovis* genome resolve phylogenetic relationships and strain variation

Piecemeal analysis of *Mycobacterium bovis* genomes and conventional genotyping methods have not lend to a comprehensive resolution of its genetic diversity to explain the wide range of disease phenotypes caused by this zoonotic pathogen. Conventional genotyping methods target a small hypervariable region on the genome of *M. bovis* and provide anonymous biallelic information insufficient to develop *M. bovis* phylogeny. Genome-wide single nucleotide polymorphisms (SNPs) studies in *M. tuberculosis* have been shown to have sufficient resolution to develop trait-allele interactions. Using the high throughput iPLEX$^{TM}$ Massarray (Sequenom), we interrogated the *M. bovis* genome for 350 loci including geneic (n =306) and intergeneic (n =44) regions for SNPs. A collection of 75 *M. bovis* isolates associated with bovine tuberculosis outbreaks in the US between 1990-2009 and isolated from a variety of mammalian hosts – cattle (*n*=25), deer (*n*=6), elk (*n*=10), elephant (*n*=2), swine (*n*=7), and humans (*n*=24) were used for the study. Sixty one *M. tuberculosis* isolates from human, primates, birds, and elephants were also included in the analysis. Based on 206 variant SNPs between the *M. bovis* strains, five major clusters consistent with epidemiologic and other strain-typing information were identified. 49/51 human *M. tuberculosis* isolates were identical at 350 loci. This SNP based phylogeny provides new insights into the evolution of *M. bovis* and a gateway for studying strain genotype-disease phenotype correlations that were undertaken in an *in vitro* infection model of the disease with 4 virulent *M. bovis* strains isolated from human (*n*=1), cattle (*n*=2) and deer (*n*=1). Their virulence based on survival in macrophages and relative gene expression profile of various virulence genes were investigated at different time points post-infection. The results reveal a differential survival of 4 strains in the macrophage model, with a differential relative gene expression profile for 6 six virulence-associated genes mce4C, PE6, speE, mmpL12. Thus we conclude that *M. bovis* isolates from diverse geographic origins and host species represent an array of genetic profiles that may potentially relate to their phenotypic variation.

## INTRODUCTION

Bovine tuberculosis is a disease of significant economic importance in the developed world affecting animal productivity and trade of animal products (158). The introduction of milk pasteurization and "test and slaughter" cattle control programs in the early 1900s were successful in eradicating bovine tuberculosis in most developed nations (158). However, in some countries like the UK, the USA, and New Zealand, *Mycobacterium bovis* infections in wildlife serves as reservoir for the pathogen with severe consequences for livestock in those countries. Tuberculosis in wildlife poses serious difficulties for control and eradication of this insidious infection as it appears to be the underlying theme for the maintenance and periodic spillover of the infection into domestic animals (184, 187). *Mycobacterium bovis* is a zoonosis and a major concern in the pastoral settings of the developing world where the animal-human interface is close, and HIV prevalence is high. A recent study of all human tuberculosis cases in the USA from 1995 through 2005 estimated that only 1.4% of cases were being caused by *Mycobacterium bovis (96)* . In San Diego, California, over 45% of all culture-confirmed tuberculosis cases in children and 8% of all tuberculosis cases were found to be due to *Mycobacterium bovis* (212). *Mycobacterium bovis* as a source of human infection is likely under reported due to cultivation medium components used to isolate the organism from sputum or other sources. *Mycobacterium bovis* does not efficiently use glycerol as carbon source that is commonly used in culture media for *Mycobacterium tuberculosis* growth and needs supplementation with pyruvate.

Differentiation of genetic variants has become an indispensible tool to study the evolution, epidemiology, and ecology of pathogenic organisms and to gain insights into host-pathogen interactions (28, 108). *Mycobacterium bovis* belongs to the *Mycobacterium Tuberculosis* Complex (MTC) group of organisms that are characterized by 99.9% nucleotide sequence similarity and carry identical 16S rRNA and show restricted allelic variation in their structural genes (170, 224). In the post genomic era, single nucleotide polymorphisms (SNPs) have emerged as a robust tool for delineating phylogenetic relationships between closely related strains of pathogenic bacteria including *Mycobacterium tuberculosis* (68, 85, 86) . Besides being a rich source of genetic variation, SNPs are easy to assay which makes them amenable to large-scale population genetic studies (85, 86). A 2009 study by Garcia Pelayo *et al* (76) reported ~700 SNPs by comparative genomic analysis of the virulent *Mycobacterium bovis* UK strain AF2122/97 and the vaccine strain *Mycobacterium bovis*-BCG Pasteur (the parent strain, *M. bovis* Nocard, originally obtained from a cow with tuberculous mastitis in France) and used the information to distinguish *Mycobacterium bovis* isolates of French and British lineages.

*Mycobacterium bovis* expresses two immunodominant antigens MPB70 and MPB83. MPB83 is differentially expressed in vitro; however, there is evidence suggesting its upregulation in vivo with *M. tuberculosis* infection. Cattle are known to recognize MPB83 yet humans rarely develop an immune response to this anitgen – the mechanisms of which are unclear (253). The precise functions of these proteins a not yet deciphered, however their striking difference in expression and putative function in interaction with the host suggests that they might play a role in host preference (93). In *Mycobacterium bovis* the most frequently recognized antigens that elicit host immune response are cell-wall associated proteins, PE/PPE protein family members, secreted proteins and conserved hypothetical proteins (42, 93). Immunomodulatory phenolic glycolipds are are differentially produced in *Mycobacterium bovis* and *Mycobacterium tuberculosis* (208). Transcriptome differences are expected to play a role in the differing ecotypes of the *Mycobacterium tuberculosis* complex that have closely related genomes with distinct host preferences (79). However whether differential gene expression plays a role in strain variation within a particular ecotype of the *Mycobacterium tuberculosis* complex has not been explored.

In the present study, we applied SNP genotyping analysis using 350 of the 700 SNP loci described (76), to develop a population genetic framework among *M. bovis* and *M. tuberculosis* organisms. We then demonstrated strain-specific variations in intramacrophage survival and gene expression, among different phylogenetic lineages of *Mycobacterium bovis*.

## MATERIALS AND METHODS

**Bacterial isolates.** A collection of 75 *M. bovis* isolates associated with bovine tuberculosis outbreaks in the US between 1990-2009 and isolated from a variety of mammalian hosts – cattle (*n* =25), deer (*n* =6), elk (*n* =10), elephant (*n* =2), swine (*n* =7), humans (*n* =24) and environmental (*n*=1) were used for the study. Sixty-one *M. tuberculosis* isolates from human (*n* =51), primates (*n* =7), avian (*n* =1), and elephants (*n* =2) were also included in the analysis. The 75 *M. bovis* strains and 61 *M. tuberculosis* strains are shown in Table 2.1, along with brief epidemiological information about these isolates. Some of these *M. bovis* isolates are derived from slaughterhouse surveillance cases within the US and known to be traced back to various states in Mexico. All these isolates have been characterized by spoligotyping and were available from the APHIS-USDA culture collections (isolates # 1-67) and Public Health Research Institute Center (PHRI), Newark, NJ (isolates # 68-136). The DNA for these strains was isolated at APHIS-USDA, Iowa, and PHRI, Newark, NJ, using standard DNA extraction protocols for mycobacteria (4) and shipped to lab. The whole genomic DNA samples were amplified in the lab using the Qiagen repli-G kit (Qiagen Inc., Valencia, CA) and stored at -80°C until further use.

**SNP selection and identification.** Based on a recent genome-wide analysis of the sequenced *Mycobacterium bovis* (AF2122/97) and *Mycobacterium bovis*-BCG Pasteur strains a total of 782 SNPs were identified by Garcia Pelayo *et al*, 2009 (12). These 782 sites identified included transitions, transversions, insertion or deletions, and block substitutions (where a block of > 1bp replaces another). Of these 782, a set of most discriminatory target loci were selected to include SNPs located within open reading frames (*n* =44) and intergeneic (*n* =306) regions. These SNP sites were selected to index variability across the whole *Mycobacterium bovis* genome (Figure 2.1a). The information on these SNP positions is available through their study as it occurs in the *Mycobacterium bovis*- BCG genome, with their genomic position, locus and gene/intergenic presence identified. Using this information, SNPs were located and verified in the genome of the sequenced *Mycobacterium bovis* strain AF2122/97 and the *Mycobacterium tuberculosis* strains H37Rv and CDC 1551 available freely through the public database of the National Center for Biological Information (NCBI, *www.ncbi.nlm.nih.gov*)

**Single nucleotide polymorphism based genotyping.** Genotyping was performed using the iPLEX$^{TM}$ chemistry on the Massarray genotyping platform (Sequenom Inc., San Diego, CA) available at the BioMedical Genomics Center, University of Minnesota. During the iPLEX$^{TM}$ reaction, oligonucloetide primers anneal directly adjacent to the SNP of interest. SNPs were queried using oligonucleotides that anneal 1-bp upstream of the base of interest; allele-specific extension products were then analyzed via matrix-assisted laser desorption ionization mass spectrometry (MALDI-TOF) to identify the base at each SNP position across the panel of strains. Allele specific extension products are then produced by single base extension of the oligonucleotide with terminator nucleotides, each of unique mass. Multiplexed iPLEX$^{TM}$ assays of between 1 to 8 assays per iPLEX$^{TM}$ were designed to detect 350 single nucleotide base changes using the Sequenom Assay Design v.3.0.2.0 package. Allele specific products resulting from iPLEX$^{TM}$ reaction were desalted through the addition of an anion-exchange resin and then analyzed by MALDI-TOF mass spectrometry. Genotypes were assigned in real-time and then evaluated using the SpectroCALLER and SpectroACQUIRE software (Sequenom Inc. San Diego, CA) respectively.

**Phylogenetic analysis**. The 206 variant SNP calls were concatenated into string of single characters resulting in a single 206-bp sequence for each strain. Sequence alignment and phylogenetic analysis was carried out using MEGA 4.1 software (233) (http://www.megasoftware.net/).

***In vitro* macrophage infection assay**. The *in vitro* infection studies were carried out at the BSL-3 facility of the NADC-USDA laboratory in Ames, Iowa. The mouse macrophage cell line J774A.1 was procured from ATCC$^{TM}$ (Manassas, VA). Cells were propagated in GIBCO$^{R}$-DMEM (Invitrogen, Carlsbad, CA) containing 10% FBS at 37$^{o}$C in 5% $CO_2$ and subsequently seeded at ~1 x $10^6$ cells/flask in 25cm2 flasks. The four *Mycobacterium bovis* strains used for infection were originally isolated in 2004 from Texas beef cattle, in 2005 from human in California, in 2008 from dairy cattle in California and in 2009 from a deer in Minnesota and represented diverse spoligotypes and MIRU profiles (Table 2.2). Bacterial suspensions were grown by the APHIS-USDA laboratory, Ames, Iowa and consisted of mid-log-phase *Mycobacterium bovis* grown in Middlebrook 7H9 liquid media supplemented with 10% oleic acid albumin- dextrose complex (OADC) (Becton Dickinson Co., Sparks, MD) plus 0.05% Tween 80 (Sigma Chemical Co., St. Louis, MO) grown for ~10 days at 37$^{o}$C. To harvest bacilli from the culture media, cells were pelleted by centrifugation and the pellet re-suspended in phosphate-buffered saline solution (PBS, 0.01 M, pH 7.2) to the appropriate concentration. Bacterial suspensions were used immediately without freezing at a multiplicity of infection (MOI) of 5:1 (bacteria: cells) for all infections. Following infection for 2 hr at 37$^{o}$C in 5% $CO_2$, macrophages were washed three times with fresh pre-warmed serum-free GIBCO$^{R}$-DMEM (Invitrogen) to remove non-adherent bacteria and treated amikacin at 200ug/mL to kill any extracellular bacteria. The cultures were subsequently grown in GIBCO$^{R}$-DMEM (Invitrogen) with 2% serum and harvested for transcriptional analysis at 0 minute, 30 minute, 2 hour, 24 hour and 48 hours post infection, in triplicate for each time point.

**Nucleic acid extraction**. Total RNA from infected macrophages (0 & 30 min., 2, 24 & 48 hrs p.i,) was extracted using TRIzol reagent (Invitrogen Inc., Carlsbad, CA) per the manufacturer's instructions. Samples were homogenized in a mini bead-beater (Biospec, Bartlesville, OK) with 0.3 ml of 0.1 mm sterile RNase-free zirconium beads for 4 min. followed by RNA extraction. Samples were stored at -80$^{o}$C until further use.

**Survival assays**. Macrophages were lysed with 0.1% Triton-X 100 (Sigma Chemical Co) in sterile water for 10 minutes and the lysate was serially diluted and inoculated on Middlebrook 7H11 media. Plates were incubated at 37$^{o}$C and growth was evaluated at 3weeks and 6weeks p.i. for cfu/ml estimation.

**Quantitative Real-time PCR**. Genomic DNA removal and cDNA synthesis was carried out using the Quantitect Rev.Transcriptase kit (Qiagen, Valencia, CA). The cDNA was stored at -20$^{o}$C until further use. Selected genes were amplified using the one-step SYBR-green based quantitative real-time PCR (Roche Inc, Indianapolis, IN) analysis in Roche LightCycler 480 II (Roche Inc.). Primers (Table 2.3) were designed using web-based tools, Primer3 http://frodo.wi.mit.edu/primer3/. The following cycle program was used: denaturation at 95°C for 15 min. and PCR at 95°C for 10 s to denature, 65°C for 15 s to anneal primer, 72°C for 22 s to extend by polymerization for 45 cycles. Test and control samples were normalized using the house keeping gene, *gyrA*, and relative expression was calculated by $2^{-\Delta\Delta CT}$ method(132). Results are reported as fold change. Each sample was analyzed in duplicate.

## RESULTS

**SNP diversity analysis**: Three hundred and fifty loci on the *M. bovis* genome were genotyped and identified 206 (Figure 2.1a b) to be variable among the total 136 isolates studied that included 75 *M. bovis* and 61 *M. tuberculosis* isolates. Information on these 350 loci was also obtained for four of the previously sequenced strains including *M. bovis* AF2122/97, *M. bovis* BCG-Pasteur, *M. tuberculosis* H37Rv, and *M. tuberculosis* CDC1551. Between the 75 *M. bovis* isolates alone, 202 SNPs were identified, out of which 118 SNPs (Table 2.1) were variable between the disease-associated *M. bovis* isolates. Of these 118 variant SNPs, 91 were genic SNPs and 27 were in the intergenic region. A second set of 84 genic SNPs (Figure 2.2) were able to distinguish isolates of the attenuated vaccine lineage of the *M. bovis* strain BCG from the virulent isolates. A set of 9 isolates previously genotyped as *M. bovis* using IS*6110* profiling and spoligotyping were submitted to the study. However in the SNP analysis they were identical to the

BCG-Pasteur vaccine strain. Hence, further probed these 9 isolates for the presence / absence of the region of difference 1 (RD1). Using methods described by Talbot et al (232), these 9 isolates were confirmed to have the RD1 missing, thus confirming them as the attenuated *M. bovis* BCG strains.

Forty-nine of the 51 *M. tuberculosis* isolates from human hosts were identical at all the 350 loci examined and clustered in a single clade. The two variant human *M. tuberculosis* isolates (Table 2.2) used in the study (68: 18463 and #87: 24282) were submitted as human *M. bovis,* and classified as spoligotypes SB0228 and SB0242 and carried 3 copies of IS*6110*, respectively. These 2 isolates were variant from the other human *M. tuberculosis* isolates at 11 of the 350 typed loci that included the genic SNPs of *katG* codon 463 and *Mb1794c* (*n*=2) codons 72 and 132, and eight SNPs that were in the intergenic region (IGR1, IGR14-15, IGR17-21). These two isolates also lacked the *M. bovis* signature SNP of the gene *pncA* codon 57. We further probed the presence of the region of difference 9 (RD9 loci: *Rv2073c*) in these two isolates thus further differentiating them as *M. tuberculosis* and not *M. bovis*. Ten *M. tuberculosis* isolates derived from animal hosts had nearly identical SNP profile to those of the human isolates except at 19 loci. These included 5 genic SNPs of the genes *katG* codon 463*, oxyR* codon 78*, fadD9* codon 600*, Mb1794* (*n*=2) codon 72 and codon 132 and fourteen intergenic SNPs (IGR1, IGR14-15, IGR17-27).

***M. bovis* phylogeny**: A consensus phylogenetic tree was derived using the Maximum Parsimony algorithm based on the Hasegawa-Kishino-Yano (orK2P) model with 1000 bootstrap replicates. The 206 variant SNPs resolved 136 isolates along with the 4 sequenced strains of AF2122/97, BCG-Pasteur, H37Rv and CDC155, 1 into five major genetic clusters or “SNP-cluster groups”; 4 groups of *M. bovis* isolates and one cluster that included all the *M. tuberculosis* isolates (Figure 2.3). There were three principal virulent *M. bovis* SNP-cluster groups variant at 118 loci (genic and intergenic) that included isolates from both animal and human hosts. However, the variation observed in the intergenic SNPs was not lineage specific. The fourth group that exclusively clustered

9 human *M. bovis* isolates along with the vaccine strain BCG-Pasteur differed at 84 genic loci from the virulent isolates (Figure 2.2). Our analysis included the sequenced *M. bovis* strain from the UK, AF2122/97, which shared genetic signatures of the first *M. bovis* SNP-cluster group. The two sequenced strains of *M. tuberculosis* - CDC1551 and H37Rv clustered with the fifth cluster group exclusive to the *M. tuberculosis* isolates in the analysis. Isolate strains from Michigan (MI, *n*=5), Minnesota (MN, *n*=5) and Hawaii (HI, n=*7*) clustered within their respective SNP-cluster groups. Isolates from states other than MI, MN and HI carried a diverse genetic profile as evidenced by their distribution across all 3 *M. bovis* SNP-cluster groups. All elk (*n*=10) isolates from a variety of geographic locations including Missouri, Montana, Nebraska, New York, Wisconsin and Kansas which were isolated between 1992-2009 in clustered in the SNP-cluster group 3. The fourth SNP-cluster group of *M. bovis* isolates is unique in that it only includes *M. bovis* BCG strains from humans and these shared the SNP genotype of the sequenced vaccine strain BCG-Pasteur thus allowing for differentiation of these isolates from the virulent *M. bovis* isolates.

**Analysis of synonymous, non-synonymous and intergeneic SNPs***:* Among the 206 SNPs, identified both intergenic (*n*=27) and genic (*n*=179) SNPs and were distributed evenly around the genome (Figure 2.1b). Of the 179 genic SNPs 59 resulted in synonymous changes and 120 were non-synonymous mutations. The ratio of synonymous SNPs to non-synonymous SNPs was 1:2.

**Variation in spoligotyping, variable number tandem repeat (VNTR) and IS*6110* restriction fragment length polymorphism (RFLP) profiles of strains***:* All isolates were previously characterized (Table 2.2) by spoligotyping and VNTR (APHIS-USDA culture collections) or IS*6110* RFLP profiling and spoligotyping (PHRI culture collections). The relationship between phylogenetic lineages of these isolates to their spoligotyping/VNTR/RFLP profiles was examined. *M. bovis* isolates with common spoligotype patterns or VNTR/RFLP profiles clustered together. However each of the 3 SNP-cluster groups was represented by more than one spoligotype or VNTR/RFLP profile. Similarly, the 49 human *M. tuberculosis* isolates from human that were identical by their SNP profile had diverse IS*6110* and spoligotype profiles. These human *M. tuberculosis* isolates that had the identical SNP genotype in this study were isolated between 1992-2010 from mainly from the New York City and New Jersey areas. Seven out of the 10 *M. tuberculosis* isolates from animal hosts had unique unregistered spoligotypes and variant VNTR profiles.

**Survival in macrophage infection model**: Four *Mycobacterium bovis* strains (Table 2.2) representing human, cattle and deer hosts from 3 different geographic locations (MN, CA and TX) were challenged in a mouse macrophage infection model to look for differences in survival (Figure 2.4). Strain#1, the 2009 MN deer isolate and strain#3 the 2008 CA dairy cattle isolate share their SNP genotypes, whereas Strain#2, the 2005 CA human isolate and Strain#4 the 2004 TX beef cattle isolate have the same SNP genotype (i.e. two isolates each shared a SNP genotype). Strain#1 and strain#3 that fell within the same SNP-cluster group in our analysis differ at 2 out of 11 VNTR loci typed and differ by 1 spacer band in their spoligotype however cluster within the same clade when analyzed for phylogeny based on these two typing methods. Strain#2 and Strain#4 have identical spoligotypes, and VNTR profiles. Strain 1 and 3 showed a diverse survival pattern in the macrophage infection model despite having an identical SNP genotype. Three out of the four strains used in this macrophage model continue to persist within the host cells at 48hrs p.i. and yet exhibit a survival pattern that is unique to each strain.

**Comparative transcriptional analysis:** Gene expression profiles were studied for six virulence associated genes for the above four strains using qRT-PCR. The relative fold change for all the six genes are shown in Figure 2.5. The genes of interest were PE6, mce4C, mmpL12, speE, fadD9 and INO1 (Table 2.5). Their relative expression profiles were analyzed at two time points 30min and 2hr p.i as compared to the time point 0 minute. The relative fold change for the six genes ranged from 0 to 6.5. The PE family protein gene PE6 was downregulated in all strains and both time points except for strains

1 and 2 where it was upregulated at least 1 fold change at 30min p.i. The spermidine synthase gene, speE, was upregulated in strains 1 and 2 at both 30min and 2hr p.i and downregulated in strain3 and strain4. The mammalian cell entry protein gene, mce4C, was ≥ 4 fold up regulated in strain 1 and down regulated in strain3. In strain 2 it was downregulated at 30min and upregulated at 2hr whereas in strain 4 it was upregulated only at 30 min p.i (thus a highly variable profile was observed in these 2 strains). The transmembrane transport protein, mmpL12, had a similar profile as mce4C, with upregulation in strain 1 and downregulation in strain 3 and a variable profile in strains 2 & 4. INO1 that is involved in the biosynthesis of phospholipids and lies 16bp downstream of an intergenic SNP was 1.5 fold up regulated in strains1, 2 & 3 and was down regulated in strain 4. Gene fadD9 that encodes fatty acid CoA ligase was upregulated in strain1, downregulated in strains 3 & 4 and had a varying expression at 30 min (downregulated) and 2hr (upregulated) p.i for strain 2.

## DISCUSSION

**Genome wide SNPs of *M. bovis* differentiate between isolates.** In a 2009 study by Garcia Pelayo et al (76) 782 SNPs were identified across the entire genome of *M. bovis* and *M. bovis BCG.* Information was derived from their study on a subset of 350 SNPs and used this to generate a population genetic framework among outbreak-associated isolates from the Unites States. Molecular variation and outbreak tracking of *M. tuberculosis* complex isolates typically employ IS*6110* profiling, spoligotyping or MIRU-VNTR analysis. While these targets and tools are considered sufficient for molecular epidemiology, they are unable to sufficiently index population genetic structure of this genus since they represent small hypervariable regions within the genome that generally evolve at higher rates than the rest of the genome . Thus, SNPs have been used to define the extent of genetic diversity in *M. tuberculosis* and other pathogenic mycobacteria that has provided insights on the evolution, pathogenicity, and molecular epidemiology of tuberculosis globally. A previous study identified 782 SNPs between the virulent *M. bovis* strain AF2122/97 and vaccine strain BCG-Pasteur, of which 158 SNPs separated all the *M. bovis* strains of the French lineage from the *M. bovis* strains of the British lineage. Similarly, our study documents that the 206 SNPs across the genome are sufficient to resolve *M. bovis* phylogeny and genetic relatedness into 3 major lineages as compared to other typing techniques used hitherto that sets the platform for downstream studies involving phenotypic characterization of factors affecting virulence and pathogenesis. Though it is important to note that all BCG vaccine strains, including BCG-Pasteur, were derived from a *M. bovis* clinical isolate of the French lineage, while the reference AF2122/97 genome sequence is from a *M. bovis* isolate of the British lineage. Furthermore, among the 206 SNPS, the ratio of non-synonymous SNPs to synonymous SNPs was 2:1, similar to that reported from genome-wide SNP studies in *M. tuberculosis* (85, 86) and indicative of recent emergence from a population bottleneck.

**SNPs differentiate lineages of *M. bovis and M. tuberculosis.*** In the current study SNP-based phylogenetic analysis was able differentiate *M. bovis-* both virulent strains as well as the attenuated BCG strains that the conventional genotyping techniques failed to resolve. This is important in the clinical diagnosis of tuberculosis because the BCG vaccine, although considered safe, is known to cause disease in immunocompromised hosts.

Further, SNP genotyping resolved misclassification of 2 *M. tuberculosis* isolates as *M. bovis* and 9 *M. bovis*-BCG isolates identified as virulent *M. bovis* by previous typing techniques. SNPs in *oxyR* (codon 78), *katG* (codon 463) and *pncA* (codon 57) genes identified them either as *M. tuberculosis or M. bovis* respectively and 206 SNPs profile differentiated *M. tuberculosis* from *M. bovis*-BCG. Thus *M. bovis* infection and outbreaks in the US documented in humans using conventional methods have a tendency for their misclassification. This further implies that genome-wide markers such as SNP sets for differentiation among *M. tuberculosis Complex* would be useful to index zoonotic transmission of this bacterium in rural areas of the developing world, where the animal-human interface is intensifying as land use patterns are changing.

Although this study does not target all the SNPs described between virulent *M. bovis* and BCG-Pasteur, within a subset of 350 target loci, were identified 206 variant SNPs. This finding suggests that the attenuation of *M. bovis BCG* may transcend beyond large sequence polymorphisms and that examining the functional consequences of variant SNPs may aid in understanding the shortcomings of BCG as a vaccine.

**SNP based spatial and host associations.** Bovine tuberculosis is a re-emerging infectious disease in the US where the deer population is identified as a potential reservoir for *M. bovis* infections. Within the US, Michigan had one of the longest ongoing bovine tuberculosis epidemics. Our deer and cattle tuberculosis isolates from MI (*n*=5) collected between 1995-2008 and from MN (*n*=5) isolated between 2006-2009, showed clustering into distinct lineages specific to geographic origins. Thus, spatial specificity of distinct lineage in these areas is likely a result of a founder effect where upon introduction the strains evolved independently in the deer and cattle populations. Evidence suggests that the MI strain of *M. bovis* spilled over into the white-tailed deer population in the 1930s and has since been maintained in that population.

A significant observation in the study was that Hawaii (HI) isolates shared SNP genotype with isolates from other geographic locations despite there being little to no epidemiological linkage. Despite depopulation and restocking of cattle on islands of Hawaii in an attempt to eradicate bovine tuberculosis, periodic cattle infections have been detected. Epidemiological studies suggest that the wildlife reservoir of feral swine maintains the pathogen. Given that the HI feral swine isolates share their SNP genotype with cattle and deer isolates of other geographic locations like MI, TX, CA, NY, OK and Mexico it is likely that the organism was introduced into the swine population here by introducing infected deer or cattle from these other states which led to its spread rapidly within the new feral hosts.

All animal isolates identified as *M. tuberculosis* were similar at all SNP loci to the human counterparts except at 19 loci. This is likely due to intra-host adaptation changes that occurred in the animal hosts after transmission from humans or suggests that animal species are susceptible only to some subtypes of *M. tuberculosis*. The data also provides robust information on diversity among *M. bovis* isolates and documents loci that can be used to differentiate *M. tuberculosis* from *M. bovis* within animals, between animals and humans, and between *M. bovis* and BCG.

Elk *M. bovis* isolates from 6 states of the US representing a 15-yr time period (1992-2009) were the only strains to cluster in a single clade, suggesting a degree of host specificity for this genotype. These isolates were also identical for their spoligotypes and VNTR profiles suggesting a clonal spread of a single strain in this host despite geographic and temporal distance. It is likely that the particular SNP genotype is elk-adapted and highly virulent for this host or else elk may be exclusively highly susceptible to this genotype of *M. bovis*. Presence of several SNP genotypes among isolates from cattle, deer and humans suggest multiple sources of introduction of infection in these host species. Identification of SNP genotypes from Mexico in every clade suggests a high level of diversity and interspecies transmission of isolates from that location.

