Tracing the Evolutionary Histories of Leprosy and Tuberculosis

using Ancient DNA and Phylogenomics Methods

by

Tanvi Prasad Honap

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved June 2017 by the
Graduate Supervisory Committee:

Anne C. Stone, Co-Chair
Michael S. Rosenberg, Co-Chair
Josephine E. Clark-Curtiss
Johannes Krause

ARIZONA STATE UNIVERSITY

August 2017

ABSTRACT

Leprosy and tuberculosis are age-old diseases that have tormented mankind and left behind a legacy of fear, mutilation, and social stigmatization. Today, leprosy is considered a Neglected Tropical Disease due to its high prevalence in developing countries, while tuberculosis is highly endemic in developing countries and rapidly re-emerging in several developed countries. In order to eradicate these diseases effectively, it is necessary to understand how they first originated in humans and whether they are prevalent in nonhuman hosts which can serve as a source of zoonotic transmission. This dissertation uses a phylogenomics approach to elucidate the evolutionary histories of the pathogens that cause leprosy and tuberculosis, *Mycobacterium leprae* and the *M. tuberculosis* complex, respectively, through three related studies. In the first study, genomes of *M. leprae* strains that infect nonhuman primates were sequenced and compared to human *M. leprae* strains to determine their genetic relationships. This study assesses whether nonhuman primates serve as a reservoir for *M. leprae* and whether there is potential for transmission of *M. leprae* between humans and nonhuman primates. In the second study, the genome of *M. lepraemurium* (which causes leprosy in mice, rats, and cats) was sequenced to clarify its genetic relationship to *M. leprae* and other mycobacterial species. This study is the first to sequence the *M. lepraemurium* genome and also describes genes that may be important for virulence in this pathogen. In the third study, an ancient DNA approach was used to recover *M. tuberculosis* genomes from human skeletal remains from the North American archaeological record. This study informs us about the types of *M. tuberculosis* strains present in post-contact era North America. Overall, this dissertation informs us about the evolutionary histories of these

pathogens and their prevalence in nonhuman hosts, which is not only important in an anthropological context but also has significant implications for disease eradication and wildlife conservation.

ACKNOWLEDGMENTS

This dissertation would not have been possible without the unwavering support and mentorship of my committee chair, Dr. Anne C. Stone. Words cannot express how much she has come to mean to me over the past five years. She has been a wonderful advisor and role model, and I will be forever grateful that she accepted me into her laboratory, supported my research endeavors, and made me part of her academic family.

I also thank Dr. Michael Rosenberg, firstly, for all his guidance as Chair of the Evolutionary Biology Ph.D. program, and secondly, for his help and encouragement as my co-advisor. I have always enjoyed our long discussions on bioinformatics, phylogenetics, and scientific writing. I thank Dr. Josephine Clark-Curtiss for her support and insights as a microbiologist on my research and Dr. Johannes Krause for his input on my dissertation and for allowing me to visit the Max Planck Institute for the Science of Human History, Jena, which was a truly enriching experience.

I am grateful to the Graduate and Professional Students Association, ASU, for partly funding my dissertation research. I thank the School of Life Sciences for funding my graduate studies through Teaching Assistantships and the Dissertation Completion Fellowship.

I would like to acknowledge the efforts of numerous researchers involved in the Ancient Tuberculosis Project. I thank Dr. Jane Buikstra for her insights as a bioarchaeologist, Dr. Kirsten Bos and Dr. Alexander Herbig for helpful discussions regarding laboratory and bioinformatics techniques, and Åshild Vågene for being an amazing research partner on this project and hosting me during my time in Jena. Most importantly, I thank the tribes and corporations involved in this project for allowing us

access to the samples, as well as the museums and researchers involved in the sample collection process. I also thank Dr. Koichi Suzuki, Dr. David Smith, Dr. Ross Tarara, and Dr. Oscar Rojas-Espinosa for providing samples for the Leprosy Project and Dr. Andrej Benjak for his bioinformatics advice and guidance.

I am indebted to a number of people who have supported me throughout my years at ASU. I thank Dr. Melissa Wilson-Sayres for being a wonderful colleague, mentor, and role model. I thank Scott Bingham, Jason Steele, and Katherine Skerry for their help with research experiments and Wendi Simonson, Yvonne Delgado, Teresa Plaskett, and Tae O'Connor for their administrative help.

I am eternally grateful to my lab-mates and academic sisters, Maria Nieves-Colon and Genevieve Housman, for taking me under their wing when I came to ASU. I thank them for all their help with lab-work, bioinformatics, and dealing with the highs and lows of grad school, in particular, and life, in general. I thank Andrew Ozga for his friendship, help, and advice during the past two years; Kelly 'Fife' Harkins and Joanna Malukiewicz for training me in the lab and giving me invaluable advice on how to survive grad school; Halszka Glowacka, Hallie Edmonds, and Susanne Daly for providing constructive feedback on my research; Andreina Castillo, Arpan Deb, and Viraj Damle for their friendship and support as we navigated through our graduate careers at ASU. I also thank my life-long support system of Pranjali Ganoo, Gayatri Marathe, Amruta Pradhan, Viren Kalsekar, Gauri Desai, and Amruta Saraf for ensuring that I have a social circle outside of grad school.

I would not have reached this stage in my life without the support of my teachers at Abhinava Vidyalaya English Medium School, my professors at Abasaheb Garware

College and National Institute of Virology, India, my undergraduate research advisor, Dr. Neelima Deshpande, and my Masters advisor, Dr. K. Alagarasu.

I am exceedingly grateful to my extended family for supporting me throughout my life and career. I thank Rajesh, Rupa, Asmita, and Rujuta Idate for providing me with a home-away-from-home where I could spend my Thanksgivings and Christmases and return to grad school fully rejuvenated; Bhagyashree Barlingay for helping me in numerous ways during my time at ASU; my grandparents for their constant love and support, and specifically, my *Aajoba* for teaching me how to use a computer.

I would like to acknowledge the three most important people in my life for shaping me into the person I am today. I thank my parents, Prasad and Deepa Honap, for their irrevocable love and support throughout my life. I will be eternally grateful they accepted that I wanted to take the road less travelled and stood by me through thick and thin. I thank Manasi Tamhankar for steadfastly supporting me throughout the years and for our nightly conversations, without which I would not have survived grad school. Our invigorating discussions about science, history, music, and our favorite TV shows have broadened my thinking and made me a well-rounded (and more interesting) person.

Lastly, this dissertation would not have been possible without the copious amounts of coffee I consumed during the past five years and therefore, I thank Starbucks for keeping me well-caffeinated during grad school.

I apologize if I have forgotten to acknowledge anyone, but please know that I am whole-heartedly grateful for your support.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

Leprosy and tuberculosis (TB) are among the oldest known human diseases and yet, they remain a public health concern even today. The causative agent of leprosy, *Mycobacterium leprae*, was discovered by Gerhard Hansen in 1874. The advent of multi-drug therapy comprising dapsone, rifampicin, and clofazimine in the 1980s resulted in almost 16 million people being cured of leprosy by the year 2000. The global prevalence of leprosy has now been reduced to less than one case per 10,000 individuals and the disease has been nearly eradicated from the developed countries of the world. However, 200,000-250,000 new leprosy cases occur worldwide every year (WHO 2016a). Leprosy remains highly endemic in several developing countries, including India, Brazil, Madagascar, the Philippines, and the Central African Republic, and thus, it is now classified as a Neglected Tropical Disease.

The primary causative agent of TB, *Mycobacterium tuberculosis*, was discovered by Robert Koch in 1882. The invention of anti-tuberculosis drugs, such as isoniazid and rifampin, in the mid-20th century led to a rapid decrease in the number of TB cases by the 1980s. However, hopes of eradicating TB were dashed due to the rise of antibiotic-resistant *M. tuberculosis* strains and HIV-AIDS in the late 1980s. Today, factors such as a declining standard of living among lower socio-economic classes, co-morbidity among HIV-positive individuals, and development of multi- and extremely-drug resistant *M. tuberculosis* strains have been implicated in the re-emergence of TB in developed countries. Additionally, TB is highly endemic in economically developing countries such

as India, Indonesia, China, Nigeria, Pakistan and South Africa. In 2015, there were an estimated 10.4 million new TB cases worldwide, resulting in nearly 1.4 million deaths (WHO 2016b).

The measures implemented to control TB and leprosy, however, target only human cases of the disease. TB control programs in high-endemicity countries have focused on preventing incidence by using the *Mycobacterium bovis* Bacille de Calmette et Guérin (BCG) vaccine. Low-endemicity countries, such as the US, rely on early detection of TB cases and antibiotic therapy in order to combat the disease. In case of leprosy, an effective vaccine is not available, and the focus of leprosy control programs has been early detection followed by multi-drug therapy. Since neither disease has been successfully controlled despite dedicated efforts, it is necessary to explore other factors that might contribute to their continued prevalence among humans.

The countries in which TB and leprosy are highly endemic are also rich in wildlife. Human population growth has led to increased encroachment of wildlife habitats and close contact with wild animals, especially other primates. The close evolutionary relationship among different primate species increases the ease of pathogen transmission among them (Pedersen and Davies 2009). Direct exploitation of primates through the use of primates as pets, performing monkeys, for bush meat, or via interactions in zoos and sanctuaries are major sources of infectious disease transmission between humans and nonhuman primates (Wolfe et al. 2005; Wolfe et al. 1998; Wallis and Lee 1999). Nonhuman primates are highly susceptible to pathogens such as the simian immunodeficiency virus (SIV), Ebola virus, and Bacillus cereus biovar Anthracis (Calvignac-Spencer et al. 2012). Therefore, it is important to understand which pathogens

are carried by wild nonhuman primates and which of these can be transmitted to humans. Conversely, diseases can also be transmitted from humans to nonhuman primates, which can lead to a decline in nonhuman primate populations and hamper conservation efforts (Leroy et al. 2004). The presence of pathogens such as *M. tuberculosis* and *M. leprae* in nonhuman primates could explain their continued persistence among human populations due to zoonotic transmission. Additionally, these pathogens might be present in other wildlife species that could serve as reservoirs for the pathogens and/or be a source of zoonotic transmission.

Mycobacterium is a genus of phylum Actinobacteria comprising aerobic, non-sporulating bacteria that are characterized by the presence of mycolic acids in their cell envelopes. There are more than 150 recognized species in the genus, broadly divided into rapid-growing and slow-growing mycobacteria. Certain members of this genus such as the *M. tuberculosis* complex (MTBC), *M. avium* complex (MAC), and *M. leprae* are important human pathogens, and thus, their genomes are well-studied. However, infections caused by non-tuberculous mycobacteria (NTM) are reportedly increasing (Yeung et al. 2016; Tortoli 2014) and hence, recent studies have attempted to clarify the phylogenetic relationships of the NTM (Fedrizzi et al. 2017; Mignard and Flandrois 2008; Devulder, Pérouse de Montclos, and Flandrois 2005). Despite this, certain members of this genus remain uncharacterized and their evolutionary relationships within the genus are unclear.

1.2 Using a Phylogenomics Approach to Study Mycobacterial Evolution

The gene encoding 16S rRNA has been widely used for the detection of relationships among bacterial species, but may not be able to resolve relationships accurately when the species being studied show genetic identities between 94 and 100% (Enrico Tortoli 2003; Zeigler 2003; Mignard and Flandrois 2008). In general, phylogenies based on the sequences of single genes may provide insufficient resolution or support incorrect topologies due to factors such as insufficient number of characters used in the analysis, horizontal gene transfer, unrecognized paralogy, and highly variable rates of evolution (Snel, Bork, and Huynen 1999; Gontcharov, Marin, and Melkonian 2004; Charles et al. 2005). Multi-gene phylogenies based on concatenated gene sequences can improve resolution (Hillis 1996) but these topologies may still be affected by factors such as long branch attraction, and internal branches may not be resolved (Sanderson and Shaffer 2002; Gontcharov, Marin, and Melkonian 2004). To this end, whole-genome phylogenies are better equipped to produce a resolved species tree with robust support (Rokas et al. 2003); however, they require increased computational resources.

When closely related species are being studied, single-nucleotide polymorphisms (SNPs) are good candidate markers for phylogenetics because they span the entire genome including intergenic regions and show relative stability over evolutionary time (Brumfield et al. 2003; Morin et al. 2004). SNPs analyzed across entire genomes usually provide sufficient characters for phylogenetic reconstructions to resolve problems associated with character state conflict and create topologies with fine-scale resolution (Foster et al. 2009). However, compared to phylogenies based on numerous concatenated

genes, SNP-based phylogenies comprise fewer characters and hence require less computational time. Monomorphic and clonally evolving pathogens such as the MTBC, MAC, and *M. leprae* show very high genetic identity between strains (> 99%). Traditional genotyping techniques such as Mycobacterial Interspersed Repetitive Units - Variable Number of Tandem Repeats (MIRU-VNTR) genotyping cannot be used to accurately reflect phylogenetic relationships because these repeat sequences are subject to homoplasy. Thus, strains with identical MIRU-VNTR profiles may not actually be closely related (Bryant et al. 2016; Anderson et al. 2013; Kay et al. 2015). On the other hand, SNP homoplasies are extremely rare and hence, SNPs are the ideal phylogenetic markers for analyzing the evolutionary relationships between these pathogens (Comas et al. 2009). Therefore, in this dissertation, SNPs across whole-genomes are used for phylogenetic reconstruction so as to study the evolutionary relationships between different mycobacterial species.

1.3 Outstanding Issues in the Evolutionary History of Leprosy

In humans, leprosy is mainly caused by the obligate intracellular bacterium, *M. leprae*. Certain cases of leprosy are also caused by a newly discovered and closely related species, *M. lepromatosis* (Han et al. 2008). *M. leprae* and *M. lepromatosis* are estimated to have diverged 13 - 14 million years ago (Singh et al. 2015) but share a number of characteristics such as an obligate intracellular parasitic lifestyle and reduced genome sizes relative to other mycobacteria.

Apart from humans, *M. leprae* naturally infects armadillos (Truman et al. 2011; Walsh, Meyers, and Binford 1986) as well as certain nonhuman primates (Donham and

Leininger 1977; Gormus et al. 1991; Suzuki et al. 2010; Meyers et al. 1985; Gormus et al. 1988; Valverde et al. 1998). Recently, both *M. leprae* and *M. lepromatosis* have been found to infect red squirrels (Avanzi et al. 2016). However, armadillos and red squirrels are known to have originally acquired *M. leprae* due to anthroponotic transmission from humans (Monot et al. 2005; Avanzi et al. 2016). It remains unknown how and where *M. leprae* was originally introduced to humans, although it is presumed that it was introduced from a hitherto unknown animal host. The most phylogenetically basal *M. leprae* strains are found in Asia (Schuenemann et al. 2013) and the oldest skeletal evidence for leprosy is attributed to 2000 BCE India (Robbins et al. 2009), suggesting that *M. leprae* was introduced to humans in this continent. Figure 1 shows the phylogenetic relationships for the human (modern and ancient), armadillo, and red squirrel *M. leprae* strains. Using Bayesian dating analyses, it has been estimated that all *M. leprae* strains shared a common ancestor less than 5000 years ago (Schuenemann et al. 2013). Thus, it is likely that *M. leprae* jumped from another host species into humans somewhere in Asia within the past 5000 years. The continued incidence of leprosy cases, especially the higher incidence in Asian countries, could be due to the presence of unknown hosts which are capable to introducing the pathogen to humans.

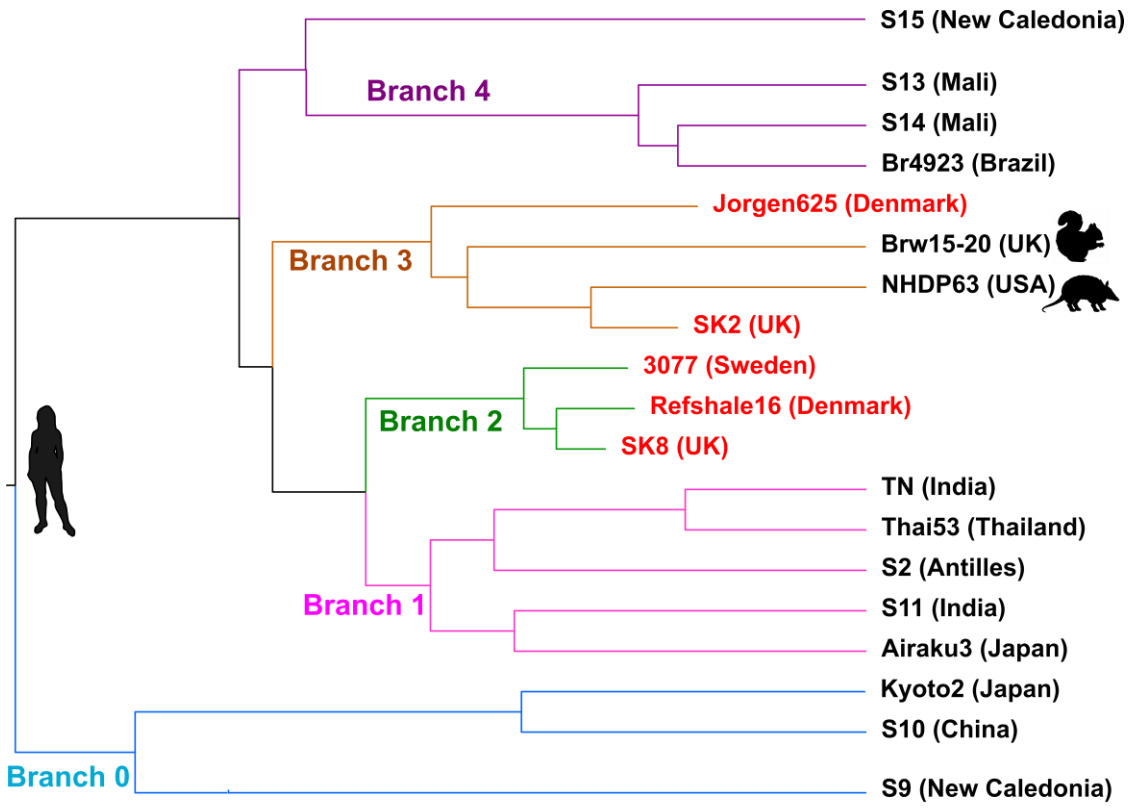

Figure 1. Phylogenetic Representation of *M. leprae* Strains. The branches are not drawn to scale. Geographic origin of the strain is given in parentheses next to its name. The five *M. leprae* branches are highlighted in different colors. The ancient human *M. leprae* strains are denoted in red. Strain Brw15-20 represents the *M. leprae* clade found in red squirrels in the UK and strain NHDP63 represents the *M. leprae* clade found in armadillos.

The recent finding that red squirrels carry *M. leprae* and *M. lepromatosis* suggests that other rodent species might be a reservoir for leprosy-causing pathogens. In rodents such as mice and rats, leprosy is caused by a different bacterial species, *M. lepraemurium* (see Rojas-Espinosa and Lovik 2001). *M. lepraemurium* also causes leprosy-like illness in cats (Hughes et al. 2004; Malik et al. 2002). However, this pathogen does not infect

humans. DNA hybridization studies suggest that *M. lepraemurium* is closely related to *M. avium* (Athwal, Deo, and Imaeda 1984) but since the genome of *M. lepraemurium* had not been sequenced, its phylogenetic placement within the genus *Mycobacterium* remained unclear.

Chapters 2 and 3 of this dissertation focus on clarifying some of the aforementioned outstanding issues in the evolutionary relationships of leprosy-causing pathogens. In Chapter 2, the genomes of *M. leprae* strains from three naturally infected nonhuman primates are sequenced. Phylogenetic analyses are used to ascertain whether nonhuman primates carry novel *M. leprae* lineages or whether they are infected by *M. leprae* strains closely related to those found in humans in these regions. Furthermore, wild nonhuman primate populations including ring-tailed lemurs from the Beza Mahafaly Special Reserve, Madagascar, and chimpanzees from Ngogo, Kibale National Park, Uganda, were screened for presence of *M. leprae* or MTBC infection. In Chapter 3, results from sequencing the genome of *M. lepraemurium* are reported. Phylogenetic analyses are conducted with the aim of clarifying the position of this species in the mycobacterial phylogeny and genes that are likely related to virulence in this species are discussed.

1.4 Outstanding Issues in the Evolutionary History of TB

TB is caused by members of the MTBC which comprises human-adapted species such as *M. tuberculosis* and *M. africanum*, animal-adapted species such as *M. microti* (voles), *M. caprae* (goats), *M. pinnipedii* (seals, sea lions), *M. bovis* (cattle), *M. orygis* (oryx), *M. mungi* (African mongooses), *M. suricattae* (meerkats), and the Dassie bacillus

(rock hyraxes), as well as *M. canettii* whose exact host range is unknown. Despite being adapted to specific hosts, members of the MTBC are capable of infecting other host species.

Genetic analyses of global MTBC strains show that the greatest diversity of strains as well as the phylogenetically basal lineages are found in Africa (Gagneux and Small 2007), suggesting that the MTBC might have evolved in this continent (Comas et al. 2010; Comas et al. 2013; Wirth et al. 2008). A recent study reconstructed ancient MTBC genomes from three pre-European contact era Peruvian individuals, and using the corresponding radiocarbon dates of the skeletal samples as calibration points, estimated that the MTBC was introduced to humans within the last 6,000 years (Bos et al. 2014). MTBC strains spread across Africa and to Europe and Asia with human population movements and diversified into seven human-adapted *M. tuberculosis* lineages that are phylogeographically associated (Gagneux et al. 2006; Firdessa et al. 2013). On the other hand, the introduction of human *M. tuberculosis* strains into animals led to the evolution of the animal-adapted lineages (Brosch et al. 2002; Hershberg et al. 2008). Figure 2 shows the phylogenetic relationships between the members of the MTBC.

The recovery of genomes of three pre-contact era MTBC strains from coastal Peru provided evidence that sometime within the past 1,200 years, *M. pinnipedii* strains were introduced to human populations living along the coast due to consumption or handling of infected seals (Bos et al. 2014). It is not known whether the seal-derived MTBC strains adapted to humans and spread to the non-coastal parts of the Americas by human-to-human transmission. Additionally, pre-contact era MTBC genomes from North America have not been recovered.

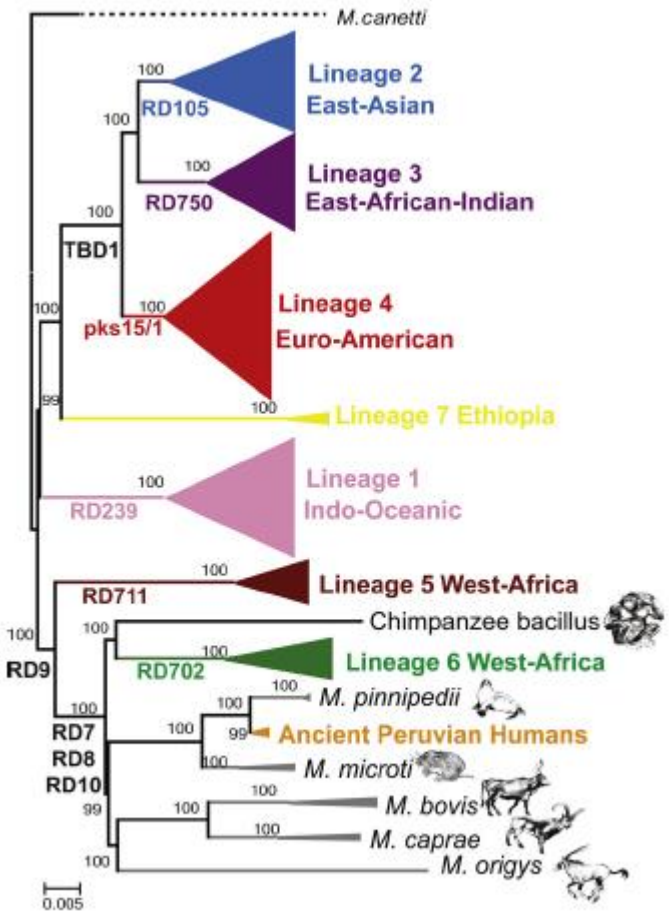

Figure 2. Phylogenetic Representation of MTBC Species. The figure was adapted from Coscolla and Gagneux (2014) and depicts a Maximum Likelihood tree modified from Bos et al. (2014). Bootstrap support estimated from 1000 replications is shown on the branches. The tree is rooted using *M. canettii*. Large Sequence Polymorphisms (LSPs) described in Brosch et al. (2002) are indicated along the branches. The scale bar indicates the number of nucleotide substitutions per site.

Therefore, it remains to be determined whether 1) pre-contact TB in this region was caused by the northward dispersal of the seal-derived MTBC strains, 2) there were other MTBC lineages present in this region, such as Asian *M. tuberculosis* strains which

may have been introduced via population movements over the Bering Strait, or 3) pre-contact TB in the North Americas was caused by an altogether different pathogen, such as *M. kansasii* which also causes clinical tuberculosis (Evans et al. 1996). Currently, the majority of the *M. tuberculosis* strains found in the Americas are of European origin (Hershberg et al. 2008; Comas et al. 2013), suggesting that pre-contact era MTBC lineages were replaced following the Age of Exploration.

Understanding which lineages of MTBC strains were present in the pre-contact New World as well as how and when they came to be replaced by European strains is important not only in an anthropological context but will also inform us about potential avenues of TB transmission in the past that may be relevant even today. Furthermore, analyzing these genome data may help us identify mutations which allow a particular strain to cross the species barrier and/or adapt to new hosts.

Chapter 4 attempts to clarify some of these outstanding questions about the origins of TB in North America by screening 66 individuals from the archaeological record for the presence of MTBC DNA. Five post-contact era *M. tuberculosis* genomes are analyzed so as to ascertain what types of strains were circulating in North America during this time.

1.5 Summary

Overall, this dissertation examines the evolutionary history of these important mycobacterial pathogens by focusing on the types of strains found in human and nonhuman hosts. This dissertation aims to elucidate these phylogenetic relationships to identify the potential for anthroponotic or zoonotic transmission. Tracing the genetic

changes that have occurred as mycobacterial pathogens cross from humans to other hosts (or vice versa) will allow us to determine whether there are clear requirements for successful cross-species transmissions and assess future zoonotic risk.

CHAPTER 2

MYCOBACTERIUM LEPRAE GENOMES FROM NATURALLY INFECTED NONHUMAN PRIMATES

2.1 Abstract

Leprosy is caused by the bacterial pathogens *Mycobacterium leprae* and *M. lepromatosis*. Apart from humans, animals such as nine-banded armadillos in the New World and red squirrels in the British Isles serve as reservoirs for leprosy. Natural leprosy has also been reported in certain nonhuman primates, but it is not known whether these occurrences are mainly due to incidental infections from humans or if host-adapted lineages of leprosy-causing pathogens exist in nonhuman primates. In this study, *M. leprae* genomes from three naturally infected nonhuman primates (a chimpanzee from Sierra Leone, a sooty mangabey from West Africa, and a cynomolgus macaque from The Philippines) were sequenced. Phylogenetic analyses show that the cynomolgus macaque *M. leprae* strain is most closely related to a human *M. leprae* strain from New Caledonia. The chimpanzee and sooty mangabey *M. leprae* strains form a new sublineage within a human *M. leprae* lineage found in West Africa. The close relationship of these two strains suggests that different nonhuman primate species may transmit *M. leprae* among themselves in the wild. Furthermore, this study aimed to assess the prevalence of *M. leprae* and the *M. tuberculosis* complex in wild nonhuman primates from countries where leprosy and/or tuberculosis are endemic. Samples were collected from ring-tailed lemurs from the Beza Mahafaly Special Reserve, Madagascar, and chimpanzees from Ngogo, Kibale National Park, Uganda, and screened using quantitative PCR assays. While the populations tested in this study did not show presence of mycobacterial pathogens,

nonhuman primates should be screened to assess the capacity for anthroponotic transmission of mycobacterial diseases in endemic areas.

2.2 Introduction

Leprosy has afflicted mankind for many millennia and remains a highly prevalent disease in economically underprivileged countries. Due to effective multi-drug therapy, the global prevalence of leprosy has been reduced to less than one case per 10,000 individuals (WHO 2016a). The disease has been almost eradicated from developed countries; however, approximately 250,000 new leprosy cases occur each year, making leprosy a Neglected Tropical Disease (WHO 2016a).

Leprosy affects the skin, mucosa of the nose and upper respiratory tract, and the peripheral nervous system. Depending upon the host's immune response, the infection can progress to either tuberculoid (paucibacillary) or lepromatous (multibacillary) leprosy. Tuberculoid leprosy is characterized by the presence of one or few hypopigmented patches with loss of sensation and thickened peripheral nerves, whereas in lepromatous leprosy, systemic lesions are seen. These lesions may become infiltrated with fluids, causing severe distortions of those parts of the body where the lesions are located, such as on the face and ears. If left untreated, it can cause permanent nerve damage, and secondary infections can lead to tissue loss resulting in disfigurement of the extremities (Britton and Lockwood 2004). The disease has a long incubation period that averages three to five years and can extend up to thirty years, which hampers early detection of cases.

In humans, leprosy is caused by the bacterial pathogens, *M. leprae* and *M. lepromatosis*, the latter of which causes a severe form of the disease called diffuse lepromatous leprosy (Gelber 2005; Vargas-Ocampo 2007). While *M. leprae* causes the majority of leprosy cases and is prevalent worldwide, *M. lepromatosis* is mainly endemic to Mexico and the Caribbean (Han et al. 2008; Han et al. 2009; Vera-Cabrera et al. 2011) although isolated cases have been reported from other countries (Han et al. 2012). *M. leprae* and *M. lepromatosis* show 88% genetic identity and are estimated to have diverged 13-14 million years ago (MYA) (Singh et al. 2015). Despite this deep divergence, they share a number of characteristics such as a reduced overall genome size (relative to other mycobacteria) of approximately 3.2 million base pair (bp), genome organization, and the inability to grow outside of a living host. This obligate parasitism is the result of a reductive evolution that occurred about 12 - 20 MYA and led to the loss of functionality of a number of genes in both *M. leprae* and *M. lepromatosis* (Gómez-Valero et al. 2007; Singh et al. 2015).

Traditionally thought to be an exclusively human pathogen, *M. leprae* has been found to naturally infect other animals. Armadillos are the only confirmed animal reservoir of *M. leprae* in the New World (Walsh, Meyers, and Binford 1986) and originally acquired the pathogen from Europeans during the Era of Exploration (Monot et al. 2005). In the southeastern US, armadillos are involved in zoonotic transmission of *M. leprae* due to human contact with infected armadillos or consumption of armadillo meat (Truman et al. 2011). Recently, red squirrel populations in the UK were found to be infected with both *M. leprae* as well as *M. lepromatosis* (Avanzi et al. 2016). The near eradication of leprosy from the human population in the UK as well as the phylogenetic

placement of the contemporary red squirrel *M. leprae* strains suggests that the squirrels were infected by human *M. leprae* strains circulating in medieval Europe before the decline of leprosy in Europe (Avanzi et al. 2016).

Apart from armadillos and red squirrels, isolated cases of naturally occurring leprosy have been observed in nonhuman primates such as chimpanzees (Donham and Leininger 1977; Leininger, Donham, and Rubino 1978; Hubbard et al. 1991; Gormus et al. 1991; Suzuki et al. 2010), sooty mangabeys (Meyers et al. 1985; Gormus et al. 1988), and cynomolgus macaques (Valverde et al. 1998). In all of these cases, the nonhuman primates were captured from the wild and imported to research facilities for experimental purposes. The animals were not experimentally infected with *M. leprae* nor did they come into contact with a known leprosy patient. All animals developed symptoms characteristic of human leprosy with varying incubation periods, and in most cases, the aetiological agent was confirmed to be *M. leprae* using microscopic or genetic analyses. However, the genomes of these nonhuman primate *M. leprae* strains had never been sequenced until now.

M. leprae is a highly clonal organism and human *M. leprae* strains show more than 99.9% genetic identity (Monot et al. 2009). Whole-genome sequencing approaches have classified *M. leprae* strains into five branches (Schuenemann et al. 2013). The most deeply diverged *M. leprae* branch contains strains from Japan (Kai et al. 2013), China, and New Caledonia (Schuenemann et al. 2013), suggesting that leprosy may have originated in Asia. To determine the relationships between nonhuman primate and human *M. leprae* strains, *M. leprae* genomes from three naturally infected nonhuman primates – a chimpanzee (Suzuki et al. 2010), a sooty mangabey (Meyers et al. 1985), and a

cynomolgus macaque (Valverde et al. 1998) were sequenced. The details regarding these three animals are given in Appendix A: Table S1.

Additionally, this study aimed to assess whether *M. leprae* and other mycobacterial pathogens are prevalent in wild nonhuman primates living in contact with human populations. Ring-tailed lemur populations from the Beza Mahafaly Special Reserve (BMSR), Madagascar, and chimpanzee populations from Ngogo, Kibale National Park, Uganda, were screened for the presence of mycobacterial infection using quantitative PCR (qPCR) assays.

2.3 Materials and Methods

Sequencing the genomes of nonhuman primate *M. leprae* strains

Sampling

A sample of genomic DNA extracted from the skin biopsy of a naturally infected female chimpanzee (*Pan troglodytes verus*) was provided by Dr. Koichi Suzuki. The *M. leprae* strain from a naturally infected sooty mangabey (*Cercocebus atys*) had been isolated by passaging in an armadillo. *M. leprae* DNA was extracted from a sample of the infected-armadillo tissue using the protocol given in Clark-Curtiss et al. (1985) and an aliquot of this DNA extract was provided by Dr. Josephine Clark-Curtiss. A sample of skin biopsy tissue stored in a formalin-fixed paraffin-embedded (FFPE) form since 1994 from a naturally infected cynomolgus macaque (*Macaca fascicularis*) was provided by Dr. David Smith and Dr. Ross Tarara. The cynomolgus macaque had been acquired from AMO Farm in The Philippines in 1990 (CITES permit 4455).

Hereafter, the chimpanzee, sooty mangabey, and cynomolgus macaque samples used in this study will be referred to as Ch4, SM1, and CM1, respectively.

DNA extraction

DNA was extracted from the CM1 tissue sample using the DNeasy Blood and Tissue Kit (Qiagen). 0.5 g of tissue was used as starting material and extraction was carried out using the manufacturer's protocol with the following modification: DNA was eluted in 100 µL AE buffer (Qiagen) that had been preheated to 65°C. The DNA extract was tested for the presence of *M. leprae* DNA using a qPCR assay targeting the *M. leprae*-specific multi-copy *RLEP* element (Truman et al. 2008).

M. leprae genome sequencing

The SM1 *M. leprae* DNA sample was converted into a paired-end fragment library and sequenced using the 454 GS-FLX Titanium sequencer (½ 70 × 75 PicoTiterPlate GS XLR70 Run) at SeqWright DNA Technology Services, TX. The Ch4 and CM1 DNA extracts were sheared to an average bp size of 300 using the M220 Focused-ultrasonicator (Covaris) and converted into double-indexed DNA libraries using a library preparation protocol modified from (Meyer and Kircher 2010). For sample CM1, two separate libraries were prepared (CM1 Library1 and CM1 Library2). Libraries were quantified using the Bioanalyzer 2100 DNA1000 assay (Agilent) and the KAPA Library Quantification kit (Kapa Biosystems).

To increase coverage, the Ch4 and CM1 libraries were target-enriched for the *M. leprae* genome using a custom MYbaits Whole Genome Enrichment kit (MYcroarray).

Specifically, biotinylated RNA baits were prepared using DNA from *M. leprae* Br4923, Thai53, and NHDP strains. 57 ng for the CH4 library, 467 ng for CM1 Library1, and 910 ng for CM1 Library2 were used for enrichment. Each library was enriched separately. Enrichment was conducted according to the MYbaits protocol with hybridization being carried out at 65°C for 24 hours. After elution, the CH4 library and the CM1 Library1 were amplified using AccuPrime *Pfx* DNA polymerase (Life Technologies) for 27 and 23 cycles, respectively, following the protocol given in Ozga et al. (2016). The enriched CM1 Library2 was amplified over two separate reactions each for 14 cycles using KAPA HiFi polymerase (Kapa Biosystems). All amplification reactions were cleaned up using the MinElute PCR Purification kit (Qiagen). Two library blank samples (PCR-grade water) were also processed into libraries and target-enriched in a similar manner to ensure that no contamination had been introduced during the process; these are referred to as LB1 and LB2. All samples (Ch4, CM1 Library1, CM1 Library2, LB1, and LB2) were sequenced over two sequencing runs on the Illumina HiSeq2500 using the Rapid PE v2 chemistry (2 ×100 bp) at the Yale Center for Genome Analysis, CT. These runs also included samples from other ongoing research projects.

Data Processing and Mapping

For sample SM1, the FASTA and QUAL files obtained from the sequencing facility were combined into a FASTQ file using the Combine FASTA and QUAL tool on the Galaxy server (https://usegalaxy.org). Reads were trimmed using AdapterRemoval v2 with default parameters (Schubert, Lindgreen, and Orlando 2016). For samples Ch4 and CM1, paired-end reads were trimmed and merged using SeqPrep

(https://github.com/jstjohn/SeqPrep) with the following modification: minimum overlap for merging = 11. Since sample CM1 had two separately sequenced libraries, paired-end reads for each library were trimmed and merged separately and then concatenated together.

For all samples including the library blanks, reads were mapped to the *M. leprae* TN genome (AL450380.1) using the Burrows Wheeler Aligner (bwa) v0.7.5 with default parameters (Li and Durbin 2009). SAMtools v0.1.19 (Li et al. 2009) was used to filter the mapped reads for a minimum Phred quality threshold of Q37 and remove PCR duplicates and reads with multiple mappings.

In order to determine the efficiency of the target enrichment, reads for samples Ch4 and CM1 were also mapped to the *Pan troglodytes* reference genome (CSAC Build 2.1.4; GCA_000001515.4) and the *Macaca fascicularis* reference genome (Washington University Macaca_fascicularis_5.0; GCA_000364345.1), respectively, using similar methodology as given above.

Comparative Data

Publicly-available Illumina reads for ancient (Jorgen625, Refshale16, SK2, SK8, and 3077) and modern (S2, S9, S10, S11, S13, S14, S15, and Airaku3) human *M. leprae* strains and a red squirrel *M. leprae* strain (Brw15-20m) were acquired from the Sequence Read Archive. Reads were processed and mapped using the same methodology as described earlier. FASTA files for the finished *M. leprae* genomes (Br4923, Kyoto2, NHDP63, and Thai53) were aligned to the *M. leprae* TN reference genome using LAST with default parameters (Kiełbasa et al. 2011). The maf-convert program was used to

covert the alignment file to a SAM file and SAMtools was used obtain a BAM file which was used for further analyses. Similarly, contigs for *M. lepromatosis* Mx1-22A (JRPY00000000.1) were aligned to the *M. leprae* TN genome using LAST with the gamma-centroid option as given in Singh et al. (2015).

Variant calling

For the BAM files obtained after processing genomes from the Illumina dataset and for the samples sequenced in this study, an mpileup file was generated using SAMtools and processed using VarScan v2.3.9 (Koboldt et al. 2012). A VCF file containing all sites was produced using the following parameters: minimum number of reads covering the position = 5, minimum of reads covering the variant allele = 3, minimum variant frequency = 0.2, minimum base quality = 30, and maximum frequency of reads on one strand = 90%. For the finished *M. leprae* genomes and *M. lepromatosis*, SAMtools (v1.3.1) mpileup and bcftools call were used to produce the VCF files. VCF files for all strains were combined using the CombineVariants tool available in the Genome Analysis Toolkit (GATK) (McKenna et al. 2010). VCFtools (Danecek et al. 2011) was used to remove insertion-deletions (InDels) and exclude positions which occurred in known repeat regions and rRNA and positions covered by the SK12 negative control sample (Schuenemann et al. 2013). The list of all positions excluded from the analyses is given in Appendix B. The SelectVariants tool in GATK was used to output a VCF file containing the sites comprising single nucleotide polymorphisms (SNPs). Positions with missing data (where one or more strains had an N) were excluded. SNP calls were manually checked for possible errors or inconsistencies with published data.

A perl script was used to generate an alignment comprising those positions where at least one of the strains had a SNP (Bergey 2012).

Phylogenetic analyses

Phylogenetic trees were constructed using the Neighbor-Joining (NJ) and Maximum Parsimony (MP) methods in MEGA7 (Kumar, Stecher, and Tamura 2016) as well as using a Bayesian approach in BEAST v1.8.4 (Drummond et al. 2012). The SNP alignment of all 22 *M. leprae* genomes and *M. lepromatosis* comprising 233,509 sites was used as input for MEGA7. The NJ tree was generated using the *p*-distance method. This method was used because the alignment did not contain invariant sites and the sequences are not deeply diverged; therefore, the *p*-distance method was considered to be the most appropriate for analyzing this dataset. Bootstrap support was estimated from 1000 replicates. The MP tree was generated using the Subtree-Pruning-Regrafting (SPR) algorithm and 1000 bootstrap replicates.

To determine divergence times of the *M. leprae* strains, a SNP alignment of all *M. leprae* strains was generated. Sites with missing data were removed, resulting in an alignment comprising 747 positions. *M. lepromatosis* was not included in this analysis. The modern human *M. leprae* strain S15 was also excluded because it contains an unusually high number of SNPs, likely related to its multi-drug resistance (Schuenemann et al. 2013). To assess whether there was a sufficient temporal signal in the data to proceed with molecular clock analysis, a regression of root-to-tip genetic distance against dates of the *M. leprae* strains was conducted using TempEst (Rambaut et al. 2016). Calibrated radiocarbon dates of the ancient *M. leprae* strains and isolation dates of the

modern *M. leprae* strains and the NJ tree generated earlier were used as input for TempEst. The R^2 value calculated in TempEst equaled 0.6212, signifying a positive correlation between genetic divergence and time for the *M. leprae* strains (Figure 3). Therefore, the data were found to be suitable for molecular clock analysis.

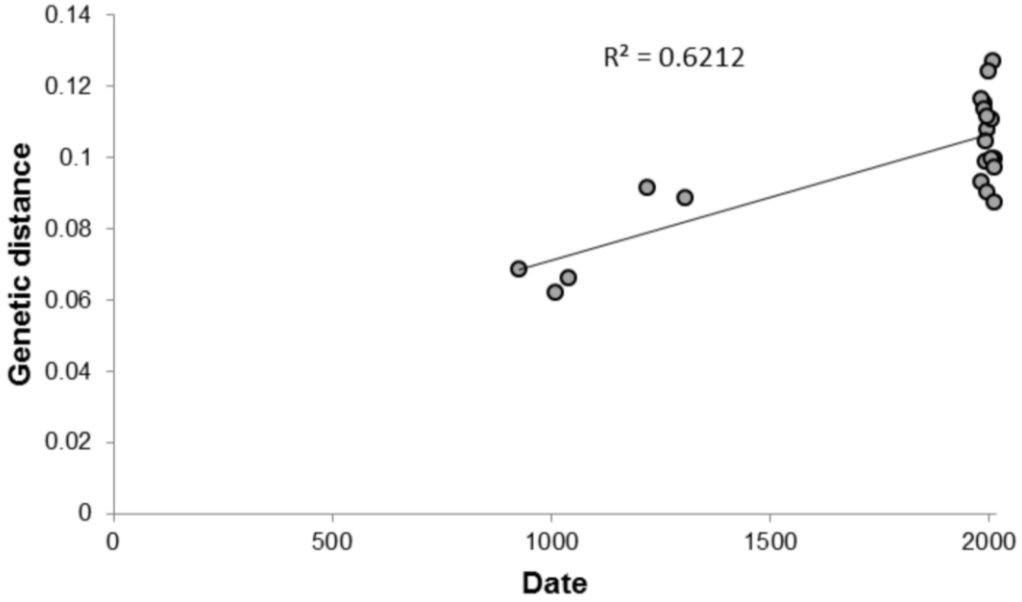

Figure 3. Scatter Plot of Date vs Genetic Distance of *M. leprae* Strains. The x-axis denotes mean date in CE (calibrated radiocarbon date for ancient strains and isolation year for modern strains). The y-axis denotes root-to-tip genetic distance for each strain.

The SNP alignment was analyzed using BEAST v1.8.4 (Drummond et al. 2012). The calibrated radiocarbon dates of the ancient strains in years before present (YBP, with present being considered as 2017), the isolation years of the modern strains, and a substitution rate of 6.87×10^{-9} substitutions per site per year as estimated by (Avanzi et al. 2016) were used as priors. Using jModelTest2 (Darriba et al. 2012), the Kimura 3-parameter model with unequal base frequencies was determined to be the best model of nucleotide substitution. A strict clock model with uniform rate across branches and a tree

model of constant population size were used. To account for ascertainment bias that might result from using only variable sites in the alignment, the number of invariant sites (number of constant As, Cs, Ts, and Gs) was included in the analysis. One Markov Chain Monte Carlo (MCMC) run was carried out with 50,000,000 iterations, sampling every 2,000 steps. The first 5,000,000 iterations were discarded as burn-in. Tracer (Rambaut et al. 2015) was used to visualize the results of the MCMC run. TreeAnnotator (Drummond et al. 2012) was used to summarize the information from the sample of trees produced onto a single target tree calculated by BEAST, with the first 2,500 trees being discarded as burn-in. Figtree (http://tree.bio.ed.ac.uk/software/figtree/) was used to visualize the Maximum Clade Credibility (MCC) tree.

SNP analysis

The VCF files for the Ch4, SM1, and CM1 samples were analyzed using snpEff v4.3 (Cingolani et al. 2012). The program was run using default parameters except that the parameter for reporting SNPs that are located upstream or downstream of protein-coding genes was set to 100 bases.

Screening wild nonhuman primates for presence of mycobacterial pathogens

Sampling

Buccal swab samples were collected from wild ring-tailed lemurs, *Lemur catta*, (n = 41) from BMSR, Madagascar, in the 2009 field season. Fruit wadge samples were collected from wild chimpanzees, *Pan troglodytes schweinfurthii*, (n = 22) from Ngogo, Kibale National Park, in the 2010 field season. Sampling was conducted according to

institutional and national guidelines. The lemur buccal swab samples were collected under CITES permit number 09US040035/9. A CITES permit was not required for collection of the chimpanzee fruit wadge samples.

DNA extractions

DNA was extracted from the buccal swab samples using a phenol-chloroform DNA extraction protocol (Sambrook and McLaughlin 2000) and from the fruit wadge samples using the DNeasy Plant Maxi Kit (Qiagen) and following the manufacturer's instructions. For each batch of DNA extractions, a negative control sample (extraction blank) was kept to ensure that no contamination was introduced during the DNA extraction process.

qPCR assays

All extracts as well as extraction blanks were tested for the presence of *M. leprae* DNA using two TaqMan qPCR assays – one targeting the multi-copy *rlep* repeat element (Truman et al. 2008) and another targeting the single-copy *fbpB* gene, which codes for the antigen 85B (Martinez et al. 2011). Similarly, all extracts were also tested using qPCR assays targeting the mycobacterial single-copy *rpoB* gene, which codes for RNA polymerase subunit B (Harkins et al. 2015) and the multi-copy insertion element IS6110, which is found in most *Mycobacterium tuberculosis* complex (MTBC) strains (McHugh, Newport, and Gillespie 1997; Klaus et al. 2010). The *rpoB* assay used in this study targets members of the MTBC as well as some closely related mycobacteria such as *M. marinum, M. avium, M. leprae, M. kansasii*, and *M. lufu* (Harkins et al. 2015). Genomic

DNA from *M. leprae* SM1 and *M. tuberculosis* H37Rv strains were used to create DNA standards for the appropriate qPCR assays. Ten-fold serial dilutions ranging from one to 100,000 copy numbers of the genome per µL were used to plot a standard curve for quantification purposes. Non-template controls (PCR-grade water) were also included on each qPCR plate. The DNA extracts, extraction blanks, and non-template control were run in triplicate whereas DNA standards were run in duplicate for each qPCR assay. qPCR reactions were run in a 20 µL total volume: 10 µL of TaqMan 2X Universal MasterMix, 0.2 µL of 10mg/mL RSA, and 2 µL of sample (DNA, standard, or non-template control). Primers and probe were added at optimized concentrations as given in Harkins et al. (2015) and Housman et al. (2015). The qPCR assays were carried out on the Applied Biosystems 7900HT thermocycler with the following conditions: 50°C for 2 minutes, 95°C for 10 minutes, and 50 cycles of amplification at 95°C for 15 seconds and 60°C for 1 minute. The results were visualized using SDS 2.3. Both amplification and multicomponent plots were used to classify the replicates of the extracts as positive or negative. An extract was considered to be positive for a qPCR assay if two or more replicates out of three were positive.

2.4 Results

Sequencing the genomes of the nonhuman primate *M. leprae* strains

Mapping analysis

The detailed summary of the sequencing results is given in Table 1. A total of 97 - 98% of the *M. leprae* genome was recovered for samples Ch4, SM1, and CM1 with mean coverage ranging from 13- to 106-fold.

Table 1. Results of Whole-genome Sequencing of Nonhuman Primate *M. leprae* Strains

Strain	Host species	Raw Reads	Processed Reads [a]	Mapped reads	Analysis-ready reads [b]	Average read length	Mean fold-coverage	Percent genome covered ≥ one-fold
Ch4	Chimpanzee	55,710,090	50,164,345	41,193,171	3,463,490	100.7	106.8	98
SM1	Sooty mangabey	697,450	526,512	349,276	293,217	279.8	25.1	98.8
CM1	Cynomolgus macaque	Library1: 17,065,716 Library2: 32,883,154	Library1: 14,101,593 Library2: 30,595,430 Total: 44,697,023	12,158,918	541,153	80.2	13.3	97.7

[a] Reads used as input for mapping after adapter trimming, merging, and removing reads less than 30 bp in length.

[b] Reads after filtering at Q37 quality threshold, removing duplicates, and removing reads with multiple mappings

For samples LB1 and LB2, approximately 6% of post-processed reads mapped to the *M. leprae* TN genome. After filtering the mapped reads, a negligible number of reads remained and less than 0.1% of the *M. leprae* genome was covered.

The efficacy of the *M. leprae* whole-genome enrichment varied for samples Ch4 and CM1. For sample Ch4, only 16.4% of post-processed reads mapped to the *Pan troglodytes* reference genome, whereas for sample CM1, 52.4% of post-processed reads mapped to the *Macaca fascicularis* reference genome. Thus, the whole-genome enrichment was more efficient for sample Ch4.

Phylogenetic analyses

Trees constructed using MP (Figure 4) and NJ (Appendix C: Figure S1) methods supported identical topologies for the *M. leprae* phylogeny. The Ch4 and the SM1 strains belong to *M. leprae* Branch 4. Within Branch 4, the Ch4 and SM1 strains are closely related to each other and form their own sublineage. On the other hand, the CM1 strain belongs to *M. leprae* Branch 0 and is most closely related to the modern human *M. leprae* strain S9 from New Caledonia.

According to the MCC tree (Figure 5), the Ch4 and SM1 strains diverged 295 YBP with a 95% Highest Posterior Density (HPD) range of 156-468 YBP. The sublineage comprising these two strains last shared a common ancestor with the Branch 4 human *M. leprae* strains 1063 YBP (95% HPD 765-1419 YBP). The CM1 strain shows a very deep divergence time of 2697 YBP (95% HPD 2011-3453 YBP) from its closest relative, *M. leprae* strain S9. The most recent common ancestor (MRCA) of all *M. leprae* strains was estimated to exist about 3590 YBP (95% HPD 2808-4606 YBP). Lastly, the estimated *M. leprae* substitution rate was 6.95×10^{-9} substitutions per site per year.

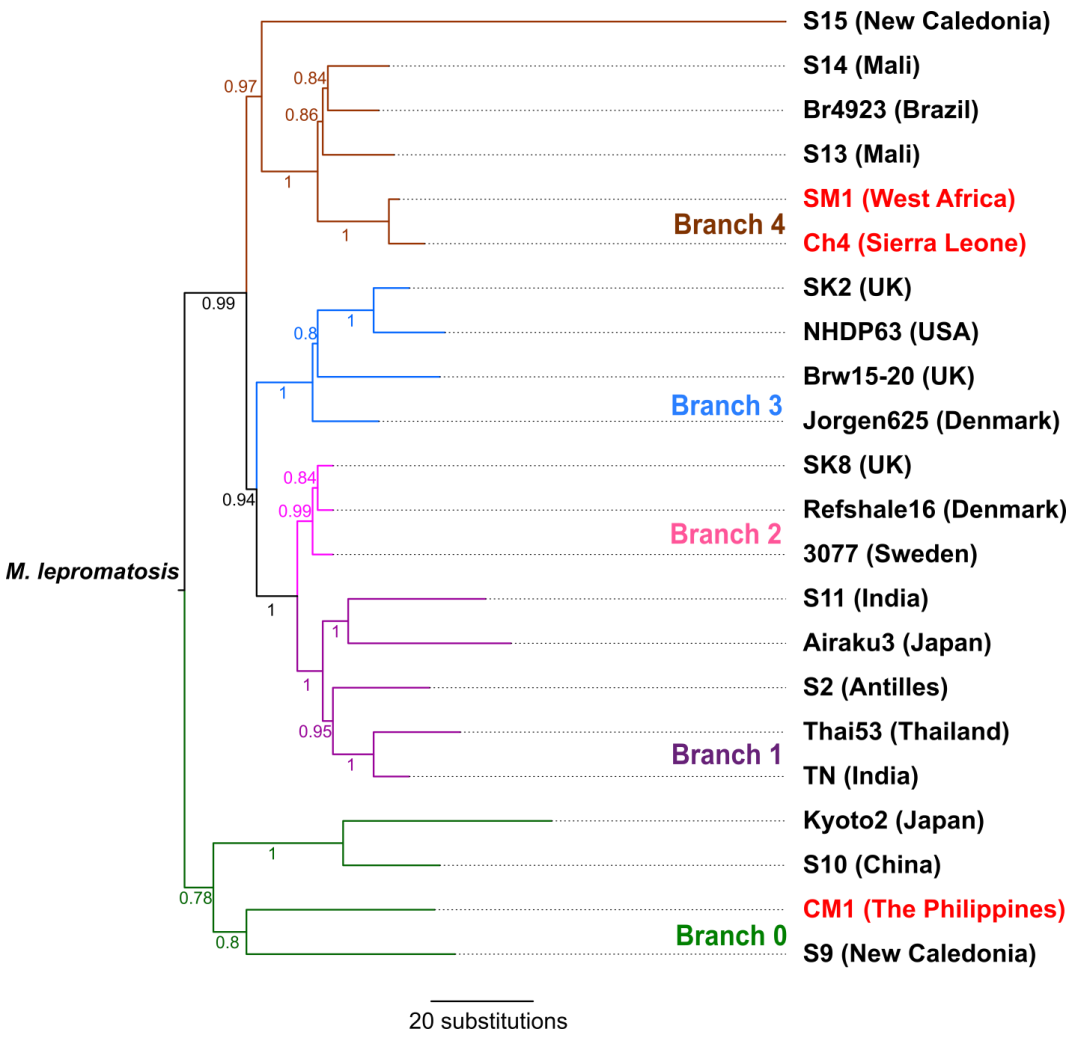

Figure 4. Maximum Parsimony Tree of *M. leprae* Strains. The tree was based on 233,509 genome-wide SNPs. *M. lepromatosis* was used as an outgroup to root the tree. The tree was generated using the SPR algorithm and bootstrap support estimated from 1000 replicates is given near each branch. The five *M. leprae* branches are highlighted. The nonhuman primate *M. leprae* strains sequenced in this study are given in red. The geographic origin is given next to the name of each strain.

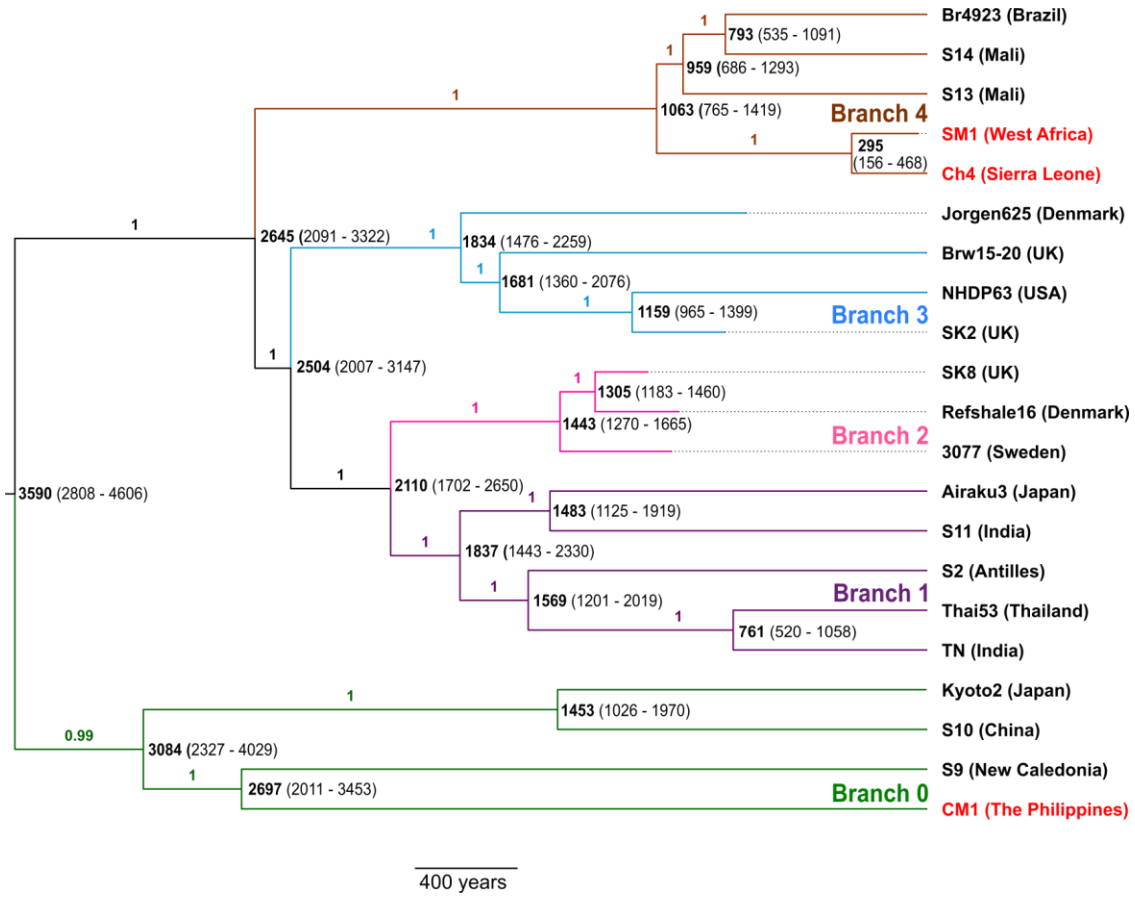

Figure 5. Maximum Clade Credibility Tree of *M. leprae* Strains. The tree was built using 747 genome-wide SNPs. The five *M. leprae* branches are highlighted. The nodes are labeled with median divergence times in years before present, with the 95% HPD given in brackets. Posterior probabilities for each branch are shown near the branches. The nonhuman primate *M. leprae* strains sequenced in this study are given in red. The geographic origin is given next to the name of each strain.

SNP-effect analysis

The Ch4, SM1, and CM1 strains showed 129, 124, and 167 total SNPs, respectively. The list of SNPs found in the nonhuman primate *M. leprae* strains and their effects are given in Appendix A: Table S2. 18 SNPs were found to be unique to the Ch4-

SM1 sublineage (i.e. they have so far not been found in any of the human *M. leprae* strains). Additionally, the Ch4, SM1, and CM1 strains showed 9, 4, and 54 unique SNPs, respectively. The summary of the SNP-effect analysis is given in Table 2.

Table 2. Summary of SNP-effect Analysis for the Nonhuman Primate *M. leprae* Strains

Type of variant	**Ch4**	**SM1**	**CM1**
missense variant in protein-coding gene	36 (4)	34 (2)	54 (20)
start loss variant in protein-coding gene	1 (0)	1 (0)	1 (0)
synonymous variant in protein-coding gene	24 (1)	24 (1)	28 (8)
variant in pseudogene	45 (3)	42 (0)	51 (10)
variant in intergenic region	23 (1)	23 (1)	33 (16)
Total	**129 (9)**	**124 (4)**	**167 (54)**

Numbers outside of the parentheses denote the total number of variants of this type. Numbers inside the parentheses denote the number of variants of this type unique to that particular strain

Screening of wild nonhuman primates for presence of mycobacterial pathogens

qPCR assays

All ring-tailed lemur samples and chimpanzee samples tested negative for *M. leprae* DNA based on the *rlep* and 85B qPCR assays. All of the samples also tested negative for the *rpoB* and IS6110 qPCR assays signifying absence of infection by pathogens belonging to the MTBC.

2.5 Discussion

Studies have shown that *M. leprae*, which was once thought to be an exclusive human pathogen, infects animals such as nine-banded armadillos in the New World (Walsh, Meyers, and Binford 1986; Truman et al. 2011) and red squirrels in the UK (Avanzi et al. 2016). Nonhuman primates including white-handed gibbons, rhesus macaques, African green monkeys, sooty mangabeys, and chimpanzees are capable of being experimentally infected with *M. leprae* resulting in symptomatic leprosy similar to that observed in humans (see Rojas-Espinosa and Lovik 2001). This study aimed at determining whether wild nonhuman primates may serve as a natural host of *M. leprae* by elucidating the phylogenetic relationships between nonhuman primate and human *M. leprae* strains. In this study, *M. leprae* genomes from three naturally infected nonhuman primates were sequenced. Furthermore, to assess the prevalence of *M. leprae* and other closely related mycobacterial pathogens in wild nonhuman primate populations, ring-tailed lemurs from BMSR, Madagascar, and chimpanzees from Ngogo, Uganda, were screened.

The Ch4 and SM1 *M. leprae* strains belong to *M. leprae* Branch 4. Strains belonging to this branch have been found in West Africa and the Caribbean where they were brought due to the slave trade (Monot et al. 2009). Strain S15, which was isolated from a human patient from New Caledonia, also falls in Branch 4. Both Ch4 and SM1 strains were found to belong to the 4O subtype (Monot et al. 2009).

The Ch4 *M. leprae* strain was isolated from a female chimpanzee captured from Sierra Leone in 1980 and held at a research facility in Japan. The chimpanzee developed symptoms of leprosy in 2009 (Suzuki et al. 2010). Since the Ch4 strain is West African in

origin, the chimpanzee was likely infected in Sierra Leone before being sent to Japan. The SM1 *M. leprae* strain was isolated from a West African sooty mangabey (originally denoted as individual A015). This mangabey was shipped from Nigeria to the US in 1975 and developed symptoms of leprosy in 1979. It is the first of two known cases of naturally occurring leprosy in sooty mangabeys (Meyers et al. 1985). The second sooty mangabey is thought to have acquired leprosy from A015 while both animals were housed together in the US (Gormus et al. 1988); however, samples from the second mangabey could not be obtained for the purposes of this study. The *M. leprae* strain isolated from A015 was reported to be partially resistant to dapsone (Meyers et al. 1985), suggesting the mangabey might have acquired leprosy directly or indirectly from a human patient who had received dapsone treatment. This study did not find SNPs known to be associated with dapsone-resistance, such as the $Thr^{53}Ile$ and $Pro^{55}Leu$ changes in the *folP1* gene (Maeda et al. 2001), in the SM1 strain.

18 SNPs were found to be unique to the Ch4-SM1 sublineage. These included seven missense variants occurring in genes coding for proteasome-related factors, glutamine-dependent NAD synthetase, acetyltransferases, and integral membrane proteins. The close relationship of the Ch4 and SM1 strains suggests that *M. leprae* might be transmitted between chimpanzees and sooty mangabeys in the wilds of Africa. The geographic ranges of chimpanzees overlap with those of sooty mangabeys (Figure 6). Chimpanzees are also known to hunt and kill other primates including mangabeys (Goodall 1986; Watts and Mitani 2000) and can acquire pathogens during predation and via consumption of bushmeat (Formenty et al. 1999).

Since *M. leprae* can be transmitted through consumption of infected animal meat, this might be one of the possible routes for transmission of *M. leprae* between nonhuman primates.

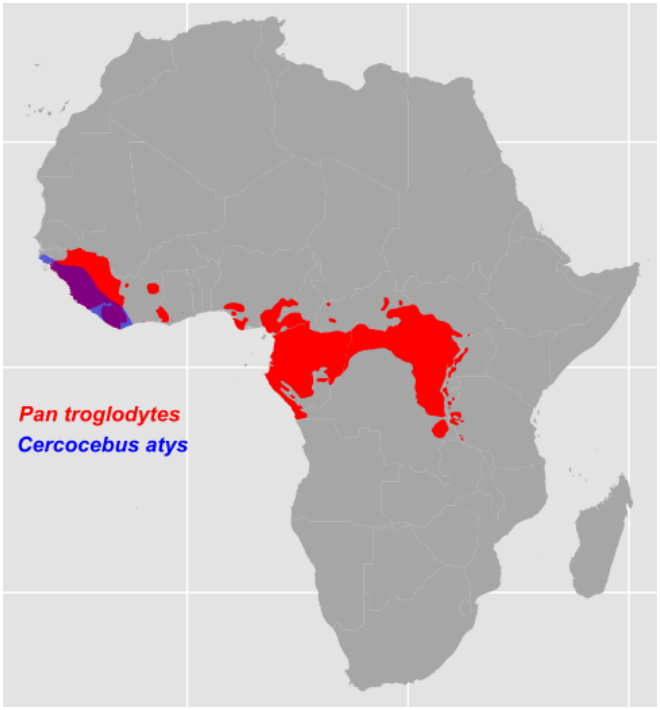

Figure 6. Map showing the Geographic Ranges of Chimpanzees and Sooty Mangabeys in Africa. The map was generated using RStudio (R Core Team 2017). The geographic range of chimpanzees (*Pan troglodytes*) is given in red and that of sooty mangabeys (*Cercocebus atys*) is given in blue. The overlap between the two ranges is shown in purple.

In Africa, interactions of humans and nonhuman primates, such as through zoos or sanctuaries, via hunting for bushmeat, or due to the use of nonhuman primates for exportation, sport, entertainment, and as family pets are major sources of pathogen transmission (Wolfe et al. 1998, Wolfe et al. 2005; Wallis and Lee 1999). Nonhuman

primates are highly susceptible to pathogens such as the simian immunodeficiency virus (SIV), Ebola virus, and *Bacillus cereus* biovar Anthracis (Calvignac-Spencer et al. 2012). In the context of mycobacterial pathogens, nonhuman primates are highly susceptible to the MTBC and may harbor novel lineages (Coscolla et al. 2013).

The results of this study support a scenario in which a human *M. leprae* sublineage was transmitted to a nonhuman primate species and has been circulating in nonhuman primates such as chimpanzees and sooty mangabeys in Africa. However, due to the paucity of *M. leprae* Branch 4 genomes, the data cannot rule out the possibility that this *M. leprae* sublineage is currently present in humans in West Africa and is not specific to nonhuman primates.

The CM1 strain belongs to *M. leprae* Branch 0. This branch also includes strains from New Caledonia, Japan and China and is the most deeply diverged branch of the *M. leprae* phylogeny (Schuenemann et al. 2013). The CM1 strain is subtype 3K similar to other strains in Branch 0 (Monot et al. 2009; Schuenemann et al. 2013). The CM1 strain has 167 SNPs, out of which 54 have so far not been found in other *M. leprae* strains. Interestingly, the CM1 strain showed presence of 54 missense variants, out of which five variants occurred in genes belonging to the ESX system. Three of these variants occur in the *ML0049* gene including a unique $Ala^{87}Thr$ change. This strain also has a unique $Glu^{273}Lys$ change in the *ML0054* gene. The *ML0049* and *ML0054* genes belong to the ESX-1 gene system encodes proteins which are major determinants of virulence in *M. leprae*, *M. tuberculosis, M. kansasii,* and *M. marinum* (Gröschel et al. 2016). They help the pathogen escape from the phagosome, thereby allowing further replication, cytolysis, necrosis, and intercellular spread (Simeone et al. 2012).

The CM1 strain was recovered from a cynomolgus macaque that had been shipped to the US from The Philippines in 1990. The animal started showing symptoms of leprosy in 1994 (Valverde et al. 1998). A sample of skin biopsy tissue from this animal had been stored using the FFPE method since 1994, from which DNA was extracted for the purposes of this study. However, FFPE preservation is known to cause fragmentation of DNA (Dedhia et al. 2007); the average length of mapped reads for sample CM1 were 80 bp, as compared to 100 bp for sample Ch4. Additionally, the efficacy of the capture was higher for sample Ch4 (82% of post-processed reads mapped to the *M. leprae* genome) as compared to that for CM1 (only 52% of post-processed reads mapping to *M. leprae*). This was not unexpected given that sample CM1 had been preserved in FFPE for over twenty years, whereas for sample Ch4, DNA had been extracted from the chimpanzee fairly recently in 2009.

Cynomolgus macaques, also known as crab-eating or long-tailed macaques, cover a broad geographic distribution in southeast Asia and have had a long history of contact with human populations (Fooden 1995). The geographic ranges of macaques overlap with human settlements, and contact between the two has increased due to human encroachment upon their habitats, hunting, and trapping activities. There is a high demand for macaques in biomedical research, pet trade, as performing animals, and as food (Jones-Engel et al. 2005). Due to their religious significance in Hinduism and Buddhism, macaques are respected in most of southeast Asia and are often included in religious festivities of the local populations. They are also a prominent species in monkey temples, which serve as popular tourist attractions. These temple settings provide ample opportunities for physical contact due to tourists feeding the monkeys as well as the

monkeys climbing on, biting, and scratching tourists (Jones-Engel et al. 2006). Such interactions significantly increase the risk for pathogen transmission between humans and macaques.