Thus, conclude that SNP-based genotyping is able to resolve misclassification within the infecting species, identify patterns of host or spatial associations, differentiate lineages, and phylogenetic structure among *M. bovis.* With increasing availability of multiple whole genome sequences, SNP identification will add considerably to phylogenetic analysis and evolutionary studies. Presented is a snapshot of the diversity and structure using 206 “informative” SNPs – further investigations should derive from comparisons of whole genome sequences of isolates derived from diverse geographic locations. It is proposed that the SNP-cluster groups identified in this study should facilitate investigation of functional and biological variation between and within the isolates of these five phylogenetic lineages.

**Variation in intracellular survival and gene expression profiles among *M. bovis* clinical isolates**. In the second objective for this study SNP-genotype and phenotype correlations were evaluated that will answer many unknowns related to strain differences in pathogenesis and virulence. For the members of *Mycobacterium Tuberculosis* Complex group of organisms, it is known that despite their genetic relatedness they exhibit spectra of phenotypic characters and host range. Strain variation in the *Mycobacterium Tuberculosis* Complex exists and has biological significance (28, 57). *Mycobacterium bovis* is a unique ecotype of the *Mycobacterium Tuberculosis* Complex group as *Mycobacterium bovis* has the widest known mammalian host range including humans (93, 163). Unlike other members of *Mycobacterium Tuberculosis* Complex group *Mycobacterium bovis* is known to survive in the animal environments for certain periods of time (58, 104, 192, 193, 261). Many studies have demonstrated strain variation in the *Mycobacterium tuberculosis* clinical isolates (5, 20, 79, 153) and other pathogenic mycobacteria (80, 105, 165, 270), however there is a paucity of such information available for clinical isolates of *Mycobacterium bovis. Mycobacterium bovis* differs from *Mycobacterium tuberculosis* in key biological properties, such as transmissibility, host range and antigenic variability, where factors unique to *Mycobacterium bovis* are expected to play a role (67). A study comparing pathogen transcriptomes of a virulent *Mycobacterium bovis* isolate to that of the attenuated vaccine strain showed 133 genes that displayed a minimum 2 fold difference in expression (22). In our assessment of the four strains (MN Deer, CA human, CA cattle and TX cattle) challenged in a mouse macrophage screen as a model for *in vitro* infection the study detected a marked difference in survivability and gene expression profiles for six virulence associated genes. Despite shared SNP profiles between strains 1 and 3 and between strains 2 and 4, their intracellular survival trends were different. In conclusion, the intracellular survival of different strains of *Mycobacterium bovis* within the host cell may not be related to their SNP genotypes and may be dependent on other bacterial or host factors like the regulation of gene expression.

In the gene expression profile, mce4C which represented the mammalian cell entry proteins family was downregulated in strain 3; a strain that did not survive in the infection model after 2hrs p.i. The mce4C gene is essential for the pathogen to gain host cell entry for the establishment of infection. The coding sequence of the mce4 operon has been described to be significantly polymorphic with a higher frequency of synonymous substitutions compared to other mce operons in *Mycobacterium tuberculosis* (196). All four strains used in this experiment did not carry the SNP at the mce4C loci, hence, it is difficult to attribute SNP differences in the mce4c loci to the inability of strain 3 to survive within macrophages. However, the SNP in the mce4c gene in the analysis of other *Mycobacterium bovis* isolates (other than the 4 strains used in this in vitro infection assay) is a non-synonymous substitution and could have a functional implication for isolates carrying this SNP, and similarly, there may be other genes with SNPs as well within the mce operon that may affect the biology of the pathogen. The gene encoding spermidine synthase that inhibits the host mycobactericidal compound NOS (nitric oxide synthase) in strains 2 and 4 carry a synonymous SNP. The contradicting expression profile of speE indicates that the SNP is not responsible for differential expression as also evidenced for strains 1 and 3 that share their SNP genotype. Similarly for genes PE6, fadD9, INO1 and mmpL12 that are implicated in the antigencity of the organism; the greatly varying regulation of gene expression profiles indicate that their SNP genotypes may not be solely responsible for this difference and warrants more mechanistic studies. For INO1, the SNP is the intergenic region of this gene that is present 16bp downstream of the SNP. The intergeneic SNP did not alter the expression of this downstream gene although there are other intergeneic SNPs identified in our study that need to be tested and could potentially influence gene regulation. Similarly, the five genes with SNPs in the coding regions represent only a subset of all the variant SNPs identified in this study. Though this subset of genic SNPs did not contribute to the variation in the gene expression profiles of the four strains studied, this study provides some of the first evidence for differential gene expression in clinical strains of *Mycobacterium bovis* compared to different host species with a diverse epidemiological background thus suggesting that strain variation in *Mycobacterium bovis* exists and maybe associated with virulence of this pathogen.

These studies were done in a mouse macrophage model, which is often questioned for its applicability in tuberculosis research since mice are not naturally infected with tuberculosis. It would be important to repeat these studies in a bovine model with inclusion of more target SNPs. Despite this shortcoming, the first evidence for strain variation was provided in clinical isolates of *Mycobacterium bovis* in an *in vitro* infection model and such genetic information specific to the pathogen would help guide in deciphering the molecular epidemiology and host-pathogen interactions and provide markers for virulence assessment that can aid in the implementation of better control programs and in vaccine research. Thus, SNP genotyping provided for the identification of genetic variation that resolved *Mycobacterium bovis* phylogeny and can be mapped to functional differences across strains which will facilitate future studies of *Mycobacterium bovis* molecular epidemiology and strain variation that relates to their virulence and pathogenesis.

## LIST OF TABLES

**Table 2.1:** Metadata on the isolates used for SNP analysis

| ***M.bovis* isolates (*n*=57) from the APHIS-USDA culture collection, Ames, IA** | | | | | |
|---|---|---|---|---|---|
| **No.** | **Isolate ID$^{\$}$** | **Host** | **State** | **Spoligotype** | **VNTR Profile** |
| 1 | HC2045T | Cattle | TX | SB0673 | 25237452534 |
| 2 | 08-5055 | Cattle | CA | SB0140 | 25215452534 |
| 3 | 08-4513 | Cattle | TX | SB0971 | 25237452534 |
| 4 | 08-2906 | Cattle | TX | SB0121 | 23326442232 |
| 5 | 08-2630 | Cattle | MN | SB0271 | 25237452534 |
| 6 | 08-2431 | Cattle | CA | SB0121 | 23326442232 |
| 7 | 08-0955 | Cattle | MI | SB0815 | 23237552533 |
| 8 | 08-0168 | Cattle | OK | SB0673 | 25237452534 |
| 9 | 07-6182 | Cattle | SD | SB0152 | 25336442635 |
| 10 | 07-5545 | Cattle | NM | SB0673 | 25237452534 |
| 11 | 07-3557 | Cattle | MI | SB0145 | 23237552533 |
| 12 | 07-3280 | Deer | MN | SB0271 | 25237452534 |
| 13 | 07-1437 | Cattle | OK | SB0327 | 25134452323 |
| 14 | 07-0608 | Cattle | MN | SB0271 | 25237452534 |
| 15 | 06-8471 | Cattle | TX | SB0121 | 23326442232 |
| 16 | 06-6855 | Cattle | MI | SB0145 | 23237552533 |
| 17 | 06-3641 | Deer | MN | SB0271 | 25237452534 |
| 18 | 06-4034 | Cattle | MI | SB0145 | 23237572533 |
| 19 | 06-2501 | Cattle | TX | SB0265 | 23335432534 |
| 20 | 04-0901 | Cattle | MX | SB0673 | 25245452534 |
| 21 | 04-3121 | Cattle | TX | SB1040 | 25237552533 |
| 22 | 03-5025 | Cattle | TX | SB0140 | 25234452534 |
| 23 | 03-2620 | Cattle | CA | SB1345 | 25336442542 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 24 | 03-0196 | Cattle | CA | SB0673 | 25237452432 |
| 25 | 95-1315 | Deer | MI | SB0145 | 23237552533 |
| 26 | 91-2299 | Deer | NY | SB1069 | 25337441535 |
| 27 | 09-4591 | Deer | MN | SB0271 | 25237452534 |
| 28 | Hbo-5 | Environmental | CA | SB1040 | 25237552533 |
| 29 | Hbo-7 | Human | CA | SB0145 | 25237472533 |
| 30 | Hbo-11 | Human | CA | SB1040 | 25238352533 |
| 31 | Hbo-13 | Human | CA | Unregistered# | 25336442642 |
| 32 | 92-3043 | Elk | NY | SB0265 | 23335432534 |
| 33 | 94-0704 | Elk | MT | SB0265 | 23335432534 |
| 34 | 94-2161 | Elk | MT | SB0265 | 23335432534 |
| 35 | 95-0059 | Elk | MO | SB1069 | 25337441535 |
| 36 | 97-2516 | Feral Swine | HI | SB0145 | 25247542533 |
| 37 | 97-3839 | Elk | WI | SB0265 | 23335432534 |
| 38 | 98-1511 | Elk | KS | SB0265 | 23335432534 |
| 39 | 99-3877 | Feral Swine | HI | SB0815 | 25247542533 |
| 40 | 00-0121 | Elk | WI | SB0265 | 23335432534 |
| 41 | 00-2550 | Elk | WI | SB0265 | 23335432534 |
| 42 | 00-5477 | Elephant | DC | SB0134 | 25432422535 |
| 43 | 00-5480 | Elephant | DC | SB0134 | 25435422535 |
| 44 | 02-1372 | Feral Swine | HI | SB0145 | 25247542533 |
| 45 | 03-5734 | Feral Swine | HI | SB0145 | 25247542533 |
| 46 | 05-5341 | Human | NY | SB0673 | 25237442534 |
| 47 | 05-5354 | Human | NY | SB0673 | 25237442534 |
| 48 | 06-4387 | Feral Swine | HI | SB0145 | 25247542533 |
| 49 | 07-6292 | Cattle | MX | SB0673 | 25237452534 |
| 50 | 09-3461 | Elk | NE | SB0265 | 23335432534 |
| 51 | 09-6071 | Elk | NE | SB0265 | 23335432534 |

| | | | | | |
|---|---|---|---|---|---|
| 52 | 07-6293 | Cattle | MX | SB0121 | 23336442535 |
| 53 | 07-7253 | Cattle | MX | SB0145 | 25237551533 |
| 54 | 07-7901 | Human | MX | SB1828 | 26336442635 |
| 55 | 07-11680 | Feral Swine | HI | SB0145 | 25247542533 |
| 56 | 08-5155 | Feral Swine | HI | SB0145 | 25247542533 |
| 57 | 08-8559 | Deer | NY | SB1069 | 25337441534 |
| ***M.tuberculosis s.s*** **(*n*=10) isolates from the APHIS-USDA culture collection, Ames, IA** | | | | | |
| 58 | 09-0453 | Primate | PA | SB1622 | 24438452534 |
| 59 | 09-0454 | Primate | PA | SB1622 | 24438452534 |
| 60 | 09-0455 | Primate | PA | SB1622 | 24438452534 |
| 61 | 09-3381 | Avian | TX | unregistered# | 44344221637 |
| 62 | 06-8534 | Monkey | WI | unregistered# | 74354421658 |
| 63 | 09-4348 | Primate | NV | unregistered# | 24257242256 |
| 64 | 05-4400 | Elephant | TX | unregistered# | 34242121527 |
| 65 | 09-8103 | Primate | SC | unregistered# | 54343421858 |
| 66 | 09-7906 | Primate | NV | unregistered# | 44332221537 |
| 67 | 97-0352 | Elephant | IL | unregistered# | 34314221639 |
| ***M. bovis*** **(*n*=9) isolates from PHRI culture collections, NJ** | | | | | |
| **No#** | **Isolate ID** | **Host & Year of isolation** | **State / City** | **Spoligotype** | **IS*6110* Bands** |
| 68 | 21540 | Human-2006 | NYC | SB0173 | 1 |
| 69 | 24489 | Human-2009 | NYC | SB1157 | 1 |
| 70 | 20701 | Human-2006 | NYC | SB0242 | 1 |
| 71 | 23244 | Human-2008 | NYC | SB0172 | 1 |
| 72 | 23396 | Human-2008 | NJ | SB0333 | 2 |
| 73 | 26515 | Human-2009 | NYC | SB0509 | 1 |
| 74 | 16862 | Human-2003 | NYC | SB0846 | 1 |

| | | | | | |
|---|---|---|---|---|---|
| 75 | 23217 | Human-2008 | NYC | SB1847 | 1 |
| 76 | 16158 | Human-2002 | Egypt* | SB1160 | 2 |
| **Isolates of *M. bovis* typed as strain BCG (*n*=9) by SNP analysis from PHRI culture collections, NJ** | | | | | |
| 77 | 20658 | Human-2005 | NYC | SB0025 | 1 |
| 78 | 21068 | Human-2006 | NYC | SB0025 | 2 |
| 79 | 24644 | Human-2009 | NYC | SB0025 | 1 |
| 80 | 20051 | Human-2005 | NY | SB0025 | 1 |
| 81 | 9682 | Human-1999 | Russia* | SB0025 | 2 |
| 82 | 9680 | Human-1999 | Russia* | SB0025 | 2 |
| 83 | 7768 | Human-1997 | NH | SB0025 | 1 |
| 84 | 22666 | Human-2007 | NYC | SB0025 | 1 |
| 85 | 20502 | Human-2005 | NYC | SB0025 | 1 |
| ***M. tuberculosis* (*n*=2) isolates typed by SNP analysis which were previously identified as *M.bovis* from PHRI culture collections, NJ** | | | | | |
| 86 | 24282 | Human-2008 | NYC | SB0228 | 3 |
| 87 | 18463 | Human-2003 | NYC | SB0242 | 3 |
| ***M. tuberculosis s.s* (*n*=49) isolates from PHRI culture collections, NJ** | | | | | |
| 88 | 6401 | Human-1997 | NJ | SB0075 | 1 |
| 89 | 6519 | Human-1997 | NJ | SB0075 | 1 |
| 90 | 7396 | Human-1997 | NYC | SB0075 | 1 |
| 91 | 8072 | Human-1998 | NJ | SB0075 | 1 |
| 92 | 9723 | Human-1999 | NJ | SB0075 | 1 |
| 93 | 10225 | Human-1999 | NJ | SB0075 | 1 |
| 94 | 10425 | Human-1999 | NJ | SB0075 | 1 |
| 95 | 13260 | Human-2001 | NJ | SB0075 | 1 |
| 96 | 14435 | Human-2002 | NYC | SB0075 | 1 |
| 97 | 17147 | Human-2003 | NYC | SB0075 | 1 |

| | | | | | |
|---|---|---|---|---|---|
| 98 | 17781 | Human-2003 | NYC | SB0075 | 1 |
| 99 | 17996 | Human-2003 | NYC | SB0075 | 1 |
| 100 | 22813 | Human-2007 | NYC | SB0075 | 1 |
| 101 | 23257 | Human-2008 | NYC | SB0075 | 1 |
| 102 | 24091 | Human-2008 | NJ | SB0075 | 1 |
| 103 | 18928 | Human-2004 | NYC | SB0030 | 1 |
| 104 | 6365 | Human-1997 | NY | SB0030 | 3 |
| 105 | 8423 | Human-1998 | NYC | SB0030 | 3 |
| 106 | 9688 | Human-1999 | NYC | SB0030 | 3 |
| 107 | 13602 | Human-2001 | NYC | SB0030 | 3 |
| 108 | 19733 | Human-2005 | NYC | SB0030 | 3 |
| 109 | 21946 | Human-2007 | NYC | SB0030 | 3 |
| 110 | 23771 | Human-2008 | NYC | SB0030 | 3 |
| 111 | 25703 | Human-2009 | NYC | SB0030 | 3 |
| 112 | 913 | Human-1992 | NYC | SB0030 | 3 |
| 113 | 5401 | Human-1996 | NJ | SB0009 | 3 |
| 114 | 9319 | Human-1998 | NJ | SB0075 | 3 |
| 115 | 9904 | Human-1999 | NJ | SB0075 | 3 |
| 116 | 6478 | Human-1997 | NJ | SB0009 | 2 |
| 117 | 9136 | Human-1998 | NYC | SB0009 | 2 |
| 118 | 12721 | Human-2000 | NJ | SB0009 | 2 |
| 119 | 13571 | Human-2001 | NYC | SB0009 | 2 |
| 120 | 18104 | Human-2003 | NYC | SB0009 | 2 |
| 121 | 19711 | Human-2005 | NYC | SB0009 | 2 |
| 122 | 22665 | Human-2007 | NYC | SB0009 | 2 |
| 123 | 26033 | Human-2010 | NYC | SB0009 | 2 |
| 124 | 11064 | Human-1997 | NJ | SB0030 | 2 |
| 125 | 24991 | Human-2009 | NYC | SB0075 | 2 |

| | | | | | |
|---|---|---|---|---|---|
| 126 | 5855 | Human-1997 | NJ | SB0075 | 2 |
| 127 | 7061 | Human-1997 | NJ | SB0075 | 2 |
| 128 | 8433 | Human-1998 | NJ | SB0075 | 2 |
| 129 | 9140 | Human-1998 | NJ | SB0075 | 2 |
| 130 | 9898 | Human-1999 | NJ | SB0075 | 2 |
| 131 | 10296 | Human-1999 | NJ | SB0075 | 2 |
| 132 | 10443 | Human-1999 | NJ | SB0075 | 2 |
| 133 | 11055 | Human-1999 | NJ | SB0075 | 2 |
| 134 | 21307 | Human-2006 | NYC | SB0075 | 2 |
| 135 | 24810 | Human-2009 | NJ | SB0075 | 2 |
| 136 | 15069 | Human-2002 | NJ | SB0075 | 2 |

[$] Isolates 1- 67 the first two digits represent the year of isolation, except for #1 (early 1990s) and # 28-31 (not known)

[#] Isolates with newly identified unregistered spoligotypes, the octal codes are (in order of appearance on the table) 676713676777600, 000000000003771, 000000000003771, 777777774413771, 777774077560731, 000000000003761, 777717607760771, 776377777760771

[*] 3 isolates that were from out of USA

**Table 2.2:** Details of the four *Mycobacterium bovis* strains used for macrophage infection studies

| No # | Strain ID | Host | Year | State | SNP-Cluster Group | Spoligotype | VNTR Profile |
|---|---|---|---|---|---|---|---|
| 1 | 09-4591 | Deer | 2009 | MN | 1 | SB0271 | 25237452534 |
| 2 | Hbo-5 | Human | 2005 | CA | 2 | SB1040 | 25237552533 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 3 | 08-5055 | Dairy Cattle | 2008 | CA | 1 | SB0140 | 25215452534 |
| 4 | 04-3121 | Beef Cattle | 2004 | TX | 2 | SB1040 | 25237552533 |

**Table 2.3**: Primers used in the qRT-PCR experiments

| Gene Name | Forward Primer Sequence | Reverse Primer Sequence |
|---|---|---|
| PE6 | 5’ GGCATCACTCCCTCAACAAT -3’ | 5’-TTATGGAACCCCCTGGTAGC -3’ |
| mce4C | 5’-CCAAGTAGCAAACACGAAGG -3’ | 5’-TTGAGGCCCGAGACATAAAC -3’ |
| speE | 5’-ATCGTCGCGGGCTACATA -3’ | 5’-GTCGAGCTGGTCCAGGAAC -3’ |
| mmpL12 | 5’-GATCGTCAAGCAAACAGTCG -3’ | 5’-GGTCACCAGGTTCCGATAGA 3’ |
| INO1 | 5’-TCGGAGAACAACACCATCAA -3’ | 5’-TGGTGTCGGCGTAGTACTTG -3’ |
| fadD9 | 5’-AACTACCGAGAGCCTGCAAA -3’ | 5’ GAGGTGCGGAATGTCTTTGT -3’ |
| gyrA (hk*) | 5’-GGTGCTCTATGCAATGTTCG -3’ | 5’-GCCGTCGTAGTTAGGGATGA -3’ |

*hk – housekeeping gene

**Table 2.4:** 118 variant SNPs between the 67 virulent *Mycobacterium bovis* isolates representing three cluster groups on the phylogenetic tree.

| **No. #** | **SNP LOCI** | **SCG-1** | **SCG-2** | **SCG-3** |
|---|---|---|---|---|
| 1. | *AtpH* | C | C | G |

|  |  |  |  |  |
|---|---|---|---|---|
| 2. | *CorA* | C | T | T |
| 3. | *Dha* | A | A | G |
| 4. | *fadD28* | T | T | G |
| 5. | *fadD9-1* | A | A | G |
| 6. | *fadD9-2* | A | G | G |
| 7. | *fadE20* | C | C | G |
| 8. | *fadE27* | A | G | G |
| 9. | *GalT* | G | G | C |
| 10. | *GlmU* | C | C | T |
| 11. | *glnA3* | A | A | G |
| 12. | *GlnB* | A | G | G |
| 13. | *GlnD* | C | C | A |
| 14. | *GlpKb* | G | C | C |
| 15. | *HisD* | G | A | A |
| 16. | *IspD* | C | C | T |
| 17. | *LpqB* | G | G | C |
| 18. | *LpqF* | A | A | G |
| 19. | *mmpL12* | C | C | T |
| 20. | *MmsA* | C | C | A |
| 21. | *NarL* | G | G | C |
| 22. | *NarU* | T | T | C |
| 23. | *NuoB* | C | C | A |
| 24. | *PE31* | T | T | C |
| 25. | *pks12* | T | T | C |
| 26. | *pks6b* | G | T | T |
| 27. | *pks7* | A | A | G |
| 28. | *PPE21* | G | A | A |
| 29. | *RecBb* | G | A | A |

| | | | | |
|---|---|---|---|---|
| 30. | *RhlE* | C | T | T |
| 31. | *SodC* | A | A | G |
| 32. | *SpeE* | A | G | G |
| 33. | *SseA* | A | G | G |
| 34. | *ThioA* | T | T | C |
| 35. | *Mb0085* | T | T | C |
| 36. | *Mb0139* | DEL | DEL | G |
| 37. | *Mb0228c* | T | T | C |
| 38. | *Mb0278c* | T | T | C |
| 39. | *Mb0353* | DEL | DEL | A |
| 40. | *Mb0378c* | A | G | G |
| 41. | *Mb0393* | C | A | A |
| 42. | *Mb0458c* | A | G | G |
| 43. | *Mb0849* | G | A | A |
| 44. | *Mb0899c* | C | T | T |
| 45. | *Mb0937* | T | T | C |
| 46. | *Mb0963* | T | T | C |
| 47. | *Mb1013* | A | G | G |
| 48. | *Mb1150c* | C | G | G |
| 49. | *Mb1365c* | A | G | G |
| 50. | *Mb1427* | G | A | A |
| 51. | *Mb1707* | G | C | C |
| 52. | *Mb1885c* | T | C | C |
| 53. | *Mb1904* | A | G | G |
| 54. | *Mb2029* | C | T | T |
| 55. | *Mb2204c* | G | T | T |
| 56. | *Mb2381c* | T | C | C |
| 57. | *Mb2410c* | C | T | T |

|  |  |  |  |  |
|---|---|---|---|---|
| 58. | *Mb2441c* | T | T | C |
| 59. | *Mb2492c* | G | G | A |
| 60. | *Mb2501c* | T | T | C |
| 61. | *Mb2507c* | G | G | A |
| 62. | *Mb2512c* | T | C | C |
| 63. | *Mb2550* | A | G | G |
| 64. | *Mb2596* | T | T | C |
| 65. | *Mb2661* | G | C | C |
| 66. | *Mb2996* | T | C | C |
| 67. | *Mb3193* | C | T | T |
| 68. | *Mb3328* | A | G | G |
| 69. | *Mb3421c* | T | T | C |
| 70. | *Mb3478* | A | C | C |
| 71. | *Mb3619c* | C | C | T |
| 72. | *Mb3718c* | T | C | C |
| 73. | *Tb39.8-1* | C | C | G |
| 74. | *Tb39.8-2* | C | C | T |
| 75. | *CysN* | T | T | T/C[*] |
| 76. | *dacB1* | A | A | A/G[*] |
| 77. | *fusA2b* | A | A | A/G[*] |
| 78. | *PPE31* | T | T | C/T[#] |
| 79. | *TypA* | T | T | C/T[#] |
| 80. | *Mb0007* | G | G | A/G[#] |
| 81. | *Mb0244* | T | T | C/T[#] |
| 82. | *Mb1072c* | T | T | T/G[*] |
| 83. | *Mb1404* | A | A | A/G[@] |
| 84. | *Mb1495* | C | C | C/T[$] |
| 85. | *Mb1794c-1* | G | G | G/A[*] |

|  |  |  |  |  |
|---|---|---|---|---|
| 86. | *Mb1794c-2* | T | T | T/C[*] |
| 87. | *Mb1860* | T | T | T/C[*] |
| 88. | *Mb2067c* | A | A | A/G[*] |
| 89. | *Mb2261* | A | A | A/G[*] |
| 90. | *Mb2439c* | C | C | T/C[#] |
| 91. | *Mb2558* | A | A | G/A[α] |
| 92-118. | IGR1, IGR2, IGR3, IGR4, IGR5, IGR6, IGR7, IGR8, IGR9, IGR10, IGR11, IGR12, IGR13, IGR14, IGR15, IGR16, IGR17, IGR18, IGR19, IGR20, IGR21, IGR22, IGR23, IGR24, IGR25, IGR26, IGR27 |  |  |  |

[*] Allele observed only in two isolates 16158 & 23217

[#] Allele observed only in five isolates 95-0059, 08-8559, 91-2299, 00-5480 & 00-5477

[@] Allele observed only in isolate 08-2906

[$] Allele observed only in isolate 16158

[@] Alllele observed only in four isolates 08-8559, 91-2299, 00-5480 & 00-5477

● No SNP-cluster group specific distribution observed for the 27 SNPs of the intergeneic region.

**Table 2.5:** Details of the six virulence-associated genes compared by qRT-PCR for their expression profiles at two time points post infection, 30min and 2hr.

| Gene ID | Gene description | Type of SNP | Implicated role in virulence |
|---|---|---|---|
| PE6 | PE family | Nonsynonymous | Source of antigenic variation, evasion of host immune response |

| | | | |
|---|---|---|---|
| mce4C | Mammalian cell entry protein | Nonsynonymous | Required for host cell entry and infection |
| mmpL12 | Transmembrane transport protein | Nonsynonymous | Transmembrane transport protein |
| fadD9 | Fatty acid CoA ligase | Nonsynonymous | Lipid metabolism |
| speE | Spemidine synthase | Synonymous | Inhibits NOS which is mycobactericidal |
| INO1 | Myo-inositol-1-phosphate synthase | Intergenic, 16bp upstream | Biosynthesis of lipids/fatty acids and signal transduction |

## FIGURES & FIGURE LEGENDS

**Figure 2.1a:** The genome wide distribution of the "350 target" SNP loci across the 4.3MB *M. bovis* genome. The figure was generated using the DNAPlotter tool from the Artemis: Genome browser and annotation tool.

**Figure 2.1b:** The genome wide distribution of the "206 variant" SNPs across the 4.3MB *M. bovis* genome. The 59 synonymous substitutions are shown in blue, the 120 non-synonymous changes are shown in red and the 27 intergeneic SNPs are shown green. The figure was generated using the DNAPlotter tool from the Artemis: Genome browser and annotation tool.

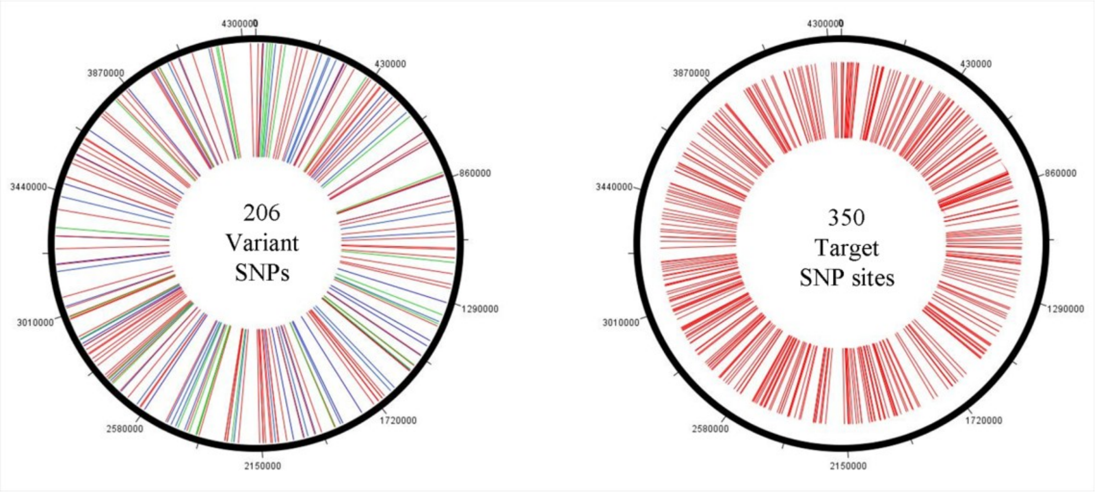


**Figure 2.2:** The genome wide distribution of the "84 genic" SNPs that separate the 67 virulent *M. bovis* isolates from the 10 attenuated BCG-lineage isolates. The synonymous[*] changes are shown in blue and the non-synonymous[#] changes are shown in red. The list of 84 genes is below the figure.

[*] Synonymous SNPs (*n*=28): *acn, embB, narK1, Mb0065-1, Mb1777, Mb2247c, oplA, pks8, fadE3, Mb2048c, Mb0023, Mb3682c, PPE12, sigE, Mb0203, Mb1808c, galE2, PE_PGRS63, tpi, Mb0600c-3, Mb1897, Mb0849-1, dnaG, gmk, pepB, PPE44, xerC* and *Mb1366.*

[#] Non-synonymous SNPs (*n*=56): *adh, lipJ, fadD24, pks13, lppP, pyrD, aroA, recG, mce4C, pks2, ndh, PPE35b, recBa, atsAb, malQ, mbtB, furA, clpX, PE_PGRS39, dnaE2, dppA, hpt, mmsB, pks12, PPE9, cyp136, Mb0271c, Mb2740c-2, Mb1780, Mb1019c, Mb2266, Mb3238 Mb1538, , Mb3763, Mb3270c, Mb3300, Mb2750, Mb2196, Mb0193, Mb3938, Mb2787, Mb3191c, Mb1098c, Mb3635c Mb1756, Mb0600c-2, Mb3865, mmpL12-2, Mb2595, Mb0474c, Mb3412c, Mb1130, Mb0300, Mb0978c, Mb1427-2* and *Mb2830.*

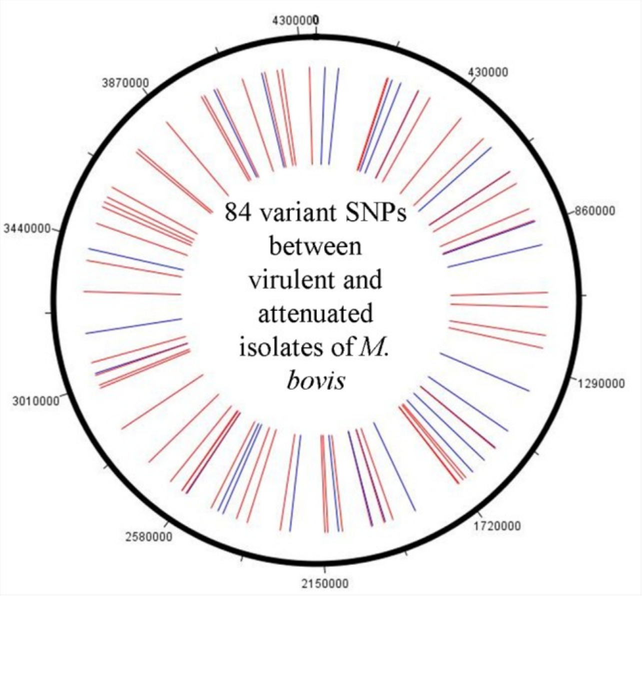


**Figure 2.3:** A consensus linear phylogenetic tree is shown here generated using the Maximum Parsimony algorithm using 1000 bootstrap replicates using the MEGA4.1 software. The tree represents the SNP genotypes of 75 *Mycobacterium bovis* (confirmed by SNP analysis) and 61 *Mycobacterium tuberculosis* (includes the 2 isolates previously identified as *M. bovis*) isolates along with the sequenced strains of virulent *M. bovis* strain AF2122/97, *M. bovis* vaccine strain BCG-Pasteur, and two *M. tuberculosis* strains H37Rv and CDC1551.The tree is rooted to the isolates of the *M. bovis* BCG-strain. Five major SNP-cluster groups, 1 through 5, indicative of the five “SNP genotypes” are identified. The first 3 are the major *M.bovis* SNP-cluster groups that include virulent isolates from various hosts and geographic locations. Cluster group 1 has all the isolates from MN, Cluster group 2 includes all the isolates of MI and HI and Cluster group 3 has all the elk isolates varying in time and geographic origins. Cluster group 5 includes the 9 human *M. bovis* isolates which cluster together with the attenuated BCG-Pasteur strain. Cluster group 4 includes all the *M. tuberculosis* isolates from animal and human hosts including the two sequenced strains. The details of the isolates that represent the five SNP-cluster groups are listed in Table 2.1.

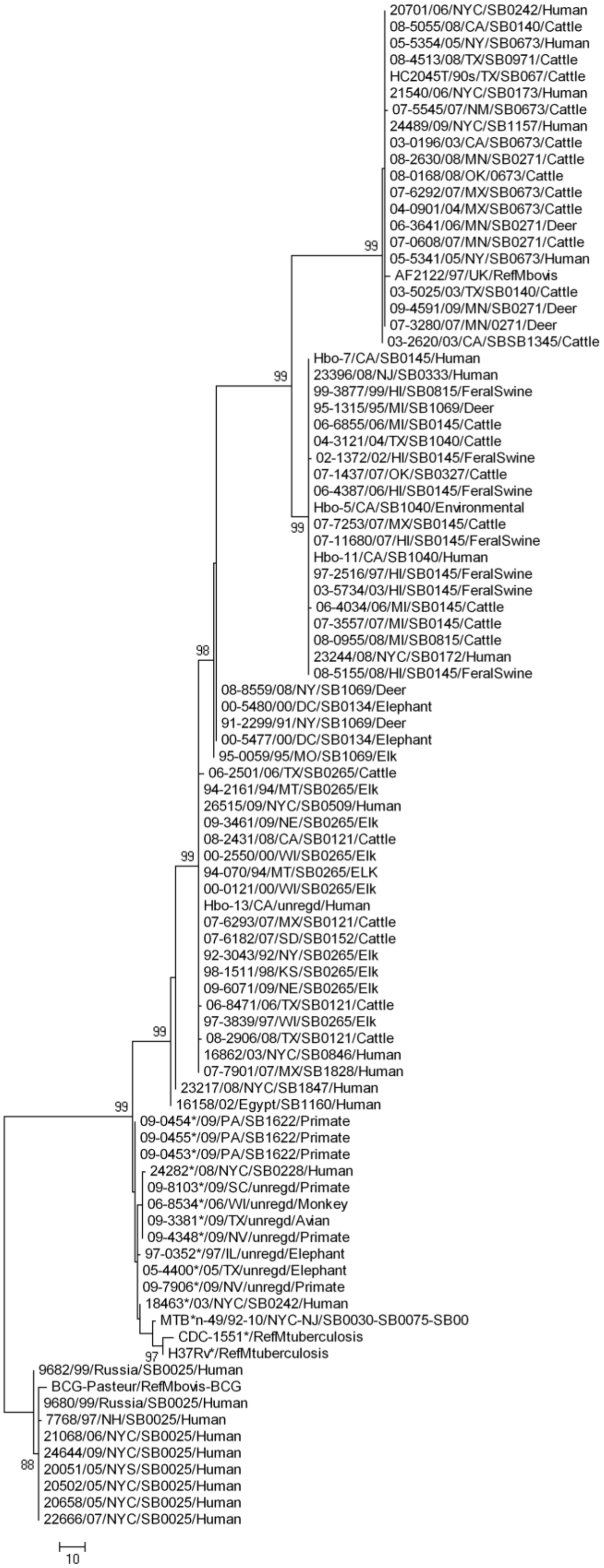

20701/06/NYC/SB0242/Human
08-5055/08/CA/SB0140/Cattle
05-5354/05/NY/SB0673/Human
08-4513/08/TX/SB0971/Cattle
HC2045T/90s/TX/SB067/Cattle
21540/06/NYC/SB0173/Human
07-5545/07/NM/SB0673/Cattle
24489/09/NYC/SB1157/Human
03-0196/03/CA/SB0673/Cattle
08-2630/08/MN/SB0271/Cattle
08-0168/08/OK/0673/Cattle
07-6292/07/MX/SB0673/Cattle
04-0901/04/MX/SB0673/Cattle
06-3641/06/MN/SB0271/Deer
07-0608/07/MN/SB0271/Cattle
05-5341/05/NY/SB0673/Human
AF2122/97/UK/RefMbovis
03-5025/03/TX/SB0140/Cattle
09-4591/09/MN/SB0271/Deer
07-3280/07/MN/0271/Deer
03-2620/03/CA/SBSB1345/Cattle
99
99
Hbo-7/CA/SB0145/Human
23396/08/NJ/SB0333/Human
99-3877/99/HI/SB0815/FeralSwine
95-1315/95/MI/SB1069/Deer
06-6855/06/MI/SB0145/Cattle
04-3121/04/TX/SB1040/Cattle
02-1372/02/HI/SB0145/FeralSwine
07-1437/07/OK/SB0327/Cattle
06-4387/06/HI/SB0145/FeralSwine
Hbo-5/CA/SB1040/Environmental
99
07-7253/07/MX/SB0145/Cattle
07-11680/07/HI/SB0145/FeralSwine
Hbo-11/CA/SB1040/Human
97-2516/97/HI/SB0145/FeralSwine
03-5734/03/HI/SB0145/FeralSwine
06-4034/06/MI/SB0145/Cattle
07-3557/07/MI/SB0145/Cattle
08-0955/08/MI/SB0815/Cattle
23244/08/NYC/SB0172/Human
08-5155/08/HI/SB0145/FeralSwine
98
08-8559/08/NY/SB1069/Deer
00-5480/00/DC/SB0134/Elephant
91-2299/91/NY/SB1069/Deer
00-5477/00/DC/SB0134/Elephant
95-0059/95/MO/SB1069/Elk
06-2501/06/TX/SB0265/Cattle
94-2161/94/MT/SB0265/Elk
26515/09/NYC/SB0509/Human
09-3461/09/NE/SB0265/Elk
08-2431/08/CA/SB0121/Cattle
99
00-2550/00/WI/SB0265/Elk
94-070/94/MT/SB0265/ELK
00-0121/00/WI/SB0265/Elk
Hbo-13/CA/unregd/Human
07-6293/07/MX/SB0121/Cattle
07-6182/07/SD/SB0152/Cattle
92-3043/92/NY/SB0265/Elk
98-1511/98/KS/SB0265/Elk
09-6071/09/NE/SB0265/Elk
06-8471/06/TX/SB0121/Cattle
97-3839/97/WI/SB0265/Elk
99
08-2906/08/TX/SB0121/Cattle
16862/03/NYC/SB0846/Human
07-7901/07/MX/SB1828/Human
23217/08/NYC/SB1847/Human
16158/02/Egypt/SB1160/Human
99
09-0454*/09/PA/SB1622/Primate
09-0455*/09/PA/SB1622/Primate
09-0453*/09/PA/SB1622/Primate
24282*/08/NYC/SB0228/Human
09-8103*/09/SC/unregd/Primate
06-8534*/06/WI/unregd/Monkey
09-3381*/09/TX/unregd/Avian
09-4348*/09/NV/unregd/Primate
97-0352*/97/IL/unregd/Elephant
05-4400*/05/TX/unregd/Elephant
09-7906*/09/NV/unregd/Primate
18463*/03/NYC/SB0242/Human
MTB*n-49/92-10/NYC-NJ/SB0030-SB0075-SB00
CDC-1551*/RefMtuberculosis
97
H37Rv*/RefMtuberculosis
9682/99/Russia/SB0025/Human
BCG-Pasteur/RefMbovis-BCG
9680/99/Russia/SB0025/Human
7768/97/NH/SB0025/Human
21068/06/NYC/SB0025/Human
24644/09/NYC/SB0025/Human
88
20051/05/NYS/SB0025/Human
20502/05/NYC/SB0025/Human
20658/05/NYC/SB0025/Human
22666/07/NYC/SB0025/Human
10

**Figure 2.4:** The differential survival pattern of the four *M.bovis* strains challenged in the mouse macrophage cell line model of *in vitro* infection is shown here. The Y-axis represents the log10 cfu/ml of the bacterial counts and X-axis represents the time points post infection. Strain1, which is a 2009 MN deer isolate shown in blue with red knots persists in the host cell up to 48hrs p.i. Strain2, is a 2005 human isolate from CA, shown here in the orange with blue knots, continues to persist in the host cell up to 48hrs p.i. Strain3 is a 2008 dairy isolate from CA, represented in the pink with black knots and was cleared by the macrophages 2hrs p.i. Strain4 is a 2004 TX beef cattle isolate, shown here in the purple with green knots, despite what looks like an initial phase of dormancy at 0min p.i, it had maximal recovery at 48hrs p.i as compared to the other 3 strains.

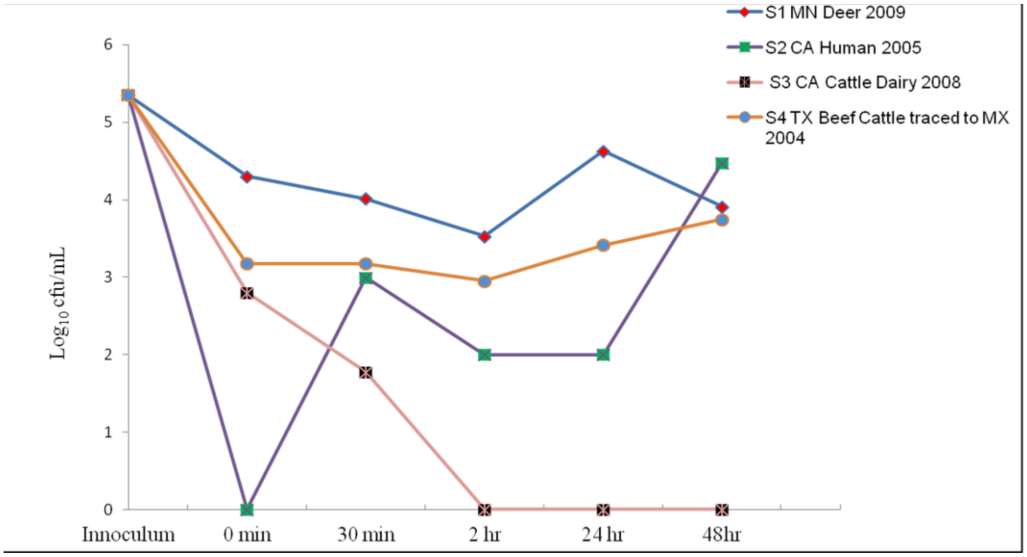
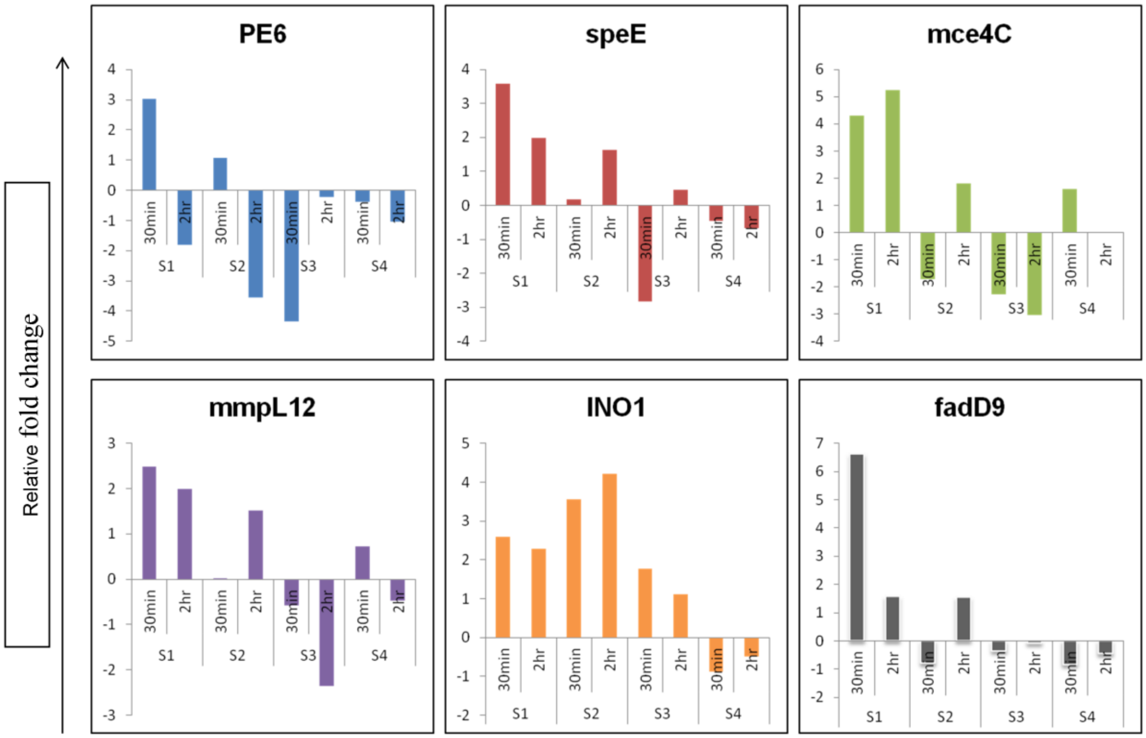


**Figure 2.5:** Relative expression profiles of six virulence-associated genes for the four *M. bovis* strains using qRT-PCR. The genes of interest include PE6, mce4C, mmpL12, speE, fadD9 and INO1 (Table 5) and their relative expression profiles at two time points post infection- 30min and 2hr compared to time point 0 minute, in an *in vitro* macrophage infection assay. The results are reported as relative fold change for the six genes and range from 0 to 6.5, plotted on the Y-axis. The two post infection time points and the four strains are shown on the X-axis. Test and control samples were normalized using the house keeping gene, gyrA, and relative expression calculated by 2-$^{\Delta\Delta}$CT method.



# Chapter 3

## Comparative genomics between two clinical isolates of *Mycobacterium bovis* from the USA reveals low genetic variability and a high degree of conservation of core genomic components.

Of the multitude of bacterial pathogens, *Mycobacterium bovis* has one of the widest known mammalian host ranges including the ability to infect humans. Despite the disease antiquity and elucidation of whole genome sequence of a UK strain almost a decade ago, the genetic basis of its host specificity and pathogenecity remain poorly understood. In this study, two *M. bovis* (MBO) strains isolated from cattle (MBO Corsentino) and elk (MBO NE elk) were sequenced using Illumina HiSeq 2000$^{TM}$ next-gen sequencing platform. The genome of *M. bovis* Corsentino comprises a circular chromosome of 4307383 bp with average G+C content of 65.4% with 4008 predicted protein coding regions. The genome of *M. bovis* NE elk comprises of a circular chromosome of 4302584 bp with average G+C content of 65.4% with 4009 predicted protein coding sequences. Genome comparisons against the UK origin reference strain AF2122/97 did not reveal any unique genes or large sequence polymorphisms. A total of 1139 and 1184 SNPs were identified in Corsentino and NE elk genomes when compared to the AF2122/97 genome, respectively. Comparison of MBO Corsentino and MBO NE elk genomes identified ~900 SNPs between them. Comparative genomics with other members of the *Mycobacterium Tuberculosis Complex* revealed a very high sequence similarity between the strains. Thus, this study provides new evidence in favor of low genetic variability in this organism suggesting variations in gene expression and post-transcriptional or post-translational regulation events as the likely sources of host specificity and phenotypic variation.

## INTRODUCTION

*Mycobacterium bovis* is a unique ecotype of the *Mycobacterium Tuberculosis* Complex group since it has the widest known mammalian host range including humans (93, 163). *M. bovis* is the causative agent of bovine tuberculosis and is responsible for worldwide annual losses to agriculture of $3 billion (77). Geographical localization of molecular types (clones, clonal complexes or groups) is emerging as a common theme for this disease at global, national and regional levels (93). Global lineages of *M. bovis* termed African 1 (unique chromosomal deletion RDAf1 and absence of spacer 30 in spoligotyping), African 2 (unique chromosomal deletion RDAf2 and absence of spacers 3 to7 in spoligotyping), European 1 (unique chromosomal deletion RDEu1) and European 2 (unique chromosomal deletion RDEu2 and absence of spacer 21 in spoligotyping) have been described, and these are based on large sequence polymorphisms rather than whole genome sequences (17, 168, 211, 221, 222). Previous study (manuscript in review, J. Clin. Microbiol; Chapter 2, this dissertation) using genome-wide single nucleotide polymorphism (SNP) typing, identified three primary lineages ("SNP-cluster" groups) of *M. bovis* in the USA. Within these lineages, three geographic sub-lineages of *M. bovis* that were localized in the states of Michigan, Minnesota and Hawaii were identified. These SNP genotypes among isolates from the 3 states were not geographically restricted. Multiple SNP genoptes were identified among isolates from Texas, New York, California suggesting a dynamic mixing and multiple introductions in these localities. Interestingly, among all isolates with a variety of host origin, elk isolates from 6 states of the US representing a 15-yr time period (1992-2009) were clonal. The absence of recombination between *M. bovis* organisms (89, 147) along with movement limitations of domesticated cattle in some states along with attempts to eradicate the disease have likely played a part in generating the molecular clones of *M. bovis* in the country. In the past decade, numerous cases of bovine tuberculosis have been reported in cattle as well as human population across the USA (95, 142) despite existing test and slaughter surveillance programs. Free ranging white tailed deer (*Odocoileus virginianus*) in the US have been recognized as the primary reservoir hosts of *M. bovis* leading to bovine tuberculosis outbreaks due to spill over at deer-cattle-human-environmental interface (183-186).

One of the key advances in the last decade in the understanding of *M. bovis* has been the elucidation of the complete genome sequence of the pathogen (77). The availability of multiple *M. tuberculosis* genomes (Cole, 1998 #77; Gordon, 1999 #208; Fleischmann, 2002 #212) and their comparative analyses has revealed novel information about associations between strains, their host populations, their evolution and interaction with the host immune system and environment, and strategies for vaccine design. Comparative genomics have identified sequence differences in the genomes of *M. bovis* BCG vaccine strains and *M. tuberculosis* laboratory and clinical strains (82, 97, 137, 202). However such studies in *M. bovis* have not been performed due to unavailability of genomes of multiple strains. The re-emergence and prevalence of bovine tuberculosis in the US warrants obtaining genetic information on local strains. Exploration of the genome sequence is expected to offer major insights on the evolution, host preference, and pathobiology of *M. bovis*.

In the present study, the whole genome sequences for two US strains are described, isolated from cattle and elk, respectively, and their comparative genomic analyses with other members of *Mycobacterium Tuberculosis* Complex.

## MATERIALS AND METHODS

**Bacterial isolates:** Two *M. bovis* isolates including one highly virulent cattle isolate from a Colorado dairy farm identified as MBO-Corsentino and one isolated from an elk in Nebraska identified as MBO NE-Elk were analyzed. Inasmuch as our genomewide SNP analysis of elk isolates representing a 10-year period and 6 different states revealed a clonal pattern, it was decided to choose a recent clone for genome analysis. The Corsentino strain was isolated from a 2010 outbreak in Colorado and identified as highly transmissible clone. Thus it was chosen to represent bovine origin strains for genome sequencing and comparative genomic analysis. The *M. bovis* genomic DNA for both the isolates was prepared using standard extraction techniques (4) at the USDA-NADC lab in Ames, Iowa and shipped on ice to laboratory.

**Sequencing:** Complete genome sequencing was performed using the Illumina HiSeq 2000$^{TM}$ next generation sequencing platform available to us through the BioMedical Genomics Center (BMGC) at the University of Minnesota. Briefly about 5µg of genomic DNA of each isolate was submitted for sequencing. Samples were quantified at the sequencing center using flourimetry (Pico Green assay). *Genomic library creation* steps included DNA shearing (Covaris acoustic shearing), fragment purification and end polishing, and ligation to indexed (barcoded) adaptors. The library was then size selected, size distribution validated using capillary electrophoresis, and quantified using fluorimetry (PicoGreen) and via Q-PCR. Indexed libraries were then normalized, pooled, clustered on a flow cell, and loaded onto the instrument for sequencing. Both isolates were loaded as single samples per lane on a 100-bp paired end multiplexed run. Approximately 90 million raw reads per isolate were obtained of which ~75 million passed filter reads.

***De novo* and reference guided assembly of sequences reads for a draft genome:** Short read sequences were modified to a Solexa FastQ format and stored in the online database of the Minnesota Supercomputing Institute (MSI) for easy retrieval. The open source web-based platform Galaxy (https://main.g2.bx.psu.edu/) was used for quality check and filtering of the short reads. Ambiguous bases and artifactual sequences were removed using the NGS-QC and manipulation tools from Galaxy interface. The de novo assembler Velvet http://www.ebi.ac.uk/~zerbino/velvet/ (268) was used for assembling the short reads. Given the nature of assembly process using de Bruijn graphs (197), the sequences from either MBO Corsentino or MBO NE Elk were assembled across a range of k-mers. These resulting Velvet contigs were then stored in the database. Each genome was dealt with separately. A consensus sequence from the mapping of the short reads to the reference genome *M. bovis* AF2122/97 was obtained and broken up into contigs wherever any ambiguous bases were recorded in the consensus. These contigs were then assembled with the *de novo* contigs from a given k-mer and the one with best assembly results based upon N50 score was chosen. The N50 score is a standard statistical measure that evaluates the assembly quality and indicates the scaffold length such that 50% of the assembled sequences lie in scaffolds of this size or larger (173). The scaffolds with longer N50 scores especially benefit the identification of protein-coding genes (173). In all, this meant that there were 8 new assemblies for each genome, all with fewer contigs. To check the validity of these new contigs, Maq (129) aligner was used to map the short reads back to the contigs. The reference guided assembly was done against the *M. bovis* AF2122/97 genome using MOSAIK package available at - http://bioinformatics.bc.edu/marthlab/Mosaik. The program MosaikAligner pairwise aligns each read to a specified series of reference sequences. MosaikSort resolves paired-end reads and sorts the alignments by the reference sequence coordinates. Finally, MosaikAssembler parses the sorted alignment archive and produces a multiple sequence alignment which is then saved into an assembly file format. MOSAIK produces gapped alignments using the Smith-Waterman algorithm. This is a well-known algorithm for performing local sequence alignment and for determining similar regions between two nucleotide or protein sequences. Instead of looking at the total sequence, the Smith–Waterman algorithm compares segments of all possible lengths and optimizes the similarity measure.