Cynomolgus macaques have been found to be infected with pathogens such as cercopithecine herpesvirus 1 (Engel et al. 2002), simian foamy viruses (Jones-Engel et al. 2006, 2001), MTBC (Wilbur et al. 2012), and *Plasmodium* species (Zhang et al. 2016). In the case of MTBC infection, prevalence is higher in macaques from Thailand, Indonesia, and Nepal, where tuberculosis is endemic, and lower in Gibraltar and Singapore, where tuberculosis is not endemic (Wilbur et al. 2012). The Philippines ranks first in the Western Pacific Region in terms of absolute number of leprosy cases, with about 2000 new leprosy cases reported annually (WHO 2016a). The results of this study support the hypothesis that *M. leprae* strains may be transmitted between humans and nonhuman primates especially in countries where leprosy is endemic.

Across the three nonhuman primate *M. leprae* strains, the highest number of SNPs were found in the *ML0411* gene, which is known to be the most polymorphic gene in *M. leprae* (Schuenemann et al. 2013). This gene codes for a serine-rich protein and is thought to have diversified under selective pressure imparted by the host immune system (Kai et al. 2013).

According to the dating analysis, the MRCA of all *M. leprae* strains was estimated to exist about 3590 YBP (95% HPD 2808-4606 YBP), which is in congruence with the previous estimate of 3483 YBP (95% 2401-4788 YBP) (Avanzi et al. 2016) as well as with the oldest skeletal evidence for leprosy which dates to 2000 BCE India (Robbins et al. 2009).

The estimated *M. leprae* substitution rate was 6.95×10^{-9} substitutions per site per year, which is also similar to previous estimates (Schuenemann et al. 2013; Avanzi et al. 2016).

To assess whether mycobacterial pathogens are transmitted between humans and nonhuman primates in tuberculosis- and leprosy-endemic regions, broad phylogeographic screenings of nonhuman primate populations need to be conducted. The ring-tailed lemur populations screened in this study were not necessarily expected to show prevalence of *M. leprae* infection, since successful experimental or natural transmission of *M. leprae* has not been reported in any lemur species. However, Madagascar reports approximately 1500 new leprosy cases (WHO 2016a) and 29,000 new tuberculosis cases (WHO 2015) each year. Interactions between the lemur populations at BMSR and the surrounding local human populations (Loudon et al. 2006) could lead to anthroponotic transmission of *M. leprae* and other pathogens to the lemur populations. However, the lemurs included in this study did not show evidence of infection by members of the MTBC or *M. leprae*.

Additionally, chimpanzee populations at Ngogo, Kibale National Park, in Uganda were also screened for the presence of these mycobacterial pathogens. Uganda reports about 43,000 new tuberculosis cases (WHO 2015) as well as approximately 250 new leprosy cases (WHO 2016a) annually. The ease of transmission of MTBC strains between different mammalian hosts underlies the need for screening wildlife for the presence of MTBC infection especially in tuberculosis-endemic regions. However, the chimpanzees screened in this study did not test positive for mycobacterial infection.

2.6 Summary

To the best of our knowledge, this is the first paper to report the genomes of nonhuman primate *M. leprae* strains. The phylogenetic analyses suggest that nonhuman primates may acquire *M. leprae* infection from humans as well as transmit *M. leprae* strains between themselves. In this study, wild nonhuman primate populations from Madagascar and Uganda were screened for the presence of mycobacterial infection; however, they tested negative. Further studies conducting broad phylogeographic screenings of nonhuman primates, especially in countries where leprosy is endemic, are necessary. The prevalence of leprosy-causing bacteria, *M. leprae* and *M. lepromatosis*, in nonhuman primate populations has important implications for leprosy eradication and nonhuman primate conservation strategies.

CHAPTER 3

INSIGHTS FROM THE GENOME SEQUENCE OF *MYCOBACTERIUM LEPRAEMURIUM*, THE CAUSATIVE AGENT OF MURINE LEPROSY

3.1 Abstract

Mycobacterium lepraemurium is the causative agent of murine leprosy. It causes a chronic, granulomatous disease similar to human leprosy; however, unlike human leprosy, the peripheral nerves are not impaired. Due to similar clinical manifestations of human and murine leprosy, *M. leprae* and *M. lepraemurium* were once thought to be closely related, although later studies suggested that *M. lepraemurium* might be closely related to *M. avium*. In this study, the complete genome of *M. lepraemurium* was sequenced using a combination of PacBio and Illumina sequencing. Phylogenomic analyses confirm that *M. lepraemurium* is a distinct species within the *M. avium* complex (MAC) and is not closely related to *M. leprae*. Members of the MAC cause tuberculosis-like disease in birds and other animal species as well as systemic disease in immunocompromised humans. The *M. lepraemurium* genome is 4.05 Mb in length, which is considerably smaller than other MAC genomes, and comprises 2,687 functional genes and 1,137 pseudogenes. The presence of numerous pseudogenes suggests that *M. lepraemurium* has undergone a genome reduction event. An error-prone repair homologue of the DNA polymerase III α-subunit was found to be non-functional in *M. lepraemurium*, which might contribute to pseudogene-formation due to accumulation of mutations in non-essential genes. *M. lepraemurium* can only be cultivated *in vitro* under highly stringent conditions and thus seems to be evolving towards retaining a minimal set of genes required for an obligatory intracellular lifestyle within its host, similar to *M.*

leprae. *M. lepraemurium* has retained the functionality of several genes thought to influence virulence among members of the MAC.

3.2 Introduction

Murine leprosy is a chronic, granulomatous disease caused by *Mycobacterium lepraemurium*. The disease mainly affects the skin, mucosa of the upper respiratory tract, and eyes (Dean 1903; Dean 1905; Stefansky 1903). Unlike in human leprosy, the viscera are commonly affected (Krakower and Gonzalez 1940; Kawaguchi et al. 1976) and the peripheral nerves are not affected (Rojas-Espinosa et al. 1999; Tanimura and Nishimura 1952). Murine leprosy was first reported in the early 20th century in rats in Ukraine (W. K. Stefansky 1902), following which similar cases were reported from other countries (Dean 1903; Marchoux and Sorel 1912). *M. lepraemurium* also causes a leprosy-like illness in cats, resulting in granulomatous skin lesions that often involve ulceration (Pedersen 1988). Feline leprosy occurring due to *M. lepraemurium* infection is thought to be acquired through bites from infected rodents (Lawrence and Wickham 1963). Recent studies have shown that feline leprosy is a syndrome caused by a number of mycobacterial species in addition to *M. lepraemurium*, such as *Mycobacterium* sp. Tarwin, *Mycobacterium* sp. cat, and *M. visibile* (Hughes et al. 2004; Malik et al. 2002; Fyfe et al. 2008; Foley et al. 2004).

In humans, leprosy is primarily caused by *Mycobacterium leprae* and *M. lepromatosis*, with the latter causing a severe form known as diffuse lepromatous leprosy. *M. leprae* also infects armadillos (Walsh, Meyers, and Binford 1986) and certain nonhuman primates (Donham and Leininger 1977; Gormus et al. 1991; Suzuki et al.

2010; Meyers et al. 1985; Gormus et al. 1988; Valverde et al. 1998), and both *M. leprae* and *M. lepromatosis* infect red squirrels (Avanzi et al. 2016). The finding that *M. leprae* and *M. lepromatosis* also infect rodents such as red squirrels implies that these pathogens might be closely related to that causative agent of murine leprosy.

Numerous similarities exist between human and murine leprosy including disease transmission through abrasions in the skin and the mucosal respiratory surfaces, similar spectrum of disease manifestation such as the tuberculoid and lepromatous forms, and the depression of cell-mediated immunity and lack of depression of humoral immunity seen in case of the more severe form of lepromatous leprosy (Banerjee 1979; Rojas-Espinosa 1994; Rojas-Espinosa and Lovik 2001). Early serological and microbiological studies of *M. leprae* and *M. lepraemurium* suggested that these species were closely related and hence, it was thought that murine leprosy might serve as a model for human leprosy (Walker and Sweeney 1929; Schmitt 1911; Dean 1905). *M. leprae* and *M. lepraemurium* are both slow-growing mycobacteria and are difficult to cultivate using standard microbiological media. *M. leprae* cannot be cultivated *in vitro* at all, whereas *M. lepraemurium* can be cultivated using a 1% Ogawa egg yolk medium (Mori and Kohsaka 1986) or a cell-free liquid medium (pH = 6.0 - 6.2) (Nakamura 1999).

M. leprae and *M. lepraemurium* are distinct species, but their relationships within the context of the mycobacterial phylogeny remain unclear. DNA hybridization studies have suggested that *M. lepraemurium* might be closely related to the *M. avium* complex (MAC) (Athwal, Deo, and Imaeda 1984); however, the lack of a genome sequence restricts our understanding of the biology and evolutionary history of *M. lepraemurium*.

Therefore, in this study, the genome of *M. lepraemurium* was sequenced using a combination of Single Molecule Real-Time (SMRT, Pacific Biosciences) and Illumina technology.

3.3 Materials and Methods

Bacilli culture and purification

M. lepraemurium Hawaii was grown using serial infections in BALB/c mice injected by the intraperitoneal route. At 6 months post-infection, the infected spleen and liver were harvested. The bacteria were purified following the protocol in Prabhakaran, Harris, and Kirchheimer (1976), followed by the Percoll step in Draper (1980). Isolation was conducted following previously established protocols (Wek-Rodriguez et al. 2007; Rojas-Espinosa, Wek-Rodriguez, and Arce-Paredes 2002). Purified and isolated bacilli were suspended in an aliquot of Middlebrook 7H9 broth medium (Becton Dickinson Co.) supplemented with 10% OADC (Becton Dickinson Co.) and air-dried to form a cell pellet.

DNA extraction

DNA extraction was carried out using a custom-designed protocol for mycobacterial DNA. The bacterial cell pellet was washed with 500 µL of phosphate buffer saline (PBS) prior to centrifugation at 5000 g for 10 minutes. The supernatant was discarded and the pellet was re-suspended in 1 mL of bacterial lysis buffer B1 (50 mM Tris-HCl pH 8.0; 50 mM EDTA pH 8.0; 0.5% Tween 20; 0.5% Triton-X100) containing 45 µL of proteinase K (20 mg/mL) and 20 µL of lysozyme (100 mg/mL). The mixture

was then transferred into bead-beating tubes containing 500 µL of 0.1 mm zirconia beads prior to physical disruption using the Precellys24 homogenizer at 6.5 m/s for 25 seconds. After incubating at 56°C for one hour, the mixture was centrifuged and the supernatant was transferred to a new tube. An additional incubation with 20 µL proteinase K was conducted at 56°C for 30 minutes. The mixture was incubated at 4°C for 15 minutes. RNase A (Sigma) was added and the sample was incubated 30 minutes at 37°C, followed by the addition of 350 µL of bacterial lysis buffer B2 (3M guanidine hydrochloride, 20% Tween 20), and incubated for 30 minutes at 50°C. The DNA was purified using the Qiagen Genomic-Tip/20G according to manufacturer's instructions, and eluted in 2 mL elution buffer. The DNA was precipitated using 0.7X volume of isopropanol and centrifuged at 4°C for 15 minutes. The pellet was washed twice with 200 µL 70% ethanol, air-dried, and suspended overnight in 200 µL Tris HCl buffer (pH 8.0) at room temperature under continuous shaking. The DNA was then purified using AMPure beads (Thermofisher) with 0.45X ratio. Quality of the DNA extract was checked using the Fragment Analyzer (Advanced Analytical Technologies) and the DNA was quantified using the Qubit 2.0 (Life Technologies).

Illumina sequencing

50 µL of DNA extract was sheared using the Covaris S220 Focused-ultrasonicator (Covaris) to obtain 400 bp-long DNA fragments, and purified using AMPure beads (1.8x) and the manufacturer's protocol. The sheared DNA was quantified using the dsDNA High Sensitivity assay and the Qubit 2.0 flurometer (Life Technologies). Up to 1 µg of DNA in 50 µL was used for library preparation using the Kapa Hyper prep kit (Roche)

and the PentAdapters (Pentabase) for indexing. The library was quantified using the dsDNA Broad Range assay and the Qubit 2.0 fluorometer. The library was sequenced on the Illumina HiSeq 2500 (1 × 101 bp run).

SMRT (PacBio) sequencing

5.1 µg DNA was sheared using a Covaris g-TUBE (Covaris S220) to obtain 10 kb fragments and the size distribution was checked using the Fragment Analyzer (Advanced Analytical Technologies). 4 µg of the sheared DNA was used to prepare a SMRTbell library with the PacBio SMRTbell Template Prep Kit 1 (Pacific Biosciences) according to the manufacturer's recommendations. The resulting library was size-selected using a BluePippin system (Sage Science Inc.) for molecules larger than eight kb. The recovered library was sequenced using a SMRT cell with P6/C4 chemistry and MagBeads on a PacBio RSII system (Pacific Biosciences) at 240 min movie length.

Genome assembly

The PacBio reads were processed using the HGAP2 and HGAP3 pipelines (Chin et al. 2013). Resulting contigs were compared to the nucleotide database at NCBI using BLAST (Altschul et al. 1990). The two largest contigs produced by HGAP3 v2.3.0 (which were 2.3 and 1.7 Mb in length, respectively) matched to *M. avium* sequences. These two contigs corresponded to the three largest contigs produced by HGAP2 v2.3.0 (which were 1.7, 1.6, and 0.6 Mb in length) and two shorter contigs (61 and 21 kb in length). The two HGAP3 contigs could be joined by the overlapping HGAP2 contigs, resulting in a single consensus sequence with overlapping ends, indicative of a circular

genome. To correct for possible sequence errors, Illumina reads were mapped onto the draft genome sequence using Bowtie2 (Langmead and Salzberg 2012) resulting in 35-fold coverage of non-duplicate reads. Variants were called using SAMtools mpileup (Li et al., 2009) and VarScan2 (Koboldt et al. 2012), resulting only in five single nucleotide polymorphisms (SNPs) and two short insertion-deletions (InDels).

Four percent of Illumina reads that did not map to the final genome sequence were assembled using MIRA (https://sourceforge.net/projects/mira-assembler/). The resulting 34 contigs (of which the largest was 1.9 kbp long) were compared to the nucleotide and protein databases at NCBI using BLAST. The contigs matched to *Mus musculus* or various bacteria. No evidence of a putative plasmid sequence was found.

Gene prediction

De novo gene prediction was conducted using the RAST server (Aziz et al. 2008) with the frameshift correction option. Reference-based gene prediction was conducted using RATT (Otto et al. 2011) with annotations from *M. avium* subsp. *paratuberculosis* K-10 (NC_002944.2) and *M. avium* subsp. *hominissuis* TH135 (AP012555.1). All predictions were merged, and inconsistencies and large intergenic areas were manually checked by using BLAST to compare the problematic sequences against the protein database at NCBI. Gene predictions, shorter than 100 nucleotides in length and not conserved in the genomes of other *M. avium* species, were removed. The annotated genome was submitted to GenBank (Accession No. CP021238).

Phylogenetic analyses

Publicly available genome data were acquired for 16 comparative mycobacterial species (Appendix D: Table S3). Since the *M. lepraemurium* contigs showed highest identity with *M. avium* sequences, representative genomes for all MAC species were included in this analysis.

Contigs or finished genomes of these species were aligned to the *M. avium* 104 reference genome using LAST (Kiełbasa et al. 2011) with the following parameters: -u = 0, -e = 34, and -j = 5. The maf-convert program was used to convert the alignment file to a SAM file and SAMtools was used obtain a BAM file which was used for further analyses. SAMtools (Li et al. 2009) mpileup and bcftools call were used to produce the VCF files. VCF files for all strains were combined using the CombineVariants tool available in the Genome Analysis Toolkit (GATK) (McKenna et al. 2010). The SelectVariants tool in GATK was used to output a VCF file containing the sites comprising SNPs. VCFtools (Danecek et al. 2011) was used to remove InDels, tri-allelic sites, and sites with missing data. A SNP alignment comprising a total of 460,625 sites was generated using a perl script (Bergey 2012) .

Phylogenetic trees were constructed using the Maximum Likelihood (ML) method in RAxML v7.2.8 (Stamatakis 2006) and the Neighbor-Joining (NJ) and Maximum Parsimony (MP) methods in MEGA7 (Kumar, Stecher, and Tamura 2016). The ML tree was generated using the GTR-GAMMA model and 100 bootstrap replicates. Since MAC species show high genetic identity, the NJ tree was generated using the *p*-distance method, and bootstrap support was estimated from 500 replicates.

The MP tree was generated using the Subtree-Pruning-Regrafting (SPR) algorithm and 500 bootstrap replicates.

3.4 Results and Discussion

Genome statistics

The PacBio and Illumina HiSeq data together provided 60-fold coverage of the *M. lepraemurium* genome, which was sufficient for de novo assembly. The *M. lepraemurium* genome was found to be circular, and 4,050,523 bp in length. The total GC content is 68.99%. The genome comprises 3,824 protein-coding genes, out of which 2,687 are functional genes and 1,137 are pseudogenes.

Phylogenetic analyses

According to the ML tree, *M. lepraemurium* belongs to the MAC and is more closely related to the *M. avium* clade than to the *M. intracellulare* clade (Figure 7). Thus, *M. lepraemurium* is not closely related to the causative agents of human leprosy, *M. leprae* and *M. lepromatosis*. The MP and NJ trees supported identical topologies (Appendix E: Figures S2 and S3).

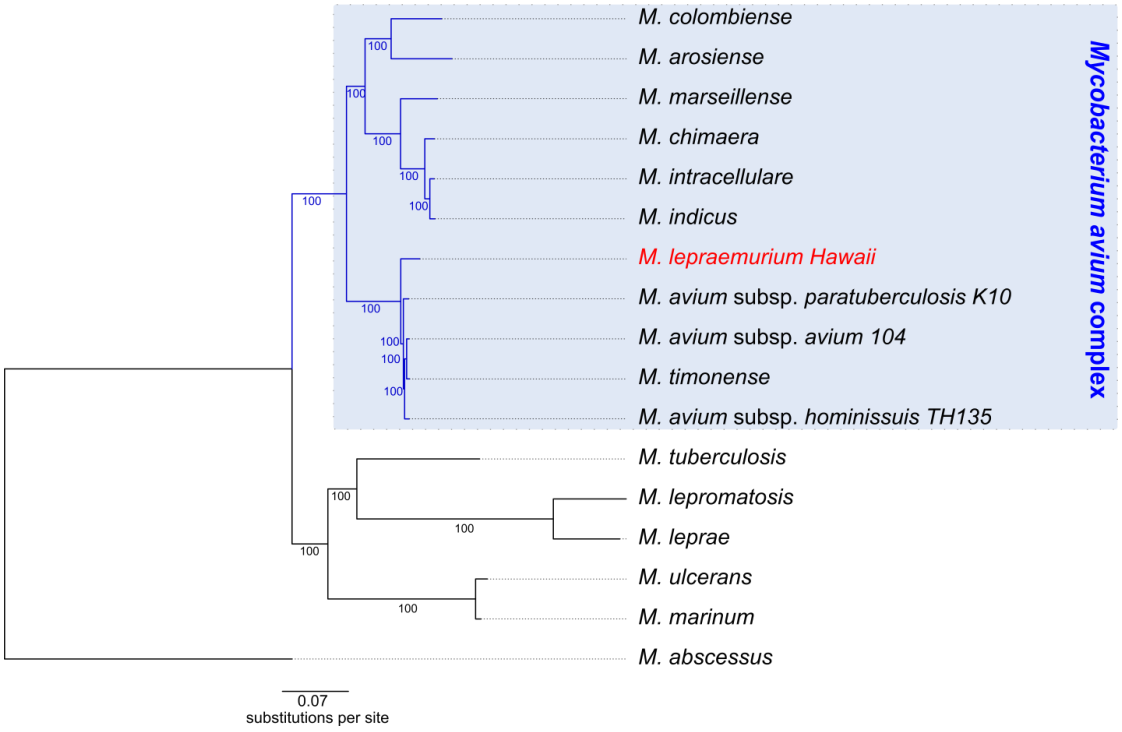

Figure 7. Maximum Likelihood tree of *M. lepraemurium* and Other Mycobacterial Species. *M. abscessus* was used as the outgroup to root the tree. The tree was based on 460,625 variable nucleotide sites and the GTR-GAMMA model. Bootstrap support estimated from 100 replicates is given below each branch. Species belonging to the *M. avium* complex are highlighted in blue and *M. lepraemurium* is denoted in red.

The MAC includes the two well-studied species, *M. avium* and *M. intracellulare*, as well as recently-defined species such as *M. chimaera*, *M. colombiense*, *M. arosiense*, *M. timonense*, *M. vulneris*, *M. marseillense*, *M. indicus*, and *M. bouchedurhonense* (see Coelho et al. 2013). Furthermore, *M. avium* comprises four subspecies, *M. avium* subsp. *avium*, *M. avium* subsp. *hominissuis*, *M. avium* subsp. *paratuberculosis*, and *M. avium* subsp. *silvaticum* (Thorel, Krichevsky, and Lévy-Frébault 1990). Members of the MAC are capable of infecting a diverse range of host species and possess a high degree of

genetic similarity. *M. avium* subsp. *avium* and *M. avium* subsp. *silvaticum* cause tuberculosis-like disease in birds (Thorel, Krichevsky, and Lévy-Frébault 1990; Dhama et al. 2011), *M. avium* subsp. *paratuberculosis* causes Johne's disease in ruminant mammals (Harris and Barletta 2001), and *M. avium* subsp. *hominissuis* causes systemic infection in immunocompromised humans especially HIV-AIDS patients (reviewed by Coelho et al. 2013). Although mice have been used as a model to study *M. avium*, *M. lepraemurium* is the first member of this complex found to be adapted to rodents. This study provides evidence that *M. lepraemurium* is a distinct species within this complex.

Genome downsizing and pseudogene formation

At 4.05 Mbp, the *M. lepraemurium* genome is the smallest genome belonging to the MAC. Within the MAC, obligatory pathogenic species such as *M. avium* subsp. *paratuberculosis* K10 (4.83 Mbp) and *M. avium* subsp. *avium* Env77 (4.58 Mbp) have smaller genomes as compared to those of opportunistic pathogens such as *M. avium* subsp. *hominissuis* TH135 (5.14 Mbp) (Cases, De Lorenzo, and Ouzounis 2003; Ignatov et al. 2012). The presence of 1,137 pseudogenes confirms that the *M. lepraemurium* genome has been downsized. To date, it is the fourth mycobacterial species known to have undergone reductive evolution; the other species include *M. leprae* and *M. lepromatosis*, which have severely downsized their genomes, as well as *M. ulcerans*, which is in a state of intermediate reductive evolution (Cole et al. 2001; Singh et al. 2015; Stinear et al. 2007). Interestingly, despite its distinct evolutionary history from *M. leprae* and *M. lepromatosis*, the *M. lepraemurium* genome seems to be evolving in a similar manner towards an obligatory intracellular parasitic lifestyle.

In *M. leprae*, loss of the DnaQ-mediated proof-reading mechanism of DNA polymerase III α-subunit has been hypothesized as the cause of pseudogene-formation (Liu et al. 2004; Cole et al. 2001). According to this study, in *M. lepraemurium*, an error-prone repair homologue of the DNA polymerase III α-subunit (MLM_3495) was found to be non-functional, which might contribute to a higher error rate leading to pseudogene formation in this species. Analysis of pseudogene families within a diverse set of prokaryotes has shown that pseudogenes are most likely to occur in ABC transporter, short-chain dehydrogenase/reductase, sugar transporter, cytochrome P450, and PE/PPE gene families (Liu et al. 2004). *M. lepraemurium* shows presence of pseudogenes in all these families.

Species-specific genes in *M. lepraemurium*

Comparison of *M. lepraemurium* and *M. leprae* genes showed that most genes which are functional in *M. leprae* also have a functional orthologue in *M. lepraemurium.* The majority of *M. lepraemurium* genes are shared with other members of the MAC, and only 27 genes are unique to *M. lepraemurium* (Appendix D: Table S4). An *M. lepraemurium*-specific gene (MLM_3300) encodes a Fic family protein, which can contribute to pathogenicity in bacteria. Fic proteins are cell filamentation proteins, which are commonly found in bacteria and are involved in post-translational modifications of proteins. Although the functions of Fic proteins are not well understood, pathogenic bacteria are known to secrete Fic proteins which act as toxins and interfere with cytoskeletal, trafficking, signaling, or translation pathways in the host cell (reviewed by

Roy and Cherfils 2015). Thus, the MLM_3300 gene might constitute a virulence factor for *M. lepraemurium*.

PE/PPE genes

Members of the *PE* and *PPE* multigene families encode the Gly-Ala-rich cell envelope proteins which are unique to mycobacteria and have been found to influence virulence (Li et al. 2004; Ramakrishnan et al. 2000). However, in general, MAC genomes show decreased numbers of *PE* and *PPE* genes as compared to *M. tuberculosis* (Li et al. 2005). In *M. lepraemurium*, three functional and two non-functional *PE* genes and 14 functional and nine non-functional *PPE* genes were identified. This relative reduction in the numbers of functional *PE/PPE* genes suggests that while they may influence virulence, they are not essential for it. This is supported by the paucity of *PE/PPE* genes in *M. leprae*, which contains only five *PE* and six *PPE* functional genes (Cole et al. 2001).

Interaction with macrophages

Mycobacteria such as the MTBC, *M. leprae*, and MAC are intracellular parasites of macrophages; however, they interact with macrophages in different ways. Upon entering the macrophage, *M. leprae* disrupts the phagolysosomal membrane and escapes into the cytoplasm where it proliferates. In contrast, after entering the macrophage, members of the MAC reside within phagosomes (Ignatov et al. 2012). These pathogens inhibit the maturation of the phagosome (by preventing the acidification of the phagosome to a pH below 6.4) and do not allow its fusion with the extremely acidic

lysosome. Studies have shown that mutations in the *PPE* gene (MAV_2928) and *PE* gene (MAV_1346) of *M. avium* cause the pathogen to be unable to inhibit maturation and acidification of phagosomes, resulting in decreased virulence (Li et al. 2010; Li et al. 2004). In *M. lepraemurium*, gene MLM_2357 (homologous to MAV_2928) and MLM_1265 (homologous to MAV_1346) are fully functional, suggesting that they may help its survival in macrophages. Additionally, the MLM_2012 gene encodes a homologue of the LppM lipoprotein, which is an important virulence factor in *M. tuberculosis*, and is also involved in the manipulation of the phagosomal maturation in macrophages (Deboosère et al. 2016).

In response to infection by mycobacteria, macrophages produce reactive oxygen species which form an integral part of the microbicidal response of macrophages. Studies have shown that the phagocytosis of *M. lepraemurium* occurs without triggering the generation of reactive oxygen species (Rojas-Espinosa et al. 1998); however, reactive oxygen species are produced during early stages of *M. lepraemurium* infection. The ability of pathogens to produce enzymes such as catalase-peroxidase, epoxide hydrolase, and superoxide dismutase (SOD), which remove reactive oxygen species, enable the survival of *M. tuberculosis* and *M. avium* within macrophages. *M. lepraemurium* shows catalase-peroxidase activity (Lygren et al. 1986). Three catalase-encoding genes were identified in *M. lepraemurium*. Among them, MLM_2092 is functional and encodes the catalase-peroxidase, whereas MLM_0454 and MLM_1574 are pseudogenes.

Additionally, four functional genes (MLM_0642, MLM_0684, MLM_1194, MLM_1485) were found to encode epoxide hydrolases, whereas two other epoxide hydrolase-encoding genes (MLM_0312 and MLM_1930) were found to be non-

functional. *M. lepraemurium* seems to produce both SODs found in *M. tuberculosis* – sodA and sodC. MLM_0123 encodes a Mn-Fe SOD (sodA) and MLM_2650 encodes a Cu-Zn SOD (sodC) precursor, whereas MLM_3522, which encodes a sodC precursor, is nonfunctional. These enzymes may help *M. lepraemurium* survive in macrophages; however, studies of *M. leprae* suggest that they are not essential for virulence, as *M. leprae* has functional sodA and sodC but non-functional catalase-peroxidase (Eiglmeier et al. 1997), and fewer peroxidoxins and epoxide hydrolases (Cole et al. 2001).

ESX gene system

The ESX system, also known as the Type VII secretory system, consists of proteins that transport selected substrates across the cell envelope and are associated with pathogenicity and host-pathogen interactions (Gröschel et al. 2016). ESX-1 is an important determinant of virulence in *M. tuberculosis* and *M. leprae*; however it is missing in most *M. avium* species, including *M. lepraemurium*. The functions of ESX-2 and ESX-4 systems are unknown; however, these systems are not essential for growth or virulence. In *M. leprae* and *M. lepromatosis*, ESX-2 is non-functional and ESX-4 is missing (Singh et al. 2015). In *M. lepraemurium*, both ESX-2 and ESX-4 are present, but are nonfunctional. The ESX-3 system is fairly conserved across all mycobacteria, including *M. lepraemurium*, and seems to fulfill an essential function in metal homeostasis. The ESX-5 system, specifically the PE/PPE proteins and EccD5, are essential in the pathogenesis of *M. tuberculosis*. In *M. lepraemurium*, the majority of ESX-5 components are functional, except for the cytochrome P450 hydroxylase (MLM_2361), which is also nonfunctional in *M. leprae*. However, in *M. leprae*, the

ESX-5 associated *pe*/*ppe* and *esx* genes are deleted, whereas these are functional genes in *M. lepraemurium*. Thus, similar to *M. tuberculosis*, the ESX-5 might influence virulence in this organism; however, the exact mechanism remains unknown.

Other virulence genes

Expression of some genes involved in polyketide synthesis (*pks* genes) is known to be upregulated in infected macrophages (Hou, Graham, and Clark-Curtiss 2002). This study shows that in *M. lepraemurium*, genes encoding *pks10* (MLM_2480), *pks11* (MLM_2477), and *pks12* (MLM_2156) are fully functional, whereas *pks2* is nonfunctional.

The mycobacterial mmpL and mmpS proteins mediate the transport of lipid metabolites to biosynthesize cell wall lipids such as mycolic acids. In *M. lepraemurium*, six *mmpS* and six *mmpL* genes are functional, whereas two *mmpS* and 11 *mmpL* genes are non-functional (Appendix D: Table S5). *M. leprae* has only two *mmpS* members and five functional *mmpL* genes. Thus, these genes may not be required for a strict intracellular lifestyle and therefore, might be undergoing pseudogenization in *M. lepraemurium*.

3.5 Summary

In this study, the 4.05 Mbp genome of *M. lepraemurium*, the causative agent of murine leprosy, was sequenced and annotated. Phylogenetic analyses confirmed that *M. lepraemurium* is a distinct species within the MAC. The presence of nearly 1100 pseudogenes suggests that *M. lepraemurium* has undergone reductive evolution. Since reductive evolution is a hallmark of pathogens that have undergone an evolutionary

bottleneck and adapted to a new environment (Gómez-Valero et al. 2007), the *M. lepraemurium* progenitor may have jumped from a different host into rodents and adapted to this new host/niche. This likely resulted in genome downsizing and losing the functionality of the majority of the genes required for survival outside of its host. However, *M. lepraemurium* seems to have retained the functionality of most of the genes required for virulence in MAC species as well as of certain genes that allow it to be grown *in vitro* under very specific conditions.

CHAPTER 4

MYCOBACTERIUM TUBERCULOSIS GENOMES FROM POST-CONTACT ERA NORTH AMERICA

4.1 Abstract

Tuberculosis (TB), caused by members of the *Mycobacterium tuberculosis* complex (MTBC), is one of the oldest known human diseases. Skeletal evidence suggests that TB was prevalent in the Americas before the arrival of Europeans whereas recent genomic evidence shows that TB cases in pre-contact era Peru were caused due to a zoonotic transfer of MTBC strains from pinnipeds, such as seals, to human populations living in the coastal regions. However, it is not known whether these pinniped-derived MTBC strains were the primary causative agents of TB in pre-contact era North America or if other lineages of the MTBC also caused TB in this region. In this study, 65 skeletal samples from pre- and post-contact era North American archaeological sites were screened for the presence of MTBC DNA using quantitative PCR assays and in-solution MTBC gene capture. Following whole-genome enrichment using in-solution hybridization capture and multiple rounds of Illumina sequencing, approximately 90% of the MTBC genome was recovered from five samples with mean coverage ranging from 5- to 26-fold. All five of these samples belong to the post-contact era archaeological sites of Cheyenne River Village (Arikara) in South Dakota; Highland Park cemetery in New York; and St. Michael, Old Hamilton, and Ekwok in Alaska. Phylogenetic analyses show that all five strains belong to the Euro-American lineage (Lineage 4). The St. Michael and Ekwok strains are closely related to Russian *M. tuberculosis* strains belonging to sublineage 4.2 (the Ural sublineage) whereas the Old Hamilton strain belongs to

sublineage L4.5 which is found in Middle Eastern or East Asian countries but rarely in the Americas or in Russia. Secondly, the Highland Park strain belongs to the sublineage comprising H37Rv-like strains which were highly prevalent in Britain during the 18th and 19th centuries as well as in the US during the early 20th century. Lastly, the Cheyenne River Village (Arikara) strain belongs to sublineage 4.4 and contains the DS6^{Quebec} deletion, which has been commonly found in strains that were brought to Canada by European fur traders. Overall, this study provides evidence for the introduction and dispersal of European *M. tuberculosis* strains to native populations in North America due to the fur trade.

4.2 Introduction

TB is one of the oldest known human diseases and remains a major public health concern with approximately 10.4 million new cases reported in 2015 (WHO 2016b). TB is caused by members of the MTBC which comprises the human-adapted *M. tuberculosis* and *M. africanum*, animal-adapted *M. microti* (voles), *M. caprae* (goats), *M. pinnipedii* (seals, sea lions), *M. bovis* (cattle), and *M. orygis* (oryx), as well as *M. canettii*. Furthermore, human *M. tuberculosis* strains are divided into seven lineages (L1 - 7) with each lineage being associated with specific geographic regions (Comas et al. 2013).

Although MTBC strains are adapted to specific hosts, cross-species transmissions occur frequently. Previous research from this group led to the recovery of three ancient MTBC genomes from approximately 1000 year-old mummies from archaeological sites in coastal Peru (Bos et al. 2014). These Peruvian MTBC strains were found to be closely related to *M. pinnipedii* which infects pinnipeds such as seals and sea lions. Today, *M.*

pinnipedii infection is restricted to pinnipeds in the southern hemisphere (Bastida et al. 1999), with occasional reports of zoonotic transfer to humans (Forshaw and Phelps 1991; Kiers et al. 2008; Thompson et al. 1993) and other animals (Moser et al. 2008; Loeffler et al. 2014). The discovery of pinniped-derived MTBC strains in ancient Peru suggested that infected pinnipeds transmitted MTBC strains to human populations living near the coast (Bos et al. 2014). The hunting of pinnipeds for meat and blubber (Orquera 2005; Orquera, Legoupil, and Piana 2011; Schiavini 1993), as well as use of their skin and bones for making artifacts and in mortuary practices (Arriaza 1996; Arriaza and Standen 2005) would have provided avenues for transmission of MTBC strains from pinnipeds to humans (Bastida, Quse, and Guichon 2011); however, the possibility of anthroponotic transfer from humans to pinnipeds is unlikely since these populations did not farm pinnipeds. It remains to be determined whether these pinniped-derived MTBC strains adapted to their human hosts and spread to non-coastal areas of the Americas or whether they were restricted to the coastal areas.

In North America, skeletal evidence of tuberculosis dates back to approximately 900 CE (reviewed by Roberts and Buikstra 2003; Stone et al. 2009), with an unpublished report from Point Hope, Alaska dating to 100 BCE - 500 CE (Dabbs 2009). The presence of partial IS6110 insertion repeat elements, which are commonly found in MTBC species (Thierry et al. 1990; McHugh, Newport, and Gillespie 1997), has been reported in individuals from the circa 11th -13th century Schild cemetery in Illinois, as well as from a 15th century Canadian ossuary sample (Braun, Collins Cook, and Pfeiffer 1998; Raff, Cook, and Kaestle 2006). However, IS6110 cannot be used to determine the phylogenetic placement of MTBC strains, and to date, MTBC genomes have not been recovered from

pre-contact era North America. Therefore, it is not known whether TB cases in pre-contact era North America were caused by spread of the pinniped-derived MTBC strains from the South or by different MTBC lineage(s) introduced via other routes.

Today, the majority of TB cases occurring in the Americas are caused by human-adapted *M. tuberculosis* L4 strains. L4 is also known as the Euro-American lineage, as it likely evolved in Europe and spread all over the world due to European migration and colonization, thereby becoming the most widespread human TB lineage (Demay et al. 2012; Stucki et al. 2016). Previous ancient DNA studies have used a metagenomics approach to reconstruct 18th century *M. tuberculosis* L4 genomes from Hungary (Kay et al. 2015; Chan et al. 2013). Using the radiocarbon dates of these individuals as calibration points, the most recent common ancestor (MRCA) of L4 strains has been estimated to have existed around 396 CE (Kay et al. 2015). This is supported by PCR-based finding of the *pks15/1* deletion specific to L4 strains (Marmiesse et al. 2004) from 2nd-4th century Britain (Müller, Roberts, and Brown 2014).

Even though MTBC strains are highly genetically identical, genome-wide single nucleotide polymorphisms (SNPs) have been used to classify L4 strains into seven sublineages – L4.1, L4.2, L4.3, L4.4, L4.5, L4.6, and L4.10 (Stucki et al. 2016; Coll et al. 2014). Some sublineages such as L4.1.2, L4.3, and L4.10 are globally distributed, others such as L4.1.1, L4.2, and L4.4 are found at intermediate frequencies, and some sublineages such as L4.1.3, L4.5, L4.6 are geographically restricted to less than ten countries all over the world (Stucki et al. 2016).

In this study, 66 individuals from various North American archaeological sites spanning the pre- and post- European contact eras were screened for the presence of

MTBC infection using quantitative PCR (qPCR) assays and in-solution gene capture techniques. Samples which passed the screening process were further analyzed using whole-genome enrichment and Illumina sequencing.

4.3 Materials and Methods

Sample processing

This study comprised 66 individuals from various North American archaeological sites spanning the pre- and post-contact eras (Appendix F: Table S6). All individuals showed characteristic symptoms of skeletal tuberculosis disease. The majority of the samples obtained were skeletal elements such as vertebrae or ribs; however, teeth or dental calculus samples were also screened. Skeletal and tooth samples were processed in a dedicated ancient DNA laboratory at Arizona State University (ASU) and dental calculus samples were processed at the University of Oklahoma Laboratories of Molecular Anthropology and Microbiome Research (LMAMR). All sample processing was conducted in accordance with established contamination control precautions and workflows (Cooper and Poinar 2000).

In case of skeletal samples, debris and dirt were removed using a sterilized Dremel tool. Surfaces of the skeletal samples and teeth were wiped with 10% bleach solution followed by distilled water, and UV-irradiated for 1 minute on each side. Skeletal samples were powdered using the 8000M Mixer/Mill (SPEX). Teeth were sliced transversally at the cementoenamel junction using a small, sterilized hand-saw and the roots were ground to a powder using a sterilized hammer. Dental calculus samples were collected using a scaler as given in Warinner et al. (2014).

DNA extraction

Three different DNA extractions protocols were used over the course of this study (Rohland and Hofreiter 2007; Dabney et al. 2013; Warinner et al. 2014) (Appendix F: Table S6). The majority of the skeletal samples and teeth were extracted at ASU using the protocol given in Dabney et al. (2013) with a minor modification - the final elution was carried out in 100 μL EBT buffer pre-heated to 65°C. Three samples had been previously extracted in 2012 using the protocol given in Rohland and Hofreiter (2007) at the University of Tuebingen, Germany. The dental calculus samples were extracted at LMAMR as given in Ozga et al. (2016). An extraction blank (negative control) was kept during each batch of extractions to check for possible contamination introduced during the extraction process. All DNA extracts and extraction blanks were quantified using the Qubit dsDNA High Sensitivity assay (Life Technologies).

Screening for MTBC DNA

The DNA extracts were screened for the presence of MTBC DNA using qPCR assays and an in-solution MTBC gene capture method.

qPCR assays

At ASU, undiluted extracts and extraction blanks were tested for MTBC DNA using three TaqMan qPCR assays. A 1:10 dilution of each extract was used to test for presence of inhibitory substances in the ancient DNA extracts. The first qPCR assay (rpoB2 assay) targets a region of the *rpoB* gene, which is a single-copy gene found in all bacteria and codes for RNA polymerase subunit B. This assay uses a TaqMan probe that binds to an MTBC-specific sequence in the gene (Harkins et al. 2015); however, due to

lack of sequence data for a number of mycobacterial species, this assay might test positive for closely-related mycobacterial species as well. The other two assays target regions of the multi-copy insertion elements IS6110 and IS1081 that are specific to the MTBC (McHugh, Newport, and Gillespie 1997; Klaus et al. 2010; Eisenach et al. 1990; Collins and Stephens 1991). Genomic DNA from *M. tuberculosis* H37Rv was used to create DNA standards for the qPCR assays. Ten-fold serial dilutions ranging from one to 1,000,000 copy numbers of the genome per µL were used to plot a standard curve for quantification purposes. Non-template controls (PCR-grade water) were also included on each qPCR plate. DNA extracts, extraction blanks, and non-template control were run in triplicate whereas DNA standards were run in duplicate for each qPCR assay. qPCR reactions were run in a 20 µL total volume: 10 µL of TaqMan 2X Universal MasterMix, 0.2 µL of 10mg/mL RSA, and 2 µL of sample (DNA, standard, or non-template control). Primers and probe were added at optimized concentrations as given in Housman et al. (2015). The qPCR assays were carried out on an Applied Biosystems 7900HT thermocycler with the following conditions: 50°C for 2 minutes, 95°C for 10 minutes, and 50 cycles of amplification at 95°C for 15 seconds and 60°C for 1 minute. The results were visualized using SDS 2.3. Both amplification and multicomponent plots were used to classify the replicates of the extracts as positive or negative. An extract was considered to be positive for a qPCR assay if two or more replicates out of three were positive.

In-solution MTBC gene capture

At ASU, DNA extracts which tested positive for one or more qPCR assays were processed into double-indexed libraries using 10-20 µL of extract and following protocols

given in Meyer and Kircher (2010) and Bos et al. (2014). Libraries were indexed using AmpliTaq Gold (Life Technologies) for 20 cycles and quantified using the DNA1000 assay on the Bioanalyzer 2100 (Agilent) and the KAPA Library Quantification kit (Kapa Biosystems). At LMAMR, the dental calculus extracts were processed into libraries as given in Ozga et al. (2016). A library blank (negative control) was processed along with the samples in each library preparation run.

All libraries, including library blanks, were target-enriched using an in-solution capture protocol at ASU. The libraries were target-enriched for five genes - the *rpoB, gyrA,* and *gyrB* genes commonly found in all mycobacterial species, and the *katG* and *mtp40* genes specific to the MTBC, as given in Bos et al. (2014). Enriched libraries were amplified to a concentration of 10^{13} copies per reaction using AccuPrime *Pfx* DNA polymerase (Life Technologies) and quantified using the Bioanalyzer 2100 (Agilent) and the KAPA Library Quantification kit (Kapa Biosystems). The libraries were pooled at equimolar concentrations and sequenced on an Illumina MiSeq using V2 chemistry (2×150 bp run).

The sequence reads were trimmed and merged using SeqPrep (https://github.com/jstjohn/SeqPrep) using default parameters except the minimum overlap for merging was modified to 11. Merged reads were mapped to the hypothetical MTBC ancestor reference (Comas et al. 2010) using the Burrows-Wheeler Aligner (bwa v0.7.5) (Li and Durbin 2009). In order to avoid mis-mapping of reads from environmental mycobacteria present in the samples, the stringency of the mapping was increased using the parameter n = 0.1, while the seed was disabled (-l = 1000) as suggested for ancient DNA (Schubert et al. 2012). SAMtools v0.1.19 (Li et al. 2009) was

used to filter the mapped reads at a minimum Phred quality threshold of Q30 and to remove PCR duplicates and reads with multiple mappings. The resulting BAM files were visually analyzed using Geneious R7 (Biomatters) and the percentage of the targeted genes covered at greater than one-fold coverage was determined. Samples for which more than 50% of all five genes were covered at least one-fold were selected for MTBC whole-genome enrichment.

Shotgun sequencing

Library preparation and sequencing

All samples that passed the cut-off for the in-solution MTBC gene capture were shotgun-sequenced to determine the percentage of endogenous MTBC DNA. At ASU, DNA extracts for these samples were processed into highly concentrated libraries using 80 μL of extract and following the protocols mentioned earlier. Prior to library preparation, the DNA extracts were treated with the USER enzyme (New England BioLabs), which contains uracil DNA glycosylase (UDG) and endonuclease VIII (endoVIII). Together, these enzymes are used to remove deaminated cytosines from ancient DNA fragments and repair the resulting abasic site (Briggs et al. 2010). At the Max Planck Institute for the Science of Human History (MPI-SHH), Germany, adapter dimers and heteroduplexes that had formed during the course of library preparation were removed using a reconditioning PCR. This was done by amplifying the libraries for two cycles using Herculase II Fusion DNA Polymerase (Agilent). The libraries were quantified using the D1000 assay on the TapeStation 4200 (Agilent), pooled at equimolar

amounts of 10 nM, and sequenced using the Illumina NextSeq 500 (1×76 bp run). Both non-UDG treated and UDG-treated libraries were sequenced.

Data analyses

Sequenced reads were trimmed and merged using EAGER (Peltzer et al. 2016). The merged reads were used as input for the MEGAN Alignment Tool (MALT) which compares the reads to a comprehensive database of all bacterial genomes available through NCBI RefSeq (Herbig et al. 2016). The MALT analysis was performed using the following parameters: minPercentIdentity = 95, minSupport = 5, topPercent = 1, BlastN mode, and SemiGlobal alignment. The resulting alignment was viewed in MEGAN6 (Huson 2016) and the number of reads assigned to the MTBC node was determined. Samples were classified as strongly or weakly positive for the MTBC based on visual analysis of whether the aligned reads were distributed randomly across the MTBC reference genome (as opposed to being accumulated at certain loci) and whether they showed high similarity (> 99%) to the MTBC reference genome.

MTBC whole-genome enrichment and sequencing

Sample selection

Based on the results of the MTBC screening process, a total of eight ancient DNA samples were selected for MTBC whole-genome enrichment and sequencing (Table 3). These included five samples (AD12, AD340, AD344, AD346, and AD351) selected based on the results of the qPCR assays and in-solution MTBC gene capture conducted at

ASU, as well as three samples (AD114, AD127, and AD128) which had been processed into libraries and screened using the in-solution gene capture in 2012 (see Bos et al. 2014).

Table 3. Samples selected for MTBC Genome Enrichment and Sequencing

Sample	**Archaeological Context**	**Date**	**Included in final analyses** [a]
AD12	Cheyenne River Village (39ST1), Arikara, South Dakota	1750 - 1775 CE	Yes
AD114	Highland Park cemetery , New York	1826 - 1863 CE	No
AD127	Highland Park cemetery , New York	1826 - 1863 CE	No
AD128	Highland Park cemetery , New York	1826 - 1863 CE	Yes
AD340	St. Michael, Alaska	1643 - 1953 CE [b]	Yes
AD344	Old Hamilton, Alaska	1681 - 1950 CE [b]	Yes
AD346	Pilot Station, Alaska	Not dated, but post-contact	No
AD351	Ekwok, Alaska	1679 - 1950 CE [b]	Yes

[a] Indicates whether enough coverage of the genome was obtained to include it for further analyses

[b] Date estimated using radiocarbon dating has a wide range, but these samples are unlikely to belong to the 20th century

MTBC whole-genome capture probe design

AT MPI-SHH, synthetic oligonucleotide probes were designed using the hypothetical MTBC ancestor genome (Comas et al. 2010) as a reference. The probes were 60 base pairs (bp) in length and had a tiling density of 5 bp. Low complexity regions were masked using DustMasker (Morgulis et al. 2006) from BLAST 2.2.31+ using standard parameters. Probes with more than 20% of masked nucleotides as well as repetitive and duplicate probes were removed, resulting in 852,164 unique probes. By randomly sampling probes, the probe set was enlarged to 968,000 probes so as to obtain the maximum number of probes that can be included on an Agilent 1,000,000-feature array.

MTBC whole-genome capture and sequencing

The UDG-treated and non-UDG treated libraries of the eight samples and corresponding library blanks were enriched for the *M. tuberculosis* genome using the aforementioned probe set and following the protocol given in Fu et al. (2013). Enriched libraries were sequenced using the Illumina HiSeq 2500 (2×76bp run) and the sequence data were analyzed using EAGER. Preliminary analyses suggested that only five samples showed enough coverage of the MTBC reference genome to be useful for further analyses. The UDG-treated enriched libraries for these samples were re-sequenced using the Illumina HiSeq 2500 (1×76 bp run) so as to obtain a targeted coverage of approximately 20-fold. Preliminary analyses of the sequence data using EAGER suggested that the two of the libraries would benefit from a further round of sequencing, and hence, these were sequenced on another 1×76 bp run on the Illumina HiSeq 2500.

Data analyses

Data from the UDG-treated enriched libraries for all five samples were preliminarily analyzed using EAGER. Reads were mapped using bwa with default parameters (except n = 0.1), duplicates were removed, and SNPs were called using the Unified Genotyper in the Genome Analysis Toolkit (GATK) (McKenna et al. 2010) with the following parameters: minimum reads covering the position = 5, minimum quality = 30, and minimum frequency to call a homozygous SNP = 0.9. SNP allele frequency histograms plotted using RStudio (R Core Team 2017) showed an unexpectedly high number of heterozygous SNPs for all samples, likely due to mis-mapping from non-MTBC environmental mycobacteria (Appendix G: Figure S4).

Filtering for MTBC-specific reads

To remove reads likely to belong to non-MTBC mycobacteria prior to mapping, a custom-designed filtering program called FINGERPRINT was used (Rosenberg 2016). This program uses *k*-mer composition profiling of the desired target (MTBC) and likely contaminants (non-MTBC mycobacteria) to score individual reads on a scale ranging from -100 to 100, based on how its *k*-mer composition compares to the target and contaminant datasets. Reads with positive values are more likely to belong to the target dataset.

For the final analyses, data from the UDG-treated libraries obtained across all Illumina runs were used. The raw sequence reads were trimmed using AdapterRemoval v2 (Schubert, Lindgreen, and Orlando 2016) with the following parameters: --trimns, --trimqualities, --minquality 20, and --minlength 30. For the paired-ended data, reads were

merged using a minimum overlap for merging equal to 11. Merged reads from the paired-ended run as well as processed reads from the single-ended runs were concatenated together. The concatenated dataset was filtered using FINGERPRINT so as to retain only those reads which scored ≥ 50. Filtered reads were mapped to the MTBC ancestor reference genome using bwa with default parameters except n = 0.1. SAMtools v0.1.19 was used to filter mapped reads at a Phred quality threshold of Q37 and to remove duplicates and reads with multiple mappings. Qualimap v2.2.1 (Garcia-Alcalde et al. 2012) was used to determine the mean coverage and the percentage of the reference covered ≥ five-fold.

Variant calling

An mpileup file was generated using SAMtools and VarScan v2.3.9 (Koboldt et al. 2012) was used to produce a VCF file containing all sites (variant as well as invariant) using the following parameters: minimum number of reads covering the position = 5, minimum of reads covering the variant allele = 3, minimum variant frequency = 0.2, minimum frequency to call a homozygous variant = 0.9, minimum base quality = 30, and maximum frequency of reads on one strand = 90%. VCFtools (Danecek et al. 2011) was used to remove insertion-deletions (InDels) and exclude positions which occurred in known repeat regions, insertion and mobile elements, phage-related genes, PE, PPE and PGRS genes, muturase and resolvase genes, REP family genes, and tRNAs and rRNAs (Appendix F: Table S7). The SelectVariants tool in GATK (McKenna et al. 2010) was used to output a VCF file containing positions comprising SNPs. Finally, the numbers of homozygous and heterozygous SNPs were determined for each sample. The

SNP allele frequency histograms for the filtered dataset are given in Appendix G: Figure S5. A comparison of the mapping statistics for the unfiltered and filtered datasets is given in Appendix F: Table S8, whereas the final summary of mapping statistics for the five North American ancient MTBC genomes given in Appendix F: Table S9.

Determining MTBC lineages

BAM files containing analysis-ready reads were visually inspected using Geneious R7 for the presence of MTBC lineage-defining SNPs as given in Coll et al. (2014). All five strains were found to belong to *M. tuberculosis* Lineage 4 (L4) based on the presence of the *pks15/1* deletion specific to this lineage (Marmiesse et al. 2004). Strains were further classified into sublineages of L4 based on the presence of sublineage-defining SNPs (Stucki et al. 2016).

Comparative data

Since the post-contact era North American strains were found to belong to L4, *M. tuberculosis* strains representing all known L4 sublineages were used for the phylogenetic analyses. In order to test hypotheses regarding the origins of the post-contact era Alaskan TB strains, such as possible introduction and dispersal via Russia, modern L4 strains originating from this region as well as strains representing other countries but belonging to similar sublineages as these five strains were highly represented in this dataset. The list of strains used in the analyses is given in Appendix F: Table S10.

For Illumina datasets, reads were processed using AdapterRemoval v2 and mapped to the MTBC reference genome using bwa with mapping stringency n = 0.1. For finished genomes, FASTA files were aligned to the MTBC reference genome using LAST with the gamma-centroid option (Kiełbasa et al. 2011). The maf-convert program was used to covert the alignment file to a SAM file and SAMtools was used obtain the BAM files. For the BAM files obtained after processing the Illumina datasets, an mpileup file was generated using SAMtools and processed using VarScan v2.3.9 (34) using the aforementioned parameters. For the finished genomes, SAMtools (v1.3.1) mpileup and bcftools call were used to produce the VCF files.

VCF files for all genomes used for the analysis were combined using the CombineVariants tool in GATK. VCFtools was used to remove Insertion-Deletions (InDels), tri-allelic sites, and exclude positions with more than 5% missing data and those given in Appendix F: Table S7. The SelectVariants tool in GATK was used to output a VCF file containing the sites where at least one of the strains has a SNP. A perl script was used to generate a FASTA alignment which included only homozygous SNPs (Bergey 2012).

Phylogenetic analyses

Trees were built using the Maximum Likelihood (ML), Maximum Parsimony (MP), and Neighbor Joining (NJ) methods using MEGA7 (Kumar, Stecher, and Tamura 2016) and a Bayesian approach using BEAST v1.8.4 (Drummond et al. 2012). The SNP alignment of all L4 genomes comprised 9,775 variable sites. The ML tree was generated using the General Time Reversible (GTR) model and 100 bootstrap replicates. The MP

tree was built using the Subtree-Pruning-Regrafting (SPR) algorithm and 500 bootstrap replicates. The NJ tree was built using the *p*-distance method and 500 bootstrap replicates.

To assess whether there was a sufficient temporal signal in the data to proceed with molecular clock analysis, a regression of root-to-tip genetic distance against dates was conducted using TempEst (Rambaut et al. 2016). Calibrated radiocarbon dates of the ancient samples and isolation dates of the modern *M. tuberculosis* strains were used as given in Appendix F: Table S10. The NJ tree was used as input for TempEst. The R^2 value calculated in TempEst equaled 0.4541 suggesting a positive correlation between genetic divergence and time for the L4 strains (Figure 8). The data were thus concluded to be adequate for molecular clock analysis. The likelihood ratio test in MEGA was used to determine whether the null hypothesis of a single molecular clock across all branches was supported. The null hypothesis of equal evolutionary rate throughout the tree was rejected at a 5% significance level (P = 0).

Estimating divergence times using BEAST

To determine divergence times of strains, a SNP alignment comprising 8,984 SNPs across the *M. tuberculosis* L4 strains was analyzed using BEAST v1.8.4 (39). The calibrated radiocarbon dates of the ancient samples in years before present (YBP, with present being considered as 2017) and the mean isolation years of the modern strains were used as priors. Samples AD340, AD344, and AD351 were excavated by Ales Hrdlicka between 1926-1938 and were not fresh graves at that time (Hrdlicka 1943), and hence, it is unlikely that they date to the 20th century. Therefore, the lower limit for the

date of these samples was constrained to 100 YBP. Using jModelTest2 (Darriba et al. 2012), the TVM model was determined to be the best model of nucleotide substitution.

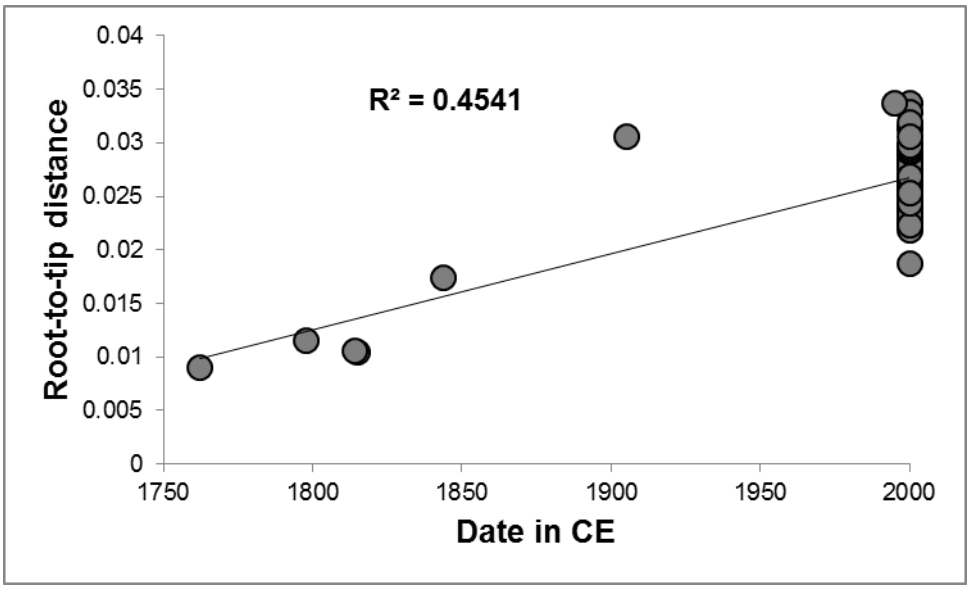

Figure 8. Linear Regression of Time vs Root-to-tip Distance for *M. tuberculosis* Lineage 4 Strains. The x-axis denotes the date of isolation (in CE) and the y-axis denotes the root-to-tip genetic distance as calculated by TempEst.

To account for ascertainment bias that might result from using only variable sites in the alignment, the number of invariant sites (number of constant As, Cs, Ts, and Gs) was included in the analysis. Lastly, an uncorrelated lognormal clock with a fixed substitution rate of 4.6×10^{-8} substitutions per site per year as estimated by Bos et al. (2014) was used as a prior. Since the modern *M. tuberculosis* strains are known to have undergone a population expansion (Bos et al. 2014; Kay et al. 2015), the analysis was conducted using a 10-step Bayesian Skyline demographic model. One Markov Chain Monte Carlo (MCMC) run was conducted at 100,000,000 iterations sampling every

10,000 steps. The first 10,000,000 iterations were discarded as burn-in. Tracer (Rambaut et al. 2015) was used to visualize the results of the run. TreeAnnotator (Drummond et al. 2012) was used to summarize the information onto a single target tree calculated in BEAST, with the initial 2,500 trees being discarded as burn-in. Figtree (http://tree.bio.ed.ac.uk/software/figtree/) was used to visualize the Maximum Clade Credibility (MCC) tree with median heights.

4.4 Results

Screening for MTBC DNA

qPCR assays

A total of 55 samples were screened using the qPCR assays. Out of these, 15 tested positive for the rpoB2 assay, 13 for the IS6110 assay, and 12 for the IS1081 assay (Appendix F: Table S6). A total of 13 extracts tested positive for more than one assay. None of the extraction blanks tested positive for any assay, signifying that no contamination had been introduced during the extraction process. A total of 18 samples which tested positive for one or more assays were selected for the in-solution MTBC gene capture.

In-solution MTBC gene capture

Libraries for these 18 samples along with those for six dental calculus libraries were screened using the in-solution gene capture. Four samples showed greater than 50% coverage of all five genes (Appendix F: Table S6). One sample did not pass the cut-offs for all genes, but showed > 60% coverage of the MTBC-specific *mtp4*0 element. All five

samples were selected for MTBC-genome enrichment. Three samples from the Highland Park cemetery, which had been screened using the in-solution capture in 2012 (see Bos et al. 2014), were also selected for MTBC-genome enrichment. Thus, a total of eight samples were selected for whole-genome enrichment and sequencing (refer Table 3). None of the dental calculus samples showed any coverage of the MTBC-specific genes *katG* and *mtp40* and were not selected for further study. The library blanks showed negligible amounts of reads mapping to the MTBC reference genome (≤ 0.1% coverage).

Shotgun sequencing and MALT analysis

The data for shotgun-sequenced UDG-treated and non-UDG treated libraries were analyzed using MALT and the reads were mapped to the MTBC ancestor reference genome to determine the percentage of endogenous DNA prior to target enrichment (Table 4). Shotgun sequencing of the UDG-treated libraries showed that the endogenous MTBC DNA content ranged from 0.032 to 0.1% signifying low prevalence of endogenous MTBC DNA and confirming the necessity of whole-genome enrichment for the *M. tuberculosis* genome. According to MALT, the number of reads assigned to the MTBC for the samples ranged from 264 to 5,424. Samples AD12 and AD346 were determined to be borderline positive for the MTBC, whereas AD340, AD344, and AD351 were determined to be strong positives for the MTBC.

Table 4. Mapping Statistics and MALT Analysis for Shotgun-Sequenced Libraries

Sample Name [a]	Raw reads	Reads post-trimming	Reads assigned to MALT	Mapped reads	Analysis-ready reads [b]	Endogenous DNA (%)
UDG-treated libraries						
AD12U	2,666,810	2,594,797	264	1,549	769	0.032
AD340U	3,997,627	3,884,932	3,011	4,394	3,326	0.088
AD344U	4,095,200	3,944,858	5,424	6,902	5,659	0.146
AD346U	4,058,362	3,899,790	476	3,832	1,435	0.039
AD351U	2,429,317	2,363,674	2,080	2,829	2,288	0.100
Non-UDG treated libraries						
AD12nU	4,711,737	4,587,951	402	5,240	2,643	0.063
AD340nU	4,052,820	3,929,877	2,769	5,551	4,010	0.106
AD344nU	501,152	485,290	514	901	686	0.143
AD346nU	4,199,809	4,057,892	408	8,773	4,258	0.111
AD351nU	4,690,218	4,558,046	4,308	8,170	5,815	0.139

[a] U denotes UDG-treated library and nU denotes non-UDG treated library

[b] Number of mapped reads retained after filtering at threshold Q37 and removing duplicates

MTBC-genome capture, sequencing, and data analyses

Authentication of ancient DNA

In order to authenticate the presence of ancient MTBC DNA, the sequence data for the enriched non-UDG treated libraries were analyzed. The number of analysis-ready reads varied from 12,458 to 261,938 (Table 5).

Table 5. Mapping Statistics for Non-UDG Treated Enriched Libraries

Sample	Raw reads	Reads post-trimming and merging	% of reads kept after processing	Mapped reads	% of reads mapped	Analysis-ready reads [a]	Average length of reads
AD12	4,406,615	4,232,561	96.06	49,764	1.18	22,939	73.56
AD128	3,313,973	3,154,935	95.21	40,707	1.3	12,458	67.19
AD340	2,886,249	2,800,507	97.03	196,175	7.01	155,664	72.62
AD344	2,292,735	2,206,367	96.24	351,126	15.92	259,320	74.72
AD351	2,944,131	2,852,249	96.88	318,693	11.18	261,938	65.99

[a] Reads obtained after filtering at Q37 quality threshold, removing duplicates, and removing reads with multiple mappings

All samples showed characteristic ancient DNA patterns such as shorter average read lengths (65-74 bp) and a bias for purines before the start of reads (Briggs, Stenzel, and Johnson 2007). Sample AD128 also showed the characteristic 5' C-to-T and corresponding 3' G-to-A mis-incorporations (Green et al. 2009; Krause et al. 2010) (Figure 9), whereas the other samples did not. The damage plots for samples AD12, AD340, AD344, and AD351 are given in Appendix G: Figures S6 - S9. The absence of C-to-T mis-incorporations in the context of mycobacterial DNA has been observed in genomes retrieved from medieval-era individuals with leprosy (Schuenemann et al. 2013) and 18th-century Hungarian individuals with tuberculosis (Kay et al. 2015); the lipid-rich mycolic acids in the cell walls of mycobacteria seem to protect the DNA from post-mortem hydrolytic damage.

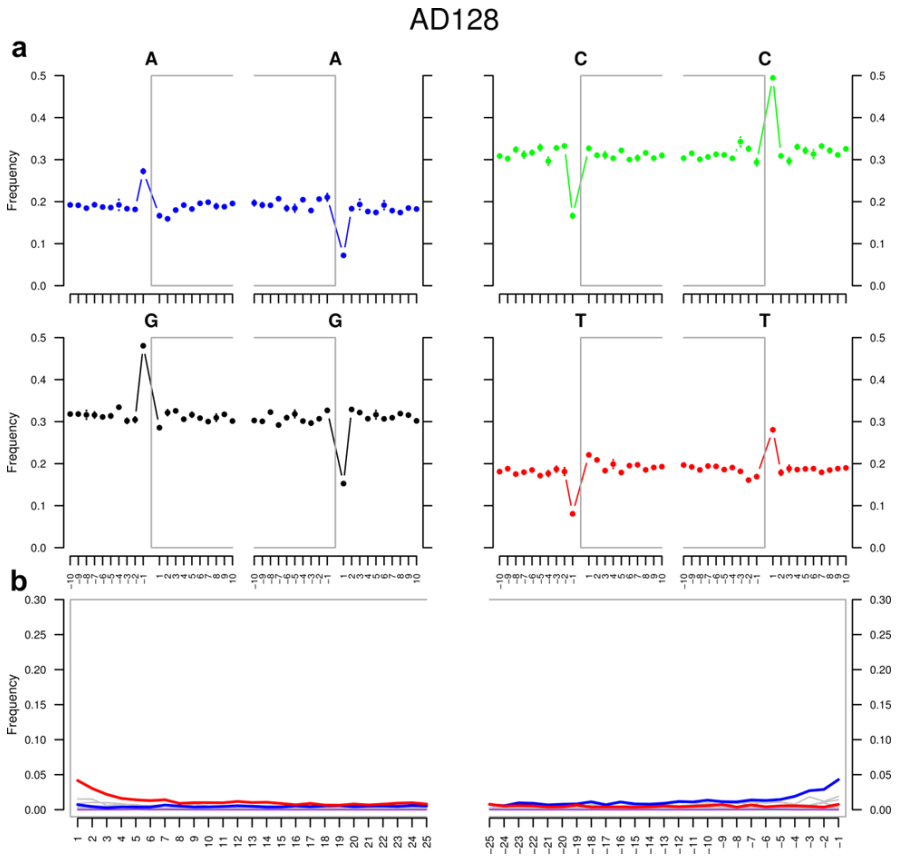

Figure 9. DNA Damage Patterns for AD128 (Enriched Library). (a) Average base frequencies at positions within individual reads (grey box) flanked by all calls from reads in neighboring sequences. (b) Frequencies of specific base substitutions at specific positions near the 5'-end (left) and 3'-end (right) occurring within reads. C-to-T changes are indicated by the red line and corresponding G-to-A changes by the blue line.

Analysis of enriched UDG-treated sequence data

Of the eight libraries that were enriched for the MTBC genome and sequenced on a paired-ended run, only five showed sufficient coverage to be considered for further analyses. These five libraries (AD12, AD128, AD340, AD344, and AD351) were deeply-sequenced to obtain mean coverage which ranged from 5- to 26-fold.

Initial mapping analyses of the data conducted using EAGER revealed that all samples showed a high number of heterozygous SNPs ranging from 190 to 2,055 (Appendix F: Table S8, Unfiltered data). Filtering the dataset for reads with FINGERPRINT scores of ≥ 50 decreased the numbers of heterozygous SNPs to realistic numbers and thus, this filtered dataset was used in further analyses (Appendix F: Table S8, Filtered data).

The number of overall SNPs in the final dataset ranged from 398 to 641. The least number of SNPs were observed in strain AD128; however, the percentage of the AD128 genome covered ≥ five-fold (which is the minimum coverage required to call a SNP) was only 53%. The sample with the highest percentage coverage at ≥ 5X was AD351 (92.9%).

Determining MTBC lineages

The five ancient North American MTBC strains belong to the human-adapted *M. tuberculosis* L4 (Euro-American lineage), based on the presence of the L4-specific *pks15/1* deletion. Additionally, strain AD12 was found to contain the $DS6^{Quebec}$ deletion (H37Rv coordinates: 1,987,457-1,998,849) that is found in modern *M. tuberculosis* strains in Quebec and other parts of Canada (Pepperell et al. 2011; Nguyen et al. 2004). Table 6 gives the L4 sublineages for these five strains.

Table 6. L4 Sublineages of Post-ontact era North American *M. tuberculosis* strains

Strain	Archaeological context	L4 Sublineage
AD12	Cheyenne River Village, South Dakota	4.4 ($DS6^{Quebec}$)
AD128	Highland Park, New York	4.10 (H37Rv-like)
AD340	St. Michael, Alaska	4.2.1 (Ural)
AD344	Old Hamilton, Alaska	4.5
AD351	Ekwok, Alaska	4.2.1 (Ural)

Phylogenetic analyses

The ML phylogeny of *M. tuberculosis* L4 strains is shown in Figure 10. Strain AD128 (Highland Park cemetery, New York) was found to be closely related to strain H37Rv. Strain AD12 (Cheyenne River Village, South Dakota) was found to belong the $DS6^{Quebec}$ sublineage of L4 strains and is closely related to strains from Canada as well as Russia. Strains AD340 and AD351 were found to belong to L4.2.1 (the Ural sublineage) and were closely also related to modern *M. tuberculosis* strains of Russian origin. Strain AD344 was found to belong to a separate branch of sublineage L4.5. These relationships were supported by the tree topologies shown by the MP and NJ method (Appendix G: Figures S10 and S11).