**Annotation, sequence analysis and comparison of genomes:** We used the Multiple Alignment of Conserved Genomic Sequence With Rearrangements (Mauve) (52) software program to align and compare the two *M. bovis* - Corsentino and NE elk genomes along with other complete genomes of the *Mycobacterium tuberculosis* complex (MTC). Mauve allows for identification and alignment of conserved genomic DNA in the presence of rearrangements and horizontal transfer. The alignment of genomes was made using the progressive Mauve algorithm that identifies successive sequences with exact similarity shared by two or more genomes and, with a distance matrix based on genomic conservation to build a tree guide. These similar regions are referred to as local regions and represent a sequence shared by two or more genomes included within the alignment. The linear regions are grouped into blocks neighboring local linearity (known as locally colinear blocks, LCB), which are separated by genomic islands. The Mauve program was also used to call for genome wide SNPs in the two genomes, each compared to the reference. For every polymorphic site in an alignment, the SNP file records the nucleotides present in each genome at that site, along with the sequence coordinates of the site in each genome. A preliminary version of RAST (7) was used to transfer the annotation from the *M. bovis* AF212297 (77) genome. The genomes from the MTC that were used for comparative genomic analyses included the *M. tuberculosis s.s* strains H37Rv and CDC 1551, along with the *M. bovis* strains AF2122/97 and the vaccine strain BCG-Pasteur. The sequences for these strains were obtained from GenBank www.ncbi.nlm.nih.gov/Genome. Artemis Comparison Tool - ACT (32), a mainstream tool for visualization, graphical presentation and annotation of sequence data was used for data representation.

**Confirmation of gaps / large sequence variations by PCR:** Gaps larger than 3,000bp and ranging up to 10,000bp observed in both the Corsentino and NE elk genomes. These gaps were confirmed for their presence / absence by PCR on respective genomic DNAs. Primers were designed using the genome sequence of *M. bovis* reference strain AF2122/97 for the corresponding gaps. Primer3 (v.0.4.0, http://frodo.wi.mit.edu/) was used to design the forward and reverse set of primers. PCR was performed with approximately 5-10ng of genomic DNA, forward and reverse primers at 0.5µM concentration and a PCR master mix containing high fidelity taq polymerase along with dNTPs and buffer (Phusion$^{TM}$, 2X, New England Biolabs) were added to the reaction tube. The final reaction volume was adjusted to 25µL by adding water. PCR amplifications were performed in a Eppendorf PCR system (Eppendorf, Hauppauge, NY), using program setting with an initial activation step of 95°C for 2 min, followed by 30 cycles of denaturation step of 30s at 95°C, annealing for 30 sec at 55°C, extension for 1 min 30s at 72°C, and ending with a final elongation step for 7 min at 72°C. The PCR products along with 100bp ladder were visualized on a 2% agarose gel using ethidium bromide staining.

## RESULTS

**Genome sequencing and annotation:** Approximately 70 million bp were included in the assembly for each of the genomes, yielding 60X high quality genome coverage. The quality control (QC) checks on the raw reads were performed using the Fastqc tool from the Galaxy interface. Phred scores (66) for the reads along with basic QC checks for both the Corsentino and NE elk *M. bovis* genomes are shown in Figure 3.1(a-d) and Figure 3.2(a-d) respectively and were of high quality. The *de novo* Velvet assembly of the *M. bovis* Corsentino genome resulted in a total of 642 contigs versus 617 contigs for the *M. bovis* NE elk genome. The Mosaik reference guided assembly resulted in a single large contig for both the genomes. The RAST server (7) was used to transfer the annotation from the *M. bovis* strain AF2122/97 used as reference. The genome of *M. bovis* Corsentino (Figure 3.3a) comprises a circular chromosome of 4307383 bp with average G+C content of 65.4% with 4008 predicted protein coding regions (Supplemental Information-Table 1). The genome of *M. bovis* NE elk comprises of a circular chromosome of 4302584 bp with average G+C content of 65.4% (Figure 3.3b) with 4009 predicted protein coding sequences (Supplemental Information- Table 2). The coding sequences have high percent similarity to the reference genome (Figure 3.3c). Each genome has a single copy of predicted 5S, 16S, and 23S rRNA genes and 48 copies of predicted tRNAs genes.

**Comparative genomic analyses:** Comparative genomic analysis was performed using the reference strain *M .bovis* AF2122/97. The draft genomes of both the US strains Corsentino and NE elk are similar in size ~4.3 Mb compared to the reference strain (4345492 bp) and with comparable G+C content of 65.4% (both) as compared to 65.63% of the reference strain. The Mauve genome alignment identified a total of 345 LCB (locally collinear blocks) grouped linearly between the *M. bovis* Corsentino and AF2122/97 genomes and 325 LCB between the *M. bovis* NE elk and AF2122/97 genomes (Figures 3.4 and 3.5). By adding the length of the LCB found in each genome and comparing the value with the corresponding genome length, it was found that these common regions covered >98.5% of each genome analyzed. The SNP calling tool from the Mauve program identified 1139 SNPs between the Corsentino and AF2122/97 genomes (Figure 3.4) and 1184 SNPs between the NE elk and AF2122/97 genomes (Figure 3.5) (SNP List: Supplemental Information-Tables 3-4). The two genomes of the US strains are highly clonal to each other and share over 99% sequence identity to the UK strain AF2122/97 (Figure 3.6). Comparative genomic analysis was also extended to include other members of the *Mycobacterium tuberculosis* complex group. These included the human *Mycobacterium tuberculosis* strains H37Rv and CDC1551 along with the vaccine / attenuated strain of *M. bovis* BCG-Pasteur. All six genomes compared share significant sequence identity. There is no evidence of extensive genomic translocations, duplications or inversions (Figure 3.7).

**PCR for gaps / sequence variations:** There were approximately 100 gaps / sequence variations identified in each of the draft genomes (Corsentino and NE elk) as compared to the reference strain AF2122/97 (Figures 3, 4). These gaps ranged from 500 bp to about 10,000 bp. We investigated the large gaps (> 3000bp) in both the genomes using conventional PCR to confirm them as assembly / sequencing errors versus real deletions or sequence variations. The gaps investigated included a 3,500 bp region missing in both the US strains along with two 10,000 bp and one 8,000 bp region missing in the Corsentino genome (*n*=3) and two 8,000 bp, one 7,000 bp and one 6,000 bp (*n*=4) regions missing in the NE elk genome. The PCR results showed that these gaps were not real and thus confirming them as likely assembly errors. We noticed fewer and smaller gaps in the reference guided assembly genomes as compared to the *de novo* assembled genomes, suggesting that the assembly methods and algorithms used in the *de novo* method need to be re-visited.

## DISCUSSION

Strain AF2122/97, a virulent cow strain from the UK, is the only reference whole genome sequence available for *M. bovis*. Here we provide the draft sequences for two *M. bovis* strains from the USA, *M. bovis* Corsentino and *M. bovis* NE elk representing two animal host species- cattle and elk.

The *M. bovis* Corsentino and NE elk genomes from the US are, as expected, homologous to each other. This reduced genetic diversity has also been observed among *M. bovis* strains in the British Isles. The cause is speculated to be result of a population bottleneck caused by bovine tuberculosis control programs that have been operating for the past many years, and the dominance of a single clonal complex could be either the result of selection (221, 223) or convergence due to host adaptation resulting from host and pathogen co-evolution over millenia. The Corsentino strain is a highly virulent cattle isolate from Colorado and NE elk strain belongs to a subset of clonal, elk adapted *M. bovis* strains (as revealed by previous study using SNP genotyping). However other than identification of SNPs we did not find any strain specific large sequence polymorphisms that were hypothesized to drive host adaptation of these strains. Comparison of the three ~ 4.3 Mb *M. bovis* genomes identified ~1100 SNPs between them. A study comparing two *M. tuberculosis* genomes (strain H37Rv and CDC 1551) detected only ~900 SNPs between them (86). Similarly, comparison of BCG and *M. bovis* genomes identified only about 350 – 700 SNPs between them, which may hold clues for attenuated profile of the BCG strain (76, 194). Thus the low number of SNPs between the genomes confirms the restricted level of structural gene sequence variation reported previously (170, 224). In this study a comparison was also made between multiple genomes of the *M. tuberculosis* complex and it was observed that all grouped in a general linear block, which validates the information on the low genetic variability within the complex. Despite their different host tropisms, the members of MTC are characterized by 99.9% or greater sequence similarity at nucleotide level. The average divergence between *M. tuberculosis* and *M. bovis* is less than 0.05% (77) and can be compared with a divergence of 1.6% between two strains of *Escherichia coli* (200, 223) . The Mauve alignment of *M. tuberculosis* strain H37Rv against the three *M. bovis* strains AF2122/97, Corsentino and NE elk identified ~ 2600 SNPs that separate the human tuberculosis strain from those of animal origin, thus further validating the reported <0.05% divergence. A global perspective of the distribution of *M. bovis* genotypes is not currently feasible due to lack of availability of data from most parts of the world.

However, one can expect to learn more about the lineages of US strains as more genomes are sequenced, with these two sequences serving as references for the mapping of further genomes generated using next generation sequencing techniques.

A study (36) comparing six MTC genomes (H37Rv, H37Ra, CDC 1551, F11, BCG-Pasteur and AF2122/97) identified the percent similarity between these genomes ranging from 96.1% to 97.8% and the variable regions observed in the genomes were mainly restricted to transposable elements, the PE-PGRS families and intergenic regions. In this study, the large variations/gaps observed in the two genomes ranging in size 3500bp to 10000 bp were noticed mainly for regions coding the PE-PPE family genes (especially the PE-PGRS sub family genes) and others that included RV1 phage proteins, repeat protein, transposase *IS*1081, ESAT-6 like and a few hypothetical proteins. Repetitive DNA sequences are abundant in many bacterial genomes, including mycobacteria and have always presented technical challenges for sequence alignment and assembly programs. Next-generation sequencing projects, with their short read lengths and high data volumes, have made the assembly of repetitive regions more challenging (242). From a computational perspective, repeats create ambiguities in alignment and assembly, which, in turn, can produce biases and errors when interpreting results (242). However we confirmed by PCR using high fidelity Taq polymerase that these were not real deletion events and hence were likely generated during genome assembly. Though we did not resolve these deletions observed in our dataset in the present study, we recommend sequencing of these regions in the future to fill the gaps and close the genome.

Genomic studies using sequence level comparisons and post genomic analysis has shown that MTC organisms evolve through clonal evolution, mutation and gene deletion (30, 162). While MTC lineages are often considered to be monomorphic, molecular typing techniques such as IS*6110*-RFLP, spoligotyping, and variable number tandem repeats - mycobacterial interspersed repetitive unit (VNTR-MIRU) reveal a certain level of genetic diversity among strains (106, 189, 230). Studies using single nucleotide polymorphisms (SNPs) (43, 57, 68, 85, 86) and large sequence polymorphisms (LSPs) (2, 28) in the last decade have helped classify *M. tuberculosis* strains into closely related groups and estimateg genetic relationships among isolates by the analysis of several variations at the genome scale level. These studies confirm the genetic diversity and genome plasticity of the mycobacterial genome (102). The tubercle bacilli are considered to be amongst the most genetically intractable microorganisms as a result of long generation times, fastidious growth requirements and contagiousness (40). Comparative genomics is expected to lead to identification of genes restricted to a given mycobacterium that may play unique biological roles, and serve as sources specific antigens or potential drug targets (40).

The goals of this study were to (i) to provide a reference *M. bovis* genome of US origin and (ii) identify and compare genomic variations between *M. bovis* isolates from different host species to identify genes responsible for host specificity. As compared to the reference strain AF2122/97, both the Corsentino and NE elk strains did not contain any unique large sequence polymorphisms or unique genes, suggesting that differences in gene expression or regulation patterns are the key players in their host tropisms (77). Data generated by this and other future sequencing projects can help identify the most informative panel of markers of genomic variability in *M. bovis*. As the cost of whole-genome sequencing continues to decrease and next-generation sequencing platforms become integrated into public health practice, combined microbial, genomic and epidemiologic approaches will become as an important and tractable first step toward molecular epidemiology and control of tuberculosis.

# LIST OF FIGURES

**Figure 3.1 (a-d). Quality check of raw reads for the sequences of *M. bovis* Corsentino obtained from Illumina$^{R}$ sequencing. The FastQC package from the Galaxy interface was used for QC analyses (www.galaxy.msi.umn.edu)**
Clockwise from top left: (a) Per base sequence quality, using Phred score, shows a high score across most sequences (> 20 = good score, higher the better) (b) Per sequence quality scores (high) (c) Per sequence GC content (~ 65% as expected for mycobacterial genome) (d) Per base GC content (shoulder peak showing presence of some contaminating sequences that need filtering prior to assembly)

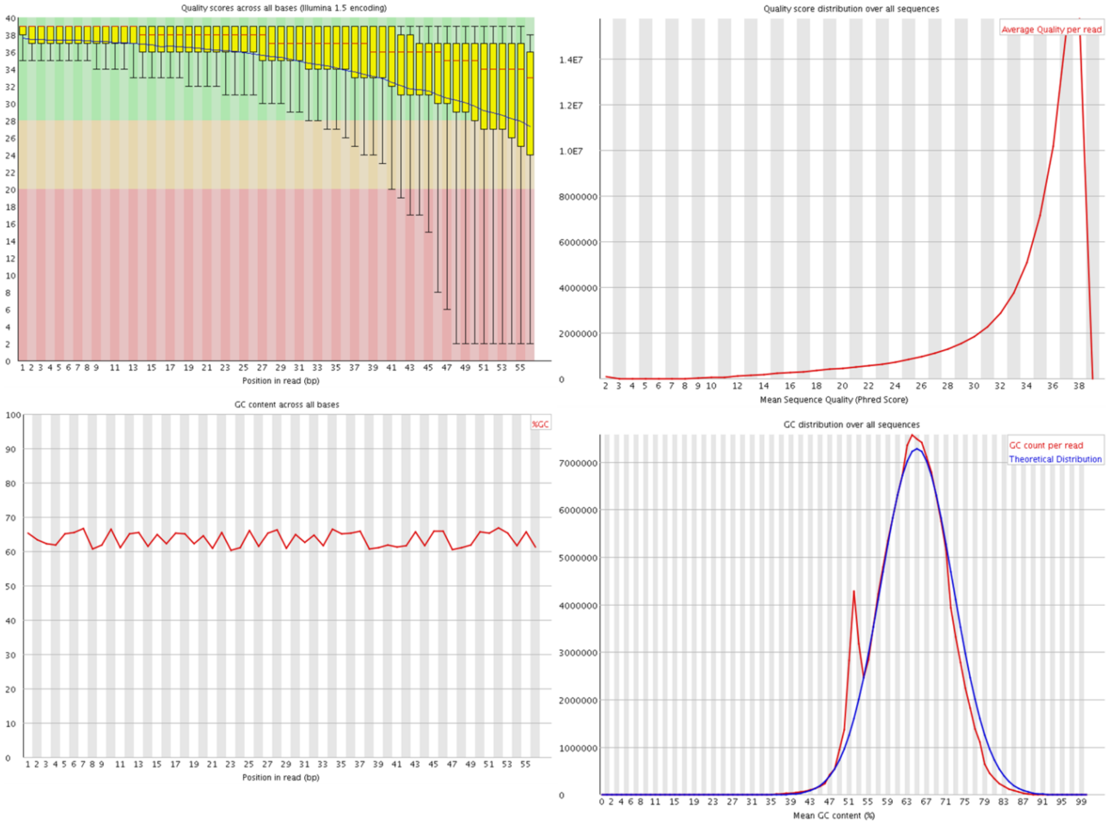


**Figure 3.2 (a-d). Quality check of raw reads for the sequences of *M. bovis* CorsentinoNE elk obtained from Illumina$^{R}$ sequencing. The FastQC package from the Galaxy interface was used for QC analyses (www.galaxy.msi.umn.edu)** Clockwise from top left: (a) Per base sequence quality, using Phred score, shows a high score across most sequences (> 20 = good score, higher the better) (b) Per sequence quality scores (high) (c) Per sequence GC content (~ 65% as expected for mycobacterial genome) (d) Per base GC content (good, negligible contaminating sequences)

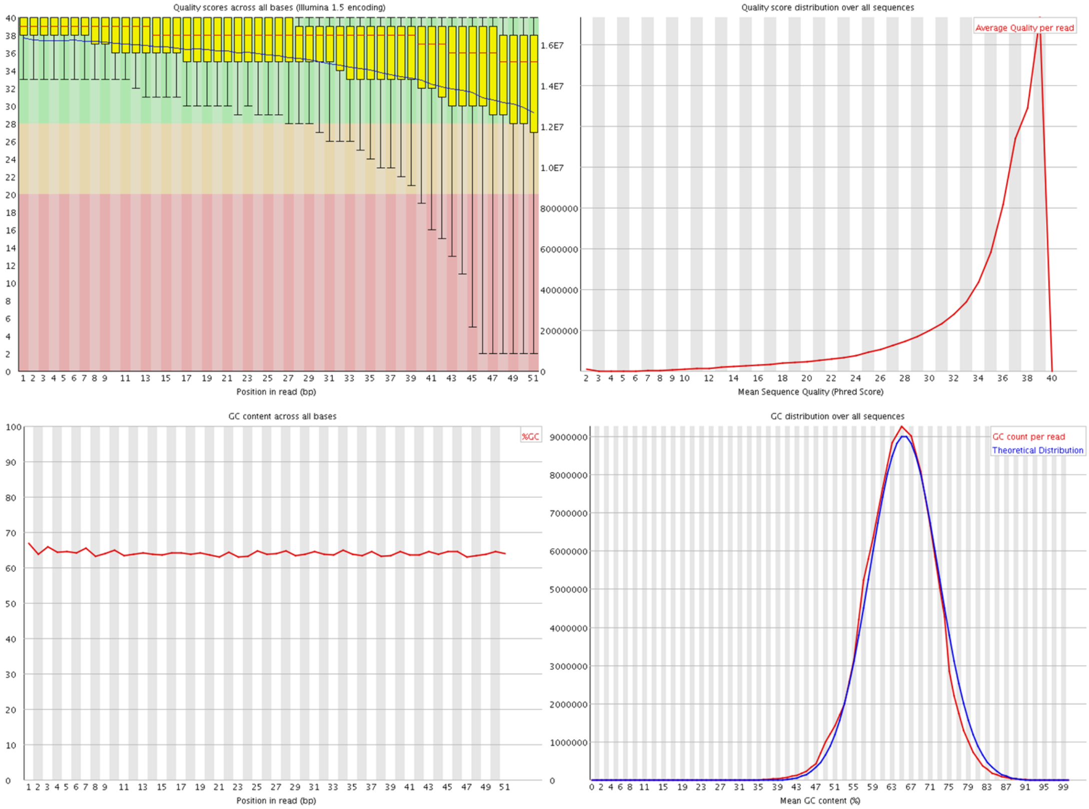
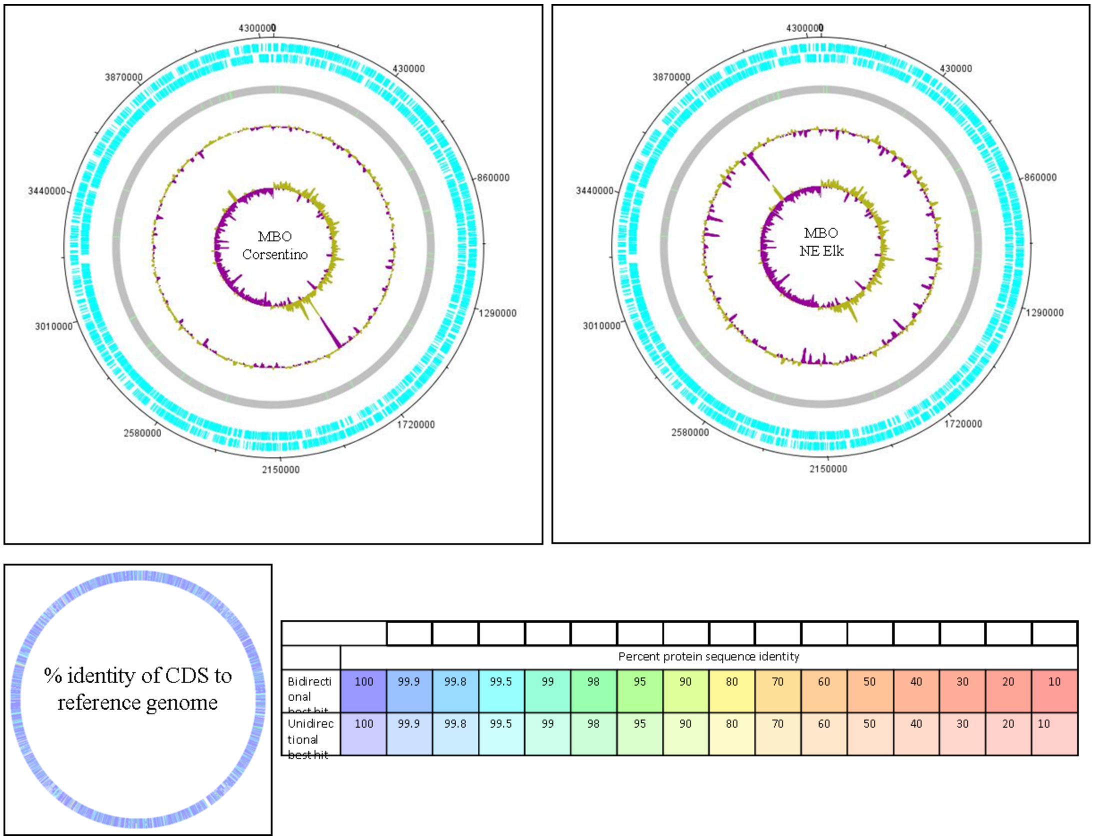


**Figure 3.3 (a-c). Graphical representation of the *M. bovis* Corsentino and NE elk genomes using the Artemis software displays the GC skew, GC plot and coding sequences of the double stranded circular chromosome.** Clockwise from top left (a) The Corsentino genome is 4307383 bp in size and (b) the NE elk genome is 4302584 bp in size. Both have a G+C content of 65.4%. (c) High percent identity (>99%) of the coding sequences (CDS) to the reference genome strain AF2122/97 from UK.



| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Percent protein sequence identity | | | | | | | | | | | | | | |
| Bidirectional best hit | 100 | 99.9 | 99.8 | 99.5 | 99 | 98 | 95 | 90 | 80 | 70 | 60 | 50 | 40 | 30 | 20 | 10 |
| Unidirectional best hit | 100 | 99.9 | 99.8 | 99.5 | 99 | 98 | 95 | 90 | 80 | 70 | 60 | 50 | 40 | 30 | 20 | 10 |

**Figure 3.4. Whole genome comparison of the *M. bovis* strain Corsentino from the USA (*de novo* assembled, using Velvet program) to the reference *M. bovis* strain AF2122/97 from the UK that identified 1139 SNPs between the two.** The gaps / sequence variations observed in the MBO Corsentino genome (represented by white gaps in the reference genome were confirmed as assembly / sequencing errors and not true deletions). The Mauve program was used for alignment and SNP calling.

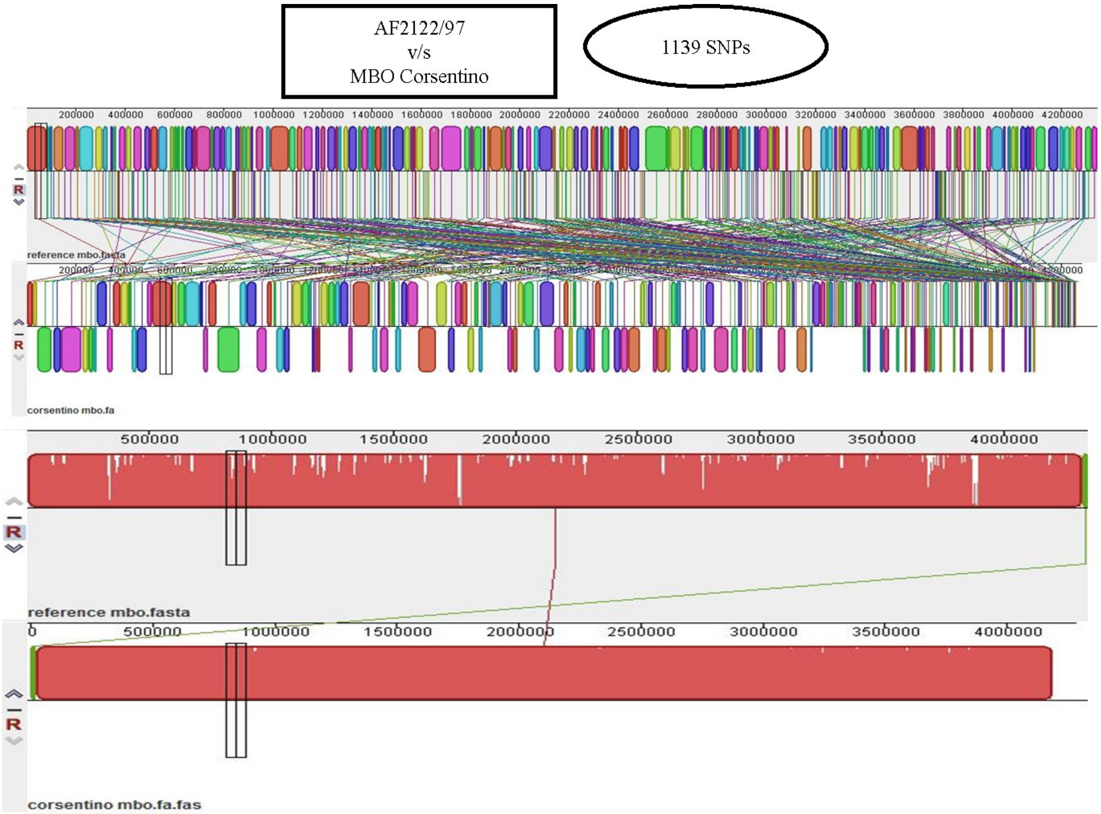


**Figure 3.5. Whole genome comparison of the *M. bovis* strain NE elk from the USA (*de novo* assembled, usingVelvet program) to the reference *M. bovis* strain AF2122/97 from the UK that identified 1184 SNPs between the two.** The gaps / sequence variations observed in the MBO NE elk genome (represented by white gaps in the reference genome were confirmed as assembly / sequencing errors and not true deletions). The Mauve program was used for alignment and SNP calling.

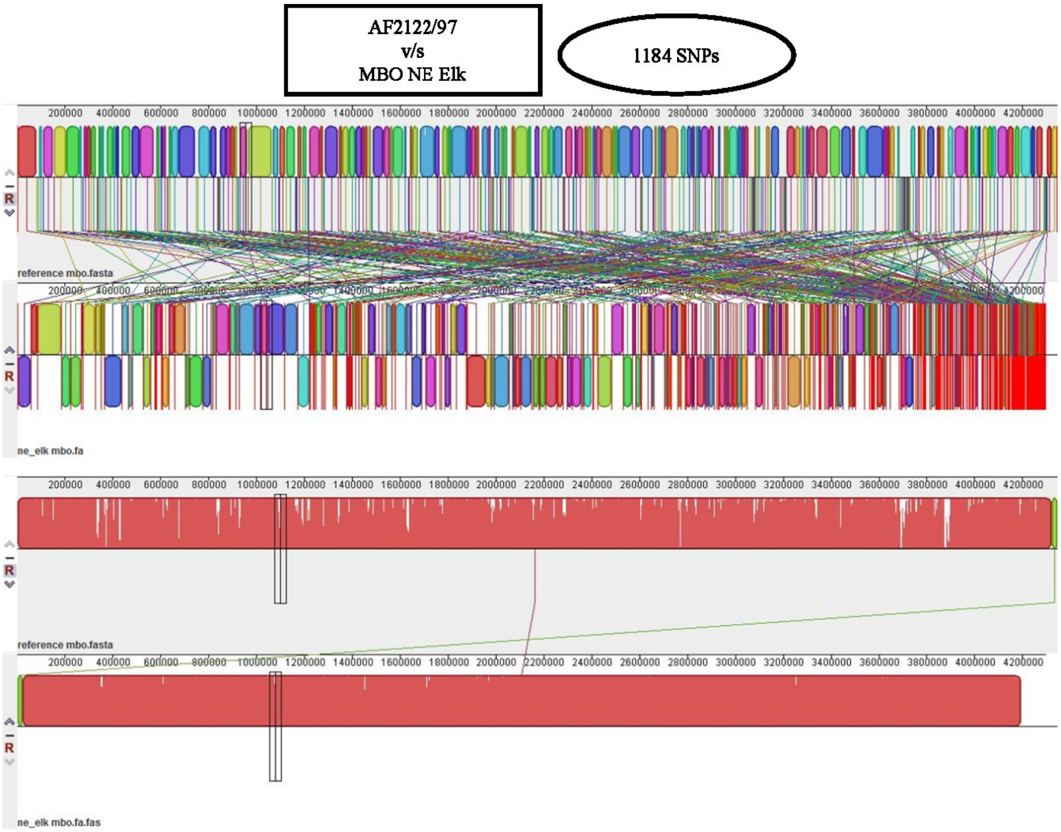


**Figure 3.6. Comparsion of the *M. bovis* Corsentino genome v/s AF2122/97 and NE elk v/s AF2122/97.** A dot plot generated by the RAST tool ( www.rast.nmpdr.org ) shows high degree of co-linearity of the genomes. The reference guided assembled sequences (Mosaik software) represent the same degree of similarity between the three *M. bovis* genomes.