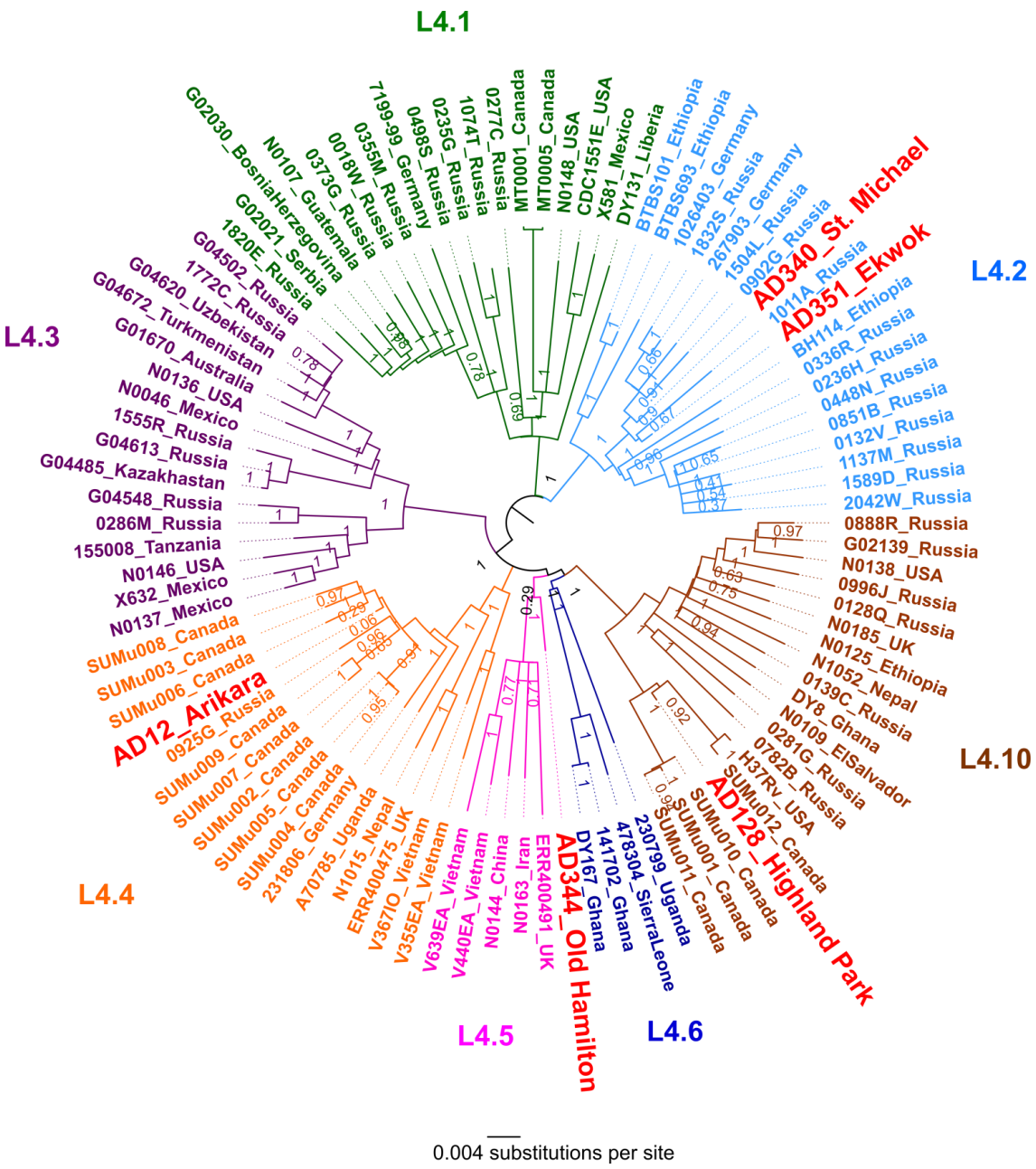

Figure 10. Maximum Likelihood Tree of 98 *M. tuberculosis* Lineage 4 Strains. The tree was based on 9,775 variable sites and built using the GTR model. Bootstrap support estimated from 100 replicates is given near the branches. The L4 sublineages are color-coded and the strains generated in this study are denoted in red. Geographic origin is given next to each strain.

Dating analysis

Based on the BEAST analysis, strain AD12 (Cheyenne River Village, Arikara) diverged from the clade comprising three modern Canadian-origin L4.4 strains (SUMu003, SUMu006, and SUMu008) 358 YBP (324-383 YBP 95% HPD). The MRCA of all L4.4 strains containing the $DS6^{Quebec}$ deletion dates to 791 YBP (735-844 YBP 95% HPD). Strain AD128 (Highland Park cemetery, New York) diverged from the branch comprising strains H37Rv and SUMu012 approximately 342 YBP (293-388 YBP 95% HPD). Strain AD344 belongs to a separate branch of sublineage L4.5 and diverged from the other strains in this sublineage about 967 YBP (913-1027 YBP 95% HPD). Two of the Alaska strains, strain AD340 (St. Michael) and AD351 (Ekwok) lie within L4.2.1 and last shared a common ancestor about 521 YBP (487-559 YBP 95% HPD). Lastly, the MRCA of all L4 strains was estimated to exist 1347 YBP (1288-1403 YBP 95% HPD), which is fairly in agreement with previous estimates of the MRCA for this lineage (Kay et al. 2015).

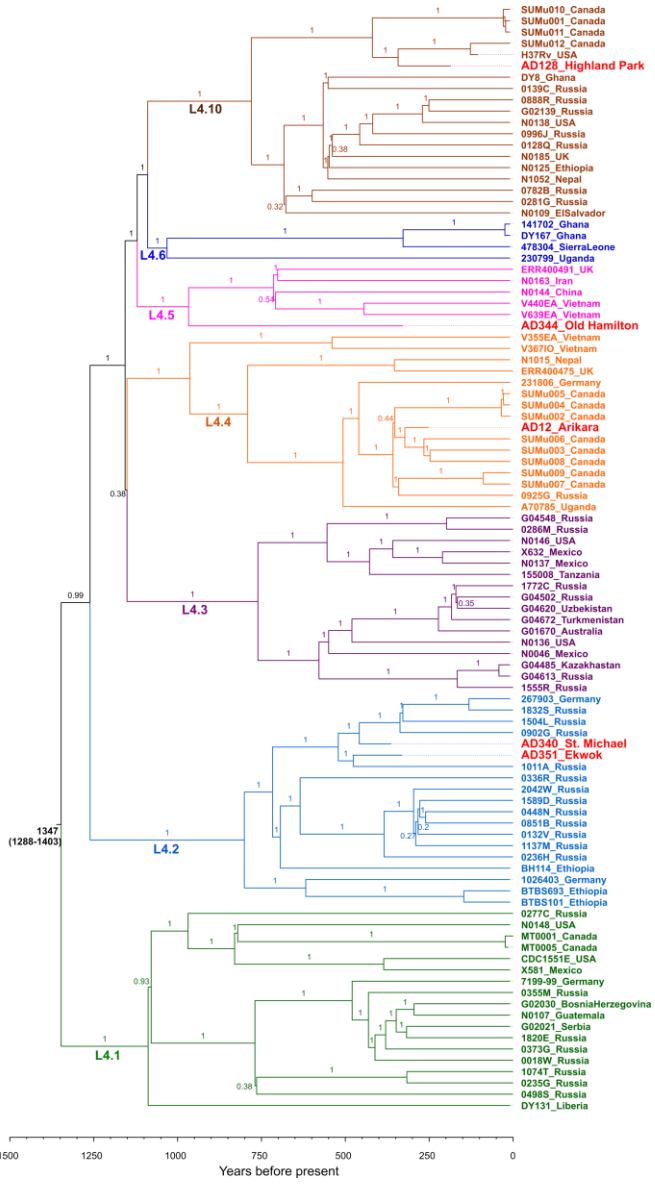

Figure 11. Maximum Clade Credibility Tree of 98 *M. tuberculosis* Lineage 4 Strains. The tree is based on 9,775 variable nucleotide positions. Branches are color-coded based on sublineages. Geographic origin of each strain is given next to the name of the strain. Names of the genomes sequenced as part of this study are given in red. Posterior probability values are shown on the appropriate branches. The median estimate of the MRCA of all L4 strains and the 95% HPD are also shown in years before present (with present being considered as 2017).

4.5 Discussion

The aim of this study was to recover ancient MTBC genomes from North American pre- and post-contact era archaeological sites to determine the type of pre-contact MTBC lineages prevalent in this region and to assess how rapidly these were replaced by European-origin *M. tuberculosis* strains. Despite screening 66 individuals, none of the pre-contact era samples showed sufficient MTBC DNA preservation to enable genome reconstruction. However, nearly-complete *M. tuberculosis* genomes were recovered from five post-contact era individuals. These ancient *M. tuberculosis* strains were found to belong to *M. tuberculosis* L4 (Euro-American lineage) confirming they were brought to the Americas after European contact.

Archaeological context of the post-contact era North American *M. tuberculosis* strains

Cheyenne River Village, South Dakota

Individual AD12 belonged to the Cheyenne River Village (39ST1) site, in Arikara, South Dakota (Figure 12). European contact with the Arikara has been documented from 1706 CE onwards (Jantz and Owsley 1994). Initially, the Arikara benefitted from European contact due to establishment of trade relations leading to increased prosperity; however ultimately, disease transmission from Europeans to the Arikara decimated their numbers (Lawrence et al. 2010). Several disseminated skeletal TB cases have been identified from Arikara sites (Palkovich 1981).

Out of these, the Cheyenne River Village site is located on the right bank of the Missouri River in South Dakota and has been attributed to the later Bad River 2 phase of the post-contact Coalescent tradition (1770-1790 CE) (Jantz 1972).

Highland Park, New York

Individual AD128 belonged to the skeletal collection excavated from Highland Park cemetery in Rochester, New York (Figure 12). The cemetery comprised the burials of European-origin individuals of low socio-economic status who died in the Monroe County Poorhouse between 1826 and 1863 CE (Steegman 1991).

Figure 12. Map showing the Cheyenne River Village and Highland Park Archaeological Sites.

Native Alaskan archaeological sites

Individuals AD340, AD344, and AD351 belonged to the Native Alaskan post-contact era sites of St. Michael, Old Hamilton, and Ekwok, respectively (Figure 13). St.

Michael is located on the east coast of St. Michael Island in Norton Sound. It was the northernmost Russian settlement in Alaska and comprised a Russian trading post that was built in the 19th century (Griffin 1996). The site of Old Hamilton is located near St. Michael and was an Eskimo village (called Aunguamut). It served as a landing area and supply station for the early riverboats (Griffin 1996; Orth 1971). Lastly, Ekwok is located farther away from the coast, along the Nushagak River, and is the oldest continuously occupied Yup'ik Eskimo village on the river.

Figure 13. Map showing the Locations of St. Michael, Old Hamilton, and Ekwok in Alaska.

Radiocarbon dating analyses

The Cheyenne River Village site has been dated to 1750-1775 CE (Jantz and Owsley 1994). The Highland Park cemetery was used for burials between 1826-1863 CE (Steegman 1991). As these sites have compact date ranges, samples AD12 and AD128

were not radiocarbon dated. Instead, the dates mentioned here were directly used in the dating analyses.

Vertebral body fragments of AD340, AD344, and AD351 were sent to Beta Analytic Inc., for radiocarbon dating and nitrogen stable isotope analyses. However, these samples posed a problem for radiocarbon dating because the diet of the individuals at these sites had a large marine component comprising marine mammals, birds, ocean and anadromous fish, and marine invertebrates (McCartney and Veltre 1999). The marine diet has been associated with inaccuracies in radiocarbon dating from skeletal elements due to the 'Old Carbon' effect (Stuiver, Pearson, and Braziunas 1986). This occurs due to upwelling of ^{14}C-depleted deep water which can result in marine organisms having ^{14}C ages 600-1000 years older than the apparent ages of terrestrial material. A regional correction (ΔR) was estimated using the proportion of marine food in the diet for all three individuals, as given in (Arneborg et al. 1999). OxCal v4.3.1 (Bronk Ramsey 2009) was used to determine a mixed marine/terrestrial calibration curve based on these proportions and to calibrate the radiocarbon dates (Table 7).

Table 7. Radiocarbon Dating Analyses for Alaskan Samples

Sample No.	Site	d15N	d13C	Conventional radiocarbon age	ΔR	Marine protein contribution (%)	Dates cal AD (95% probability)
AD340	St. Michael	+18.4	-15.9	640 ± 30 BP	486 ± 54 years	60	1643-1953
AD344	Old Hamilton	+15.8	-15.5	470 ± 30 BP	486 ± 54 years	64.7	1681-1782 (28%) and 1793-1950 (68%)
AD351	Ekwok	+13.4	-16.9	390 ± 30 BP	242 ± 50 years	48	1679-1950

For all three samples, bone elements were used for radiocarbon dating.

Distribution of L4 sublineages in post-contact era and modern North America

Modern L4 sublineages in North America

Currently, the majority of *M. tuberculosis* L4 strains prevalent in the US belong to sublineages L4.1.2, L4.3, and L4.10, whereas sublineages L4.1.1 and L4.4 are found at intermediate frequencies, and L4.5 is found very rarely (Stucki et al. 2016). In Canada, sublineages L4.1.2, L4.3, and L4.10 are dominant, whereas L4.1.1 and L4.4 are prevalent at intermediate frequencies, and L4.2 and L4.6 are found very rarely (Stucki et al. 2016; Lee et al. 2015; Pepperell et al. 2011). Recent studies have shown that the Canadian fur trade between 1710 to 1870 is known to have caused the spread of *M. tuberculosis* $DS6^{Quebec}$-type strains (Nguyen et al. 2004; Nguyen et al. 2003) from European fur traders to the Aboriginal populations living in Ontario, Saskatchewan, and Alberta (Pepperell et al. 2011).

Prevalence of H37Rv-like M. tuberculosis strains in 19th century New York

Strain AD128 (Highland Park) belongs to L4.10, which is one of the major sublineages prevalent in North America today and also has a widespread dispersal all over the world. Within L4.10, strain AD128 is closely related to strain H37Rv and the two strains diverged about 342 YBP. Strain H37Rv was first isolated in 1905 from a patient in the US (Steenken and Gardner 1946). PCR-based analyses have suggested that H37Rv-like strains were prevalent in continental Europe in the 16th-18th centuries and were geographically dispersed in the 18th-19th centuries in Britain (Müller, Roberts, and Brown 2014). Thus, our analyses suggest that H37Rv-like strains were brought to the US sometime in the past 350-400 years and may have been prevalent among European-origin individuals especially belonging to the lower socioeconomic classes.

Introduction of DS6Quebec-lineage M. tuberculosis strains to Arikara populations due to the fur trade

Strain AD12 (Cheyenne River Village, Arikara) belongs to L4.4, which is found in intermediate frequencies in North America, but more frequently in China, southeast Asia, and Australia as well as in certain countries in Europe and Africa (Stucki et al. 2016). Within this sublineage, strain AD12 is closely related to modern L4.4 strains containing the DS6Quebec deletion; these include strains from Canada, Russia, Germany, UK, Nepal, and Uganda. L4.4 also comprises strains from Vietnam; however, these do not show the DS6Quebec deletion (Figure 14). The DS6Quebec strains are thought to have been introduced to Quebec, Canada, by European fur traders in the 17th-18th centuries. Further dispersal of these strains to Aboriginal populations in Canada likely occurred

between 1730-1870 due to the westward expansion of the fur trade and through social contact (Pepperell et al. 2011). Our results suggest that these DS6^{Quebec} strains were also introduced to the Arikara peoples of South Dakota by the mid-18th century likely through contact with European fur traders.

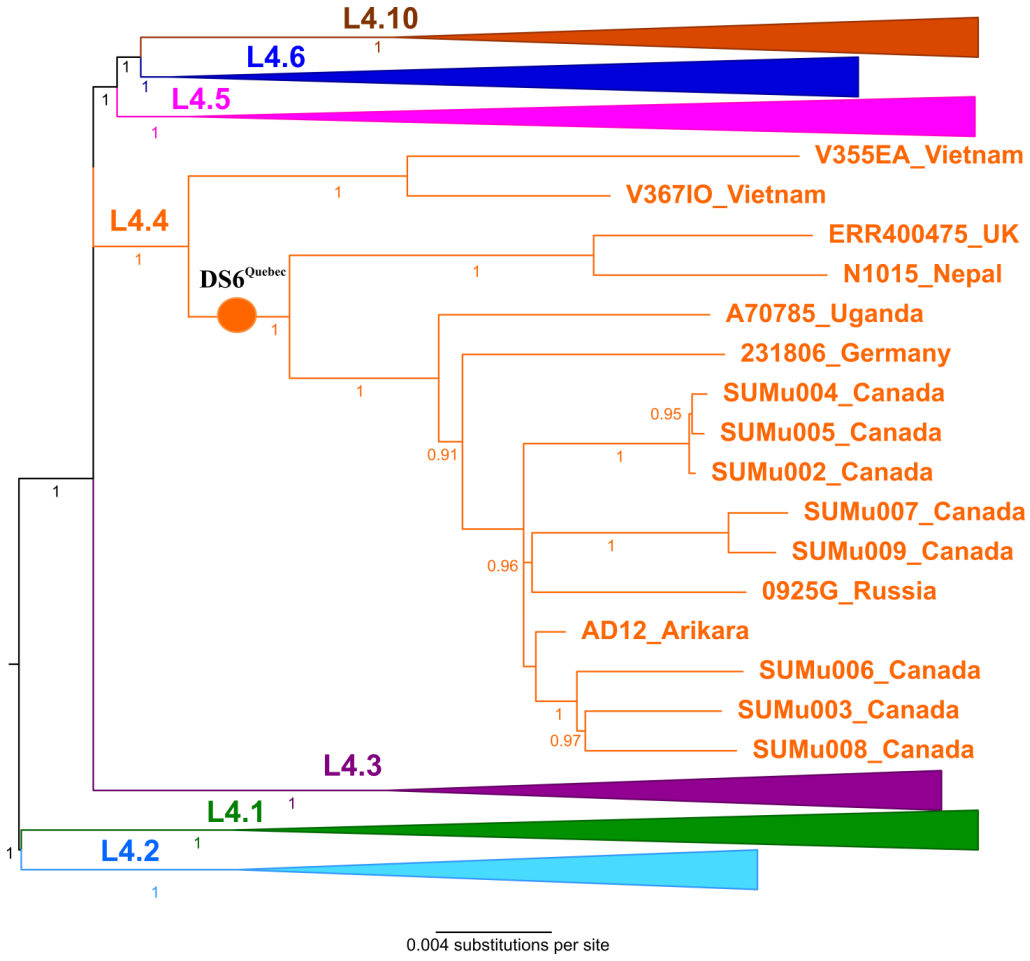

Figure 14. Position of the DS6^{Quebec} Deletion within the *M. tuberculosis* Lineage 4 Strains. The deletion is found in strains belonging to sublineage L4.4 as defined by Coll et al. (2014). The ML tree shown here is the same as that in Figure 10, except that certain sublineages have been collapsed to save space.

Introduction of Russian M. tuberculosis strains to Native Alaskan populations

Strains AD340 (St. Michael) and AD351 (Ekwok) belong to sublineage L.4.2.1 (known as the Ural sublineage) and are closely related to modern strains of Russian origin. A recent study of nearly 1000 modern Russian *M. tuberculosis* strains revealed that the L4 sublineages prevalent in the country include L4.1.2, L4.3, L4.2.1, and L4.4.1.1 (Casali et al. 2014). L.4.2.1 strains are mostly prevalent in modern-day Russia and China and to a lesser extent in certain countries in Africa and Europe (Stucki et al. 2016).

Russian contact with the Native Aleut began around 1741 (Smith and Veltre 2010) and expanded considerably due to the demand for fur and sea otter pelts. The fur trade played a vital role in the development of Siberia, the Russian Far East and the Russian colonization of the Americas. Tuberculosis deaths among the Aleut were documented as early as 1770 (Fortuine 1989; Fortuine 2005). Strains AD340 and AD351 last shared a common ancestor about 500 years ago. Thus, our analyses support the introduction of Russian-origin *M. tuberculosis* strains were introduced to the Alaskan populations in the latter half of the 18th century.

The AD344 strain (Old Hamilton) belongs to sublineage L4.5 but lies on its own branch within this sublineage. Interestingly, the Old Hamilton and St. Michael strains are not closely related, despite the geographic proximity of these sites. Strains from sublineage L4.5 have been isolated from Middle Eastern and East Asian countries including Iran, China, and Vietnam, but this sublineage is rarely found in the Americas or Russia (Stucki et al. 2016).

However, only a few L4.5 genomes have been sequenced. Genome data for more L4.5 strains may help clarify the origins and phylogenetic relationships of strain AD344 within L4.5.

M. tuberculosis strains in post-contact era and modern-day Alaska

Interestingly, all post-contact era Native Alaskan L4 strains belong to sublineages that are not commonly found in North America today. Today, Alaska ranks first among the US states in terms of TB incidence. In 2015, the number of reported TB cases was 68, which is equivalent to a case rate of 9.2 per 100,000 individuals and is significantly higher than the rest of the US (3 per 100,000) (Department of Public Health and Social Services, State of Alaska, 2015). Modern *M. tuberculosis* strains from Alaska have been genotyped using epidemiological techniques used for detection of TB cases, such as using spoligotyping and MIRU-VNTR. However, to the best of our knowledge, these modern Alaskan TB strains have not been classified using the SNP-based typing scheme.

Spoligotype and MIRU-VNTR genotype data cannot be used to ascertain the SNP-based sublineages due to convergence of spoligotypes and MIRU-genotypes across different sublineages (Kay et al. 2015; Anderson et al. 2013). Since MTBC strains exhibit high genetic identity and very less horizontal gene exchange, SNP homoplasies are extremely rare; hence, SNPs are the ideal phylogenetic markers for pathogens such as the MTBC (Stucki et al. 2016; Comas et al. 2009). Therefore, this study underlies a need to obtain SNP-genotype data for *M. tuberculosis* strains currently prevalent in Alaska, especially in the geographically isolated areas, so as to determine the relationships between post-contact era and currently prevalent *M. tuberculosis* strains in this region.

4.6 Summary

This is the first study to report post-contact era *M. tuberculosis* genomes from North America and underlies the role of the fur trade in the introduction and dispersal of *M. tuberculosis* strains in the northern part of North America. This study found that Russian *M. tuberculosis* L4 strains were introduced to Native Alaskan populations in the post-contact era. Secondly, L4 strains belonging to the $DS6^{Quebec}$ lineage, which were dispersed from European fur traders in Quebec to Aboriginal populations, were also introduced to Arikara populations in South Dakota by the latter half of the 18^{th} century. Thirdly, *M. tuberculosis* strains in 19^{th} century New York were found to be of European origin and similar to the H37Rv strain that was widely prevalent in the UK. Thus, this study provides evidence for the diversity of L4 strains that were brought to North America after the 15^{th} century. Although this study could not recover pre-contact era North American MTBC genomes, the availability of these in the future coupled with the post-contact era genomes generated here, will help answer questions about the pattern and timing of the replacement of the pre-contact *M. tuberculosis* strains in North America.

CHAPTER 5

CONCLUSION

The aim of this dissertation was to answer outstanding questions regarding the evolutionary histories of the pathogens causing leprosy and TB, two diseases that have afflicted human populations for millennia and continue to be a major public health concern in developing countries (WHO 2016a, WHO 2016b). An important factor behind the continued incidence of leprosy and TB in developing countries is the persistence and propagation of the pathogens in reservoir hosts. In countries such as the US, where leprosy is not a public health concern, the majority of leprosy cases in native-born individuals, are due to zoonotic transmission from armadillos which serve as reservoir for *M. leprae* (Truman et al. 2011). Furthermore, recent research has shown that a severe form of leprosy is caused by a novel bacterial species *M. lepromatosis* (Han et al. 2008) and red squirrels serve as a reservoir for *M. leprae* and *M. lepromatosis* at least in the UK (Avanzi et al. 2016).

The aim of Chapter 2 was to assess whether nonhuman primates may serve as a reservoir for *M. leprae*. To test this hypothesis, *M. leprae* genomes from three naturally infected nonhuman primates were sequenced using whole-genome enrichment and next-generation sequencing technology. Phylogenetic analyses suggest that nonhuman primates may acquire *M. leprae* from humans as well as transmit *M. leprae* strains between themselves. A novel *M. leprae* sublineage was discovered, which might be specific to nonhuman primates in Africa. However, the lack of genomic data for human *M. leprae* strains especially from Africa, where leprosy is endemic in several countries, prevents us from conclusively determining whether this sublineage is restricted to

nonhuman primates or is also prevalent in humans. As part of this study, wild nonhuman primate populations from Madagascar and Uganda were screened for the presence of *M. leprae* and other mycobacterial pathogens; however, they tested negative. Nonetheless, this study underlies a need for conducting broad phylogeographic screenings of nonhuman primates, especially in countries where leprosy is endemic. Future studies on the prevalence of *M. leprae* and *M. lepromatosis* in nonhuman hosts have important implications for leprosy eradication and wildlife conservation strategies.

In Chapter 3, the genome of *M. lepraemurium*, the causative agent of murine leprosy, was sequenced and annotated. The aim of this study was to test the hypothesis that *M. lepraemurium*, which infects mice, rats, and cats, might be closely related to the pathogens causing human leprosy. Phylogenetic analyses confirmed that *M. lepraemurium* is not closely related to *M. leprae* or *M. lepromatosis*; rather, it is a distinct species within the MAC. Despite its lack of phylogenetic proximity to *M. leprae* and *M. lepromatosis*, *M. lepraemurium* is undergoing reductive evolution similar to that found in these two species. Reductive evolution has been thought to occur primarily due to a change in lifestyle of a microorganism such as from a free-living to a host-associated life or from a wide host range to a specific host (Gómez-Valero et al. 2007). These changes in lifestyle may produce a relaxation of the natural selection pressure, resulting in individuals accumulating detrimental or loss-of-function mutations. Based on the results of this study, it can be hypothesized that the *M. lepraemurium* progenitor underwent an evolutionary bottleneck (possibly a host switch) and after adapting to this new lifestyle started losing the functionality of the majority of genes required for survival outside of its

host. However, *M. lepraemurium* seems to have retained the functionality of most of the genes required for virulence in MAC species.

In Chapter 4, 66 individuals from the North American archaeological record showing symptoms characteristic of skeletal TB were screened for the presence of MTBC DNA. The aim of this study was to recover MTBC genomes from pre- and post-contact era North America to determine the types of pre-contact era strains present and to assess the timing of their replacement by European-origin *M. tuberculosis* L4 strains. Previous research from this laboratory showed that zoonotic transmission from pinnipeds (such as seals) introduced MTBC strains to the coastal areas of Peru during pre-Columbian times (Bos et al. 2014). The recovery of MTBC genomes from pre-contact era North American sites could not be achieved in this study, likely due to lack of preservation of MTBC DNA in these individuals. However, this study is ongoing in the laboratory and the availability of these data in future will help clarify how far the seal-derived MTBC strains were dispersed throughout the Americas as well as whether other types of MTBC strains were introduced during pre-contact times. Five post-contact era *M. tuberculosis* genomes from South Dakota, Alaska, and New York were analyzed. Phylogenetic analyses suggest *M. tuberculosis* L4 strains were introduced from multiple sources to Native Alaskan populations. The post-contact era Alaskan strains are related to modern *M. tuberculosis* strains commonly found Russia and south-east Asia. Secondly, strains belonging to the $DS6^{Quebec}$ lineage, which were introduced to Native populations of Canada by European fur traders (Pepperell et al. 2011), were also prevalent in the native populations of South Dakota. Interestingly, the post-contact era Alaskan strains do not comprise the L4 sublineages that are commonly found in Canada or the US today. Since Alaska has a

disproportionately higher number of TB cases as compared to the continental US, future work on this ongoing study could include generating whole-genome data for modern Alaskan *M. tuberculosis* strains so as to place these in a phylogenetic context with the post-contact era Alaskan strains. These data will help determine whether the post-contact era L4 sublineages continue to persist in Alaska today or whether they have been replaced by other L4 sublineages commonly found in the Americas. Lastly, *M. tuberculosis* strains in 19th century New York were found to be of European origin and similar to the H37Rv strain that was highly prevalent in the UK (Müller, Roberts, and Brown 2014). Thus, this study provides evidence for the diversity of L4 strains that were brought to North America post-contact.

In summary, this dissertation has enhanced our understanding of how the pathogens causing leprosy and TB have evolved over time. This work also helped assess the prevalence of these pathogens in nonhuman hosts, which will help identify reservoir hosts and inform us about the strategies necessary to control these diseases in highly endemic regions.

REFERENCES

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10. doi:10.1016/S0022-2836(05)80360-2.

Anderson, J., L. G. Jarlsberg, J. Grindsdale, D. Osmond, M. Kawamura, P. C. Hopewell, and M. Kato-Maeda. 2013. "Sublineages of Lineage 4 (Euro-American) Mycobacterium Tuberculosis Differ in Genotypic Clustering." *The International Journal of Tuberculosis and Lung Disease* 17 (7): 885–91. doi:10.5588/ijtld.12.0960.

Arneborg, Jette, Bullet Jan Heinemeier, Bullet Niels Lynnerup, Bullet L Henrik Nielsen, Bullet Niels Rud, and Bullet E Árný Sveinbjörnsdóttir. 1999. "Change of Diet of the Greenland Vikings Determined from Stable Carbon Isotope Analysis and 14 C Dating of Their Bones." *Radiocarbon* 41 (2): 157–68.

Arriaza, B. 1996. "Preparation of the Dead in Coastal Andean Preceramic Populations." *Human Mummies* Springer: 131–40.

Arriaza, B.T., and V.G. Standen. 2005. "Differential Mortuary Treatment among the Andean Chinchorro Fishers: Social Inequalities or In Situ Regional Cultural Evolution?" *Curr Anthropol* 46: 662–71.

Athwal, R S, S S Deo, and T Imaeda. 1984. "Deoxyribonucleic Acid Relatedness among Mycobacterium Leprae, Mycobacterium Lepraemurium, and Selected Bacteria by Dot Blot and Spectrophotometric Deoxyribonucleic Acid Hybridization Assays." *International Journal of Systematic Bacteriology* 34 (4): 371–75.

Avanzi, Charlotte, Andrej Benjak, Karen Stevenson, Victor R Simpson, Philippe Busso, Joyce Mcluckie, Chloé Loiseau, et al. 2016. "Red Squirrels in the British Isles Are Infected With Leprosy Bacilli." *Science* 354 (6313): 744–48.

Aziz, Ramy K, Daniela Bartels, Aaron A Best, Matthew DeJongh, Terrence Disz, Robert A Edwards, Kevin Formsma, et al. 2008. "The RAST Server: Rapid Annotations Using Subsystems Technology." *BMC Genomics* 9 (1): 75. doi:10.1186/1471-2164-9-75.

Banerjee, DK. 1979. "Functional Activity of T Lymphocytes in Murine Leprosy Infection." *Lepr. India* 51: 553–54.

Bastida, R, V Quse, and R Guichon. 2011. "Tuberculosis in Hunter-Gatherer Groups of Patagonia and Tierra Del Fuego: New Alternatives of Disease." *Revista Argentina de Antropología Biológica* 13 (1): 83–95.

Bastida, Ricardo, Julio Loureiro, Viviana Quse, Amelia Bernardelli, Diego Rodríguez, and Enrique Costa. 1999. "Tuberculosis in a Wild Subantarctic Fur Seal from Argentina." *Journal of Wildlife Diseases* 35 (4). Wildlife Disease Association : 796–98. doi:10.7589/0090-3558-35.4.796.

Bergey, Christina. 2012. "Vcf-Tab-to-Fasta."

Bos, K I, K M Harkins, A Herbig, M Coscolla, N Weber, I Comas, S a Forrest, et al. 2014. "Pre-Columbian Mycobacterial Genomes Reveal Seals as a Source of New World Human Tuberculosis." *Nature* 514 (7523): 494–97. doi:10.1038/nature13591.

Braun, Mark, Della Collins Cook, and Susan Pfeiffer. 1998. "DNA fromMycobacterium tuberculosisComplex Identified in North American, Pre-Columbian Human Skeletal Remains." *Journal of Archaeological Science* 25 (3): 271–77. doi:10.1006/jasc.1997.0240.

Briggs, Adrian W, Udo Stenzel, Matthias Meyer, Johannes Krause, Martin Kircher, and Svante Pääbo. 2010. "Removal of Deaminated Cytosines and Detection of in Vivo Methylation in Ancient DNA." *Nucleic Acids Research* 38 (6). Oxford University Press: e87. doi:10.1093/nar/gkp1163.

Briggs, AW, U Stenzel, and PLF Johnson. 2007. "Patterns of Damage in Genomic DNA Sequences from a Neandertal." *Proceedings of the National Academy of Sciences of the United States of America* 104 (37): 14616–21.

Britton, Warwick J., and Diana N J Lockwood. 2004. "Leprosy." *Lancet* 363 (9416): 1209–19. doi:10.1016/S0140-6736(04)15952-7.

Bronk Ramsey, C. 2009. "Bayesian Analysis of Radiocarbon Dates." *Radiocarbon* 51 (1): 337–60. doi:10.2458/azu_js_rc.v51i1.3494.

Brosch, R., S. V. Gordon, M. Marmiesse, P. Brodin, C. Buchrieser, K. Eiglmeier, T. Garnier, et al. 2002. "A New Evolutionary Scenario for the Mycobacterium Tuberculosis Complex." *Proceedings of the National Academy of Sciences* 99 (6): 3684–89. doi:10.1073/pnas.052548299.

Brumfield, R T, P Beerli, D a Nickerson, and S V Edwards. 2003. "The Utility of Single Nucleotide Polymorphisms in Inferences of Population History." *Trends in Ecology and Evolution* 18 (1): 249– 256. doi:doi:10.1016/S0169-5347(03)00018-1.

Bryant, Josephine M, Virginie C Thibault, David G E Smith, Joyce Mcluckie, Ian Heron, Iker A Sevilla, Franck Biet, et al. 2016. "Phylogenomic Exploration of the Relationships between Strains of Mycobacterium Avium Subspecies Paratuberculosis." *BMC Genomics* 17 (79). doi:10.1186/s12864-015-2234-5.

Calvignac-Spencer, S., S.A.J. Leendertz, T.R. Gillespie, and F.H. Leendertz. 2012. “Wild Great Apes as Sentinels and Sources of Infectious Disease.” *Clinical Microbiology and Infection* 18 (6). Blackwell Publishing Ltd: 521–27. doi:10.1111/j.1469-0691.2012.03816.x.

Casali, Nicola, Vladyslav Nikolayevskyy, Yanina Balabanova, Simon R Harris, Olga Ignatyeva, Irina Kontsevaya, Jukka Corander, et al. 2014. “Evolution and Transmission of Drug-Resistant Tuberculosis in a Russian Population.” *Nature Genetics* 46 (3): 279–86. doi:10.1038/ng.2878.

Cases, Ildefonso, Victor De Lorenzo, and Christos A Ouzounis. 2003. “Transcription Regulation and Environmental Adaptation in Bacteria.” *Trends in Microbiology*. Elsevier. doi:10.1016/S0966-842X(03)00103-3.

Chan, Jacqueline Z.-M., Martin J. Sergeant, Oona Y.-C. Lee, David E. Minnikin, Gurdyal S. Besra, Ilidkó Pap, Mark Spigelman, Helen D. Donoghue, and Mark J. Pallen. 2013. “Metagenomic Analysis of Tuberculosis in a Mummy.” *The New England Journal of Medicine* 369 (3): 289–90. doi:10.1056/NEJMc1302295.

Charles, Lauren, Ignazio Carbone, Keith G Davies, David Bird, Mark Burke, Brian R Kerry, and Charles H Opperman. 2005. “Phylogenetic Analysis of Pasteuria Penetrans by Use of Multiple Genetic Loci.” *Journal of Bacteriology* 187 (16). American Society for Microbiology: 5700–5708. doi:10.1128/JB.187.16.5700-5708.2005.

Chin, Chen-Shan, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, et al. 2013. “Nonhybrid, Finished Microbial Genome Assemblies from Long-Read SMRT Sequencing Data.” *Nature Methods* 10 (6): 563–69. doi:10.1038/nmeth.2474.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, Douglas M Ruden, and Xiangyi Lu. 2012. “A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain w1118; Iso-2; Iso-3.” *Fly* 6 (2): 1–13.

Clark-Curtiss, J. E., W. R. Jacobs, M. A. Docherty, and L. R. Ritchie. 1985. “Molecular Analysis of DNA and Construction of Genomic Libraries of Mycobacterium Leprae.” *Journal of Bacteriology* 161 (3): 1093–1102.

Coelho, Ana Cláudia, Maria de Lurdes Pinto, Ana Matos, Manuela Matos, and Maria dos Anjos Pires. 2013. “Mycobacterium Avium Complex in Domestic and Wild Animals.” In *Insights from Veterinary Medicine*, edited by Rita Payan-Carreira. Rijeka: InTech. doi:10.5772/54323.

Cole, S T, K Eiglmeier, J Parkhill, K D James, N R Thomson, P R Wheeler, N Honoré, et al. 2001. “Massive Gene Decay in the Leprosy Bacillus.” *Nature* 409 (6823): 1007–11. doi:10.1038/35059006.

Coll, Francesc, Ruth McNerney, José Afonso Guerra-Assunção, Judith R. Glynn, João Perdigão, Miguel Viveiros, Isabel Portugal, Arnab Pain, Nigel Martin, and Taane G. Clark. 2014. “A Robust SNP Barcode for Typing Mycobacterium Tuberculosis Complex Strains.” *Nature Communications* 5 (September). Nature Publishing Group: 4812. doi:10.1038/ncomms5812.

Collins, D.M., and D.M. Stephens. 1991. “Identification of an Insertion Sequence, IS1081, in Mycobacterium Bovis.” *FEMS Microbiol Lett* 67: 11–15.

Comas, Inaki, Jaidip Chakravartti, Peter M Small, James Galagan, Stefan Niemann, Kristin Kremer, Joel D Ernst, and Sebastien Gagneux. 2010. “Human T Cell Epitopes of Mycobacterium Tuberculosis Are Evolutionarily Hyperconserved.” *Nature Genetics* 42 (6): 498–503. doi:10.1038/ng.590.

Comas, Inaki, Mireia Coscolla, Tao Luo, Sonia Borrell, Kathryn E Holt, Midori Kato-Maeda, Julian Parkhill, et al. 2013. “Out-of-Africa Migration and Neolithic Coexpansion of Mycobacterium Tuberculosis with Modern Humans.” *Nature Genetics* 45 (10). Nature Publishing Group: 1176–82. doi:10.1038/ng.2744.

Comas, Inaki, Susanne Homolka, Stefan Niemann, and Sebastien Gagneux. 2009. “Genotyping of Genetically Monomorphic Bacteria: DNA Sequencing in Mycobacterium Tuberculosis Highlights the Limitations of Current Methodologies.” Edited by Anastasia P. Litvintseva. *PLoS ONE* 4 (11). Sinauer Associates: e7815. doi:10.1371/journal.pone.0007815.

Cooper, Alan, and H N Poinar. 2000. “Ancient DNA: Do It Right or Not at All.” *Science (New York, N.Y.)*. doi:10.1126/science.289.5482.1139b.

Coscolla, Mireia, and Sebastien Gagneux. 2014. “Consequences of Genomic Diversity in Mycobacterium Tuberculosis.” *Seminars in Immunology* 26 (6): 431–44. doi:10.1016/j.smim.2014.09.012.

Coscolla, Mireia, Astrid Lewin, Sonja Metzger, Kerstin Maetz-Rennsing, Sébastien Calvignac-Spencer, Andreas Nitsche, Pjotr Wojtek Dabrowski, et al. 2013. “Novel Mycobacterium Tuberculosis Complex Isolate from a Wild Chimpanzee.” *Emerging Infectious Diseases* 19 (6): 969–76. doi:10.3201/eid1906.121012.

Dabbs, G.R. 2009. “Resuscitating the Epidemiological Model of Differential Diagnosis: Tuberculosis at Prehistoric Point Hope, Alaska.” *Paleopathology Association Newsletter* 148: 11–24.

Dabney, Jesse, Michael Knapp, Isabelle Glocke, Marie-Theres Gansauge, Antje Weihmann, Birgit Nickel, Cristina Valdiosera, et al. 2013. “Complete Mitochondrial Genome Sequence of a Middle Pleistocene Cave Bear Reconstructed from Ultrashort DNA Fragments.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (39): 15758–63. doi:10.1073/pnas.1314445110.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, et al. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics* 27 (15). Oxford University Press: 2156–58. doi:10.1093/bioinformatics/btr330.

Darriba, Diego, Guillermo L Taboada, Ramón Doallo, and David Posada. 2012. “jModelTest 2: More Models, New Heuristics and Parallel Computing.” *Nature Methods* 9 (8). Nature Research: 772–772. doi:10.1038/nmeth.2109.

Dean, George. 1903. “A Disease of the Rat Caused by an Acid-Fast Bacillus.” *Centralbl. F. Bakteriol.* 34: 222–24.

Dean, George. 1905. “Further Observations on a Leprosy-like Disease of the Rat.” *Journal of Hygiene* 5 (1): 99–112.

Deboosère, Nathalie, Raffaella Iantomasi, Christophe J. Queval, Ok Ryul Song, Gaspard Deloison, Samuel Jouny, Anne Sophie Debrie, et al. 2016. “LppM Impact on the Colonization of Macrophages by Mycobacterium Tuberculosis.” *Cellular Microbiology*, January. doi:10.1111/cmi.12619.

Dedhia, Pratiksha, Shivraj Tarale, Gargi Dhongde, Rashmi Khadapkar, and Bibhu Das. 2007. “Evaluation of DNA Extraction Methods and Real Time PCR Optimization on Formalin-Fixed Paraffin-Embedded Tissues.” *Asian Pacific Journal of Cancer Prevention : APJCP* 8 (1): 55–59.

Demay, Christophe, Benjamin Liens, Thomas Burguière, Véronique Hill, David Couvin, Julie Millet, Igor Mokrousov, Christophe Sola, Thierry Zozio, and Nalin Rastogi. 2012. “SITVITWEB – A Publicly Available International Multimarker Database for Studying Mycobacterium Tuberculosis Genetic Diversity and Molecular Epidemiology.” *Infection, Genetics and Evolution* 12 (4): 755–66. doi:10.1016/j.meegid.2012.02.004.

Department of Health and Social Services, State of Alaska. 2015 “Tuberculosis in Alaska 2015 Annual Report.”

Devulder, G, M Pérouse de Montclos, and J P Flandrois. 2005. “A Multigene Approach to Phylogenetic Analysis Using the Genus Mycobacterium as a Model.” *International Journal of Systematic and Evolutionary Microbiology* 55 (Pt 1): 293–302. doi:10.1099/ijs.0.63222-0.

Dhama, Kuldeep, Mahesh Mahendran, Ruchi Tiwari, Shambhu Dayal Singh, Deepak Kumar, Shoorvir Singh, and Pradeep Mahadev Sawant. 2011. "Tuberculosis in Birds: Insights into the Mycobacterium Avium Infections." *Veterinary Medicine International* 2011. Hindawi Publishing Corporation: 712369. doi:10.4061/2011/712369.

Donham, K J, and J R Leininger. 1977. "Spontaneous Leprosy-like Disease in a Chimpanzee." *The Journal of Infectious Diseases* 136 (1): 132–36.

Draper, P. 1980. "Purification of Mycobacterium Leprae." In *Report of the Fifth Meeting of the Scientific Working Group on the Immunology of Leprosy, TDR/IMMLEP ‑SWG 5/80.3*.

Drummond, Alexei J., Marc A. Suchard, Dong Xie, and Andrew Rambaut. 2012. "Bayesian Phylogenetics with BEAUti and the BEAST 1.7." *Molecular Biology and Evolution* 29 (8). Oxford University Press: 1969–73. doi:10.1093/molbev/mss075.

Eiglmeier, K, H Fsihi, B Heym, and S T Cole. 1997. "On the Catalase-Peroxidase Gene, katG, of Mycobacterium Leprae and the Implications for Treatment of Leprosy with Isoniazid." *FEMS Microbiology Letters* 149 (2): 273–78.

Eisenach, K.D., M.D. Cave, J.H. Bates, and J.T. Crawford. 1990. "Polymerase Chain Reaction Amplification of a Repetitive DNA Sequence Specific for Mycobacterium Tuberculosis." *The Journal of Infectious Diseases* 161: 977–81.

Engel, Gregory A, Lisa Jones-Engel, Michael A Schillaci, Komang Gde Suaryana, Artha Putra, Agustin Fuentes, and Richard Henkel. 2002. "Human Exposure to Herpesvirus B-Seropositive Macaques, Bali, Indonesia." *Emerging Infectious Diseases* 8 (8): 789–95. doi:10.3201/eid0805.010467.

Evans, S A, A Colville, A J Evans, A J Crisp, and I D Johnston. 1996. "Pulmonary Mycobacterium Kansasii Infection: Comparison of the Clinical Features, Treatment and Outcome with Pulmonary Tuberculosis." *Thorax* 51 (12): 1248–52. doi:10.1136/thx.51.12.1243.

Fedrizzi, Tarcisio, Conor J Meehan, Antonella Grottola, Elisabetta Giacobazzi, Fregni Serpini, Sara Tagliazucchi, Anna Fabio, et al. 2017. "Genomic Characterization of Nontuberculous Mycobacteria." *Scientific Reports* 7 (March). Nature Publishing Group: 1–19. doi:10.1038/srep45258.

Firdessa, Rebuma, Stefan Berg, Elena Hailu, Esther Schelling, Balako Gumi, Girume Erenso, Endalamaw Gadisa, et al. 2013. "Mycobacterial Lineages Causing Pulmonary and Extrapulmonary Tuberculosis, Ethiopia." *Emerging Infectious Diseases* 19 (3): 460–63. doi:10.3201/eid1903.120256.

Foley, J. E., T. L. Gross, N. Drazenovich, F. Ramiro-Ibanez, and E. Anacleto. 2004. "Clinical, Pathological, and Molecular Characterization of Feline Leprosy Syndrome in the Western USA." *Veterinary Dermatology* 15 (s1). Blackwell Publishing Ltd: 16–17. doi:10.1111/j.1365-3164.2004.00410_5-2.x.

Fooden, J. 1995. "Systematic Review of Southeast Asia Long-Tail Macaques, Macaca Fascicularis Raffles (1821)." *Fieldiana Zoology* 64: 1–44.

Formenty, Pierre, Christophe Boesch, Monique Wyers, Claudia Steiner, Franca Donati, Frédéric Dind, Francine Walker, and BL Guenno. 1999. "Ebola Virus Outbreak among Wild Chimpanzees Living in a Rain Forest of Cote d'Ivoire." *Journal of Infectious Diseases* 179 (Suppl 1). Oxford University Press: S120–26. doi:10.1086/514296.

Forshaw, D., and G. R. Phelps. 1991. "Tuberculosis in a Captive Colony of Pinnipeds." *Journal of WildUfe DIseases* 27 (2). Wildlife Disease Association: 288–95. doi:10.1136/bjo.75.4.229.

Fortuine, Robert. 1989. *Chills and Fevers: Health and Disease in the Early History of Alaska*. University of Alaska Press.

Fortuine, Robert. 2005. "Must We All Die? Alaska's Enduring Struggle with Tuberculosis." University of Alaska Press. doi:10.1353/bhm.2006.0080.

Foster, Jeffrey T, Stephen M Beckstrom-Sternberg, Talima Pearson, James S Beckstrom-Sternberg, Patrick S G Chain, Francisco F Roberto, Jonathan Hnath, Tom Brettin, and Paul Keim. 2009. "Whole-Genome-Based Phylogeny and Divergence of the Genus Brucella." *Journal of Bacteriology* 191 (8): 2864–70. doi:10.1128/JB.01581-08.

Fu, Qiaomei, Matthias Meyer, Xing Gao, Udo Stenzel, Hernán a Burbano, Janet Kelso, and Svante Pääbo. 2013. "DNA Analysis of an Early Modern Human from Tianyuan Cave, China." *Proceedings of the National Academy of Sciences of the United States of America* 110 (6): 2223–27. doi:10.1073/pnas.1221359110.

Fyfe, JA, C McCowan, C R O'Brien, M Global, C Birch, P Revill, V R D Barrs, et al. 2008. "Molecular Characterization of a Novel Fastidious Mycobacterium Causing Lepromatous Lesions of the Skin, Subcutis, Cornea, and Conjunctiva of Cats Living in Victoria, Australia." *Journal of Clinical Microbiology* 46 (2): 618–26. doi:10.1128/JCM.01186-07.

Gagneux, Sebastien, Kathryn DeRiemer, Tran Van, Midori Kato-Maeda, Bouke C de Jong, Sujatha Narayanan, Mark Nicol, et al. 2006. "Variable Host-Pathogen Compatibility in Mycobacterium Tuberculosis." *Proceedings of the National Academy of Sciences of the United States of America* 103 (8). National Academy of

Sciences: 2869–73. doi:10.1073/pnas.0511240103.

Gagneux, Sebastien, and Peter M. Small. 2007. "Global Phylogeography of Mycobacterium Tuberculosis and Implications for Tuberculosis Product Development." *Lancet Infectious Diseases* 7 (5): 328–37. doi:10.1016/S1473-3099(07)70108-1.

Garcia-Alcalde, F., K. Okonechnikov, J. Carbonell, L. M. Cruz, S. Gotz, S. Tarazona, J. Dopazo, T. F. Meyer, and A. Conesa. 2012. "Qualimap: Evaluating next-Generation Sequencing Alignment Data." *Bioinformatics* 28 (20). Oxford University Press: 2678–79. doi:10.1093/bioinformatics/bts503.

Gelber, RH. 2005. *Leprosy (Hansen's Disease)*. *Harrison's Principles of Internal Medicine*.

Gómez-Valero, Laura, Eduardo P C Rocha, Amparo Latorre, and Francisco J Silva. 2007. "Reconstructing the Ancestor of Mycobacterium Leprae: The Dynamics of Gene Loss and Genome Reduction." *Genome Research* 17 (8): 1178–85. doi:10.1101/gr.6360207.

Gontcharov, A. A., Birger Marin, and Michael Melkonian. 2004. "Are Combined Analyses Better Than Single Gene Phylogenies? A Case Study Using SSU rDNA and rbcL Sequence Comparisons in the Zygnematophyceae (Streptophyta)." *Molecular Biology and Evolution* 21 (3). Oxford University Press: 612–24. doi:10.1093/molbev/msh052.

Goodall, Jane. 1986. *The Chimpanzeee of Gombe: Patterns of Behaviour*.

Gormus, B J, R H Wolf, G B Baskin, S Ohkawa, P J Gerone, G P Walsh, W M Meyers, C H Binford, and W E Greer. 1988. "A Second Sooty Mangabey Monkey with Naturally Acquired Leprosy: First Reported Possible Monkey-to-Monkey Transmission." *International Journal of Leprosy and Other Mycobacterial Diseases* 56 (1): 61–65.

Gormus, B J, K Y Xu, P L Alford, D R Lee, G B Hubbard, J W Eichberg, and W M Meyers. 1991. "A Serologic Study of Naturally Acquired Leprosy in Chimpanzees." *International Journal of Leprosy and Other Mycobacterial Diseases* 59 (3): 450–57.

Green, Richard E, Adrian W Briggs, Johannes Krause, Kay Prüfer, Hernán a Burbano, Michael Siebauer, Michael Lachmann, and Svante Pääbo. 2009. "The Neandertal Genome and Ancient DNA Authenticity." *The EMBO Journal* 28 (17): 2494–2502. doi:10.1038/emboj.2009.222.

Griffin, Dennis. 1996. "A Culture in Transition: A History of Acculturation and Settlement Near the Mouth of the Yukon River, Alaska." *Arctic Anthropology* 33

(1): 98–115.

Gröschel, Matthias I., Fadel Sayes, Roxane Simeone, Laleh Majlessi, and Roland Brosch. 2016. “ESX Secretion Systems: Mycobacterial Evolution to Counter Host Immunity.” *Nature Reviews Microbiology* 14 (11). Nature Publishing Group: 677–91. doi:10.1038/nrmicro.2016.131.

Han, Xiang Y., Yiel Hea Seo, Kurt C. Sizer, Taylor Schoberle, Gregory S. May, John S. Spencer, Wei Li, and R. Geetha Nair. 2008. “A New Mycobacterium Species Causing Diffuse Lepromatous Leprosy.” *American Journal of Clinical Pathology* 130 (6): 856–64. doi:10.1309/AJCPP72FJZZRRVMM.

Han, Xiang Y., Kurt Clement Sizer, Jesús S. Velarde-Félix, Luis O. Frias-Castro, and Francisco Vargas-Ocampo. 2012. “The Leprosy Agents Mycobacterium Lepromatosis and Mycobacterium Leprae in Mexico.” *International Journal of Dermatology* 51 (8): 952–59. doi:10.1111/j.1365-4632.2011.05414.x.

Han, Xiang Y, Kurt C Sizer, Erika J Thompson, Juma Kabanja, Jun Li, Peter Hu, Laura Gómez-Valero, and Francisco J Silva. 2009. “Comparative Sequence Analysis of Mycobacterium Leprae and the New Leprosy-Causing Mycobacterium Lepromatosis.” *Journal of Bacteriology* 191 (19): 6067–74. doi:10.1128/JB.00762-09.

Harkins, Kelly M, Jane E Buikstra, Tessa Campbell, Kirsten I Bos, Eric D Johnson, Johannes Krause, Anne C Stone, et al. 2015. “Screening Ancient Tuberculosis with qPCR : Challenges and Opportunities.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 370: 1–10. doi:10.1098/rstb.2013.0622.

Harris, N B, and R G Barletta. 2001. “Mycobacterium Avium Subsp. Paratuberculosis in Veterinary Medicine.” *Clinical Microbiology Reviews* 14 (3). American Society for Microbiology: 489–512. doi:10.1128/CMR.14.3.489-512.2001.

Herbig, Alexander, Frank Maixner, Kirsten I. Bos, Albert Zink, Johannes Krause, and Daniel H. Huson. 2016. “MALT: Fast Alignment and Analysis of Metagenomic DNA Sequence Data Applied to the Tyrolean Iceman.” *bioRxiv*. doi:10.1101/050559.

Hershberg, Ruth, Mikhail Lipatov, Peter M. Small, Hadar Sheffer, Stefan Niemann, Susanne Homolka, Jared C. Roach, et al. 2008. “High Functional Diversity in Mycobacterium Tuberculosis Driven by Genetic Drift and Human Demography.” *PLoS Biology* 6 (12): 2658–71. doi:10.1371/journal.pbio.0060311.

Hillis, David M. 1996. “Inferring Complex Phylogenies.” *Nature*. doi:10.1038/383130a0.

Hou, J. Y., J. E. Graham, and J. E. Clark-Curtiss. 2002. “Mycobacterium Avium Genes Expressed during Growth in Human Macrophages Detected by Selective Capture of Transcribed Sequences (SCOTS) Mycobacterium Avium Genes Expressed during Growth in Human Macrophages Detected by Selective Capture of Transcribed Seq.” *Infection and Immunity* 70 (7): 3714–26. doi:10.1128/IAI.70.7.3714.

Housman, Genevieve, Joanna Malukiewicz, Vanner Boere, Adriana D. Grativol, Luiz Cezar M Pereira, Ita de Oliveira e Silva, Carlos R. Ruiz-Miranda, Richard Truman, and Anne C. Stone. 2015. “Validation of qPCR Methods for the Detection of Mycobacterium in New World Animal Reservoirs.” *PLoS Neglected Tropical Diseases* 9 (11): 1–13. doi:10.1371/journal.pntd.0004198.

Hrdlicka, Ales. 1943. “Alaska Diary.”

Hubbard, G. B., D. R. Lee, J. W. Eichberg, B. J. Gormus, K. Xu, and W. M. Meyers. 1991. “Spontaneous Leprosy in a Chimpanzee (Pan Troglodytes).” *Veterinary Pathology* 28 (6): 546–48. doi:10.1177/030098589102800617.

Hughes, M S, G James, M J Taylor, J McCarroll, S D Neill, S C a Chen, D H Mitchell, D N Love, and R Malik. 2004. “PCR Studies of Feline Leprosy Cases.” *Journal of Feline Medicine and Surgery* 6 (4): 235–43. doi:10.1016/j.jfms.2003.09.003.

Huson, D.H. 2016. “MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequening Data.” *PLos Computational Biology* 12 (6): e1004957. doi:10.1371/journal. pcbi.1004957.

Ignatov, Dmitriy, Elena Kondratieva, Tatyana Azhikina, and Alexander Apt. 2012. “Mycobacterium Avium-Triggered Diseases: Pathogenomics.” *Cellular Microbiology*. Blackwell Publishing Ltd. doi:10.1111/j.1462-5822.2012.01776.x.

Jantz, R. 1972. “Cranial Variation and Microevolution in Arikara Skeletal Populations.” *Plains Anthropologist* 1 (7): 20–35.

Jantz, R, and D Owsley. 1994. “Growth and Dental Development in Arikara Children.” In *Skeletal Biology in the Great Plains: Migration, Warfare, Health, and Subsistence*. Smithsonian Institution Press, Washington, DC.

Jones-Engel, Lisa, Gregory A. Engel, John Heidrich, Mukesh Chalise, Narayan Poudel, Raphael Viscidi, Peter A. Barry, Jonathan S. Allan, Richard Grant, and Randy Kyes. 2006. “Temple Monkeys and Health Implications of Commensalism, Kathmandu, Nepal.” *Emerging Infectious Diseases* 12 (6): 900–906. doi:10.3201/eid1206.060030.

Jones-Engel, Lisa, Gregory A. Engel, Michael A Schillaci, Rosnany Babo, and Jeffery Froehlich. 2001. “Detection of Antibodies to Selected Human Pathogens among

Wild and Pet Macaques (Macaca Tonkeana) in Sulawesi, Indonesia." *American Journal of Primatology* 54 (3). John Wiley & Sons, Inc.: 171–78. doi:10.1002/ajp.1021.

Jones-Engel, Lisa, M.a. Schillaci, Gregory Engel, Umar Paputungan, and J.W. Froehlich. 2005. "Characterizing Primate Pet Ownership in Sulawesi: Implications for Disease Transmission." *Commensalism and Conflict: The Human Primate Interface. Special Topics in Primatology* 4: 97–221.

Kai, M., N. Nakata, M. Matsuoka, T. Sekizuka, M. Kuroda, and M. Makino. 2013. "Characteristic Mutations Found in the ML0411 Gene of Mycobacterium Leprae Isolated in Northeast Asian Countries." *Infection, Genetics and Evolution* 19 (October): 200–204. doi:10.1016/j.meegid.2013.07.014.

Kawaguchi, Y, M Matsuoka, K Kawatsu, J Y Homma, and C Abe. 1976. "Susceptibility to Murine Leprosy Bacilli of Nude Mice." *The Japanese Journal of Experimental Medicine* 46 (3): 167—180.

Kay, Gemma L., Martin J. Sergeant, Zhemin Zhou, Jacqueline Z.-M. Chan, Andrew Millard, Joshua Quick, Ildikó Szikossy, et al. 2015. "Eighteenth-Century Genomes Show That Mixed Infections Were Common at Time of Peak Tuberculosis in Europe." *Nature Communications* 6 (April). Nature Publishing Group: 6717. doi:10.1038/ncomms7717.

Kiełbasa, Szymon M, Raymond Wan, Kengo Sato, Paul Horton, and Martin C Frith. 2011. "Adaptive Seeds Tame Genomic Sequence Comparison." *Genome Research* 21 (3). Cold Spring Harbor Laboratory Press: 487–93. doi:10.1101/gr.113985.110.

Kiers, Albert, A. Klarenbeek, B. Mendelts, D. Van Soolingen, and G. Koëter. 2008. "Transmission of Mycobacterium Pinnipedii to Humans in a Zoo with Marine Mammals." *International Journal of Tuberculosis and Lung Disease* 12 (12). International Union Against Tuberculosis and Lung Disease: 1469–73.

Klaus, Haagen D., Alicia K. Wilbur, Daniel H. Temple, Jane E. Buikstra, Anne C. Stone, Marco Fernandez, Carlos Wester, and Manuel E. Tam. 2010. "Tuberculosis on the North Coast of Peru: Skeletal and Molecular Paleopathology of Late Pre-Hispanic and Postcontact Mycobacterial Disease." *Journal of Archaeological Science* 37 (10): 2587–97. doi:10.1016/j.jas.2010.05.019.

Koboldt, Daniel C, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher A Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. 2012. "VarScan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing." *Genome Research* 22 (3). Cold Spring Harbor Laboratory Press: 568–76. doi:10.1101/gr.129684.111.

Krakower, C, and L.M. Gonzalez. 1940. “Mouse Leprosy.” *Arch. Pathol.* 30: 308–29.

Krause, Johannes, Adrian W. Briggs, Martin Kircher, Tomislav Maricic, Nicolas Zwyns, Anatoli Derevianko, and Svante Pääbo. 2010. “A Complete mtDNA Genome of an Early Modern Human from Kostenki, Russia.” *Current Biology* 20 (3): 231–36. doi:10.1016/j.cub.2009.11.068.

Kumar, Sudhir, Glen Stecher, and Koichiro Tamura. 2016. “MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets.” *Molecular Biology and Evolution* 33 (7). Oxford University Press: 1870–74. doi:10.1093/molbev/msw054.

Langmead, Ben, and Steven L Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nat Methods* 9 (4): 357–59. doi:10.1038/nmeth.1923.

Lawrence, Diana M, Brian M Kemp, Jason Eshleman, Richard L Jantz, Meradeth Snow, Debra George, David Glenn Smith, and David Glenn Smith3. 2010. “Mitochondrial DNA of Protohistoric Remains of an Arikara Population from South Dakota: Implications for the Macro-Siouan Language Hypothesis.” *Source: Human Biology* 82 (2). Wayne State University Press: 157–78.

Lawrence, WE, and N Wickham. 1963. “Cat Leprosy: Infection by a Bacillus Resembling Mycobacterium Lepraemurium.” *Australian Veterinary Journal* 39: 390–393.

Lee R.S., Radomski N., Proulx J.F., Levadee I., Shapiro J.B., McIntosh F., Soualhine H., Menzies D., and Behr M.A. 2015. “Population Genomics of Mycobacterium Tuberculosis in the Inuit.” *Proceedings of the National Academy of Sciences of the United States of America* 2000 (4): 1–6. doi:10.1073/pnas.1507071112.

Leininger, J. R., K. J. Donham, and M. J. Rubino. 1978. “Leprosy in a Chimpanzee: Morphology of the Skin Lesions and Characterization of the Organism.” *Veterinary Pathology* 15 (3): 339–46. doi:10.1177/030098587801500308.

Leroy, Eric M., Pierre Rouquet, Pierre Formenty, Sandrine Souquière, Annelisa Kilbourne, Jean-Marc Froment, Magdalena Bermejo, et al. 2004. “Multiple Ebola Virus Transmission Events and Rapid Decline of Central African Wildlife.” *Science (New York, N.Y.)* 303 (5656): 387–90. doi:10.1126/science.1092528.

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 (16). Oxford University Press: 2078–79. doi:10.1093/bioinformatics/btp352.

Li, H, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with

Burrows-Wheeler Transform." *Bioinformatics* 25 (14). Oxford University Press: 1754–60. doi:10.1093/bioinformatics/btp324.

Li, L., J. P. Bannantine, Q. Zhang, A. Amonsin, B. J. May, D. Alt, N. Banerji, S. Kanjilal, and V. Kapur. 2005. "The Complete Genome Sequence of Mycobacterium Avium Subspecies Paratuberculosis." *Proceedings of the National Academy of Sciences* 102 (35): 12344–49. doi:10.1073/pnas.0505662102.

Li, Yong-Jun, Lia Danelishvili, Dirk Wagner, Mary Petrofsky, and Luiz E Bermudez. n.d. "Identification of Virulence Determinants of Mycobacterium Avium That Impact on the Ability to Resist Host Killing Mechanisms." doi:10.1099/jmm.0.012864-0.

Li, Yongjun, Elizabeth Miltner, Martin Wu, Mary Petrofsky, and Luiz E. Bermudez. 2004. "A Mycobacterium Avium PPE Gene Is Associated with the Ability of the Bacterium to Grow in Macrophages and Virulence in Mice." *Cellular Microbiology* 7 (4). Blackwell Science Ltd: 539–48. doi:10.1111/j.1462-5822.2004.00484.x.

Liu, Yang, Paul M Harrison, Victor Kunin, and Mark Gerstein. 2004. "Comprehensive Analysis of Pseudogenes in Prokaryotes: Widespread Gene Decay and Failure of Putative Horizontally Transferred Genes." *Genome Biology* 5 (9): R64. doi:10.1186/gb-2004-5-9-r64.

Loeffler, Scott H., Geoffrey W. de Lisle, Mark A. Neill, Desmond M. Collins, Marian Price-Carter, Brent Paterson, and Kevin B. Crews. 2014. "The Seal Tuberculosis Agent, Mycobacterium Pinnipedii, Infects Domestic Cattle in New Zealand: Epidemiologic Factors and DNA Strain Typing." *Journal of Wildlife Diseases* 50 (2): 180–87. doi:10.7589/2013-09-237.

Loudon, J E, M L Sauther, K D Fish, and M Hunter-Ishikawa. 2006. "Three Primates - One Reserve: Applying a Holistic Approach to Understand the Dynamics of Behavior, Conservation, and Disease amongst Ring-Tailed Lemurs, Verreaux's Sifaka, and Humans at Beza Mahafaly Special Reserve, Madagascar." *Ecological and Environmental Anthropology* 2 (2).

Lygren, S. T., O. Closs, H. Bercouvier, and L. G. Wayne. 1986. "Catalases, Peroxidases, and Superoxide Dismutases in Mycobacterium Leprae and Other Mycobacteria Studied by Crossed Immunoelectrophoresis and Polyacrylamide Gel Electrophoresis." *Infection and Immunity* 54 (3): 666–72.

Maeda, S, M Matsuoka, N Nakata, M Kai, Y Maeda, K Hashimoto, H Kimura, K Kobayashi, and Y Kashiwabara. 2001. "Multidrug Resistant Mycobacterium Leprae from Patients with Leprosy." *Antimicrobial Agents and Chemotherapy* 45 (12). American Society for Microbiology: 3635–39. doi:10.1128/AAC.45.12.3635-3639.2001.

Malik, R, M S Hughes, G James, P Martin, D I Wigney, P J Canfield, S C a Chen, D H Mitchell, and D N Love. 2002. “Feline Leprosy: Two Different Clinical Syndromes.” *Journal of Feline Medicine and Surgery* 4 (1): 43–59. doi:10.1053/jfms.2001.0151.

Marchoux, E, and F Sorel. 1912. “Recherches Sur La Lèpre: La Lèpre Des Rats (Lepra Murium).” *Ann. Inst. Pasteur* 56: 778–801.

Marmiesse, M., Priscille Brodin, Carmen Buchrieser, Christina Gutierrez, Nathalie Simoes, Veronique Vincent, Philippe Glaser, Stewart T Cole, and Roland Brosch. 2004. “Macro-Array and Bioinformatic Analyses Reveal Mycobacterial ‘Core’ Genes, Variation in the ESAT-6 Gene Family and New Phylogenetic Markers for the Mycobacterium Tuberculosis Complex.” *Microbiology* 150 (2): 483–96. doi:10.1099/mic.0.26662-0.

Martinez, Alejandra Nóbrega, Marcelo Ribeiro-Alves, Euzenir Nunes Sarno, and Milton Ozório Moraes. 2011. “Evaluation of qPCR-Based Assays for Leprosy Diagnosis Directly in Clinical Specimens.” Edited by Mehmet Ali Ozcel. *PLoS Neglected Tropical Diseases* 5 (10). Morgan Kaufmann Publishers: e1354. doi:10.1371/journal.pntd.0001354.

McCartney, Allen P., and Douglas W. Veltre. 1999. “Aleutian Island Prehistory: Living in Insular Extremes.” *World Archaeology* 30 (3). Taylor & Francis Group : 503–15. doi:10.1080/00438243.1999.9980426.

McHugh, T D, L E Newport, and S H Gillespie. 1997. “IS6110 Homologs Are Present in Multiple Copies in Mycobacteria Other than Tuberculosis-Causing Mycobacteria.” *Journal of Clinical Microbiology* 35 (7). American Society for Microbiology: 1769–71.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Research* 20 (9). Cold Spring Harbor Laboratory Press: 1297–1303. doi:10.1101/gr.107524.110.

Meyer, Matthias, and Martin Kircher. 2010. “Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing.” *Cold Spring Harbor Protocols*, no. 6. Cold Spring Harbor Laboratory Press: pdb.prot5448. doi:10.1101/pdb.prot5448.

Meyers, Wayne M, Gerald P Walsh, Harriet L Brown, Chapman H Binford, George D Imes, Ted L Hadfield, Charles J Schlagel, et al. 1985. “Leprosy in a Mangabey Monkey - Naturally Acquired Infection.” *International Journal of Leprosy and Other Mycobacterial Diseases* 53 (1): 1–14.

Mignard, Sophie, and Jean-Pierre Flandrois. 2008. “A Seven-Gene, Multilocus, Genus-Wide Approach to the Phylogeny of Mycobacteria Using Supertrees.” *International Journal of Systematic and Evolutionary Microbiology* 58 (Pt 6): 1432–41. doi:10.1099/ijs.0.65658-0.