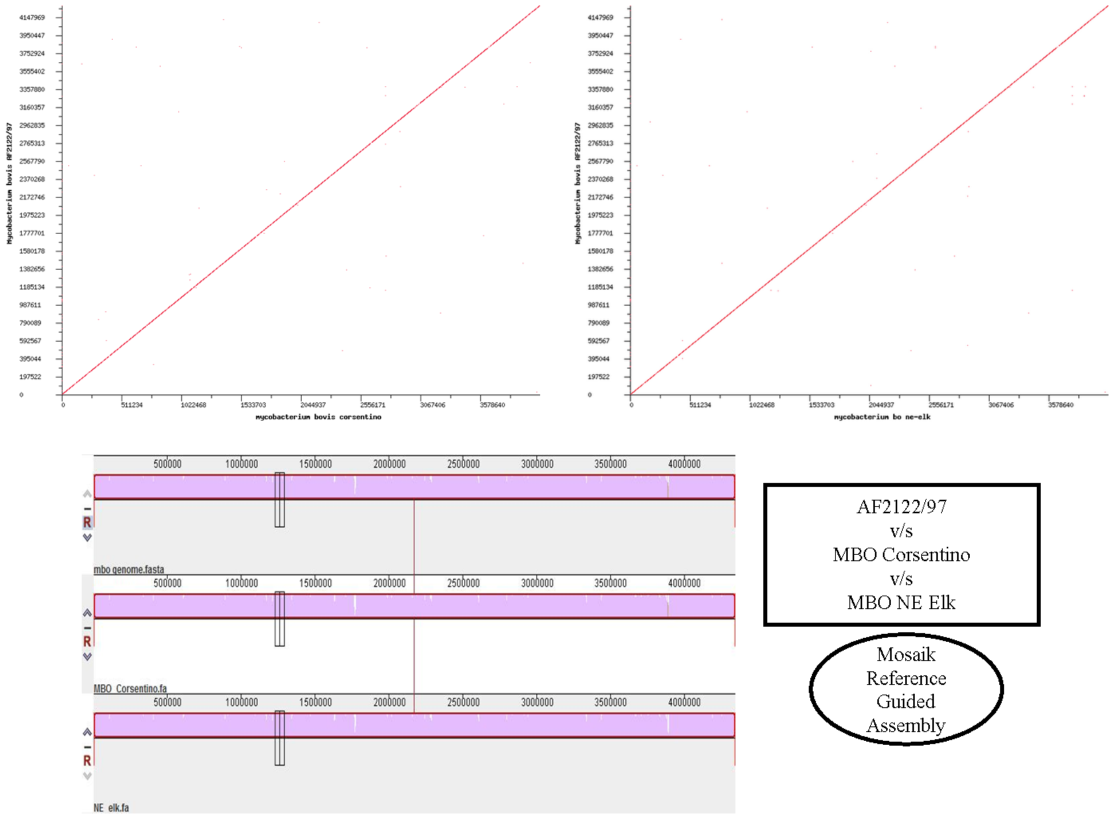
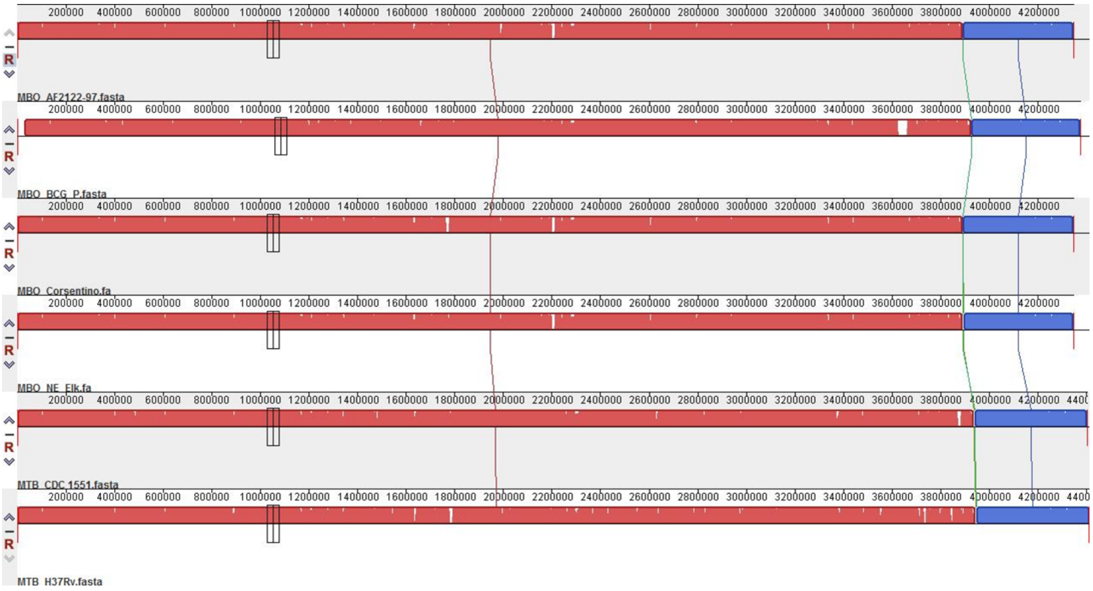


**Figure 3.7: Whole genome comparison of six strains from the *M. tuberculosis* complex group of organisms reveals high clonality and sequence identity between different ecotypes. The analysis was done using the Mauve software.** The genomes from top to bottom are – *M. bovis* strain AF2122/97 (reference strain), *M. bovis* BCG-Pasteur, *M. bovis* Corsentino, *M. bovis* NE elk, *M. tuberculosis* CDC 1551 and *M. tuberculosis H37Rv*. There is no evidence of genomic translocations, inversions or duplications.



# Chapter 4
# In vivo transcriptional profiling of a *Mycobacterium bovis* infection

Bovine tuberculosis caused by *Mycobacterium bovis* is a major and economically important disease of livestock with a zoonotic potential. The overall goal of this project was to explore the biology via transcriptome profiling of *M. bovis* during its infection cycle within the bovine host. This study aimed to decipher mechanisms of pathogenecity and to identify virulence markers of this damaging pathogen. The study also aimed to understand the molecular mechanisms governing the host response to *M. bovis* infection. Mediastinal lymph node tissues from two *M. bovis* infected cattle and two age matched control cattle were obtained, that displayed the characteristic granulomatous pathology of bovine tuberculosis.Total RNA was extracted and enriched for bacterial mRNA using commercially available kits. The enriched samples were submitted for next-gen sequencing employing the Illumina RNA-Seq Platform for transcriptomics profiling. The reads obtained from sequencing were assembled against the bacterial reference genome of *M. bovis* strain AF2122/97 and the bovine genome (*Bos taurus*) to build the gene expression profiles of the bacteria as well as the host. However the enrichment protocol used failed, leading to poor qualityof bacterial sequences and no significant gene expression profile could be obtained for the host sequences. Re-evaluation and standardization of RNA extraction techniques are sought for future studies.

## INTRODUCTION

Bovine tuberculosis is an established zoonotic disease which affects cattle worldwide with major economic losses. Many wildlife reservoirs of its causative agent, *Myocbacterium bovis,* have also been identified globally that continue to impact the disease surveillance and control programs implemented in many countries.

The disease etiology and host immune response of bovine tuberculosis is similar to that of human tuberculosis that is caused by *Mycobacterium tuberculosis* (252, 253).

*M. bovis* is transmitted primarily via aerosolised respiratory secretions that contain infectious bacilli, with the natural site of infection being the respiratory tract (136). Following an initial exposure, the pathogen is encountered by host alveolar macrophages, which serve as key effector cells in activating the innate and adaptive immune responses required to determine the outcome of infection (103). Infectious bacilli are phagocytosed by host macrophages upon exposure where they persist, resulting in lengthy subclinical phases of infection that can lead to immunopathology and disease dissemination. Macrophage recognition of mycobacteria occurs through the interaction of mycobacterial pathogen-associated molecular patterns (PAMPs) with host pathogen recognition receptors (PRRs), such as the Toll-like receptors (TLRs) that are expressed on the macrophage cell surface (91). PRR activation induces signaling pathways resulting in the production of endogenous NF-κB-inducible cytokines that promote an adaptive immune response characterized by the release of proinflammatory interferon-gamma (IFN-γ) from T cells and natural killer (NK) cells (44). In turn, IFN-γ induces microbicidal activity in infected macrophages and enhances the expression of the major histocompatibility complex (MHC) class I and II molecules necessary for the presentation of mycobacterial antigens on the macrophage surface to T cells (70). These molecular mechanisms culminate in the formation of granulomas-organized complexes of immune cells comprised of lymphocytes, non-infected macrophages and neutrophils that contain mycobacterial-infected macrophages and prevent the dissemination of bacilli to other organs and tissues although in most cases the pathogen is not eliminated by the host (91, 136). The persistence of mycobacteria within granulomas is the hallmark of tuberculosis infection. This latent infection can progress to active tuberculosis whenever the host immunity is compromised. Survival within the granuloma through the subversion of host immune response is achieved through a diverse set of molecular mechanisms.

With the recent availability of a complete *Bos taurus* genome sequence (63) and the *M. bovis* genome sequence (77) coupled with the continuing development of high-throughput genomic technologies, an analysis of the transcriptional changes induced during infection can be undertaken. Several studies have mainly focused on the host response primarily in the macrophage infection model (112, 136, 151). The overall goal of this project was to explore the biology of granuloma via transcriptome profiling of *M. bovis* during its infection cycle as it resides in the granuloma. This study was aimed at deciphering mechanisms of dormancy, pathogenecity and virulence markers of *M. bovis* and associated host response. Analysis of host and bacterial transcriptomes during infection can provide valuable insights into the molecular mechanisms that underlie the disease latency and are expected to augment current diagnostic tests, surveillance and control programs.

## MATERIALS AND METHODS

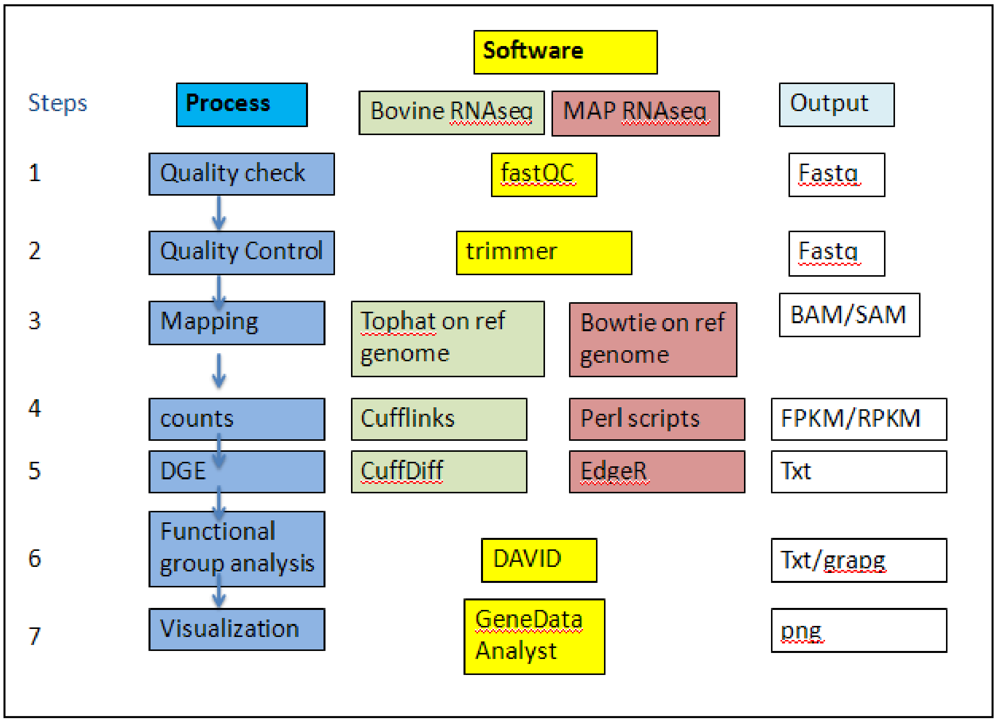
**Experimental design for animal infections.** Animal infections were carried out by the Mycobacteriology Research Division of the National Animal Disease Center, (NADC), USDA, Ames, IA, in October 2010. The USDA IACUC study approval number was #3930 and the IBC approval number was #0327. Male Holstein steers of ~1 year of age were obtained from a TB-free source (Van Voorst) and housed at the NADC animal housing. The animals infections were carried out under biosecurity level (BSL) 3 conditions. Bacterial strains used for infection included *M .bovis* Ravenel, a laboratory strain believed to have an attenuated profile in clinical infections and *M. bovis* 95-1315- a virulent field strain isolated in 1995 from cattle and deer and associated with outbreaks in Michigan. *M. bovis* Ravenel was prepared by Catherine Vilcheze at the Howard Hughes Medical Institute, Albert Einstein College of Medicine. *M. bovis* 95-1315 was prepared at NADC. The treatment group of animals (*n*=5) was challenged with $10^5$ cfu *M. bovis* strain 95-1315 through aerosol inoculation. A group of animals (*n*=5) as non-infected controls was also maintained. However, just prior to euthanasia (~ 2 weeks) the non-infected control group of animals received $10^8$ cfu *M. bovis* strain Ravenel by aerosol route. Approximately 3 months post challenge, the calves were euthanized and various tissues were collected at necropsy for further analyses. This lab received the samples of the mediastinal lymph nodes from these experimentally infected animals, saved in RNAlater$^{TM}$ , (Life Technologies, Carlsbad, CA) and shipped on dry ice. Two samples each (*n*=2) from the infected group and non-infected controls were processed in the lab for transcriptomic analyses using RNA-Seq.

**RNA extraction and enrichment.** Total RNA was extracted from the tissue samples (*n*=4) using the TRIzol method, (Gibco$^{TM}$, Life Technologies), following the manufacturer's protocol. Briefly, the tissue samples were homogenized and bead-beaten using zirconium beads (for effective dissociation of the mycobacterial cell wall) and mixed with the TRIzol reagent. Chloroform was added to the TRIzol for phase separation, during which RNA remains exclusively dissolved in the uppermost aqueous phase. Finally the RNA was precipitated using isopropyl alcohol. The quality and quantity of the total RNA was estimated using the Nanodrop$^{TM}$ 1000 Spectrophotometer,( ThermoScientifc, Asheville, NC ) and stored at -80$^{0}$C until further use. Since the total RNA extract would have over representation of host RNA, the following protocol was followed to enrich for bacterial RNA. Microb*Enrich*$^{TM}$ kit from Ambion, (Life Technologies, Grand Island, NY), was employed for enrichment of bacterial RNA from the mixture of mammalian and prokaryotic RNA. This was followed by application of Microb*Express*$^{TM}$ kit from Ambion, (Life Technologies), for further enrichment of bacterial mRNAs.

**Transcriptome Sequencing.** The samples were submitted for next-gen sequencing at the BioMedical Genomics Center (BMGC), Univeristy of Minnesota employing the Illumina$^{TM}$ Hi-Seq 2000 Platform. The samples were pooled in one lane of a 100bp PE run on the Hiseq2000. Insert length was 200 nt (library size 320 bp).

**Analytical approach.** *Bos taurus* and *M. bovis* reference genomes available through the NCBI database (ncbi.nlm.nih.gov/genome) were used for read mapping and annotation. TopHat v.2.0.4 and Cufflinks available through the Galaxy interface (galaxy.msi.umn.edu) were used for host Bovine RNAseq analysis including mapping, data normalization and gene differential analysis. This suite is a well known for RNAseq analysis. TopHat is a fast splice junction mapper for RNA-Seq reads to mammalian-sized genomes using ultra high-throughput short read aligner. Cufflinks assembles transcripts, estimates their abundances and tests for differential expression and regulation in RNA-

Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonius set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one, taking into account biases in library preparation protocols. Bowtie (ultra fast short read aligner) was used for *M. bovis* bacterial RNAseq mapping and edgeR bioconductor script was used for DGE inference. DAVID tools (Database for Annotation, Visualization and Integrated Discovery) version 6.7, a free resource available at http://david.abcc.ncifcrf.gov/, was used to provide the functional interpretation of gene lists derived from the analyses. The RNA-Seq analysis workflow is summarized in the flow chart below-



| | | Software | | |
|---|---|---|---|---|
| Steps | Process | Bovine RNAseq | MAP RNAseq | Output |
| 1 | Quality check | fastQC | | Fastq |
| 2 | Quality Control | trimmer | | Fastq |
| 3 | Mapping | Tophat on ref genome | Bowtie on ref genome | BAM/SAM |
| 4 | counts | Cufflinks | Perl scripts | FPKM/RPKM |
| 5 | DGE | CuffDiff | EdgeR | Txt |
| 6 | Functional group analysis | DAVID | | Txt/grapg |
| 7 | Visualization | GeneData Analyst | | png |

## RESULTS

**Sequencing read statistics. Table 4.1**: The lane summary is shown below.

| HiSeq2000 | Read+ | Total Read Number | Read Length | Total bases Sequenced | Reads Avg. Quality Score# | Total No. of Passed Reads | Total Bases in Passed reads | Average Quality Score in Passed Reads | PF Ratio * |
|---|---|---|---|---|---|---|---|---|---|
| L5 | R1 | 58612569 | 100 | 5861256900 | 32.6 | 55134903 | 5513490300 | 34.1 | 94.07% |
|  | R2 | 58612569 | 100 | 5861256900 | 32.7 | 55134903 | 5513490300 | 33.8 | 94.07% |

*PFRatio: pass-fail ratio; pass or fail is determined by CASAVA software based on reads quality scores, labeled as Y/N in the reads name.

+In Paired-End library each DNA fragment is sequenced from the left end and the right end thus generating two read files labeled with the suffix R1 and R2

#Phred nucleotide base quality score of 30 = 1 error per 1000 nucleotide bases. (20:1 per 100)

**Table 4.2:** The demultiplex summary is shown below-

| L3 Read | Oligo | Index # | SampleID, Group | Total Read Number | Pass Filter Read Number | PFRatio% |
|---|---|---|---|---|---|---|
| R1 | ATCACG | Index 1 | 1, infected | 13293675 | 12693923 | 95.49 |
|  | CGATGT | Index 2 | 2, non-infected | 14282371 | 13612162 | 95.31 |
|  | TTAGGC | Index 3 | 3, non-infected | 17280851 | 16308841 | 94.38 |
|  | TGACCA | Index 4 | 4, infected | 12052582 | 11396493 | 94.56 |
| R2 | ATCACG | Index 1 | 1, infected | 13293675 | 12693923 | 95.49 |
|  | CGATGT | Index 2 | 2, non-infected | 14282371 | 13612162 | 95.31 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | TTAGGC | Index 3 | 3, non-infected | 17280851 | 16308841 | 94.38 |
| | TGACCA | Index 4 | 4, infected | 12052582 | 11396493 | 94.56 |

**Data Quality.** The program FASTQC, from the Galaxy interface was used for data analysis. Strong data quality was observed. All Phred scores (average and individual) were > 30. Base contents bias was caused by Primer extension. Sample demonstration is depicted in Figure 4.1.

**RNA-Seq mapping to reference genomes**.

**Host Bovine RNAseq**: Program 1: TopHat - (used for sequence mapping) results are summarized in Table 3 below. Parameters included: Inner distance: 42, Mismatch number: 2, Max aligned pairs: 40. Program 2: Cufflinks - (assembly of aligned RNA-Seq reads into transcripts; abundance estimate; DGE test). Reference genome: *Bos taurus* (UCSC format) with use of reference transcriptome.

The results of mapping were-

- Over 80% of the paired end reads mapped to the Bovine reference genome (*B. taurus*)
- After Cufflinks assembly, 13,754 annotated transcripts were recovered (Supplemental Information Table-1).

**Table 4.3: Host (bovine) reads mapping summary**

| Samples | Total Reads | Reads Mapped |
|---|---|---|
| 1 | 21049567 | 0.829 |
| 2 | 22443192 | 0.824 |
| 3 | 25523645 | 0.782 |
| 4 | 17551300 | 0.770 |

**Pathogen *M. bovis* RNAseq**: Program**:** Bowtie - (sequence mapping), Reference genome: *M. bovis* AF2122_97. Parameters included: Mismatch number: 2, Max aligned pairs: 1.

Only about 10% paired reads were mapped on the *M. bovis* reference genome. The other 90% were still the bovine host sequences. This implies that the bacterial RNA enrichment protocol failed.

**Table 4.4: Pathogen (*M. bovis)* read mapping summary**

| Index | Oligo | mapped Reads | Reads Mapped % |
|---|---|---|---|
| 1 | ATCACG | 18 | ~0 |
| 2 | GGATGT | 6 | ~0 |
| 3 | TTAGGC | 16 | ~0 |
| 4 | TGACCA | 6 | ~0 |

**Differential gene expression.**

Bovine host RNA-Seq. Cell types were stratified as expected (group of infected samples animal ID#591 and ID#6096;  group of non-infected control animal ID#5439, ID#1149, see Figure 2).

Program used**:** Cufflinks (Cuffdiff).

The differential expression was determined by a q-value cut-off of 0.05, with parameters as follows-

- Q-value is the statistical p-value corrected for multiple testing.
- Cuffdiff uses the standard FDR (false discovery rate) correction to compute the q values.
- Minimum align reads: 10
- Two groups with 2 replictes in each group.

Two-group comparisons were carried out as infections versus non infection controls.

Table 1 (Supplemental Information) shows number of differentially expressed and regulated genes with q-value cut-off= 0.05. In summary, only 20 DEGs (Table 5) were identified as per the cut off value.

**Table 4.5: List of 20 host differentially expressed genes**

| # | Genbank Accession ID | Locus | Gene | q Value | Fold Change (log2) |
|---|---|---|---|---|---|
| 1 | NM_001098865 | chr13:74377094-74379388 | secretory leukocyte peptidase inhibitor | 0.002219 | 1.79769e+308 |
| 2 | NM_001083800 | chr17:74229420-74454505 | immunoglobulin lambda-like polypeptide 1 | 1.09E-05 | 1.64144 |
| 3 | NM_174745 | chr18:23262980-23291598 | matrix metallopeptidase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase) | 0.044278 | 2.67239 |
| 4 | NM_001034039 | chr19:37634830-37651585 | collagen, type I, alpha 1 | 2.09E-12 | 2.38904 |
| 5 | NM_001076831 | chr2:7740061-7779695 | collagen, type III, alpha 1 | 0 | 2.26512 |
| 6 | NM_001242573 | chr29:27770023-27773911 | mammary serum amyloid A3.2 | 0.045818 | 3.13317 |
| 7 | NM_001075942 | chr3:1711077-1722465 | cellular repressor of E1A-stimulated genes 1 | 0.000338 | 2.2064 |
| 8 | NM_001034435 | chr3:21489959-21502825 | cathepsin K | 0.037161 | 1.5204 |
| 9 | NM_001033615 | chr3:21518292-21541221 | cathepsin S | 0.000146 | 1.32481 |
| 10 | NM_174520 | chr4:12018222-12054892 | collagen, type I, alpha 2 | 6.98E-13 | 2.49162 |

| | | | | | |
|---|---|---|---|---|---|
| 11 | NM_001075375 | chr4:72320291-72392246 | sorting nexin 10 | 0.004765 | 2.29537 |
| 12 | NM_001109795 | chr5:108241814-108290194 | alpha-2-macroglobulin | 0.007723 | 1.0818 |
| 13 | NM_001078159 | chr5:48056540-48065376 | Lysozyme | 1.86E-12 | 2.90518 |
| 14 | NM_001113172 | chr6:94105448-94110095 | chemokine (C-X-C motif) ligand 9 | 0.040323 | 1.79435 |
| 15 | NM_001040469 | chr7:16318283-16354276 | complement component 3 | 6.08E-05 | 2.35968 |
| 16 | NM_174464 | chr7:62635246-62657995 | secreted protein, acidic, cysteine-rich (osteonectin) | 9.56E-05 | 1.65798 |
| 17 | NM_201527 | chr9:99818143-99828056 | superoxide dismutase 2, mitochondrial | 0.021158 | 2.06121 |
| 18 | NM_174035 | chrUn.004.237:19315-52246 | cytochrome b-245, beta polypeptide | 0.025325 | 1.17081 |
| 19 | NM_001163778 | chrUn.004.3:728356-797518 | fibronectin 1 | 3.63E-12 | 2.17457 |
| 20 | NM_001080219 | chrUn.004.323:87209-97318 | chitinase 3-like 1 (cartilage glycoprotein39) | 0.019307 | 4.33496 |

The functional annotation analysis / pathway analysis for the differentially expressed host genes (*n*=20) did not  reveal any significant / reportable results.

## DISCUSSION / FUTURE DIRECTIONS

Both the host and pathogen response during infection contribute to the balance that determines the outcome of  tuberculosis infection. Many previous studies have shown that *M. bovis* infection is associated with supression of host immune response genes (112, 136, 151) and upregulation of certain bacterial genes especially those  involved in the stress response and lipid metabolism (115). The level of intracellular expression of mycobacterial stress-response genes upon infection has been shown to reflect the extent of immune pressure exerted by the host immune response. The outcome following infection by pathogenic mycobacteria is determined by a complex and dynamic host-pathogen interaction in which the phenotype of the pathogen and the immune status of the host play a role (115).

The present study was undertaken with a primary goal to characterize the *M. bovis* transcriptome during the course of infection in its primary bovine host in an attempt to decipher molecular mechanisms employed by the pathogen in an established infection. This along with simultaneous descritption of  the host transcriptomic profile was expected to identify novel transcriptional markers of bovine tuberculosis. The availability of novel high-throughput DNA sequencing methods has provided the opportunity to study transcriptomes of several species (250).  To our knowledge, no one has used deep sequencing to study pathogen transcription in infected tissues. However failure of the bacterial RNA enrichment protocol during sample processing is the most likely reason behind the poor quality read data obtained for *M. bovis* by the next-gen sequencing method. The sample processing also seems to have had an adverse effect on the quality of the host RNA, as evidenced by the host differential gene expression (DGE) profile. No information was obtained on genes involved in disease pathogenesis that have been widely reported in the mycobacterial literature. The other probable reason responsible for loss of information might be attributable to the time between harvesting of the mediastinal lymph nodes post euthanasia and the time in the lab for RNA extraction, including storage and preservation techniques that might have been breached. It has been reported that the total RNA from host and bacterial origin contains only ~0.04% mycobacterial RNA (9), thus implying that the bacterial RNA requires many-fold enrichment from such mixed samples. *M. bovis,* being an intracellular infection, requires a high degree of enrichment of bacterial RNA. Alternately there were few bacteria in the granuloma leading to poor overall RNA yields that may have led to this failure. Also, high levels of cell death, necrosis and pyrotopsis within granulomas may have lead to pathogen and host RNA degradation.

It has also been recommended to seperate the host and pathogen cells by differential lysis procedures (140). The differential lysis approach involves the physical separation of bacteria from host cells before RNA preparation. To leave the bacteria intact while performing efficient lysis of the host cells or tissue, detergents and lysis conditions need to be selected individually for each pathogen (6). This may be the reason why so few intracellular pathogens were enriched by differential lysis. This additional step could be critical in work with highly unstable bacterial RNA and also with small amounts of material. Several other experimental approaches have also been proposed like the enrichment of bacterial RNA by cDNA-RNA subtractive hybridization (130), the DECAL method (1), and hybridization-based positive cDNA selection (selective capture of transcribed sequences, or SCOTS) (Graham, 1999).

The efficacy of any transcriptome profiling technique critically depends upon the availability of RNA samples that precisely reflect the real ratios of individual bacterial mRNAs in the infected host tissues. This is very challenging, given the paucity of bacterial mRNA compared to the amounts of mammalian RNA in the samples. Prokaryotic mRNAs also present challenges owing to their short half lives, limited message polyadenylation, and a scarcity of starting material, particularly in terms of lower abundance of class messages (Graham, 1999). These issues can impede the identification of relevant host interaction-mediated gene expression by direct examination of bacterial mRNA.  However if this is truly a biological issue due to very little bacterial activity inside a granuloma it is suggested to increase the starting material (pooling of samples), selectively the center of granulomas. In conclusion, it is suggested to use bacterial mRNA enrichment methods like SCOTS prior to the sequencing along with the revision and standardization of processing techniques and protocols to enable repetition of this study in the future.

# LIST OF FIGURES AND FIGURE LEGENDS

**Figure 4.1. QC analysis of reads – a sample demonstration.**

A sample representation of the QC analysis of the RNA-Seq reads using Illumina HiSeq platform. The FastQC package from the Galaxy interface was used for QC analyses (www.galaxy.msi.umn.edu). (a) Per base sequence quality, using Phred score, shows a high score across most sequences (> 20 = good score, higher the better) (b) Per base sequence content – base content biases are caused by primer extension. Based on QC results, we trimmed off the first 9 bases and the last 21 of each read before further analysis.

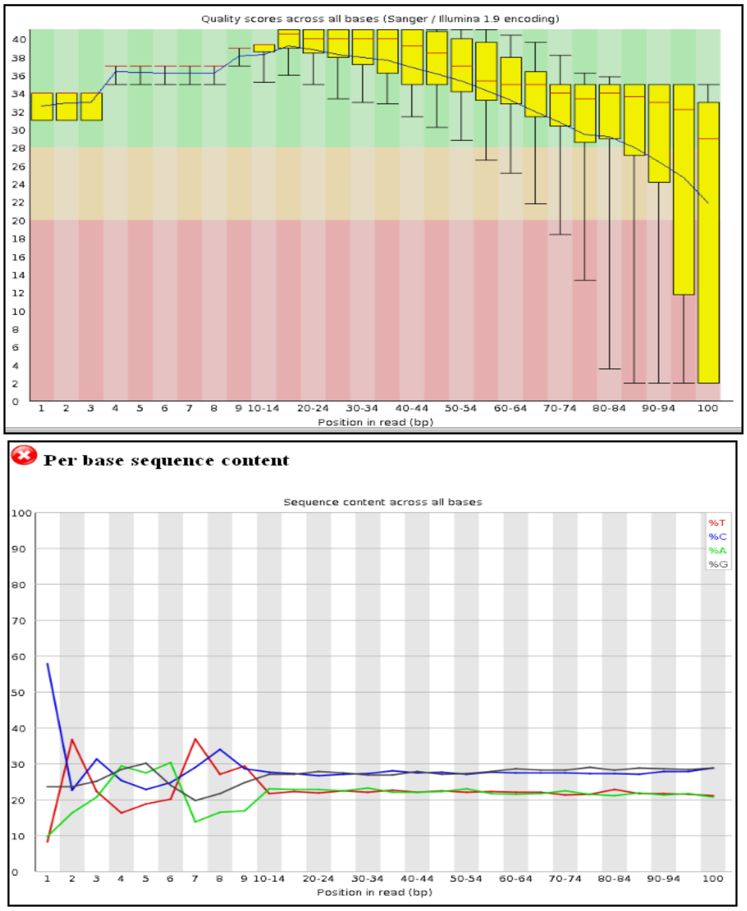
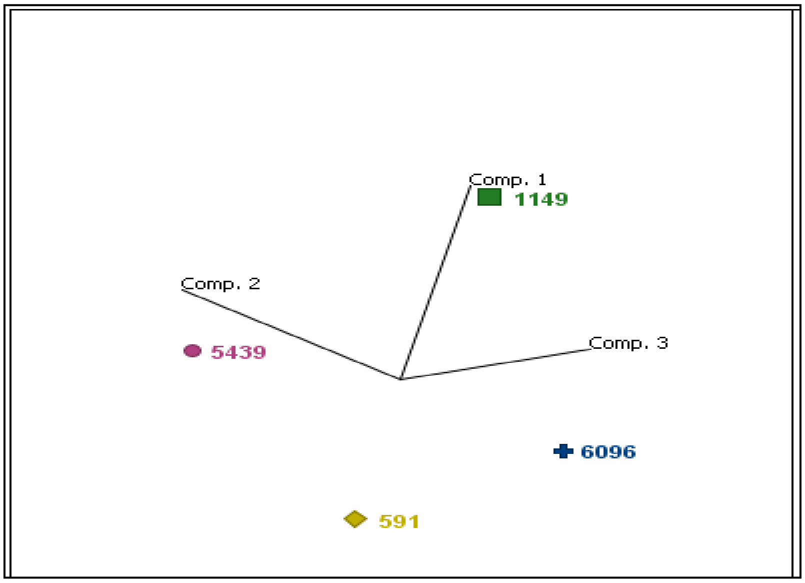


**Figure 4.2 The figure represents that the cell types were stratified as expected (group of infected samples #591, #6096 versus group of non-infection controls #5439, #1149)**



# CONCLUDING REMARKS

The dissertation presented here provides a study on bovine tuberculosis, with specific focus on the molecular sub-typing and phylogenetic analysis for identification of genetic variants. The practical implication of such information is fundamental to the understanding of genoptype-disease phenotype associations of the pathogen. In addition the studies highlight the importance of whole genome plus transcriptome analysis of *M. bovis* strains to identify the most informative panel of markers of genetic variability. As the cost of next-generation sequencing continues to decrease and sequencing platforms become integrated into public health practice, combined microbial genomic and epidemiologic approaches described in the research manuscripts can become an important and tractable first step toward a systems approach to tuberculosis control.

In the first study, a nationwide collection of *Mycobacterium bovis* strains using 206 single nucleotide polymorphism (SNP) markers were analyzed. Phylogenetic analysis identified five SNP cluster groups (SCGs) of which three were unique to *M. bovis* isolates, one SCG clustered to all *M. tuberculosis* isolates and the fifth SCG included all the *M. bovis*- BCG isolates. Data from this phylogenetic analysis provides evidence that SNPs can be used to a derive temporo-spatial population structure for *M. bovis* isolates. The data also suggests host susceptibility is driven by specific genotypes as was evidenced by clonality among the elk isolates. Additionally, isolates representing unique genetic signatures (SNP profiles) were further assessed for intracellular competence in an *in vitro* macrophage survival model and by differential gene expression profiling of six virulence-associated genes. With this study, an *in vitro* screening approach for studying trait-allele associations has been established. Overall, the conclusion is that *M. bovis* isolates from diverse geographic and host origins represent an array of genetic profiles that could potentially relate to their phenotypic properties. Despite several limitations these studies were unique and provided evidence that strain variation among *M. bovis* is real and warrants further investigation. Studies such as this can provide a better understanding of the underlying principles that determine the strain characteristics and relative disease patterns and pathogenesis of bovine tuberculosis.

Next, the whole genome sequencing of two *M. bovis* strains from the US was undertaken with a goal to characterize local strains and perform comparative genomic analysis to discover unique genetic variants and single nucleotide polymorphisms markers. To date there was only one complete genome sequence of a UK strain of *M. bovis* (strain AF2122/97). Based on our SNP phylogeny we decided to improve resolution in the genomes of the US strains. Two strains were sequenced to explore the genetic diversity of this organism and help epidemiologic studies associated with the ongoing micro-epidemic in this country. Draft genomes were assembled of a cattle strain (MBO Corsentino) and elk strain (MBO NE elk) of *M. bovis*, from the US. Comparative genomic analysis revealed high sequence similarity between these two genomes and the UK strain AF2122/97, as well as with other genomes of *Mycobacterium Tuberuclosis Complex*. We hypothesized that host adaptation was driven by species-specific genetic signatures in *M. bovis* genomes but none were identified. With the availability of these genomic sequences, transcriptomic profiling may prove that variation in gene expression patterns likely drive phenotypic variation and host adaptation of *M. bovis*. Further, the data identified >1000 SNPs that distinguish these strains from each other, which can further enhance the identification of SNP markers that enable molecular epidemiologic investigations as described in our study 1. In conclusion, the first reference genome of *M. bovis* isolates of US origin is provided here.

Finally, attempt was made to describe the bacterial as well as host transcriptomics profiles during infection with *M. bovis* in its natural host. The overall goal was to explore the biology of *M. bovis* strains to derive information related to their gene and protein expression profiles in an *in vivo* model, to provide better understanding of genotypes that are associated with enhanced survival attributes and help decipher mechanisms of pathogenicity of these strains. However, failed enrichment protocols (during RNA extraction and sample processing) did not provide with high-resolution information to infer any significant biological pathways that may be operational inside a granuloma.

Future studies should address these challenges concerning sample preparation, apply better bacterial mRNA enrichment methods such as differential centrifugation or application of genome directed primers or selective capture of transcribed sequences prior to sequencing.

The work presented in this dissertation adds a small amount to the understanding of bovine tuberculosis functional genomics. These studies on the genetic and genomic variations are likely to be useful in the improvement of detection, tracking, and control of bovine tuberculosis in the United States.

# BIBLIOGRAPHY

1. **Alland, D., I. Kramnik, T. R. Weisbrod, L. Otsubo, R. Cerny, L. P. Miller, W. R. Jacobs, Jr., and B. R. Bloom.** 1998. Identification of differentially expressed mRNA in prokaryotic organisms by customized amplification libraries (DECAL): the effect of isoniazid on gene expression in Mycobacterium tuberculosis. Proc Natl Acad Sci U S A **95:**13227-32.
2. **Alland, D., D. W. Lacher, M. H. Hazbon, A. S. Motiwala, W. Qi, R. D. Fleischmann, and T. S. Whittam.** 2007. Role of large sequence polymorphisms (LSPs) in generating genomic diversity among clinical isolates of Mycobacterium tuberculosis and the utility of LSPs in phylogenetic analysis. J Clin Microbiol **45:**39-46.
3. **Alonso-Rodriguez, N., M. Martinez-Lirola, M. L. Sanchez, M. Herranz, T. Penafiel, C. Bonillo Mdel, M. Gonzalez-Rivera, J. Martinez, T. Cabezas, L. F. Diez-Garcia, E. Bouza, and D. Garcia de Viedma.** 2009. Prospective universal application of mycobacterial interspersed repetitive-unit-variable-number tandem-repeat genotyping to characterize Mycobacterium tuberculosis isolates for fast identification of clustered and orphan cases. J Clin Microbiol **47:**2026-32.
4. **Amaro, A., E. Duarte, A. Amado, H. Ferronha, and A. Botelho.** 2008. Comparison of three DNA extraction methods for Mycobacterium bovis, Mycobacterium tuberculosis and Mycobacterium avium subsp. avium. Lett Appl Microbiol **47:**8-11.
5. **Armstrong, J. A., and P. D. Hart.** 1975. Phagosome-lysosome interactions in cultured macrophages infected with virulent tubercle bacilli. Reversal of the usual nonfusion pattern and observations on bacterial survival. J Exp Med **142:**1-16.
6. **Azhikina, T., T. Skvortsov, T. Radaeva, A. Mardanov, N. Ravin, A. Apt, and E. Sverdlov.** A new technique for obtaining whole pathogen transcriptomes from infected host tissues. Biotechniques **48:**139-44.
7. **Aziz, R. K., D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. A. Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke, and O. Zagnitko.** 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics **9:**75.
8. **Baess, I.** 1979. Deoxyribonucleic acid relatedness among species of slowly-growing mycobacteria. Acta Pathol Microbiol Scand B **87:**221-6.
9. **Banaiee, N., W. R. Jacobs, Jr., and J. D. Ernst.** 2006. Regulation of Mycobacterium tuberculosis whiB3 in the mouse lung and macrophages. Infect Immun **74:**6449-57.
10. **Baumler, D. J., L. M. Banta, K. F. Hung, J. A. Schwarz, E. L. Cabot, J. D. Glasner, and N. T. Perna.** Using comparative genomics for inquiry-based learning to dissect virulence of Escherichia coli O157:H7 and Yersinia pestis. CBE Life Sci Educ **11:**81-93.

11. **Bayjanov, J. R., D. Molenaar, V. Tzeneva, R. J. Siezen, and S. A. van Hijum.** PhenoLink - a web-tool for linking phenotype to ~omics data for bacteria: application to gene-trait matching for Lactobacillus plantarum strains. BMC Genomics **13:**170.

12. **Beckloff, N., S. Starkenburg, T. Freitas, and P. Chain.** Bacterial genome annotation. Methods Mol Biol **881:**471-503.

13. **Behr, M. A.** 2002. BCG--different strains, different vaccines? Lancet Infect Dis **2:**86-92.

14. **Behr, M. A.** 2001. Comparative genomics of BCG vaccines. Tuberculosis (Edinb) **81:**165-8.

15. **Behr, M. A., M. A. Wilson, W. P. Gill, H. Salamon, G. K. Schoolnik, S. Rane, and P. M. Small.** 1999. Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science **284:**1520-3.

16. **Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, E. C. M. Chiara, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, et al.** 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature **456:**53-9.

17. **Berg, S., M. C. Garcia-Pelayo, B. Muller, E. Hailu, B. Asiimwe, K. Kremer, J. Dale, M. B. Boniotti, S. Rodriguez, M. Hilty, L. Rigouts, R. Firdessa, A. Machado, C. Mucavele, B. N. Ngandolo, J. Bruchfeld, L. Boschiroli, A. Muller, N. Sahraoui, M. Pacciarini, S. Cadmus, M. Joloba, D. van Soolingen, A. L. Michel, B. Djonne, A. Aranaz, J. Zinsstag, P. van Helden, F. Portaels, R. Kazwala, G. Kallenius, R. G. Hewinson, A. Aseffa, S. V. Gordon, and N. H. Smith.** African 2, a clonal complex of Mycobacterium bovis epidemiologically important in East Africa. J Bacteriol **193:**670-8.

18. **Berget, I., E. Heir, J. Petcovic, and K. Rudi.** 2007. Discriminatory power, typability, and accuracy of single nucleotide extension microarrays. J AOAC Int **90:**802-9.
19. **Bertone, P., V. Stolc, T. E. Royce, J. S. Rozowsky, A. E. Urban, X. Zhu, J. L. Rinn, W. Tongprasit, M. Samanta, S. Weissman, M. Gerstein, and M. Snyder.** 2004. Global identification of human transcribed sequences with genome tiling arrays. Science **306:**2242-6.
20. **Betts, J. C.** 2002. Transcriptomics and proteomics: tools for the identification of novel drug targets and vaccine candidates for tuberculosis. IUBMB Life **53:**239-42.
21. **Bibb, L. A., and G. F. Hatfull.** 2002. Integration and excision of the Mycobacterium tuberculosis prophage-like element, phiRv1. Mol Microbiol **45:**1515-26.
22. **Blanco, F. C., J. Nunez-Garcia, C. Garcia-Pelayo, M. Soria, M. V. Bianco, M. Zumarraga, P. Golby, A. A. Cataldi, S. V. Gordon, and F. Bigi.** 2009. Differential transcriptome profiles of attenuated and hypervirulent strains of Mycobacterium bovis. Microbes Infect **11:**956-63.
23. **Bouakaze, C., C. Keyser, A. Gonzalez, W. Sougakoff, N. Veziris, H. Dabernat, B. Jaulhac, and B. Ludes.** Matrix-assisted laser desorption ionization-time of flight mass spectrometry-based single nucleotide polymorphism genotyping assay using iPLEX gold technology for identification of Mycobacterium tuberculosis complex species and lineages. J Clin Microbiol **49:**3292-9.
24. **Bradic, M., J. Costa, and I. M. Chelo.** Genotyping with Sequenom. Methods Mol Biol **772:**193-210.
25. **Braslavsky, I., B. Hebert, E. Kartalov, and S. R. Quake.** 2003. Sequence information can be obtained from single DNA molecules. Proc Natl Acad Sci U S A **100:**3960-4.
26. **Brookes, A. J.** 1999. The essence of SNPs. Gene **234:**177-86.
27. **Brosch, R., S. V. Gordon, T. Garnier, K. Eiglmeier, W. Frigui, P. Valenti, S. Dos Santos, S. Duthoy, C. Lacroix, C. Garcia-Pelayo, J. K. Inwald, P. Golby, J. N. Garcia, R. G. Hewinson, M. A. Behr, M. A. Quail, C. Churcher, B. G. Barrell, J. Parkhill, and S. T. Cole.** 2007. Genome plasticity of BCG and impact on vaccine efficacy. Proc Natl Acad Sci U S A **104:**5596-601.
28. **Brosch, R., S. V. Gordon, M. Marmiesse, P. Brodin, C. Buchrieser, K. Eiglmeier, T. Garnier, C. Gutierrez, G. Hewinson, K. Kremer, L. M. Parsons, A. S. Pym, S. Samper, D. van Soolingen, and S. T. Cole.** 2002. A new evolutionary scenario for the Mycobacterium tuberculosis complex. Proc Natl Acad Sci U S A **99:**3684-9.
29. **Brosch, R., S. V. Gordon, A. Pym, K. Eiglmeier, T. Garnier, and S. T. Cole.** 2000. Comparative genomics of the mycobacteria. Int J Med Microbiol **290:**143-52.

30. **Brosch, R., A. S. Pym, S. V. Gordon, and S. T. Cole.** 2001. The evolution of mycobacterial pathogenicity: clues from comparative genomics. Trends Microbiol **9:**452-8.

31. **Calmette, A., and C. Guérin.** 1909. Sur quelques propriétés du bacille tuberculeux d'origine, cultivé sur la bile de boeuf glycérinée. C. R. Acad. Sci. Paris**:**149716-718.

32. **Carver, T., M. Berriman, A. Tivey, C. Patel, U. Bohme, B. G. Barrell, J. Parkhill, and M. A. Rajandream.** 2008. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. Bioinformatics **24:**2672-6.

33. **Castellanos, E., A. Aranaz, L. de Juan, J. Alvarez, S. Rodriguez, B. Romero, J. Bezos, K. Stevenson, A. Mateos, and L. Dominguez.** 2009. Single nucleotide polymorphisms in the IS900 sequence of Mycobacterium avium subsp. paratuberculosis are strain type specific. J Clin Microbiol **47:**2260-4.

34. **Caws, M., G. Thwaites, S. Dunstan, T. R. Hawn, N. T. Lan, N. T. Thuong, K. Stepniewska, M. N. Huyen, N. D. Bang, T. H. Loc, S. Gagneux, D. van Soolingen, K. Kremer, M. van der Sande, P. Small, P. T. Anh, N. T. Chinh, H. T. Quy, N. T. Duyen, D. Q. Tho, N. T. Hieu, E. Torok, T. T. Hien, N. H. Dung, N. T. Nhu, P. M. Duy, N. van Vinh Chau, and J. Farrar.** 2008. The influence of host and bacterial genotype on the development of disseminated disease with Mycobacterium tuberculosis. PLoS Pathog **4:**e1000034.

35. **Chagne, D., R. N. Crowhurst, M. Troggio, M. W. Davey, B. Gilmore, C. Lawley, S. Vanderzande, R. P. Hellens, S. Kumar, A. Cestaro, R. Velasco, D. Main, J. D. Rees, A. Iezzoni, T. Mockler, L. Wilhelm, E. Van de Weg, S. E. Gardiner, N. Bassil, and C. Peace.** Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. PLoS One **7:**e31745.

36. **Chaves, D., A. Sandoval, L. Rodriguez, J. C. Garcia, S. Restrepo, and M. M. Zambrano.** [Comparative analysis of six Mycobacterium tuberculosis complex genomes]. Biomedica **30:**23-31.

37. **Chen, G., C. Wang, and T. Shi.** Overview of available methods for diverse RNA-Seq data analyses. Sci China Life Sci **54:**1121-8.

38. **Chepelev, I., G. Wei, Q. Tang, and K. Zhao.** 2009. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. Nucleic Acids Res **37:**e106.

39. **Cock, P. J., C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice.** The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res **38:**1767-71.

40. **Cole, S. T.** 1998. Comparative Mycobacterial Genomics. Current Opinion in Microbiology **1:**567-571.

41. **Cole, S. T., and B. G. Barrell.** 1998. Analysis of the genome of Mycobacterium tuberculosis H37Rv. Novartis Found Symp **217:**160-72; discussion 172-7.

42. **Cole, S. T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry, 3rd, F. Tekaia, K. Badcock, D.**

**Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. A. Quail, M. A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J. E. Sulston, K. Taylor, S. Whitehead, and B. G. Barrell.** 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature **393:**537-44.

43. **Comas, I., J. Chakravartti, P. M. Small, J. Galagan, S. Niemann, K. Kremer, J. D. Ernst, and S. Gagneux.** Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. Nat Genet **42:**498-503.
44. **Cooper, A. M.** 2009. Cell-mediated immune responses in tuberculosis. Annu Rev Immunol **27:**393-422.
45. **Cordero, F., M. Beccuti, S. Donatelli, and R. A. Calogero.** Large Disclosing the Nature of Computational Tools for the Analysis of Next Generation Sequencing Data. Curr Top Med Chem.
46. **Cordes, D. O., J. A. Bullians, D. E. Lake, and M. E. Carter.** 1981. Observations on tuberculosis caused by Mycobacterium bovis in sheep. N Z Vet J **29:**60-2.
47. **Corona, G., and G. Toffoli.** 2004. High throughput screening of genetic polymorphisms by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Comb Chem High Throughput Screen **7:**707-25.
48. **Coros, A., E. DeConno, and K. M. Derbyshire.** 2008. IS6110, a Mycobacterium tuberculosis complex-specific insertion sequence, is also present in the genome of Mycobacterium smegmatis, suggestive of lateral gene transfer among mycobacterial species. J Bacteriol **190:**3408-10.
49. **Cousins, D., S. Williams, E. Liebana, A. Aranaz, A. Bunschoten, J. Van Embden, and T. Ellis.** 1998. Evaluation of four DNA typing techniques in epidemiological investigations of bovine tuberculosis. J Clin Microbiol **36:**168-78.
50. **Cousins, D. V.** 2001. Mycobacterium bovis infection and control in domestic livestock. Rev Sci Tech **20:**71-85.
51. **Coussens, P. M., A. Jeffers, and C. Colvin.** 2004. Rapid and transient activation of gene expression in peripheral blood mononuclear cells from Johne's disease positive cows exposed to Mycobacterium paratuberculosis in vitro. Microb Pathog **36:**93-108.
52. **Darling, A. C., B. Mau, F. R. Blattner, and N. T. Perna.** 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res **14:**1394-403.
53. **David, L., W. Huber, M. Granovskaia, J. Toedling, C. J. Palm, L. Bofkin, T. Jones, R. W. Davis, and L. M. Steinmetz.** 2006. A high-resolution map of transcription in the yeast genome. Proc Natl Acad Sci U S A **103:**5320-5.

54. **de la Rua-Domenech, R.** 2006. Human Mycobacterium bovis infection in the United Kingdom: Incidence, risks, control measures and review of the zoonotic aspects of bovine tuberculosis. Tuberculosis (Edinb) **86:**77-109.

55. **Diamond, J.** 2002. Evolution, consequences and future of plant and animal domestication. Nature **418:**700-7.

56. **Dormans, J., M. Burger, D. Aguilar, R. Hernandez-Pando, K. Kremer, P. Roholl, S. M. Arend, and D. van Soolingen.** 2004. Correlation of virulence, lung pathology, bacterial load and delayed type hypersensitivity responses after infection with different Mycobacterium tuberculosis genotypes in a BALB/c mouse model. Clin Exp Immunol **137:**460-8.

57. **Dos Vultos, T., O. Mestre, J. Rauzier, M. Golec, N. Rastogi, V. Rasolofo, T. Tonjum, C. Sola, I. Matic, and B. Gicquel.** 2008. Evolution and diversity of clonal bacteria: the paradigm of Mycobacterium tuberculosis. PLoS One **3:**e1538.

58. **Duffield, B. J., and D. A. Young.** 1985. Survival of Mycobacterium bovis in defined environmental conditions. Vet Microbiol **10:**193-7.

59. **Dunne, W. M., Jr., L. F. Westblade, and B. Ford.** Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory. Eur J Clin Microbiol Infect Dis.

60. **Eisen, J. A.** 2000. Assessing evolutionary relationships among microbes from whole-genome analysis. Curr Opin Microbiol **3:**475-80.

61. **Eisen, J. A.** 1995. The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. J Mol Evol **41:**1105-23.

62. **Ekdahl, M. O., B. L. Smith, and D. F. Money.** 1970. Tuberculosis in some wild and feral animals in New Zealand. N Z Vet J **18:**44-5.

63. **Elsik, C. G., R. L. Tellam, K. C. Worley, R. A. Gibbs, D. M. Muzny, G. M. Weinstock, D. L. Adelson, E. E. Eichler, L. Elnitski, R. Guigo, D. L. Hamernik, S. M. Kappes, H. A. Lewin, D. J. Lynn, F. W. Nicholas, A. Reymond, M. Rijnkels, L. C. Skow, E. M. Zdobnov, L. Schook, J. Womack, T. Alioto, S. E. Antonarakis, A. Astashyn, C. E. Chapple, H. C. Chen, J. Chrast, F. Camara, O. Ermolaeva, C. N. Henrichsen, W. Hlavina, Y. Kapustin, B. Kiryutin, P. Kitts, F. Kokocinski, M. Landrum, D. Maglott, K. Pruitt, V. Sapojnikov, S. M. Searle, V. Solovyev, A. Souvorov, C. Ucla, C. Wyss, J. M. Anzola, D. Gerlach, E. Elhaik, D. Graur, J. T. Reese, R. C. Edgar, J. C. McEwan, G. M. Payne, J. M. Raison, T. Junier, E. V. Kriventseva, E. Eyras, M. Plass, R. Donthu, D. M. Larkin, J. Reecy, M. Q. Yang, L. Chen, Z. Cheng, C. G. Chitko-McKown, G. E. Liu, L. K. Matukumalli, J. Song, B. Zhu, D. G. Bradley, F. S. Brinkman, L. P. Lau, M. D. Whiteside, A. Walker, T. T. Wheeler, T. Casey, J. B. German, D. G. Lemay, N. J. Maqbool, A. J. Molenaar, S. Seo, P. Stothard, C. L. Baldwin, R. Baxter, C. L. Brinkmeyer-Langford, W. C. Brown, C. P. Childers, T. Connelley, S. A. Ellis, K. Fritz, E. J. Glass, C. T. Herzig, A. Iivanainen, K. K. Lahmers, A. K. Bennett, C. M. Dickens, J. G. Gilbert, D. E. Hagen, H. Salih,**

**J. Aerts, A. R. Caetano, et al.** 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science **324:**522-8.

64. **Evans, J. T., E. G. Smith, A. Banerjee, R. M. Smith, J. Dale, J. A. Innes, D. Hunt, A. Tweddell, A. Wood, C. Anderson, R. G. Hewinson, N. H. Smith, P. M. Hawkey, and P. Sonnenberg.** 2007. Cluster of human tuberculosis caused by Mycobacterium bovis: evidence for person-to-person transmission in the UK. Lancet **369:**1270-6.

65. **Ewing, B., and P. Green.** 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res **8:**186-94.

66. **Ewing, B., L. Hillier, M. C. Wendl, and P. Green.** 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res **8:**175-85.

67. **Feizabadi, M. M., I. D. Robertson, D. V. Cousins, and D. J. Hampson.** 1996. Genomic analysis of Mycobacterium bovis and other members of the Mycobacterium tuberculosis complex by isoenzyme analysis and pulsed-field gel electrophoresis. J Clin Microbiol **34:**1136-42.