Monot, Marc, Nadine Honore, Thierry Garnier, Romulo Araoz, Jean-Yves Coppée, Celine Lacroix, Samba Sow, et al. 2005. “On the Origin of Leprosy.” *Science* 308 (5724): 1040–42.

Monot, Marc, Nadine Honoré, Thierry Garnier, Nora Zidane, Diana Sherafi, Alberto Paniz-Mondolfi, Masanori Matsuoka, et al. 2009. “Comparative Genomic and Phylogeographic Analysis of Mycobacterium Leprae.” *Nature Genetics* 41 (12): 1282–89. doi:10.1038/ng.477.

Morgulis, Aleksandr, E Michael Gertz, Alejandro A Schäffer, and Richa Agarwala. 2006. “A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences.” *Journal of Computational Biology* 13 (5): 1028–40. doi:10.1089/cmb.2006.13.1028.

Mori, Tatsuo, and K Kohsaka. 1986. “Identification of Cat Leprosy Bacillus Grown in Mice.” *International Journal of Leprosy* 54 (4): 584–95.

Morin, Phillip A, Gordon Luikart, Robert K Wayne, and S N P Working Group. 2004. “SNPs in Ecology, Evolution and Conservation.” *TRENDS in Ecology & Evolution* 19 (4): 208–16.

Moser, I., W. M. Prodinger, H. Hotzel, R. Greenwald, K. P. Lyashchenko, D. Bakker, D. Gomis, et al. 2008. “Mycobacterium Pinnipedii: Transmission from South American Sea Lion (Otaria Byronia) to Bactrian Camel (Camelus Bactrianus Bactrianus) and Malayan Tapirs (Tapirus Indicus).” *Veterinary Microbiology* 127 (3–4): 399–406. doi:10.1016/j.vetmic.2007.08.028.

Müller, Romy, Charlotte A. Roberts, and Terence A. Brown. 2014. “Genotyping of Ancient Mycobacterium Tuberculosis Strains Reveals Historic Genetic Diversity.” *Proceedings. Biological Sciences / The Royal Society* 281 (1781): 20133236. doi:10.1098/rspb.2013.3236.

Nakamura, M. 1999. “For the Growth of Mycobacterium Lepraemurium in Cell-Free Liquid Medium, the Key Essential Factor May Be the pH (Optimal 6.0-6.2) of the Culture Medium, rather than the Presence of Alpha-Ketoglutaric Acid.” *Nihon Hansenbyo Gakkai Zasshi = Japanese Journal of Leprosy* 68 (3): 157–63.

Nguyen, Dao, Paul Brassard, Dick Menzies, Louise Thibert, Rob Warren, Serge Mostowy, and Marcel Behr. 2004. “Genomic Characterization of an Endemic Mycobacterium Tuberculosis Strain: Evolutionary and Epidemiologic Implications.”

Journal of Clinical Microbiology 42 (6). American Society for Microbiology (ASM): 2573–80. doi:10.1128/JCM.42.6.2573-2580.2004.

Nguyen, Dao, Jean-François Proulx, Jennifer Westley, Louise Thibert, Serge Dery, and Marcel A. Behr. 2003. "Tuberculosis in the Inuit Community of Quebec, Canada." *American Journal of Respiratory and Critical Care Medicine* 168 (11). American Thoracic Society: 1353–57. doi:10.1164/rccm.200307-910OC.

Orquera, L.A. 2005. "Mid-Holocene Littoral Adaptation at the Southern End of South America." *Quaternary International* 132: 107–15.

Orquera, L.A., D. Legoupil, and E.L. Piana. 2011. "Littoral Adaptation at the Southern End of South America." *Quaternary International* 239: 61–69.

Orth, D. J. 1971. "Dictionary of Alaska Place Names: Geological Survey Professional Paper 567." In . US Department of the Interior.

Otto, Thomas D, Gary P Dillon, Wim S Degrave, and Matthew Berriman. 2011. "RATT: Rapid Annotation Transfer Tool." *Nucleic Acids Research* 39 (9): e57. doi:10.1093/nar/gkq1268.

Ozga, Andrew T., Maria A. Nieves-Colón, Tanvi P. Honap, Krithivasan Sankaranarayanan, Courtney A. Hofman, George R. Milner, Cecil M. Lewis, Anne C. Stone, and Christina Warinner. 2016. "Successful Enrichment and Recovery of Whole Mitochondrial Genomes from Ancient Human Dental Calculus." *American Journal of Physical Anthropology* 160 (2): 220–28. doi:10.1002/ajpa.22960.

Palkovich, Ann M. 1981. "Tuberculosis Epidemiology in Two Arikara Skeletal Samples: A Study of Disease Impact." *JE Buikstra, Prehistoric Tuberculosis in the Americas, Northwestern University Archaelogical Program, Evanston*, 161–75.

Pedersen, Amy B., and T. Jonathan Davies. 2009. "Cross-Species Pathogen Transmission and Disease Emergence in Primates." *EcoHealth* 6 (4): 496–508. doi:10.1007/s10393-010-0284-3.

Pedersen, NC. 1988. "Atypical Mycobacteriosis." *Feline Infectious Diseases*, 197–200.

Peltzer, Alexander, Günter Jäger, Alexander Herbig, Alexander Seitz, Christian Kniep, Johannes Krause, and Kay Nieselt. 2016. "EAGER: Efficient Ancient Genome Reconstruction." *Genome Biology* 17 (1): 60. doi:10.1186/s13059-016-0918-z.

Pepperell, Caitlin S, Julie M Granka, David C Alexander, Marcel A Behr, Linda Chui, Janet Gordon, Jennifer L Guthrie, et al. 2011. "Dispersal of Mycobacterium Tuberculosis via the Canadian Fur Trade." *Proceedings of the National Academy of Sciences of the United States of America* 108 (16). National Academy of Sciences:

6526–31. doi:10.1073/pnas.1016708108.

Prabhakaran, K, E B Harris, and W F Kirchheimer. 1976. “Binding of 14C Labeled Dopa by Mycobacterium Leprae in Vitro.” *International Journal of Leprosy* 44 (1–2): 58–64.

R Core Team. 2017. “R: A Language and Environment for Statistical Computing.”

Raff, Jennifer, Della Collins Cook, and Frederika Kaestle. 2006. “Tuberculosis in the New World: A Study of Ribs from the Schild Mississippian Population, West-Central Illinois.” In *Memorias Do Instituto Oswaldo Cruz*, 101:25–27. Fundação Oswaldo Cruz. doi:10.1590/S0074-02762006001000005.

Ramakrishnan, Lalita, Nancy A Federspiel, and Stanley Falkow. 2000. “Granuloma-Specific Expression of Mycobacterium Virulence Proteins from the Glycine-Rich PE-PGRS Family.” *Science* 288 (5470): 1436–39. doi:10.1126/science.288.5470.1436.

Rambaut, Andrew, Tommy T. Lam, Luiz Max Carvalho, and Oliver G. Pybus. 2016. “Exploring the Temporal Structure of Heterochronous Sequences Using TempEst (Formerly Path-O-Gen).” *Virus Evolution* 2 (1). Oxford University Press. doi:10.1093/ve/vew007.

Rambaut, Andrew, Marc A Suchard, D Xie, and Alexei J Drummond. 2015. “Tracer v1. 6. 2014.”

Robbins, Gwen, V Mushrif Tripathy, V N Misra, R K Mohanty, V S Shinde, Kelsey M Gray, and Malcolm D Schug. 2009. “Ancient Skeletal Evidence for Leprosy in India (2000 B.C.).” *PloS One* 4 (5): e5669. doi:10.1371/journal.pone.0005669.

Roberts, C. A., and J. E. Buikstra. 2003. *The Bioarchaeology of Tuberculosis. A Global View on a Reemering Disease*. *Gainesville, Fl: University Press of Florida*. University Press of Florida.

Rohland, Nadin, and Michael Hofreiter. 2007. “Ancient DNA Extraction from Bones and Teeth.” *Nature Protocols* 2 (7). Nature Publishing Group: 1756–62. doi:10.1038/nprot.2007.247.

Rojas-Espinosa, O. 1994. “Active Humoral Immunity in the Absence of Cell-Mediated Immunity in Murine Leprosy: Lastly an Explanation.” *International Journal of Leprosy and Other Mycobacterial Diseases* 62: 143–47.

Rojas-Espinosa, O, E Becerril-Villanueva, K. Wek-Rodriguez, P Arce-Paredes, and E Reyes-Maldonado. 2005. “Palsy of the Rear Limbs in Mycobacterium Lepraemurium-Infected Mice Results from Bone Damage and Not from Nerve

Involvement." *Clinical and Experimental Immunology* 140 (3): 436–42. doi:10.1111/j.1365-2249.2005.02776.x.

Rojas-Espinosa, O, and M Lovik. 2001. "Mycobacterium Leprae and Mycobacterium Lepraemurium Infections in Domestic and Wild Animals." *Rev.Sci.Tech.* 20 (0253–1933): 219–51.

Rojas-Espinosa, O, Rodriguez K Wek, J A Vargas Hernandez, and Paredes P Arce. 1999. "Do Antibodies to Phospholipid Antigens Play Any Role in Murine Leprosy?" *Int.J.Lepr.Other Mycobact.Dis.* 67 (4): 453–59.

Rojas-Espinosa, Oscar, Veronica Camarena-servin, Iris Estrada-garcia, Patricia Arce-paredes, and Kendy Wek-rodriguez. 1998. "Mycobacterium Lepraemurium , a Well-Adapted Parasite of Macrophages: I. Oxygen Metabolites." *International Journal of Leprosy* 66 (3): 365–73.

Rojas-Espinosa, Oscar, K Wek-Rodriguez, and Patricia Arce-Paredes. 2002. "The Effect of Exogenous Peroxidase on the Evolution of Murine Leprosy." *Int.J.Lepr.Other Mycobact.Dis.* 70 (0148–916X): 191–200.

Rokas, Antonis, Barry L. Williams, Nicole King, and Sean B. Carroll. 2003. "Genome-Scale Approaches to Resolving Incongruence in Molecular Phylogenies." *Nature* 425 (6960): 798–804. doi:10.1038/nature02053.

Rosenberg, Michael S. 2016. "FINGERPRINT: Computational Filtering of Targeted Sequences from Environmental Contaminants." In *Joint Meeting of the Society for Molecular Biology and Evolution & Genetic Society of Australia, Gold Coast, Australia*.

Roy, Craig R, and Jacqueline Cherfils. 2015. "Structure and Function of Fic Proteins." *Nature Reviews Microbiology* 13 (10). Nature Publishing Group: 631–40. doi:10.1038/nrmicro3520.

Sambrook, Joseph., and Russell L McLaughlin. 2000. *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press.

Sanderson, Michael J, and H Bradley Shaffer. 2002. "Troubleshooting Molecular Phylogenetic Analyses." *Ann Rev. Ecol. Evol. Syst.* 33: 49–72. doi:10.1146/annurev.ecolsys.33.010802.150509.

Schiavini, A. 1993. "Los Lobos Marinos Como Recurso Para Cazadores-Recolectores Marinos: El Caso de Tierra Del Fuego." *Latin American Antiquity*, 346–66.

Schmitt, LS. 1911. "On the Relation between Rat and Human Leprosy." *The University Press*.

Schubert, Mikkel, Aurelien Ginolhac, Stinus Lindgreen, John F Thompson, Khaled A S Al-Rasheid, Eske Willerslev, Anders Krogh, and Ludovic Orlando. 2012. "Improving Ancient DNA Read Mapping against Modern Reference Genomes." *BMC Genomics* 13 (1): 178. doi:10.1186/1471-2164-13-178.

Schubert, Mikkel, Stinus Lindgreen, and Ludovic Orlando. 2016. "AdapterRemoval v2: Rapid Adapter Trimming, Identification, and Read Merging." *BMC Research Notes* 9 (88). BioMed Central. doi:10.1093/bioinformatics/bts187.

Schuenemann, Verena J, Pushpendra Singh, Thomas A Mendum, Ben Krause-Kyora, Günter Jäger, Kirsten I Bos, Alexander Herbig, et al. 2013. "Genome-Wide Comparison of Medieval and Modern Mycobacterium Leprae." *Science* 341 (6142): 179–83. doi:10.1126/science.1238286.

Simeone, Roxane, Alexandre Bobard, Juliane Lippmann, Wilbert Bitter, Laleh Majlessi, Roland Brosch, and Jost Enninga. 2012. "Phagosomal Rupture by Mycobacterium Tuberculosis Results in Toxicity and Host Cell Death." Edited by Sabine Ehrt. *PLoS Pathogens* 8 (2). Public Library of Science: e1002507. doi:10.1371/journal.ppat.1002507.

Singh, Pushpendra, Andrej Benjak, Verena J Schuenemann, Alexander Herbig, Charlotte Avanzi, Philippe Busso, Kay Nieselt, Johannes Krause, Lucio Vera-Cabrera, and Stewart T Cole. 2015. "Insight into the Evolution and Origin of Leprosy Bacilli from the Genome Sequence of Mycobacterium Lepromatosis." *Proceedings of the National Academy of Sciences of the United States of America* 112 (14): 4459–64. doi:10.1073/pnas.1421504112.

Smith, Melvin A., and Douglas W. Veltre. 2010. "Historical Overview of Archaeological Research in the Aleut Region of Alaska." *Human Biology* 82 (December 2010). Wayne State University Press Human Biology: 487–506. doi:10.3378/027.082.0502.

Snel, Berend, Peer Bork, and Martijn A. Huynen. 1999. "Genome Phylogeny Based on Gene Content." *Nature Genetics* 21 (1). Nature Publishing Group: 108–10. doi:10.1038/5052.

Stamatakis, A. 2006. "RAxML-VI-HPC: Maximum Likelihood-Based Phylogenetic Analyses with Thousands of Taxa and Mixed Models." *Bioinformatics* 22 (21): 2688–2690.

Steegman, A. 1991. "Stature in an Early Mid-19th Century Poorhouse Population: Highland Park, Rochester, New York." *American Journal of Physical Anthropology* 85 (3): 261–68.

Steenken, W., and LU. Gardner. 1946. "History of H37Rv Strain of Tubercle Bacillus." *The American Review of Tuberculosis* 54 (July): 62–66.

Stefansky, W.K. 1903. “Eine Lepräähnliche Erkrankung Der Haut Un Der Lymphdrüsen Bei Wanderratten.” *Centralbl. Bakteriol.* 33: 481–87.

Stefansky, W.K. 1902. “Zabolevanija U Krys, Vyzvannyja Kislotoupornoj Palotsjkoj.” *Russkij Vratsj* 47: 1726–27.

Stinear, Timothy P., Torsten Seemann, Sacha Pidot, Wafa Frigui, Gilles Reysset, Thierry Garnier, Guillaume Meurice, et al. 2007. “Reductive Evolution and Niche Adaptation Inferred from the Genome of Mycobacterium Ulcerans, the Causative Agent of Buruli Ulcer.” *Genome Research* 17 (2): 192–200. doi:10.1101/gr.5942807.

Stone, Anne C, Alicia K Wilbur, Jane E Buikstra, and Charlotte a Roberts. 2009. “Tuberculosis and Leprosy in Perspective.” *American Journal of Physical Anthropology* 140 Suppl (January): 66–94. doi:10.1002/ajpa.21185.

Stucki, David, Daniela Brites, Leïla Jeljeli, Mireia Coscolla, Qingyun Liu, Andrej Trauner, Lukas Fenner, et al. 2016. “Mycobacterium Tuberculosis Lineage 4 Comprises Globally Distributed and Geographically Restricted Sublineages.” *Nature Genetics* 48 (12). Nature Research: 1535–43. doi:10.1038/ng.3704.

Stuiver, Minze, G.W. Pearson, and Tom Braziunas. 1986. “Radiocarbon Age Calibration of Marine Samples back to 9000 Cal Yr BP.” *Radiocarbon* 28 (2B): 980–1021. doi:10.2458/azu_js_rc.28.1011.

Suzuki, Koichi, Toshifumi Udono, Michiko Fujisawa, Kazunari Tanigawa, Gen’ichi Idani, and Norihisa Ishii. 2010. “Infection during Infancy and Long Incubation Period of Leprosy Suggested in a Case of a Chimpanzee Used for Medical Research.” *Journal of Clinical Microbiology* 48 (9): 3432–34. doi:10.1128/JCM.00017-10.

Tanimura, T, and S Nishimura. 1952. “Studies on the Pathology of Murine Leprosy.” *International Journal of Leprosy and Other Mycobacterial Diseases* 20 (0020–7349 (Print)): 83–94.

Thierry, D., M.D. Cave, K.D. Eisenach, J.T. Crawford, J.H. Bates, B. Gicquel, and J.L. Guesdon. 1990. “IS6110, an IS-like Element of Mycobacterium Tuberculosis Complex.” *Nucleic Acids Research* 18: 188.

Thompson, Philip J., Debby V. Cousins, Beth L. Gow, Desmond M. Collins, Bruce H. Williamson, and Harvey T. Dagnia. 1993. “Seals, Seal Trainers, and Mycobacterial Infection.” *American Review of Respiratory Disease* 147 (1). American Lung Association: 164–67. doi:10.1164/ajrccm/147.1.164.

Thorel, MF, M Krichevsky, and VV Lévy-Frébault. 1990. “Numerical Taxonomy of

Mycobactin-Dependent Mycobacteria, Emended Description of Mycobacterium Avium, and Description of Mycobacterium Avium Subsp. Avium Subsp. Nov., Mycobacterium Avium Subsp. Paratuberculosis Subsp. Nov., and Mycobacterium Avium Subsp. S." *International Journal of Systematic Bacteriology* 40: 254–60. doi:10.1099/00207713-40-3-254.

Tortoli, E. 2014. "Microbiological Features and Clinical Relevance of New Species of the Genus Mycobacterium." *Clinical Microbiology Reviews* 27 (4). American Society for Microbiology: 727–52. doi:10.1128/CMR.00035-14.

Tortoli, E. 2003. "Impact of Genotypic Studies on Mycobacterial Taxonomy: The New Mycobacteria of the 1990s." *Clinical Microbiology Reviews* 16 (2). American Society for Microbiology (ASM): 319–54. doi:10.1128/cmr.16.2.319-354.2003.

Truman, Richard W, P Kyle Andrews, Naoko Y Robbins, Linda B Adams, James L Krahenbuhl, and Thomas P Gillis. 2008. "Enumeration of Mycobacterium Leprae Using Real-Time PCR." *PLoS Neglected Tropical Diseases* 2 (11): e328. doi:10.1371/journal.pntd.0000328.

Truman, Richard W, Pushpendra Singh, Rahul Sharma, Philippe Busso, Jacques Rougemont, Alberto Paniz-Mondolfi, Adamandia Kapopoulou, et al. 2011. "Probable Zoonotic Leprosy in the Southern United States." *The New England Journal of Medicine* 364 (17): 1626–33. doi:10.1056/NEJMoa1010536.

Valverde, Celia R, Don Canfield, Ross Tarara, Maria I Esteves, and Bobby J Gormus. 1998. "Spontaneous Leprosy in a Wild-Caught Cynomolgus Macaque." *International Journal of Leprosy* 66 (2): 140–48.

Vargas-Ocampo, Francisco. 2007. "Diffuse Leprosy of Lucio and Latapí: A Histologic Study." *Leprosy Review* 78 (3): 248–60.

Vera-Cabrera, Lucio, Wendy G Escalante-Fuentes, Minerva Gomez-Flores, Jorge Ocampo-Candiani, Philippe Busso, Pushpendra Singh, and Stewart T Cole. 2011. "Case of Diffuse Lepromatous Leprosy Associated with Mycobacterium Lepromatosis." *Journal of Clinical Microbiology* 49 (12). American Society for Microbiology: 4366–68. doi:10.1128/JCM.05634-11.

Walker, Ernest Linwood, and Marion A. Sweeney. 1929. "The Identity of Human Leprosy and Rat Leprosy." *Journal of Preventive Medicine* 3 (4). George Williams Hooper Foundation for Med. Research, Univ. of California, San Francisco.

Wallis, J, and D.R. Lee. 1999. "Primate Conservation: The Prevention of Disease Transmission." *Space Science Reviews* 96 (November). Kluwer Academic Publishers-Plenum Publishers: 317–30. doi:10.1023/A.

Walsh, G P, W M Meyers, and C H Binford. 1986. “Naturally Acquired Leprosy in the Nine-Banded Armadillo: A Decade of Experience 1975-1985.” *Journal of Leukocyte Biology* 40 (5). Society for Leukocyte Biology: 645–56.

Warinner, C., J. Hendy, C. Speller, E. Cappellini, R. Fischer, C. Trachsel, J. Arneborg, et al. 2014. “Direct Evidence of Milk Consumption from Ancient Human Dental Calculus.” *Scientific Reports* 4 (November). Nature Publishing Group: 7104. doi:10.1038/srep07104.

Watts, David P., and John C. Mitani. 2000. “Infanticide and Cannibalism by Male Chimpanzees at Ngogo, Kibale National Park, Uganda.” *Primates* 41 (4). Springer-Verlag: 357–65. doi:10.1007/BF02557646.

Wek-Rodriguez, Kendy, Mayra Silva-Miranda, Patricia Arce-Paredes, and Oscar Rojas-Espinosa. 2007. “Effect of Reactive Oxygen Intermediaries on the Viability and Infectivity of Mycobacterium Lepraemurium.” *International Journal of Experimental Pathology* 88 (3). Wiley-Blackwell: 137–45. doi:10.1111/j.1365-2613.2007.00524.x.

WHO. 2015. “Glocal Tuberculosis Report.” Vol. 1. doi:10.1017/CBO9781107415324.004.

WHO. 2016a. *Global Leprosy Strategy 2016-2020*.

WHO. 2016b. “Global Tuberculosis Report 2016.” doi:ISBN 978 92 4 156539 4.

Wilbur, Alicia K., Gregory A. Engel, Aida Rompis, I. G A A Putra, Benjamin P Y H Lee, Nantiya Aggimarangsee, Mukesh Chalise, et al. 2012. “From the Mouths of Monkeys: Detection of Mycobacterium Tuberculosis Complex DNA From Buccal Swabs of Synanthropic Macaques.” *American Journal of Primatology* 74 (7): 676–86. doi:10.1002/ajp.22022.

Wirth, Thierry, Falk Hildebrand, Caroline Allix-Béguec, Florian Wölbeling, Tanja Kubica, Kristin Kremer, Dick Van Soolingen, et al. 2008. “Origin, Spread and Demography of the Mycobacterium Tuberculosis Complex.” Edited by Mark Achtman. *PLoS Pathogens* 4 (9). W.H.O: e1000160. doi:10.1371/journal.ppat.1000160.

Wolfe, Nathan D., Peter Daszak, A. Marm Kilpatrick, and Donald S. Burke. 2005. “Bushmeat Hunting, Deforestation, and Prediction of Zoonotic Disease Emergence.” *Emerging Infectious Diseases* 11 (12): 1822–27. doi:10.3201/eid1112.040789.

Wolfe, Nathan D, Ananias A Escalante, William B Karesh, Annelisa Kilbourn, Andrew Spielman, and Altaf A Lal. 1998. “Wild Primate Populations in Emerging Infectious Disease Research: The Missing Link?” *Emerging Infectious Diseases*. Centers for

Disease Control and Prevention. doi:10.3201/eid0402.980202.

Yeung, Man Wah, Edwin Khoo, Sarah K. Brode, Frances B. Jamieson, Hiroyuki Kamiya, Jeffrey C. Kwong, Liane Macdonald, Theodore K. Marras, Kozo Morimoto, and Beate Sander. 2016. "Health-Related Quality of Life, Comorbidities and Mortality in Pulmonary Nontuberculous Mycobacterial Infections: A Systematic Review." *Respirology*. doi:10.1111/resp.12767.

Zeigler, Daniel R. 2003. "Gene Sequences Useful for Predicting Relatedness of Whole Genomes in Bacteria." *International Journal of Systematic and Evolutionary Microbiology* 53 (6): 1893–1900. doi:10.1099/ijs.0.02713-0.

Zhang, Xinjun, Khamisah Abdul Kadir, Leslie Fabiola Quintanilla-Zariñan, Jason Villano, Paul Houghton, Hongli Du, Balbir Singh, and David Glen Smith. 2016. "Distribution and Prevalence of Malaria Parasites among Long-Tailed Macaques (Macaca Fascicularis) in Regional Populations across Southeast Asia." *Malaria Journal* 15 (450). BioMed Central. doi:10.1186/s13071-016-1527-0.

APPENDIX A

SUPPLEMENTARY TABLES FOR CHAPTER 2 (TABLES S1 – S2)

Consult Attached Tables using Microsoft Excel.

APPENDIX B

LIST OF POSITIONS IN THE *M. LEPRAE* GENOME EXCLUDED FROM THE PHYLOGENETIC ANALYSES

Consult Attached File using a Text Editor or Microsoft Excel.

APPENDIX C

SUPPLEMENTARY FIGURES FOR CHAPTER 2 (FIGURE S1).

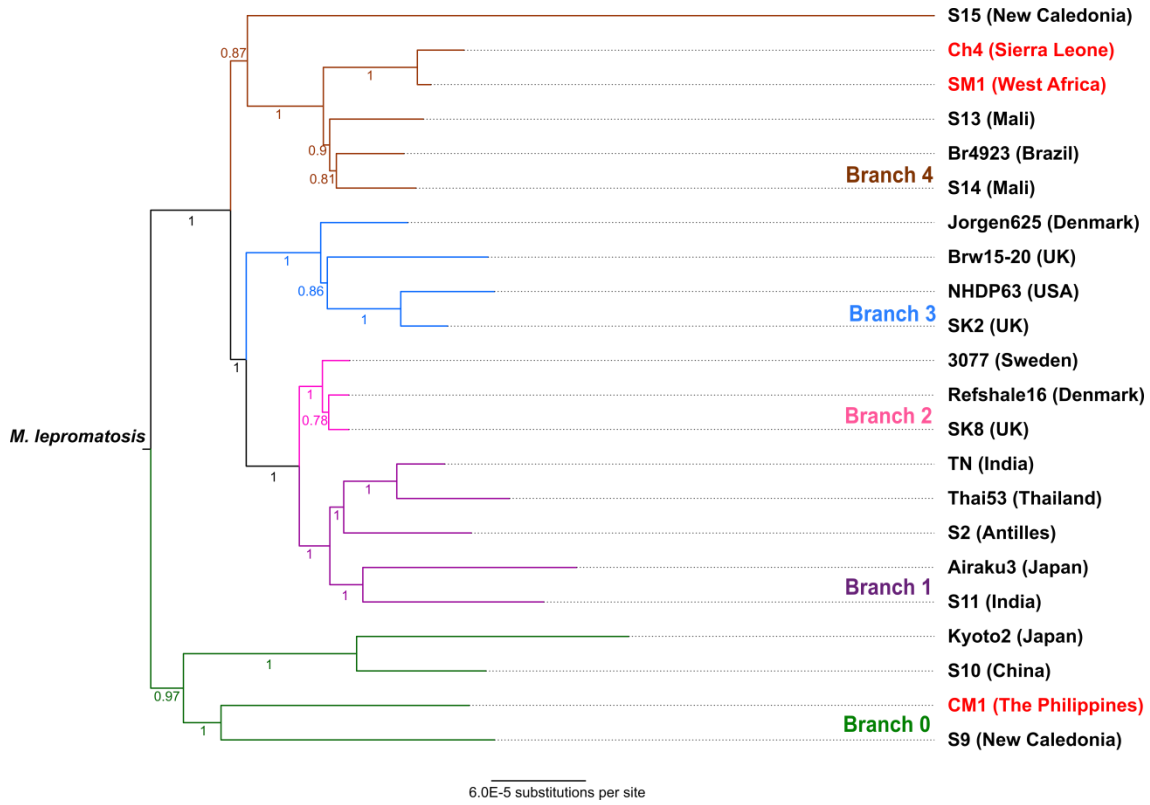

Figure S1. Neighbor Joining tree of all *M. leprae* strains based on an alignment comprising 233,509 genome-wide SNPs. *M. lepromatosis* was used as the outgroup to root the tree. The tree was built using the *p*-distance method. Bootstrap support estimated from 1,000 replicates is given on each branch. The five *M. leprae* branches are highlighted. The nonhuman primate *M. leprae* strains sequenced in this study are denoted in red. Geographic origin is given next to the name of each strain.

APPENDIX D

SUPPLEMENTARY TABLES FOR CHAPTER 3 (TABLES S3 – S5)

Table S3. Publicly available mycobacterial genomes used for phylogenetic analyses

Species	**NCBI Accession Number**
M. avium subsp. *hominissuis* TH135	AP012555.1
M. avium subsp. *paratuberculosis* K10	NC_002944.2
M. avium subsp. *avium* 104	NC_008595.1
M. marseillense DSM 45437	MVHX00000000.1
M. timonense CCUG 56329	MVIL00000000.1
M. arosiense DSM 45069	MVHG00000000.1
M. chimaera AH16	CP012885.2
M. colombiense CECT 3035	CP020821.1
M. indicus pranii MTCC 9506	NC_018612.1
M. intracellulare ATCC 13950	NC_016946.1
M. leprae TN	AL450380.1
M. lepromatosis Mx1-22A	JRPY01000001
M. tuberculosis H37Rv	NC_000962.3
M. marinum E11	HG917972.2
M. ulcerans Agy99	CP000325.1
M. abscessus	NC_010397.1

Table S4. *M. lepraemurium* - specific genes

Locus tag	Description	Functional
MLM_0980	FAD-binding mono-oxygenase	Protein-coding
MLM_0981	TetR family transcriptional regulator	Protein-coding
MLM_1065	Helix-turn-helix XRE-family transcriptional regulator	Pseudogene
MLM_1066	Hypothetical protein	Protein-coding
MLM_1327	Putative extradiol dioxygenase	Protein-coding
MLM_1462	Uncharacterized protein	Pseudogene
MLM_1829	Restriction endonuclease	Pseudogene
MLM_1829A	Hypothetical protein	Protein-coding
MLM_2063	Short-chain dehydrogenase	Pseudogene
MLM_2065	Putative short-chain dehydrogenase	Pseudogene
MLM_2701	Putative LysR-family transcriptional regulator	Protein-coding
MLM_2702	Hydroxymethylglutaryl-coA lyase	Protein-coding
MLM_2703	Acyl dehydratase	Protein-coding
MLM_2704	4-hydroxybutyrate:acetyl-coA coA transferase	Protein-coding
MLM_2705	Pyruvate oxidase	Protein-coding
MLM_2706	Nitroreductase	Protein-coding
MLM_3070	Uncharacterized protein	Pseudogene
MLM_3071	Uncharacterized protein	Pseudogene
MLM_3074	Uncharacterized protein	Pseudogene
MLM_3077	DUF58 domain-containing protein	Pseudogene

MLM_3078	Mobile element protein	Pseudogene
MLM_3079	moxR-like ATPases	Pseudogene
MLM_3080	Uncharacterized protein	Pseudogene
MLM_3300	Fic family protein	Protein-coding
MLM_3510	Uncharacterized protein	Pseudogene
MLM_3526	Uncharacterized protein	Pseudogene
MLM_3527	Uncharacterized protein	Pseudogene

Table S5. *mmpS* and *mmpL* genes in *M. lepraemurium*

Locus tag	Description	Functionality
MLM_2755	mmpS1	Protein-coding
MLM_3163	mmpS2	Protein-coding
MLM_1982	mmpS3	Protein-coding
MLM_2607	mmpS4	Protein-coding
MLM_3862	mmpS protein homologous to MAH_4110	Protein-coding
MLM_3920	mmpS protein homologous to MAH_4169	Protein-coding
MLM_0064	mmpS protein homologous to MAH_0105	Pseudogene
MLM_3086	mmpS protein homologous to MAP_3050c	Pseudogene
MLM_3085	mmpL2	Pseudogene
MLM_0413	mmpL3	Protein-coding
MLM_0065	mmpL4	Pseudogene
MLM_2756	mmpL4	Protein-coding
MLM_2609	mmpL4_2	Protein-coding
MLM_2608	mmpL4_3	Protein-coding
MLM_3919	mmpL4_6	Pseudogene
MLM_3861	mmpL4_7	Pseudogene
MLM_2999	mmpL5	Pseudogene
MLM_3511	mmpL5	Pseudogene
MLM_3164	mmpL6	Pseudogene
MLM_2750	mmpL10	Protein-coding
MLM_0409	mmpL11	Protein-coding

MLM_1218	mmpL13	Pseudogene
MLM_2843	mmpL protein homologous to MAH_1462	Pseudogene
MLM_3116	mmpL protein homologous to MAH_3317	Pseudogene
MLM_4004	mmpL protein homologous to MAH_4604	Pseudogene

APPENDIX E

SUPPLEMENTARY FIGURES FOR CHAPTER 3 (FIGURES S2 – S3)

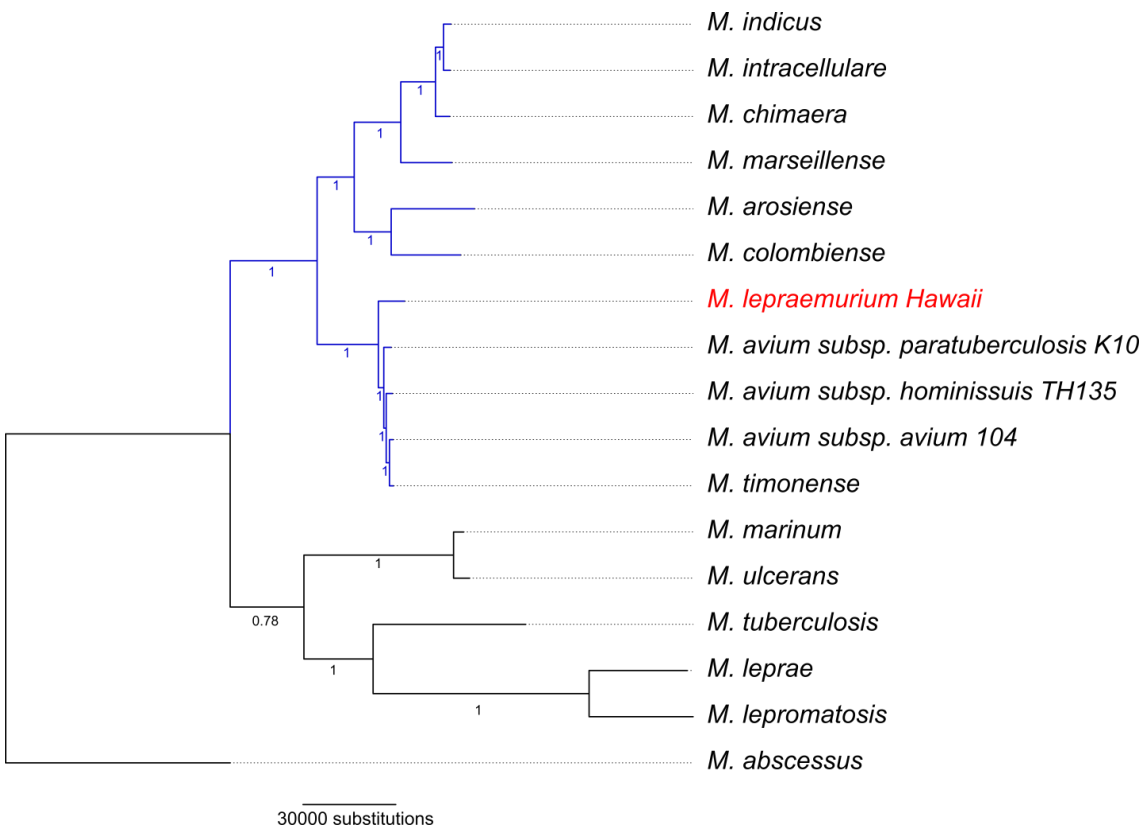

Figure S2. Maximum-Parsimony tree of *M. lepraemurium* and other mycobacterial species. *M. abscessus* was used as the outgroup to root the tree. The tree was based on 460,625 variable nucleotide sites and built using the SPR algorithm. Bootstrap support estimated from 500 replicates is given below each branch. Species belonging to the *M. avium* complex are highlighted in blue and *M. lepraemurium* is denoted in red.

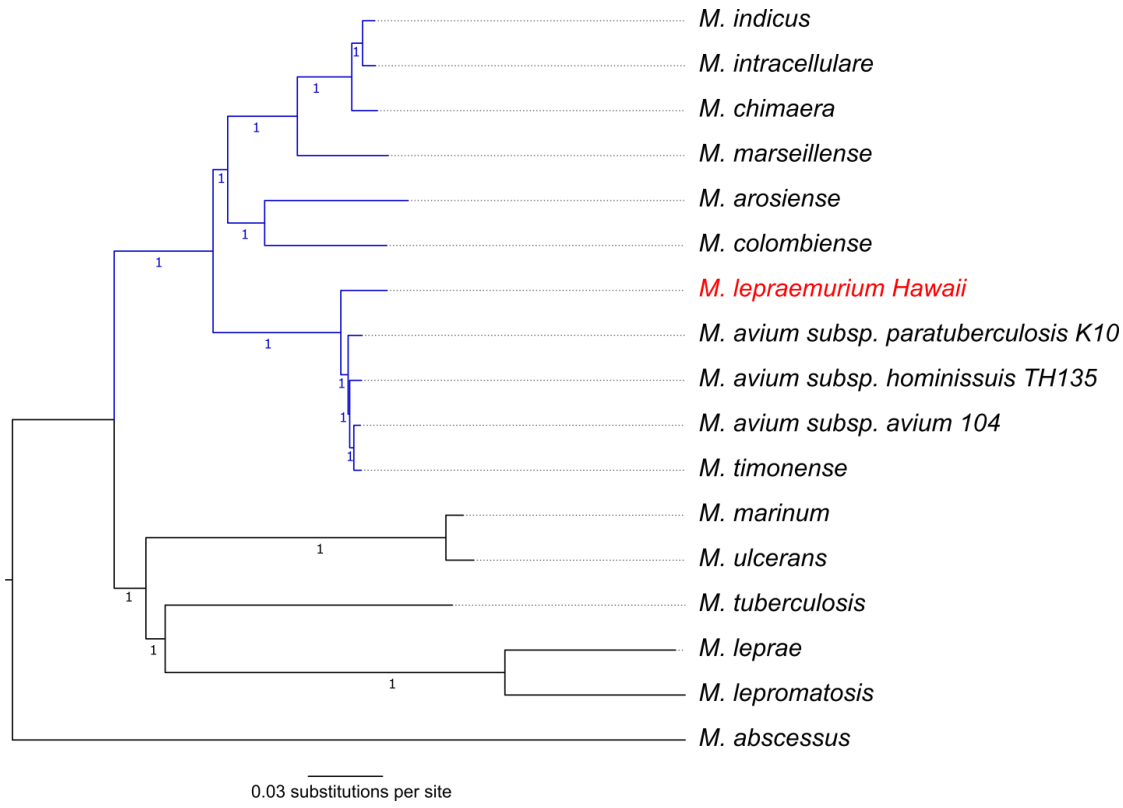

Figure S3. Neighbor-Joining tree of *M. lepraemurium* and other mycobacterial species. *M. abscessus* was used as the outgroup to root the tree. The tree was based on 460,625 variable nucleotide sites and built using the *p*-distance method. Bootstrap support estimated from 500 replicates is given below each branch. Species belonging to the *M. avium* complex are highlighted in blue and *M. lepraemurium* is denoted in red.

APPENDIX F

SUPPLEMENTARY TABLES FOR CHAPTER 4 (TABLES S6 – S10)

Consult Attached File using Microsoft Excel.

APPENDIX G

SUPPLEMENTARY FIGURES FOR CHAPTER 4 (FIGURES S4 – S11)

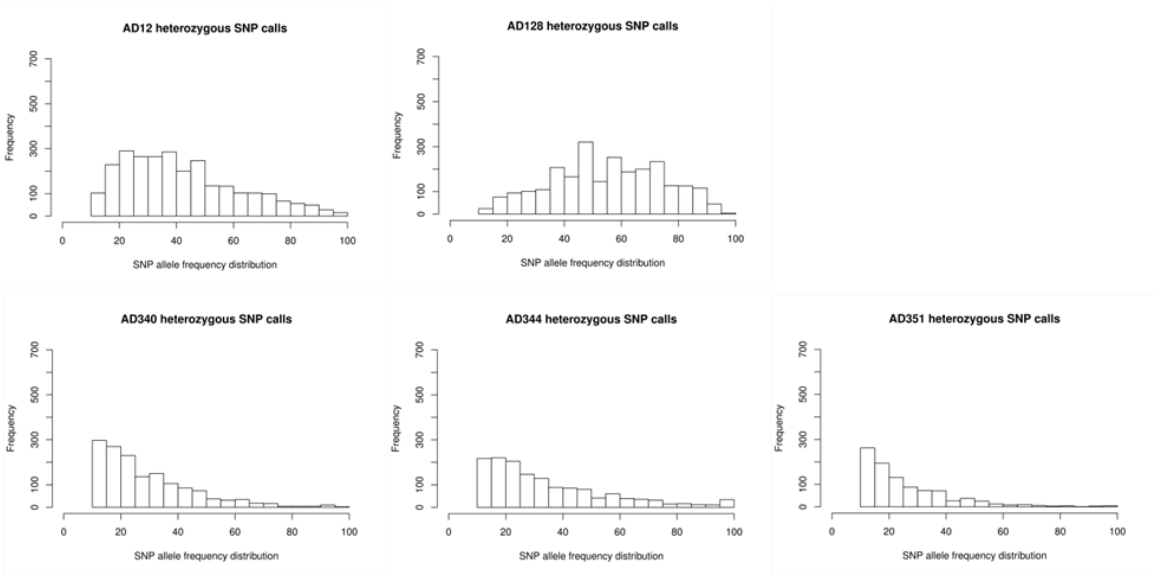

Figure S4. Histograms of SNP allele frequency distributions for heterozygous SNPs called in the unfiltered dataset.

Heterozygous SNPs were called if the SNP allele was covered by 10 - 90% of reads. The x axis denotes the frequency of reads covering the SNP allele (given in percentage) and the y axis denotes the number of SNP calls corresponding to the particular frequency.

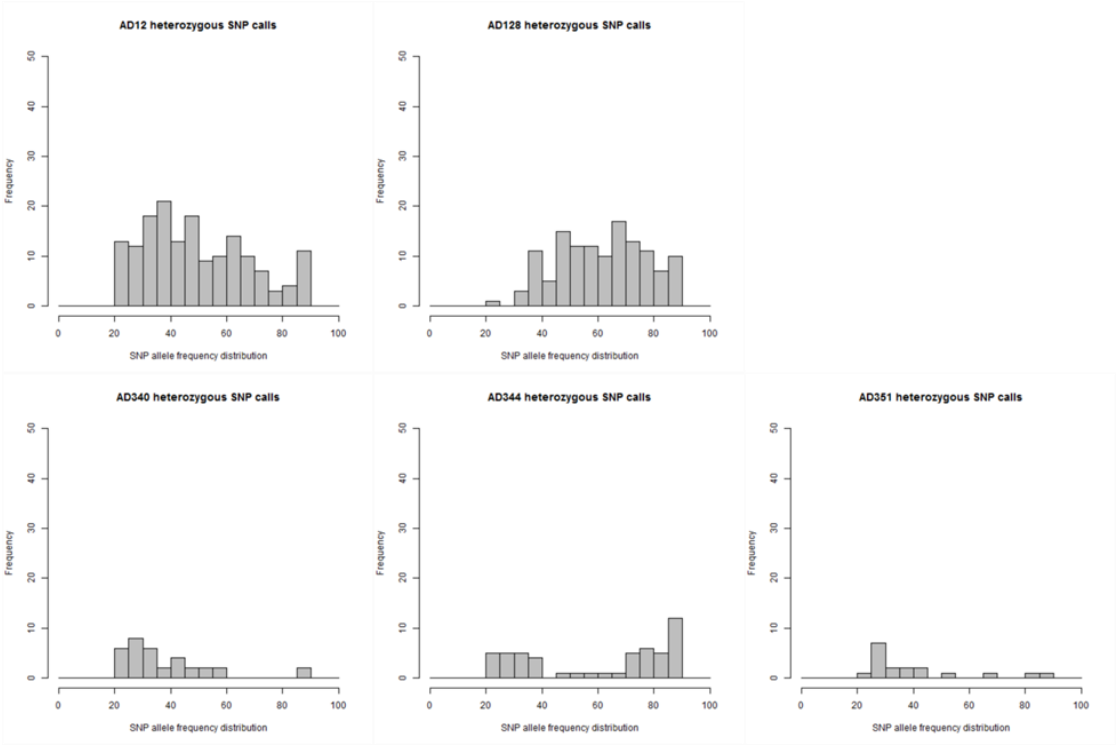

Figure S5. Histograms of SNP allele frequency distributions for heterozygous SNPs called in the filtered dataset.

Heterozygous SNPs were called if the SNP allele was covered by 20 - 90% of reads. The x axis denotes the frequency of reads covering the SNP allele (given in percentage) and the y axis denotes the number of SNP calls corresponding to the particular frequency.

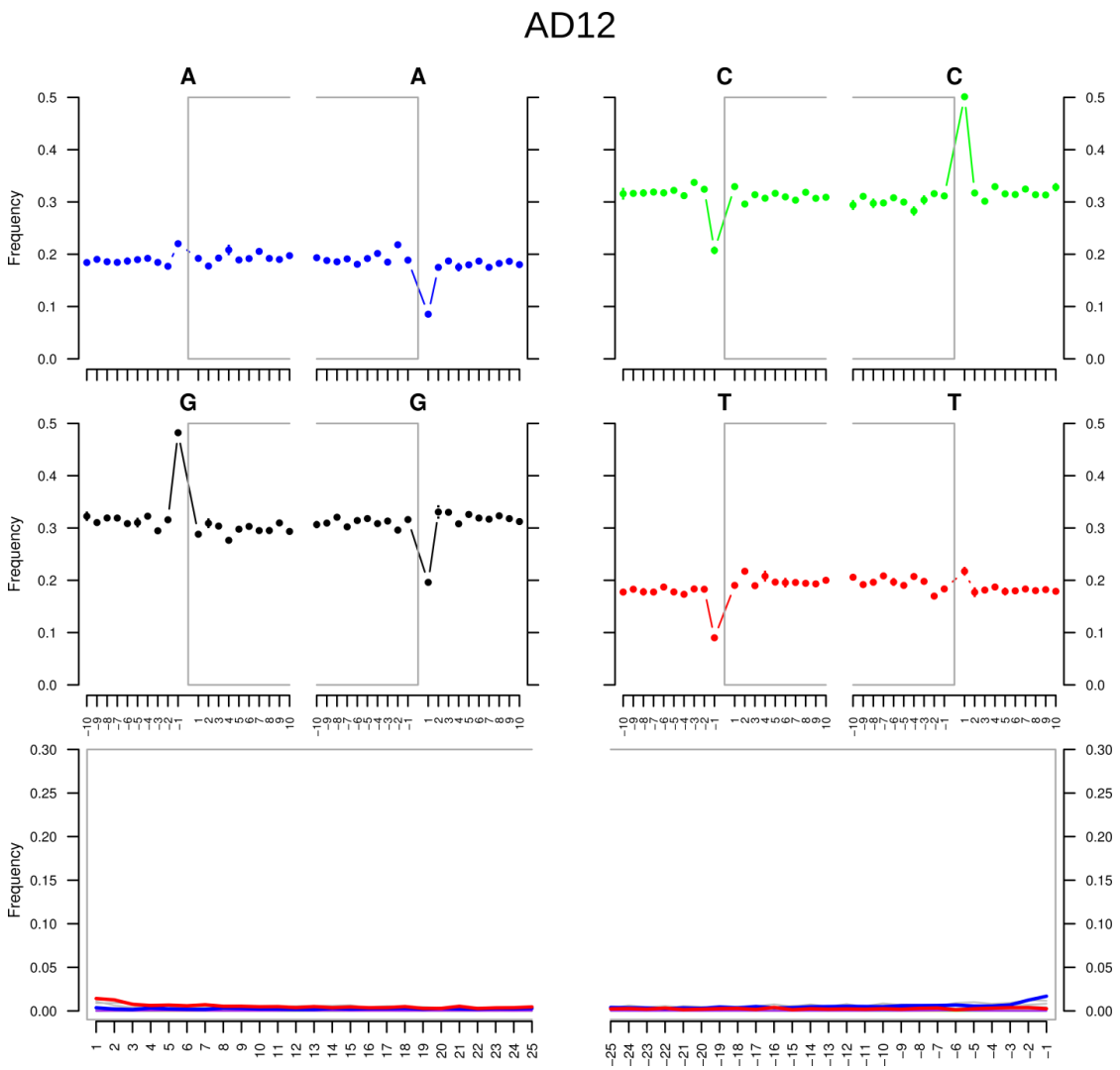

Figure S6. DNA damage patterns for AD12 (enriched non-UDG treated library).

a) Average base frequencies at positions within individual reads (grey box) flanked by all calls from reads in neighboring sequences. b) Frequencies of specific base substitutions at specific positions near the 5'-end (left) and 3'-end (right) occurring within reads. C-to-T changes are indicated by the red line and corresponding G-to-A changes by the blue line.

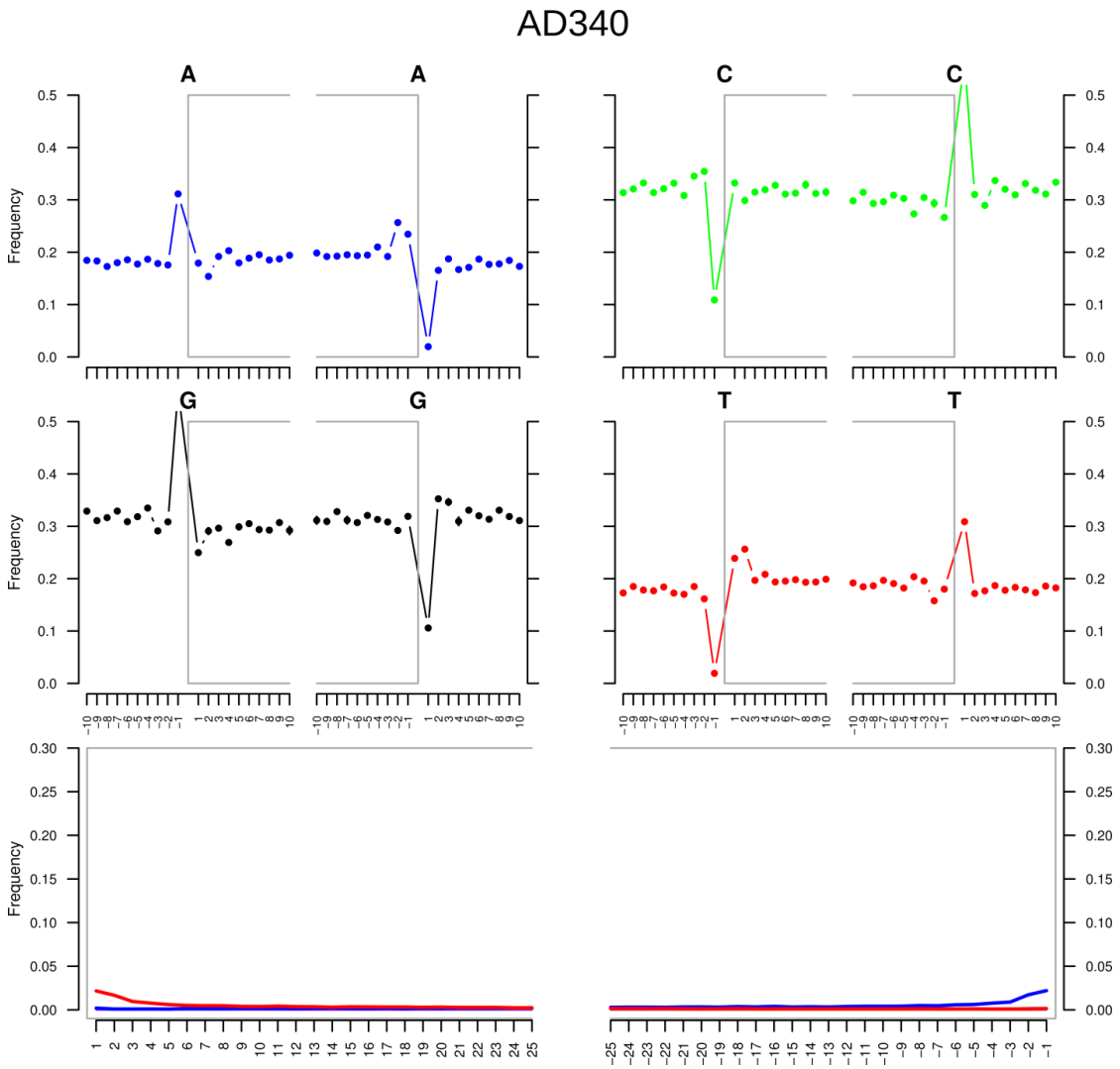

Figure S7. DNA damage patterns for AD340 (enriched non-UDG treated library).

a) Average base frequencies at positions within individual reads (grey box) flanked by all calls from reads in neighboring sequences. b) Frequencies of specific base substitutions at specific positions near the 5'-end (left) and 3'-end (right) occurring within reads. C-to-T changes are indicated by the red line and corresponding G-to-A changes by the blue line.

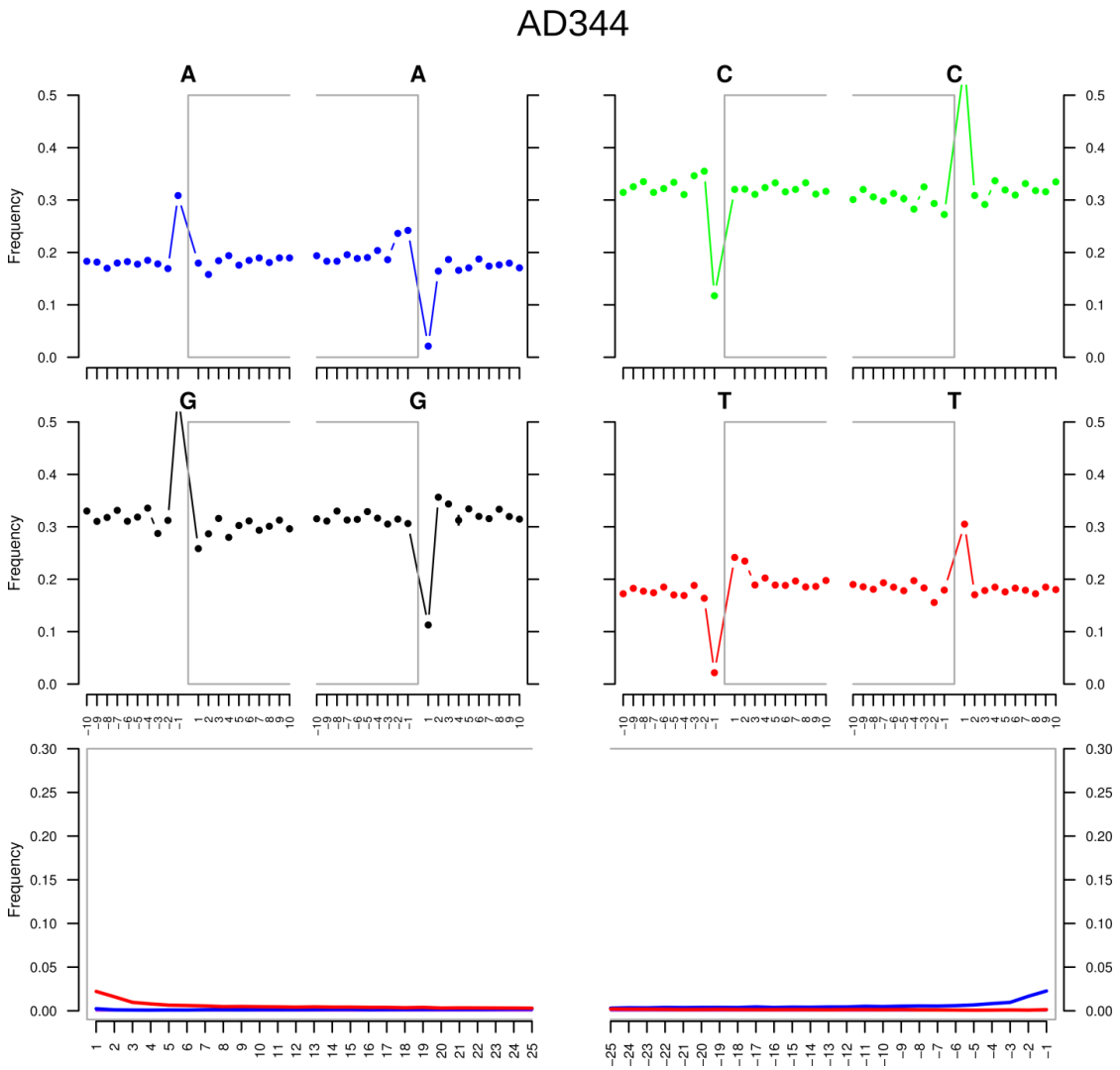

Figure S8. DNA damage patterns for AD344 (enriched non-UDG treated library). Average base frequencies at positions within individual reads (grey box) flanked by all calls from reads in neighboring sequences. b) Frequencies of specific base substitutions at specific positions near the 5'-end (left) and 3'-end (right) occurring within reads. C-to-T changes are indicated by the red line and corresponding G-to-A changes by the blue line.

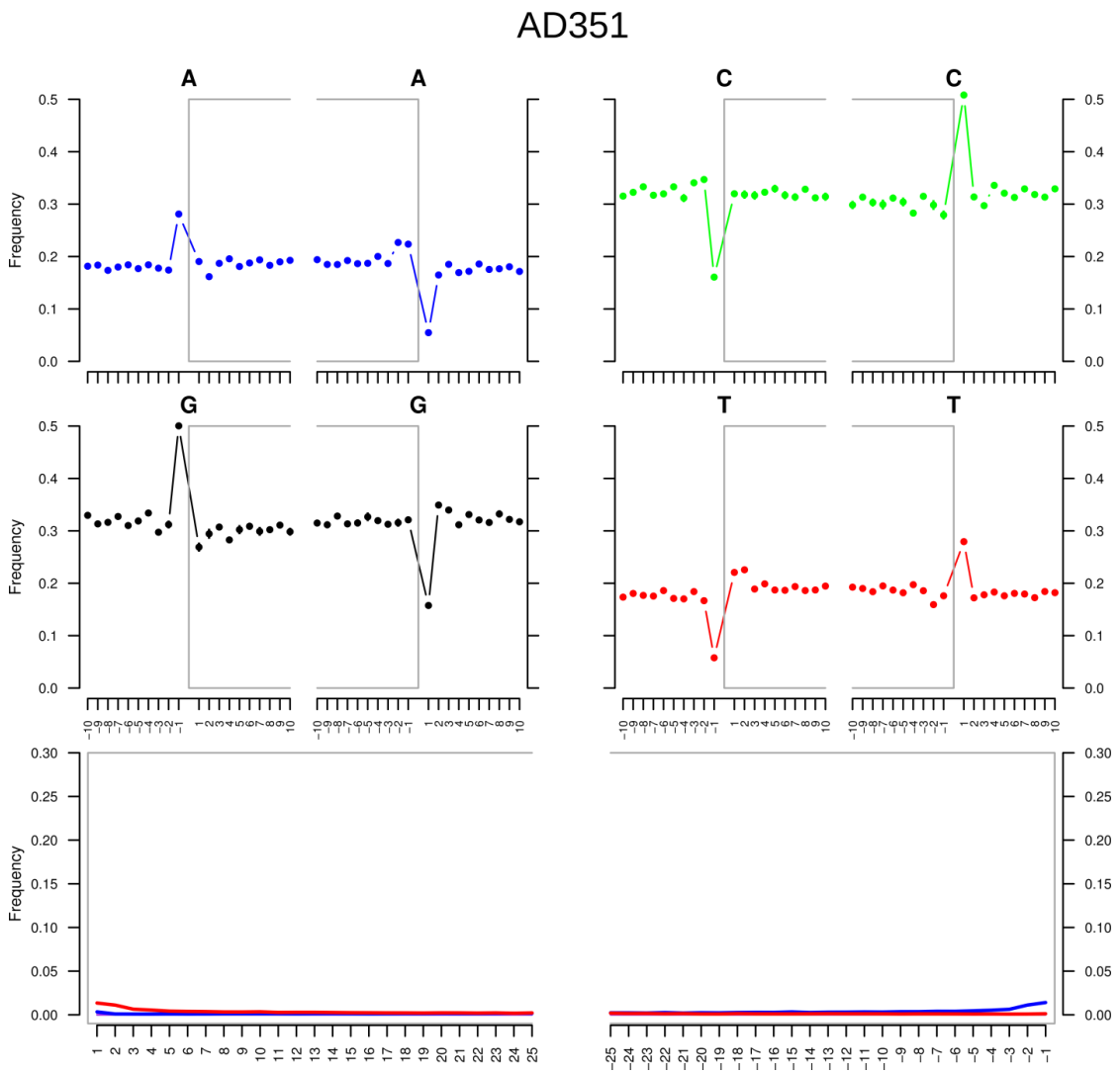

Figure S9. DNA damage patterns for AD351 (enriched non-UDG treated library).

a) Average base frequencies at positions within individual reads (grey box) flanked by all calls from reads in neighboring sequences. b) Frequencies of specific base substitutions at specific positions near the 5'-end (left) and 3'-end (right) occurring within reads. C-to-T changes are indicated by the red line and corresponding G-to-A changes by the blue line.

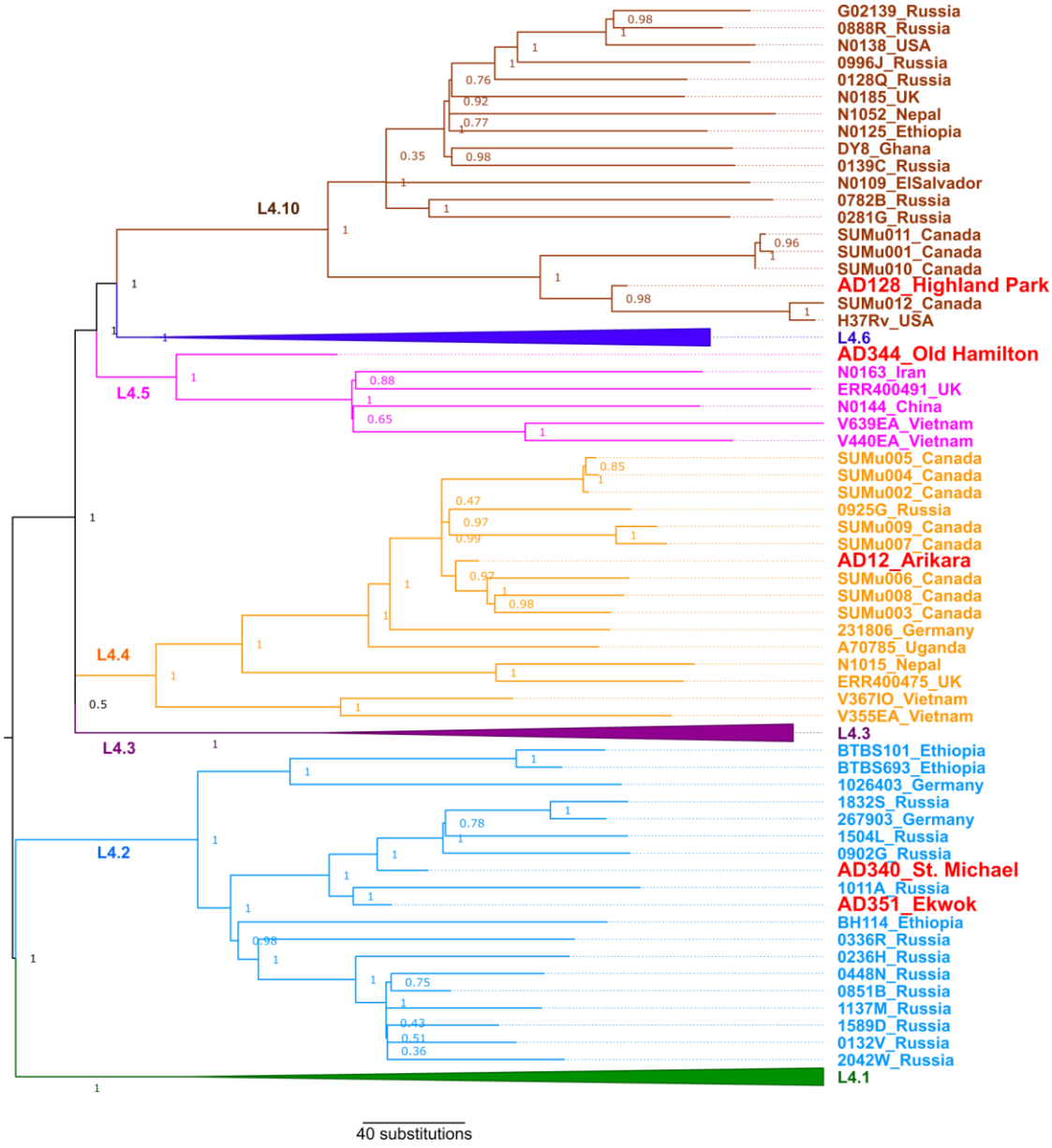

Figure S10: Maximum Parsimony tree of MTBC strains built using 9,775 variable nucleotide positions across 98 *M. tuberculosis* L4 strains. The tree was generated using the SPR algorithm and bootstrap support was estimated from 500 replicates. The *M. tuberculosis* L4 strains are color-coded by sublineages; certain sublineages are collapsed to save space. The ancient North American L4 genomes sequenced in this study are shown in red.

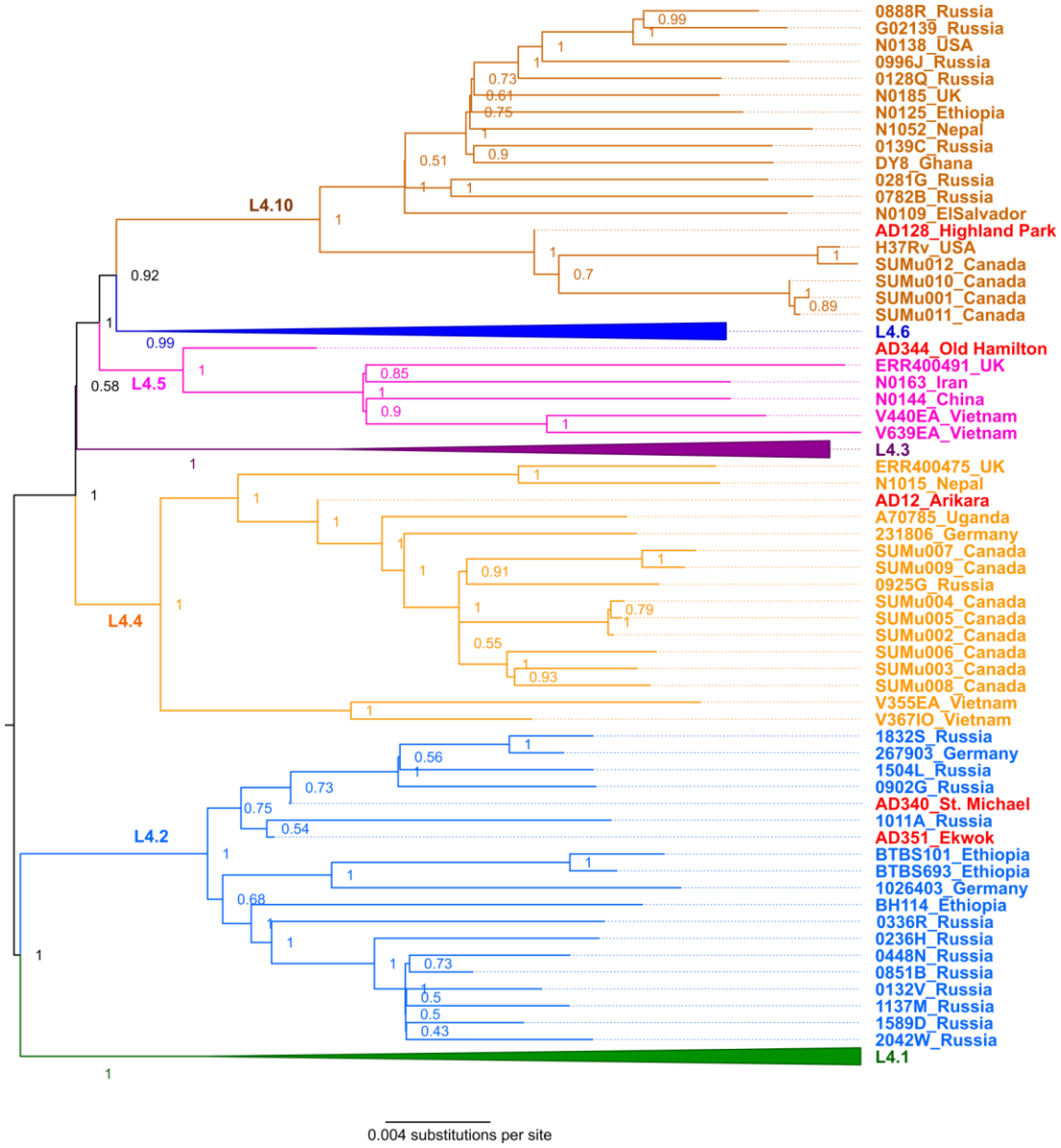

Figure S11: Neighbor Joining tree of MTBC strains built using 9,775 variable nucleotide positions across 98 *M. tuberculosis* L4 strains. The tree was generated using the *p*-distance method and bootstrap support was estimated from 500 replicates. The *M. tuberculosis* L4 strains are color-coded by sublineages; certain sublineages are collapsed to save space. The ancient North American L4 genomes sequenced in this study are shown in red.