68. **Filliol, I., A. S. Motiwala, M. Cavatore, W. Qi, M. H. Hazbon, M. Bobadilla del Valle, J. Fyfe, L. Garcia-Garcia, N. Rastogi, C. Sola, T. Zozio, M. I. Guerrero, C. I. Leon, J. Crabtree, S. Angiuoli, K. D. Eisenach, R. Durmaz, M. L. Joloba, A. Rendon, J. Sifuentes-Osornio, A. Ponce de Leon, M. D. Cave, R. Fleischmann, T. S. Whittam, and D. Alland.** 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 Bacteriol **188:**759-72.

69. **Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, and et al.** 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science **269:**496-512.

70. **Flynn, J. L., J. Chan, K. J. Triebold, D. K. Dalton, T. A. Stewart, and B. R. Bloom.** 1993. An essential role for interferon gamma in resistance to Mycobacterium tuberculosis infection. J Exp Med **178:**2249-54.

71. **Gabriel, S., and L. Ziaugra.** 2004. SNP genotyping using Sequenom MassARRAY 7K platform. Curr Protoc Hum Genet **Chapter 2:**Unit 2 12.

72. **Gabriel, S., L. Ziaugra, and D. Tabbaa.** 2009. SNP genotyping using the Sequenom MassARRAY iPLEX platform. Curr Protoc Hum Genet **Chapter 2:**Unit 2 12.

73. **Galindo, R. C., P. Ayoubi, V. Naranjo, C. Gortazar, K. M. Kocan, and J. de la Fuente.** 2009. Gene expression profiles of European wild boar naturally infected with Mycobacterium bovis. Vet Immunol Immunopathol **129:**119-25.

74. **Gan, Q., I. Chepelev, G. Wei, L. Tarayrah, K. Cui, K. Zhao, and X. Chen.** Dynamic regulation of alternative splicing and chromatin structure in Drosophila gonads revealed by RNA-seq. Cell Res **20:**763-83.

75. **Gao, Q., K. E. Kripke, A. J. Saldanha, W. Yan, S. Holmes, and P. M. Small.** 2005. Gene expression diversity among Mycobacterium tuberculosis clinical isolates. Microbiology **151:**5-14.

76. **Garcia Pelayo, M. C., S. Uplekar, A. Keniry, P. Mendoza Lopez, T. Garnier, J. Nunez Garcia, L. Boschiroli, X. Zhou, J. Parkhill, N. Smith, R. G. Hewinson, S. T. Cole, and S. V. Gordon.** 2009. A comprehensive survey of single nucleotide polymorphisms (SNPs) across Mycobacterium bovis strains and M. bovis BCG vaccine strains refines the genealogy and defines a minimal set of SNPs that separate virulent M. bovis strains and M. bovis BCG strains. Infect Immun **77:**2230-8.

77. **Garnier, T., K. Eiglmeier, J. C. Camus, N. Medina, H. Mansoor, M. Pryor, S. Duthoy, S. Grondin, C. Lacroix, C. Monsempe, S. Simon, B. Harris, R. Atkin, J. Doggett, R. Mayes, L. Keating, P. R. Wheeler, J. Parkhill, B. G. Barrell, S. T. Cole, S. V. Gordon, and R. G. Hewinson.** 2003. The complete genome sequence of Mycobacterium bovis. Proc Natl Acad Sci U S A **100:**7877-82.

78. **Gatewood, M. L., P. Bralley, M. R. Weil, and G. H. Jones.** RNA-Seq and RNA immunoprecipitation analyses of the transcriptome of Streptomyces coelicolor identify substrates for RNase III. J Bacteriol **194:**2228-37.

79. **Golby, P., K. A. Hatch, J. Bacon, R. Cooney, P. Riley, J. Allnutt, J. Hinds, J. Nunez, P. D. Marsh, R. G. Hewinson, and S. V. Gordon.** 2007. Comparative transcriptomics reveals key gene expression differences between the human and bovine pathogens of the Mycobacterium tuberculosis complex. Microbiology **153:**3323-36.

80. **Gollnick, N. S., R. M. Mitchell, M. Baumgart, H. K. Janagama, S. Sreevatsan, and Y. H. Schukken.** 2007. Survival of Mycobacterium avium subsp. paratuberculosis in bovine monocyte-derived macrophages is not affected by host infection status but depends on the infecting bacterial genotype. Vet Immunol Immunopathol **120:**93-105.

81. **Gomez-Lozano, M., R. L. Marvig, S. Molin, and K. S. Long.** Genome-wide identification of novel small RNAs in Pseudomonas aeruginosa. Environ Microbiol.

82. **Gordon, S. V., R. Brosch, A. Billault, T. Garnier, K. Eiglmeier, and S. T. Cole.** 1999. Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. Mol Microbiol **32:**643-55.

83. **Grange, J. M.** 2001. Mycobacterium bovis infection in human beings. Tuberculosis (Edinb) **81:**71-7.

84. **Greenwald, J. W., C. J. Greenwald, B. J. Philmus, T. P. Begley, and D. C. Gross.** RNA-seq analysis reveals that an ECF sigma factor, AcsS, regulates achromobactin biosynthesis in Pseudomonas syringae pv. syringae B728a. PLoS One **7:**e34804.

85. **Gutacker, M. M., B. Mathema, H. Soini, E. Shashkina, B. N. Kreiswirth, E. A. Graviss, and J. M. Musser.** 2006. Single-nucleotide polymorphism-based

population genetic analysis of Mycobacterium tuberculosis strains from 4 geographic sites. J Infect Dis **193:**121-8.

86. **Gutacker, M. M., J. C. Smoot, C. A. Migliaccio, S. M. Ricklefs, S. Hua, D. V. Cousins, E. A. Graviss, E. Shashkina, B. N. Kreiswirth, and J. M. Musser.** 2002. Genome-wide analysis of synonymous single nucleotide polymorphisms in Mycobacterium tuberculosis complex organisms: resolution of genetic relationships among closely related microbial strains. Genetics **162:**1533-43.

87. **Gutierrez, M. C., S. Brisse, R. Brosch, M. Fabre, B. Omais, M. Marmiesse, P. Supply, and V. Vincent.** 2005. Ancient origin and gene mosaicism of the progenitor of Mycobacterium tuberculosis. PLoS Pathog **1:**e5.

88. **Guttman, M., M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, J. L. Rinn, E. S. Lander, and A. Regev.** Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol **28:**503-10.

89. **Hacker, J., G. Blum-Oehler, I. Muhldorfer, and H. Tschape.** 1997. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol Microbiol **23:**1089-97.

90. **Haddad, N., M. Masselot, and B. Durand.** 2004. Molecular differentiation of Mycobacterium bovis isolates. Review of main techniques and applications. Res Vet Sci **76:**1-18.

91. **Harding, C. V., and W. H. Boom.** Regulation of antigen presentation by Mycobacterium tuberculosis: a role for Toll-like receptors. Nat Rev Microbiol **8:**296-307.

92. **Hazbon, M. H., and D. Alland.** 2004. Hairpin primers for simplified single-nucleotide polymorphism analysis of Mycobacterium tuberculosis and other organisms. J Clin Microbiol **42:**1236-42.

93. **Hewinson, R. G., H. M. Vordermeier, N. H. Smith, and S. V. Gordon.** 2006. Recent advances in our knowledge of Mycobacterium bovis: a feeling for the organism. Vet Microbiol **112:**127-39.

94. **Hiller, N. L., R. A. Eutsey, E. Powell, J. P. Earl, B. Janto, D. P. Martin, S. Dawid, A. Ahmed, M. J. Longwell, M. E. Dahlgren, S. Ezzo, H. Tettelin, S. C. Daugherty, T. J. Mitchell, T. A. Hillman, F. J. Buchinsky, A. Tomasz, H. Lencastre, R. Sa-Leao, J. C. Post, F. Z. Hu, and G. D. Ehrlich.** Differences in genotype and virulence among four multidrug-resistant Streptococcus pneumoniae isolates belonging to the PMEN1 clone. PLoS One **6:**e28850.

95. **Hilty, M., C. Diguimbaye, E. Schelling, F. Baggi, M. Tanner, and J. Zinsstag.** 2005. Evaluation of the discriminatory power of variable number tandem repeat (VNTR) typing of Mycobacterium bovis strains. Vet Microbiol **109:**217-22.

96. **Hlavsa, M. C., P. K. Moonan, L. S. Cowan, T. R. Navin, J. S. Kammerer, G. P. Morlock, J. T. Crawford, and P. A. Lobue.** 2008. Human tuberculosis due to Mycobacterium bovis in the United States, 1995-2005. Clin Infect Dis **47:**168-75.

97. **Ho, T. B., B. D. Robertson, G. M. Taylor, R. J. Shaw, and D. B. Young.** 2000. Comparison of Mycobacterium tuberculosis genomes reveals frequent deletions in a 20 kb variable region in clinical isolates. Yeast **17:**272-82.

98. **Homolka, S., S. Niemann, D. G. Russell, and K. H. Rohde.** Functional genetic diversity among Mycobacterium tuberculosis complex clinical isolates: delineation of conserved core and lineage-specific transcriptomes during intracellular survival. PLoS Pathog **6:**e1000988.

99. **Huard, R. C., L. C. Lazzarini, W. R. Butler, D. van Soolingen, and J. L. Ho.** 2003. PCR-based method to differentiate the subspecies of the Mycobacterium tuberculosis complex on the basis of genomic deletions. J Clin Microbiol **41:**1637-50.

100. **Hutchison, C. A., 3rd.** 2007. DNA sequencing: bench to bedside and beyond. Nucleic Acids Res **35:**6227-37.

101. **Idury, R. M., and M. S. Waterman.** 1995. A new algorithm for DNA sequence assembly. J Comput Biol **2:**291-306.

102. **Inwald, J., J. Hinds, J. Dale, S. Palmer, P. Butcher, R. G. Hewinson, and S. V. Gordon.** 2002. Microarray-based comparative genomics: genome plasticity in Mycobacterium bovis. Comp Funct Genomics **3:**342-4.

103. **Iwasaki, A., and R. Medzhitov.** Regulation of adaptive immunity by the innate immune system. Science **327:**291-5.

104. **Jackson, R., G. W. de Lisle, and R. S. Morris.** 1995. A study of the environmental survival of Mycobacterium bovis on a farm in New Zealand. N Z Vet J **43:**346-52.

105. **Kabara, E., C. C. Kloss, M. Wilson, R. J. Tempelman, S. Sreevatsan, H. Janagama, and P. M. Coussens.** A large-scale study of differential gene expression in monocyte-derived macrophages infected with several strains of Mycobacterium avium subspecies paratuberculosis. Brief Funct Genomics **9:**220-37.

106. **Kamerbeek, J., L. Schouls, A. Kolk, M. van Agterveld, D. van Soolingen, S. Kuijper, A. Bunschoten, H. Molhuizen, R. Shaw, M. Goyal, and J. van Embden.** 1997. Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. J Clin Microbiol **35:**907-14.

107. **Kanduma, E., T. D. McHugh, and S. H. Gillespie.** 2003. Molecular methods for Mycobacterium tuberculosis strain typing: a users guide. J Appl Microbiol **94:**781-91.

108. **Kaser, M., J. Hauser, and G. Pluschke.** 2009. Single nucleotide polymorphisms on the road to strain differentiation in Mycobacterium ulcerans. J Clin Microbiol **47:**3647-52.

109. **Kato-Maeda, M., J. T. Rhee, T. R. Gingeras, H. Salamon, J. Drenkow, N. Smittipat, and P. M. Small.** 2001. Comparing genomes within the species Mycobacterium tuberculosis. Genome Res **11:**547-54.

110. **Kaufmann, S. H.** 1993. Immunity to intracellular bacteria. Annu Rev Immunol **11:**129-63.

111. **Keating, L. A., P. R. Wheeler, H. Mansoor, J. K. Inwald, J. Dale, R. G. Hewinson, and S. V. Gordon.** 2005. The pyruvate requirement of some members of the Mycobacterium tuberculosis complex is due to an inactive pyruvate kinase: implications for in vivo growth. Mol Microbiol **56:**163-74.

112. **Killick, K. E., J. A. Browne, S. D. Park, D. A. Magee, I. Martin, K. G. Meade, S. V. Gordon, E. Gormley, C. O'Farrelly, K. Hokamp, and D. E. MacHugh.** Genome-wide transcriptional profiling of peripheral blood leukocytes from cattle infected with Mycobacterium bovis reveals suppression of host immune genes. BMC Genomics **12:**611.

113. **Kim, S., and A. Misra.** 2007. SNP genotyping: technologies and biomedical applications. Annu Rev Biomed Eng **9:**289-320.

114. **Kleinnijenhuis, J., M. Oosting, L. A. Joosten, M. G. Netea, and R. Van Crevel.** Innate immune recognition of Mycobacterium tuberculosis. Clin Dev Immunol **2011:**405310.

115. **Koo, M. S., S. Subbian, and G. Kaplan.** Strain specific transcriptional response in Mycobacterium tuberculosis infected macrophages. Cell Commun Signal **10:**2.

116. **Kristoffersen, S. M., C. Haase, M. R. Weil, K. D. Passalacqua, F. Niazi, S. K. Hutchison, B. Desany, A. B. Kolsto, N. J. Tourasse, T. D. Read, and O. A. Okstad.** Global mRNA decay analysis at single nucleotide resolution reveals segmental and positional degradation patterns in a Gram-positive bacterium. Genome Biol **13:**R30.

117. **Kruh, N. A., J. Troudt, A. Izzo, J. Prenni, and K. M. Dobos.** Portrait of a pathogen: the Mycobacterium tuberculosis proteome in vivo. PLoS One **5:**e13938.

118. **Kumar, R., M. L. Lawrence, J. Watt, A. M. Cooksey, S. C. Burgess, and B. Nanduri.** RNA-seq based transcriptional map of bovine respiratory disease pathogen "Histophilus somni 2336". PLoS One **7:**e29435.

119. **Kunst, F., N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni, V. Azevedo, M. G. Bertero, P. Bessieres, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M. Braun, S. C. Brignell, S. Bron, S. Brouillet, C. V. Bruschi, B. Caldwell, V. Capuano, N. M. Carter, S. K. Choi, J. J. Codani, I. F. Connerton, A. Danchin, and et al.** 1997. The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature **390:**249-56.

120. **Kwan, C. K., and J. D. Ernst.** HIV and tuberculosis: a deadly human syndemic. Clin Microbiol Rev **24:**351-76.

121. **Kwok, P. Y.** 2001. Methods for genotyping single nucleotide polymorphisms. Annu Rev Genomics Hum Genet **2:**235-58.

122. **Kwok, P. Y., and X. Chen.** 2003. Detection of single nucleotide polymorphisms. Curr Issues Mol Biol **5:**43-60.

123. **LaFramboise, T.** 2009. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Res **37:**4181-93.

124. **Lambros, M. B., P. M. Wilkerson, R. Natrajan, N. Patani, V. Pawar, R. Vatcheva, M. Mansour, M. Laschet, B. Oelze, N. Orr, S. Muller, and J. S. Reis-Filho.** High-throughput detection of fusion genes in cancer using the Sequenom MassARRAY platform. Lab Invest **91:**1491-501.

125. **Lewinsohn, D. A., A. S. Heinzel, J. M. Gardner, L. Zhu, M. R. Alderson, and D. M. Lewinsohn.** 2003. Mycobacterium tuberculosis-specific CD8+ T cells preferentially recognize heavily infected cells. Am J Respir Crit Care Med **168:**1346-52.

126. **Lewis, K. N., R. Liao, K. M. Guinn, M. J. Hickey, S. Smith, M. A. Behr, and D. R. Sherman.** 2003. Deletion of RD1 from Mycobacterium tuberculosis mimics bacille Calmette-Guerin attenuation. J Infect Dis **187:**117-23.

127. **Li, C., Y. Zhang, R. Wang, J. Lu, S. Nandi, S. Mohanty, J. Terhune, Z. Liu, and E. Peatman.** RNA-seq analysis of mucosal immune responses reveals signatures of intestinal barrier disruption and pathogen entry following Edwardsiella ictaluri infection in channel catfish, Ictalurus punctatus. Fish Shellfish Immunol **32:**816-27.

128. **Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin.** 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25:**2078-9.

129. **Li, H., J. Ruan, and R. Durbin.** 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res **18:**1851-8.

130. **Li, M. S., I. M. Monahan, S. J. Waddell, J. A. Mangan, S. L. Martin, M. J. Everett, and P. D. Butcher.** 2001. cDNA-RNA subtractive hybridization reveals increased expression of mycocerosic acid synthase in intracellular Mycobacterium bovis BCG. Microbiology **147:**2293-305.

131. **Liu, W., L. Fang, M. Li, S. Li, S. Guo, R. Luo, Z. Feng, B. Li, Z. Zhou, G. Shao, H. Chen, and S. Xiao.** Comparative genomics of Mycoplasma: analysis of conserved essential genes and diversity of the pan-genome. PLoS One **7:**e35698.

132. **Livak, K. J., and T. D. Schmittgen.** 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods **25:**402-8.

133. **LoBue, P. A., D. A. Enarson, and C. O. Thoen.** Tuberculosis in humans and animals: an overview. Int J Tuberc Lung Dis **14:**1075-8.

134. **Luo, H., L. Pang, and J. Xie.** [Biosynthesis and regulation of mycolic acids in Mycobacterium tuberculosis--a review]. Wei Sheng Wu Xue Bao **52:**146-51.

135. **MacHugh, D. E., E. Gormley, S. D. Park, J. A. Browne, M. Taraktsoglou, C. O'Farrelly, and K. G. Meade.** 2009. Gene expression profiling of the host response to Mycobacterium bovis infection in cattle. Transbound Emerg Dis **56:**204-14.

136. **Magee, D. A., M. Taraktsoglou, K. E. Killick, N. C. Nalpas, J. A. Browne, S. D. Park, K. M. Conlon, D. J. Lynn, K. Hokamp, S. V. Gordon, E. Gormley, and D. E. MacHugh.** Global gene expression and systems biology analysis of bovine monocyte-derived macrophages in response to in vitro challenge with Mycobacterium bovis. PLoS One **7:**e32034.

137. **Mahairas, G. G., P. J. Sabo, M. J. Hickey, D. C. Singh, and C. K. Stover.** 1996. Molecular analysis of genetic differences between Mycobacterium bovis BCG and virulent M. bovis. J Bacteriol **178:**1274-82.

138. **Maher, C. A., C. Kumar-Sinha, X. Cao, S. Kalyana-Sundaram, B. Han, X. Jing, L. Sam, T. Barrette, N. Palanisamy, and A. M. Chinnaiyan.** 2009. Transcriptome sequencing to detect gene fusions in cancer. Nature **458:**97-101.

139. **Manca, C., L. Tsenova, A. Bergtold, S. Freeman, M. Tovey, J. M. Musser, C. E. Barry, 3rd, V. H. Freedman, and G. Kaplan.** 2001. Virulence of a Mycobacterium tuberculosis clinical isolate in mice is determined by failure to induce Th1 type immunity and is associated with induction of IFN-alpha /beta. Proc Natl Acad Sci U S A **98:**5752-7.

140. **Mangan, A., I. Monahan, and P. Butcher.** 2002. Gene expression during host-pathogen interactions: approaches to bacterial mRNA extraction and labelling for microarray analysis. Methods in Microbiology**:**137-151.

141. **Marguerat, S., and J. Bahler.** RNA-seq: from technology to biology. Cell Mol Life Sci **67:**569-79.

142. **Mathema, B., N. Kurepina, G. Yang, E. Shashkina, C. Manca, C. Mehaffy, H. Bielefeldt-Ohmann, S. Ahuja, D. A. Fallows, A. Izzo, P. Bifani, K. Dobos, G. Kaplan, and B. N. Kreiswirth.** Epidemiologic consequences of microvariation in Mycobacterium tuberculosis. J Infect Dis **205:**964-74.

143. **McClelland, M., and R. K. Wilson.** 1998. Comparison of sample sequences of the Salmonella typhi genome to the sequence of the complete Escherichia coli K-12 genome. Infect Immun **66:**4305-12.

144. **McDonough, K. A., Y. Kress, and B. R. Bloom.** 1993. Pathogenesis of tuberculosis: interaction of Mycobacterium tuberculosis with macrophages. Infect Immun **61:**2763-73.

145. **McEvoy, C. R., R. Cloete, B. Muller, A. C. Schurch, P. D. van Helden, S. Gagneux, R. M. Warren, and N. C. Gey van Pittius.** Comparative analysis of Mycobacterium tuberculosis pe and ppe genes reveals high sequence variation and an apparent absence of selective constraints. PLoS One **7:**e30593.

146. **McEvoy, C. R., A. A. Falmer, N. C. Gey van Pittius, T. C. Victor, P. D. van Helden, and R. M. Warren.** 2007. The role of IS6110 in the evolution of Mycobacterium tuberculosis. Tuberculosis (Edinb) **87:**393-404.

147. **McFadden, J.** 1996. Recombination in mycobacteria. Mol Microbiol **21:**205-11.

148. **McPherson, J. D.** 2009. Next-generation gap. Nat Methods **6:**S2-5.

149. **Meade, K. G., E. Gormley, M. B. Doyle, T. Fitzsimons, C. O'Farrelly, E. Costello, J. Keane, Y. Zhao, and D. E. MacHugh.** 2007. Innate gene repression

associated with Mycobacterium bovis infection in cattle: toward a gene signature of disease. BMC Genomics **8:**400.

150. **Meade, K. G., E. Gormley, C. O'Farrelly, S. D. Park, E. Costello, J. Keane, Y. Zhao, and D. E. MacHugh.** 2008. Antigen stimulation of peripheral blood mononuclear cells from Mycobacterium bovis infected cattle yields evidence for a novel gene expression program. BMC Genomics **9:**447.

151. **Meade, K. G., E. Gormley, S. D. Park, T. Fitzsimons, G. J. Rosa, E. Costello, J. Keane, P. M. Coussens, and D. E. MacHugh.** 2006. Gene expression profiling of peripheral blood mononuclear cells (PBMC) from Mycobacterium bovis infected cattle after in vitro antigenic stimulation with purified protein derivative of tuberculin (PPD). Vet Immunol Immunopathol **113:**73-89.

152. **Means, T. K., S. Wang, E. Lien, A. Yoshimura, D. T. Golenbock, and M. J. Fenton.** 1999. Human toll-like receptors mediate cellular activation by Mycobacterium tuberculosis. J Immunol **163:**3920-7.

153. **Mehaffy, C., A. Hess, J. E. Prenni, B. Mathema, B. Kreiswirth, and K. M. Dobos.** Descriptive proteomic analysis shows protein variability between closely related clinical isolates of Mycobacterium tuberculosis. Proteomics **10:**1966-1984.

154. **Mehra, S., B. Pahar, N. K. Dutta, C. N. Conerly, K. Philippi-Falkenstein, X. Alvarez, and D. Kaushal.** Transcriptional reprogramming in nonhuman primate (rhesus macaque) tuberculosis granulomas. PLoS One **5:**e12266.

155. **Meijer, A. H., F. J. Verbeek, E. Salas-Vidal, M. Corredor-Adamez, J. Bussman, A. M. van der Sar, G. W. Otto, R. Geisler, and H. P. Spaink.** 2005. Transcriptome profiling of adult zebrafish at the late stage of chronic tuberculosis due to Mycobacterium marinum infection. Mol Immunol **42:**1185-203.

156. **Metzker, M. L.** Sequencing technologies - the next generation. Nat Rev Genet **11:**31-46.

157. **Meyer, K., and P. M. Ueland.** Use of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry for multiplex genotyping. Adv Clin Chem **53:**1-29.

158. **Michel, A. L., B. Muller, and P. D. van Helden.** Mycobacterium bovis at the animal-human interface: a problem, or not? Vet Microbiol **140:**371-81.

159. **Miller, J. R., S. Koren, and G. Sutton.** Assembly algorithms for next-generation sequencing data. Genomics **95:**315-27.

160. **Monot, M., N. Honore, T. Garnier, N. Zidane, D. Sherafi, A. Paniz-Mondolfi, M. Matsuoka, G. M. Taylor, H. D. Donoghue, A. Bouwman, S. Mays, C. Watson, D. Lockwood, A. Khamesipour, Y. Dowlati, S. Jianping, T. H. Rea, L. Vera-Cabrera, M. M. Stefani, S. Banu, M. Macdonald, B. R. Sapkota, J. S. Spencer, J. Thomas, K. Harshman, P. Singh, P. Busso, A. Gattiker, J. Rougemont, P. J. Brennan, and S. T. Cole.** 2009. Comparative genomic and phylogeographic analysis of Mycobacterium leprae. Nat Genet **41:**1282-9.

161. **Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold.** 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods **5:**621-8.

162. **Mostowy, S., and M. A. Behr.** 2005. The origin and evolution of Mycobacterium tuberculosis. Clin Chest Med **26:**207-16, v-vi.

163. **Mostowy, S., J. Inwald, S. Gordon, C. Martin, R. Warren, K. Kremer, D. Cousins, and M. A. Behr.** 2005. Revisiting the evolution of Mycobacterium bovis. J Bacteriol **187:**6386-95.

164. **Mostowy, S., A. G. Tsolaki, P. M. Small, and M. A. Behr.** 2003. The in vitro evolution of BCG vaccines. Vaccine **21:**4270-4.

165. **Motiwala, A. S., H. K. Janagama, M. L. Paustian, X. Zhu, J. P. Bannantine, V. Kapur, and S. Sreevatsan.** 2006. Comparative transcriptional analysis of human macrophages exposed to animal and human isolates of Mycobacterium avium subspecies paratuberculosis with diverse genotypes. Infect Immun **74:**6046-56.

166. **Mukhopadhyay, S., S. Nair, and S. Ghosh.** Pathogenesis in tuberculosis: transcriptomic approaches to unraveling virulence mechanisms and finding new drug targets. FEMS Microbiol Rev **36:**463-85.

167. **Mulcahy, G. M., Z. C. Kaminski, E. A. Albanese, R. Sood, and M. Pierce.** 1996. IS6110-based PCR methods for detection of Mycobacterium tuberculosis. J Clin Microbiol **34:**1348-9.

168. **Muller, B., M. Hilty, S. Berg, M. C. Garcia-Pelayo, J. Dale, M. L. Boschiroli, S. Cadmus, B. N. Ngandolo, S. Godreuil, C. Diguimbaye-Djaibe, R. Kazwala, B. Bonfoh, B. M. Njanpop-Lafourcade, N. Sahraoui, D. Guetarni, A. Aseffa, M. H. Mekonnen, V. R. Razanamparany, H. Ramarokoto, B. Djonne, J. Oloya, A. Machado, C. Mucavele, E. Skjerve, F. Portaels, L. Rigouts, A. Michel, A. Muller, G. Kallenius, P. D. van Helden, R. G. Hewinson, J. Zinsstag, S. V. Gordon, and N. H. Smith.** 2009. African 1, an epidemiologically important clonal complex of Mycobacterium bovis dominant in Mali, Nigeria, Cameroon, and Chad. J Bacteriol **191:**1951-60.

169. **Munroe, F. A., I. R. Dohoo, and W. B. McNab.** 2000. Estimates of within-herd incidence rates of Mycobacterium bovis in Canadian cattle and cervids between 1985 and 1994. Prev Vet Med **45:**247-56.

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

171. **Myers, E. W.** 1995. Toward simplifying and accurately formulating fragment assembly. J Comput Biol **2:**275-90.

172. **Nagalakshmi, U., K. Waern, and M. Snyder.** RNA-Seq: a method for comprehensive transcriptome analysis. Curr Protoc Mol Biol **Chapter 4:**Unit 4 11 1-13.

173. **Namiki, T., T. Hachiya, H. Tanaka, and Y. Sakakibara.** MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res.

174. **Naranjo, V., C. Gortazar, M. Villar, and J. de la Fuente.** 2007. Comparative genomics and proteomics to study tissue-specific response and function in natural Mycobacterium bovis infections. Anim Health Res Rev **8:**81-8.

175. **Naranjo, V., M. Villar, M. P. Martin-Hernando, D. Vidal, U. Hofle, C. Gortazar, K. M. Kocan, J. Vazquez, and J. de la Fuente.** 2007. Proteomic and transcriptomic analyses of differential stress/inflammatory responses in mandibular lymph nodes and oropharyngeal tonsils of European wild boars naturally infected with Mycobacterium bovis. Proteomics **7:**220-31.

176. **Neafsey, D. E., S. F. Schaffner, S. K. Volkman, D. Park, P. Montgomery, D. A. Milner, Jr., A. Lukens, D. Rosen, R. Daniels, N. Houde, J. F. Cortese, E. Tyndall, C. Gates, N. Stange-Thomann, O. Sarr, D. Ndiaye, O. Ndir, S. Mboup, M. U. Ferreira, L. Moraes Sdo, A. P. Dash, C. E. Chitnis, R. C. Wiegand, D. L. Hartl, B. W. Birren, E. S. Lander, P. C. Sabeti, and D. F. Wirth.** 2008. Genome-wide SNP genotyping highlights the role of natural selection in Plasmodium falciparum population divergence. Genome Biol **9:**R171.

177. **Newton, S. M., R. J. Smith, K. A. Wilkinson, M. P. Nicol, N. J. Garton, K. J. Staples, G. R. Stewart, J. R. Wain, A. R. Martineau, S. Fandrich, T. Smallie, B. Foxwell, A. Al-Obaidi, J. Shafi, K. Rajakumar, B. Kampmann, P. W. Andrew, L. Ziegler-Heitbrock, M. R. Barer, and R. J. Wilkinson.** 2006. A deletion defining a common Asian lineage of Mycobacterium tuberculosis associates with immune subversion. Proc Natl Acad Sci U S A **103:**15594-8.

178. **Nguyen, D., P. Brassard, D. Menzies, L. Thibert, R. Warren, S. Mostowy, and M. Behr.** 2004. Genomic characterization of an endemic Mycobacterium tuberculosis strain: evolutionary and epidemiologic implications. J Clin Microbiol **42:**2573-80.

179. **Nishi, J. S., T. Shury, and B. T. Elkin.** 2006. Wildlife reservoirs for bovine tuberculosis (Mycobacterium bovis) in Canada: strategies for management and research. Vet Microbiol **112:**325-38.

180. **Nissum, M., D. Preuss, A. Harig, U. Lieberwirth, C. Betz, S. Neumann, E. Deravanessian, M. Bock, L. Wehmeier, and T. Bonk.** 2002. High-throughput genetic screening using matrix-assisted laser desorption/ionization mass spectrometry. Psychiatr Genet **12:**109-17.

181. **Nowrousian, M.** Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. Eukaryot Cell **9:**1300-10.

182. **Nowrousian, M., J. E. Stajich, M. Chu, I. Engh, E. Espagne, K. Halliday, J. Kamerewerd, F. Kempken, B. Knab, H. C. Kuo, H. D. Osiewacz, S. Poggeler, N. D. Read, S. Seiler, K. M. Smith, D. Zickler, U. Kuck, and M. Freitag.** De novo assembly of a 40 Mb eukaryotic genome from short sequence reads:

Sordaria macrospora, a model organism for fungal morphogenesis. PLoS Genet **6:**e1000891.

183. **O'Brien, D. J., S. D. Fitzgerald, T. J. Lyon, K. L. Butler, J. S. Fierke, K. R. Clarke, S. M. Schmitt, T. M. Cooley, and D. E. Derry.** 2001. Tuberculous lesions in free-ranging white-tailed deer in Michigan. J Wildl Dis **37:**608-13.

184. **O'Brien, D. J., S. M. Schmitt, J. S. Fierke, S. A. Hogle, S. R. Winterstein, T. M. Cooley, W. E. Moritz, K. L. Diegel, S. D. Fitzgerald, D. E. Berry, and J. B. Kaneene.** 2002. Epidemiology of Mycobacterium bovis in free-ranging white-tailed deer, Michigan, USA, 1995-2000. Prev Vet Med **54:**47-63.

185. **O'Brien, D. J., S. M. Schmitt, S. D. Fitzgerald, and D. E. Berry.** Management of bovine tuberculosis in Michigan wildlife: current status and near term prospects. Vet Microbiol **151:**179-87.

186. **Okafor, C. C., D. L. Grooms, C. S. Bruning-Fann, J. J. Averill, and J. B. Kaneene.** Descriptive epidemiology of bovine tuberculosis in michigan (1975-2010): lessons learned. Vet Med Int **2011:**874924.

187. **Olea-Popelka, F. J., J. Phelan, P. W. White, G. McGrath, J. D. Collins, J. O'Keeffe, M. Duggan, D. M. Collins, D. F. Kelton, O. Berke, S. J. More, and S. W. Martin.** 2006. Quantifying badger exposure and the risk of bovine tuberculosis for cattle herds in county Kilkenny, Ireland. Prev Vet Med **75:**34-46.

188. **Otal, I., A. B. Gomez, K. Kremer, P. de Haas, M. J. Garcia, C. Martin, and D. van Soolingen.** 2008. Mapping of IS6110 insertion sites in Mycobacterium bovis isolates in relation to adaptation from the animal to human host. Vet Microbiol **129:**333-41.

189. **Otal, I., C. Martin, V. Vincent-Levy-Frebault, D. Thierry, and B. Gicquel.** 1991. Restriction fragment length polymorphism analysis using IS6110 as an epidemiological marker in tuberculosis. J Clin Microbiol **29:**1252-4.

190. **Palanisamy, G. S., N. DuTeau, K. D. Eisenach, D. M. Cave, S. A. Theus, B. N. Kreiswirth, R. J. Basaraba, and I. M. Orme.** 2009. Clinical strains of Mycobacterium tuberculosis display a wide range of virulence in guinea pigs. Tuberculosis (Edinb) **89:**203-9.

191. **Palmer, K. L., P. Godfrey, A. Griggs, V. N. Kos, J. Zucker, C. Desjardins, G. Cerqueira, D. Gevers, S. Walker, J. Wortman, M. Feldgarden, B. Haas, B. Birren, and M. S. Gilmore.** Comparative genomics of enterococci: variation in Enterococcus faecalis, clade structure in E. faecium, and defining characteristics of E. gallinarum and E. casseliflavus. MBio **3:**e00318-11.

192. **Palmer, M. V., W. R. Waters, and D. L. Whipple.** 2004. Shared feed as a means of deer-to-deer transmission of Mycobacterium bovis. J Wildl Dis **40:**87-91.

193. **Palmer, M. V., and D. L. Whipple.** 2006. Survival of Mycobacterium bovis on feedstuffs commonly used as supplemental feed for white-tailed deer (Odocoileus virginianus). J Wildl Dis **42:**853-8.

194. **Pan, Y., X. Yang, J. Duan, N. Lu, A. S. Leung, V. Tran, Y. Hu, N. Wu, D. Liu, Z. Wang, X. Yu, C. Chen, Y. Zhang, K. Wan, J. Liu, and B. Zhu.** Whole-

genome sequences of four Mycobacterium bovis BCG vaccine strains. J Bacteriol **193:**3152-3.

195. **Pareek, C. S., R. Smoczynski, and A. Tretyn.** Sequencing technologies and genome sequencing. J Appl Genet **52:**413-35.

196. **Pasricha, R., A. Chandolia, P. Ponnan, N. K. Saini, S. Sharma, M. Chopra, M. V. Basil, V. Brahmachari, and M. Bose.** Single nucleotide polymorphism in the genes of mce1 and mce4 operons of Mycobacterium tuberculosis: analysis of clinical isolates and standard reference strains. BMC Microbiol **11:**41.

197. **Paszkiewicz, K., and D. J. Studholme.** De novo assembly of short sequence reads. Brief Bioinform **11:**457-72.

198. **Penn, K., and P. R. Jensen.** Comparative genomics reveals evidence of marine adaptation in Salinispora species. BMC Genomics **13:**86.

199. **Pepke, S., B. Wold, and A. Mortazavi.** 2009. Computation for ChIP-seq and RNA-seq studies. Nat Methods **6:**S22-32.

200. **Perna, N. T., G. Plunkett, 3rd, V. Burland, B. Mau, J. D. Glasner, D. J. Rose, G. F. Mayhew, P. S. Evans, J. Gregor, H. A. Kirkpatrick, G. Posfai, J. Hackett, S. Klink, A. Boutin, Y. Shao, L. Miller, E. J. Grotbeck, N. W. Davis, A. Lim, E. T. Dimalanta, K. D. Potamousis, J. Apodaca, T. S. Anantharaman, J. Lin, G. Yen, D. C. Schwartz, R. A. Welch, and F. R. Blattner.** 2001. Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. Nature **409:**529-33.

201. **Pflueger, D., S. Terry, A. Sboner, L. Habegger, R. Esgueva, P. C. Lin, M. A. Svensson, N. Kitabayashi, B. J. Moss, T. Y. MacDonald, X. Cao, T. Barrette, A. K. Tewari, M. S. Chee, A. M. Chinnaiyan, D. S. Rickman, F. Demichelis, M. B. Gerstein, and M. A. Rubin.** Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome Res **21:**56-67.

202. **Philipp, W. J., S. Nair, G. Guglielmi, M. Lagranderie, B. Gicquel, and S. T. Cole.** 1996. Physical mapping of Mycobacterium bovis BCG pasteur reveals differences from the genome map of Mycobacterium tuberculosis H37Rv and from M. bovis. Microbiology **142 (Pt 11):**3135-45.

203. **Philips, J. A., and J. D. Ernst.** Tuberculosis pathogenesis and immunity. Annu Rev Pathol **7:**353-84.

204. **Pinto, A. C., H. P. Melo-Barbosa, A. Miyoshi, A. Silva, and V. Azevedo.** Application of RNA-seq to reveal the transcript profile in bacteria. Genet Mol Res **10:**1707-18.

205. **Portevin, D., S. Gagneux, I. Comas, and D. Young.** Human macrophage responses to clinical isolates from the Mycobacterium tuberculosis complex discriminate between ancient and modern lineages. PLoS Pathog **7:**e1001307.

206. **Pym, A. S., P. Brodin, R. Brosch, M. Huerre, and S. T. Cole.** 2002. Loss of RD1 contributed to the attenuation of the live tuberculosis vaccines Mycobacterium bovis BCG and Mycobacterium microti. Mol Microbiol **46:**709-17.

207. **Raja, A.** 2004. Immunology of tuberculosis. Indian J Med Res **120:**213-32.

208. **Reed, M. B., P. Domenech, C. Manca, H. Su, A. K. Barczak, B. N. Kreiswirth, G. Kaplan, and C. E. Barry, 3rd.** 2004. A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. Nature **431:**84-7.

209. **Rehren, G., S. Walters, P. Fontan, I. Smith, and A. M. Zarraga.** 2007. Differential gene expression between Mycobacterium bovis and Mycobacterium tuberculosis. Tuberculosis (Edinb) **87:**347-59.

210. **Rengarajan, J., E. Murphy, A. Park, C. L. Krone, E. C. Hett, B. R. Bloom, L. H. Glimcher, and E. J. Rubin.** 2008. Mycobacterium tuberculosis Rv2224c modulates innate immune responses. Proc Natl Acad Sci U S A **105:**264-9.

211. **Rodriguez-Campos, S., A. C. Schurch, J. Dale, A. J. Lohan, M. V. Cunha, A. Botelho, K. De Cruz, M. L. Boschiroli, M. B. Boniotti, M. Pacciarini, M. C. Garcia-Pelayo, B. Romero, L. de Juan, L. Dominguez, S. V. Gordon, D. van Soolingen, B. Loftus, S. Berg, R. G. Hewinson, A. Aranaz, and N. H. Smith.** European 2--a clonal complex of Mycobacterium bovis dominant in the Iberian Peninsula. Infect Genet Evol **12:**866-72.

212. **Rodwell, T. C., A. J. Kapasi, M. Moore, F. Milian-Suazo, B. Harris, L. P. Guerrero, K. Moser, S. A. Strathdee, and R. S. Garfein.** Tracing the origins of Mycobacterium bovis tuberculosis in humans in the USA to cattle in Mexico using spoligotyping. Int J Infect Dis **14 Suppl 3:**e129-35.

213. **Russell, D. G.** 2001. Mycobacterium tuberculosis: here today, and here tomorrow. Nat Rev Mol Cell Biol **2:**569-77.

214. **Sahl, J. W., and D. A. Rasko.** Analysis of global transcriptional profiles of enterotoxigenic Escherichia coli isolate E24377A. Infect Immun **80:**1232-42.

215. **Sakamoto, K.** The Pathology of Mycobacterium tuberculosis Infection. Vet Pathol **49:**423-39.

216. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1992. DNA sequencing with chain-terminating inhibitors. 1977. Biotechnology **24:**104-8.

217. **Sauer, S., and I. G. Gut.** 2002. Genotyping single-nucleotide polymorphisms by matrix-assisted laser-desorption/ionization time-of-flight mass spectrometry. J Chromatogr B Analyt Technol Biomed Life Sci **782:**73-87.

218. **Shen, G. Q., K. G. Abdullah, and Q. K. Wang.** 2009. The TaqMan method for SNP genotyping. Methods Mol Biol **578:**293-306.

219. **Shendure, J., and H. Ji.** 2008. Next-generation DNA sequencing. Nat Biotechnol **26:**1135-45.

220. **Shi, M. M.** 2001. Enabling large-scale pharmacogenetic studies by high-throughput mutation detection and genotyping technologies. Clin Chem **47:**164-72.

221. **Smith, N. H.** The global distribution and phylogeography of Mycobacterium bovis clonal complexes. Infect Genet Evol **12:**857-65.

222. **Smith, N. H., S. Berg, J. Dale, A. Allen, S. Rodriguez, B. Romero, F. Matos, S. Ghebremichael, C. Karoui, C. Donati, C. Machado Ada, C. Mucavele, R. R. Kazwala, M. Hilty, S. Cadmus, B. N. Ngandolo, M. Habtamu, J. Oloya, A.

**Muller, F. Milian-Suazo, O. Andrievskaia, M. Projahn, S. Barandiaran, A. Macias, B. Muller, M. S. Zanini, C. Y. Ikuta, C. A. Rodriguez, S. R. Pinheiro, A. Figueroa, S. N. Cho, N. Mosavari, P. C. Chuang, R. Jou, J. Zinsstag, D. van Soolingen, E. Costello, A. Aseffa, F. Proano-Perez, F. Portaels, L. Rigouts, A. A. Cataldi, D. M. Collins, M. L. Boschiroli, R. G. Hewinson, J. S. Ferreira Neto, O. Surujballi, K. Tadyon, A. Botelho, A. M. Zarraga, N. Buller, R. Skuce, A. Michel, A. Aranaz, S. V. Gordon, B. Y. Jeon, G. Kallenius, S. Niemann, M. B. Boniotti, P. D. van Helden, B. Harris, M. J. Zumarraga, and K. Kremer.** European 1: a globally important clonal complex of Mycobacterium bovis. Infect Genet Evol **11:**1340-51.

223. **Smith, N. H., S. V. Gordon, R. de la Rua-Domenech, R. S. Clifton-Hadley, and R. G. Hewinson.** 2006. Bottlenecks and broomsticks: the molecular evolution of Mycobacterium bovis. Nat Rev Microbiol **4:**670-81.

224. **Sreevatsan, S., X. Pan, K. E. Stockbauer, N. D. Connell, B. N. Kreiswirth, T. S. Whittam, and J. M. Musser.** 1997. Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination. Proc Natl Acad Sci U S A **94:**9869-74.

225. **Stinear, T. P., T. Seemann, P. F. Harrison, G. A. Jenkin, J. K. Davies, P. D. Johnson, Z. Abdellah, C. Arrowsmith, T. Chillingworth, C. Churcher, K. Clarke, A. Cronin, P. Davis, I. Goodhead, N. Holroyd, K. Jagels, A. Lord, S. Moule, K. Mungall, H. Norbertczak, M. A. Quail, E. Rabbinowitsch, D. Walker, B. White, S. Whitehead, P. L. Small, R. Brosch, L. Ramakrishnan, M. A. Fischbach, J. Parkhill, and S. T. Cole.** 2008. Insights from the complete genome sequence of Mycobacterium marinum on the evolution of Mycobacterium tuberculosis. Genome Res **18:**729-41.

226. **Stothard, P., and D. S. Wishart.** 2006. Automated bacterial genome analysis and annotation. Curr Opin Microbiol **9:**505-10.

227. **Su, Z., B. Ning, H. Fang, H. Hong, R. Perkins, W. Tong, and L. Shi.** Next-generation sequencing and its applications in molecular diagnostics. Expert Rev Mol Diagn **11:**333-43.

228. **Sudheesh, P. S., A. Al-Ghabshi, N. Al-Mazrooei, and S. Al-Habsi.** Comparative pathogenomics of bacteria causing infectious diseases in fish. Int J Evol Biol **2012:**457264.

229. **Sultan, M., M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O'Keeffe, S. Haas, M. Vingron, H. Lehrach, and M. L. Yaspo.** 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science **321:**956-60.

230. **Supply, P., C. Allix, S. Lesjean, M. Cardoso-Oelemann, S. Rusch-Gerdes, E. Willery, E. Savine, P. de Haas, H. van Deutekom, S. Roring, P. Bifani, N. Kurepina, B. Kreiswirth, C. Sola, N. Rastogi, V. Vatin, M. C. Gutierrez, M. Fauville, S. Niemann, R. Skuce, K. Kremer, C. Locht, and D. van Soolingen.** 2006. Proposal for standardization of optimized mycobacterial interspersed

repetitive unit-variable-number tandem repeat typing of Mycobacterium tuberculosis. J Clin Microbiol **44:**4498-510.

231. **Syvanen, A. C.** 2005. Toward genome-wide SNP genotyping. Nat Genet **37 Suppl:**S5-10.

232. **Talbot, E. A., D. L. Williams, and R. Frothingham.** 1997. PCR identification of Mycobacterium bovis BCG. J Clin Microbiol **35:**566-9.

233. **Tamura, K., J. Dudley, M. Nei, and S. Kumar.** 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol **24:**1596-9.

234. **Tang, H., Y. Yao, L. Wang, H. Yu, Y. Ren, G. Wu, and P. Xu.** Genomic analysis of Pseudomonas putida: genes in a genome island are crucial for nicotine degradation. Sci Rep **2:**377.

235. **Taraktsoglou, M., U. Szalabska, D. A. Magee, J. A. Browne, T. Sweeney, E. Gormley, and D. E. MacHugh.** Transcriptional profiling of immune genes in bovine monocyte-derived macrophages exposed to bacterial antigens. Vet Immunol Immunopathol **140:**130-9.

236. **Thoen, C. O., Barletta, R. G.** 2004. Pathogenesis of bacteria infections animals.69-76.

237. **Tian, C., and X. Jian-Ping.** Roles of PE_PGRS family in Mycobacterium tuberculosis pathogenesis and novel measures against tuberculosis. Microb Pathog **49:**311-4.

238. **Tost, J., and I. G. Gut.** 2005. Genotyping single nucleotide polymorphisms by MALDI mass spectrometry in clinical applications. Clin Biochem **38:**335-50.

239. **Trapnell, C., L. Pachter, and S. L. Salzberg.** 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics **25:**1105-11.

240. **Trapnell, C., and S. L. Salzberg.** 2009. How to map billions of short reads onto genomes. Nat Biotechnol **27:**455-7.

241. **Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter.** Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol **28:**511-5.

242. **Treangen, T. J., and S. L. Salzberg.** Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet **13:**36-46.

243. **Tsenova, L., E. Ellison, R. Harbacheuski, A. L. Moreira, N. Kurepina, M. B. Reed, B. Mathema, C. E. Barry, 3rd, and G. Kaplan.** 2005. Virulence of selected Mycobacterium tuberculosis clinical isolates in the rabbit model of meningitis is dependent on phenolic glycolipid produced by the bacilli. J Infect Dis **192:**98-106.

244. **Tsolaki, A. G., A. E. Hirsh, K. DeRiemer, J. A. Enciso, M. Z. Wong, M. Hannan, Y. O. Goguet de la Salmoniere, K. Aman, M. Kato-Maeda, and P. M. Small.** 2004. Functional and evolutionary genomics of Mycobacterium tuberculosis: insights from genomic deletions in 100 strains. Proc Natl Acad Sci U S A **101:**4865-70.

245. **Tuefferd, M., A. de Bondt, I. Van den Wyngaert, W. Talloen, and H. Gohlmann.** Microarray profiling of DNA extracted from FFPE tissues using SNP 6.0 Affymetrix platform. Methods Mol Biol **724:**147-60.

246. **van der Sar, A. M., H. P. Spaink, A. Zakrzewska, W. Bitter, and A. H. Meijer.** 2009. Specificity of the zebrafish host transcriptome response to acute and chronic mycobacterial infection and the role of innate and adaptive immune components. Mol Immunol **46:**2317-32.

247. **Walzl, G., K. Ronacher, W. Hanekom, T. J. Scriba, and A. Zumla.** Immunological biomarkers of tuberculosis. Nat Rev Immunol **11:**343-54.

248. **Wang, J., W. Wang, Y. Liu, L. Duo, L. Huang, and X. Jiang.** 2009. The method of single-nucleotide variations detection using capillary electrophoresis and molecular beacons. Mol Biol Rep **36:**1903-8.

249. **Wang, Y., X. Zhou, J. Lin, F. Yin, L. Xu, Y. Huang, T. Ding, and D. Zhao.** Effects of Mycobacterium bovis on monocyte-derived macrophages from bovine tuberculosis infection and healthy cattle. FEMS Microbiol Lett **321:**30-6.

250. **Wang, Z., M. Gerstein, and M. Snyder.** 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet **10:**57-63.

251. **Warren, R. M., E. M. Streicher, S. L. Sampson, G. D. van der Spuy, M. Richardson, D. Nguyen, M. A. Behr, T. C. Victor, and P. D. van Helden.** 2002. Microevolution of the direct repeat region of Mycobacterium tuberculosis: implications for interpretation of spoligotyping data. J Clin Microbiol **40:**4457-65.

252. **Waters, W. R., M. V. Palmer, T. C. Thacker, W. C. Davis, S. Sreevatsan, P. Coussens, K. G. Meade, J. C. Hope, and D. M. Estes.** Tuberculosis immunity: opportunities from studies with cattle. Clin Dev Immunol **2011:**768542.

253. **Waters, W. R., A. O. Whelan, K. P. Lyashchenko, R. Greenwald, M. V. Palmer, B. N. Harris, R. G. Hewinson, and H. M. Vordermeier.** Immune responses in cattle inoculated with Mycobacterium bovis, Mycobacterium tuberculosis, or Mycobacterium kansasii. Clin Vaccine Immunol **17:**247-52.

254. **Wedlock, D. N., R. P. Kawakami, J. Koach, B. M. Buddle, and D. M. Collins.** 2006. Differences of gene expression in bovine alveolar macrophages infected with virulent and attenuated isogenic strains of Mycobacterium bovis. Int Immunopharmacol **6:**957-61.

255. **Wheelis, M. L., O. Kandler, and C. R. Woese.** 1992. On the nature of global classification. Proc Natl Acad Sci U S A **89:**2930-4.

256. **Widdison, S., M. Watson, J. Piercy, C. Howard, and T. J. Coffey.** 2008. Granulocyte chemotactic properties of M. tuberculosis versus M. bovis-infected bovine alveolar macrophages. Mol Immunol **45:**740-9.

257. **Wilkinson, M. D.** Genomics data resources: frameworks and standards. Methods Mol Biol **856:**489-511.

258. **Wilms, I., A. Overloper, M. Nowrousian, C. M. Sharma, and F. Narberhaus.** Deep sequencing uncovers numerous small RNAs on all four replicons of the plant pathogen Agrobacterium tumefaciens. RNA Biol **9**.

259. **Wirth, T., F. Hildebrand, C. Allix-Beguec, F. Wolbeling, T. Kubica, K. Kremer, D. van Soolingen, S. Rusch-Gerdes, C. Locht, S. Brisse, A. Meyer, P. Supply, and S. Niemann.** 2008. Origin, spread and demography of the Mycobacterium tuberculosis complex. PLoS Pathog **4:**e1000160.

260. **Wolf, Y. I., I. B. Rogozin, N. V. Grishin, and E. V. Koonin.** 2002. Genome trees and the tree of life. Trends Genet **18:**472-9.

261. **Wray, C.** 1975. Survival and spread of pathogenic bacteria of veterinary importance within the environment. Vet Bull **45:**543-550.

262. **Wu, R., and A. D. Kaiser.** 1968. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. J Mol Biol **35:**523-37.

263. **Yamada, K., J. Lim, J. M. Dale, H. Chen, P. Shinn, C. J. Palm, A. M. Southwick, H. C. Wu, C. Kim, M. Nguyen, P. Pham, R. Cheuk, G. Karlin-Newmann, S. X. Liu, B. Lam, H. Sakano, T. Wu, G. Yu, M. Miranda, H. L. Quach, M. Tripp, C. H. Chang, J. M. Lee, M. Toriumi, M. M. Chan, C. C. Tang, C. S. Onodera, J. M. Deng, K. Akiyama, Y. Ansari, T. Arakawa, J. Banh, F. Banno, L. Bowser, S. Brooks, P. Carninci, Q. Chao, N. Choy, A. Enju, A. D. Goldsmith, M. Gurjal, N. F. Hansen, Y. Hayashizaki, C. Johnson-Hopson, V. W. Hsuan, K. Iida, M. Karnes, S. Khan, E. Koesema, J. Ishida, P. X. Jiang, T. Jones, J. Kawai, A. Kamiya, C. Meyers, M. Nakajima, M. Narusaka, M. Seki, T. Sakurai, M. Satou, R. Tamse, M. Vaysberg, E. K. Wallender, C. Wong, Y. Yamamura, S. Yuan, K. Shinozaki, R. W. Davis, A. Theologis, and J. R. Ecker.** 2003. Empirical analysis of transcriptional activity in the Arabidopsis genome. Science **302:**842-6.

264. **Yamamoto, T., T. Takano, W. Higuchi, Y. Iwao, O. Singur, I. Reva, Y. Otsuka, T. Nakayashiki, H. Mori, G. Reva, V. Kuznetsov, and V. Potapov.** Comparative genomics and drug resistance of a geographic variant of ST239 methicillin-resistant Staphylococcus aureus emerged in Russia. PLoS One **7:**e29187.

265. **Yesilkaya, H., F. Meacci, S. Niemann, D. Hillemann, S. Rusch-Gerdes, M. R. Barer, P. W. Andrew, and M. R. Oggioni.** 2006. Evaluation of molecular-Beacon, TaqMan, and fluorescence resonance energy transfer probes for detection of antibiotic resistance-conferring single nucleotide polymorphisms in mixed Mycobacterium tuberculosis DNA extracts. J Clin Microbiol **44:**3826-9.

266. **Yi, H., Y. J. Cho, S. H. Yoon, S. C. Park, and J. Chun.** Comparative genomics of Neisseria weaveri clarifies the taxonomy of this species and identifies genetic determinants that may be associated with virulence. FEMS Microbiol Lett **328:**100-5.

267. **Zarate-Blades, C. R., C. L. Silva, and G. A. Passos.** The impact of transcriptomics on the fight against tuberculosis: focus on biomarkers, BCG vaccination, and immunotherapy. Clin Dev Immunol **2011:**192630.

268. **Zerbino, D. R., and E. Birney.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **18:**821-9.

269. **Zhang, J., R. Chiodini, A. Badr, and G. Zhang.** The impact of next-generation sequencing on genomics. J Genet Genomics **38:**95-109.

270. **Zhu, X., Z. J. Tu, P. M. Coussens, V. Kapur, H. Janagama, S. Naser, and S. Sreevatsan.** 2008. Transcriptional analysis of diverse strains Mycobacterium avium subspecies paratuberculosis in primary bovine monocyte derived macrophages. Microbes Infect **10:**1274-82.