MYCOBACTERIAL EVOLUTION AND ADAPTATION THROUGH COMPARATIVE GENOMICS

By

Evan Pierce Brenner

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Comparative Medicine and Integrative Biology – Doctor of Philosophy

2023

ABSTRACT

Members of the *Mycobacterium tuberculosis* complex (MTBC) inflict tremendous morbidity and mortality on humans and animals worldwide. These bacteria have co-evolved alongside humans and spread out of Africa with us. How environmental mycobacteria grew to become dedicated human pathogens, and how they adapted to infect domesticated livestock and beyond are unclear. Clues to their origins can be found in their genomes; gain and loss of genes, mutations, and genetic transfers between species can reflect specialization or diversification for new niches. Using whole genome sequencing projects from around the world, these hidden marks of adaptation can help explore pathways involved in host specificity, antigenic indicators of host immune subversion, and the footprints of evolution through analysis of conservation and exchange of CRISPR/Cas systems.

A genome-wide association study of MTBC genomes identified 120 loci associated with classification of *M. tuberculosis* variants *bovis* vs. others, which overlapped with an identical set associated with isolation from bovine hosts vs. others. These markers may be useful for SNP-based classification of MTBC variants, but more importantly, some are in genes involved in cholesterol and fatty acid catabolism, including genes known essential to grow on cholesterol. Adapting to new host lipid profiles may have allowed for a human host-specialist like *M. tuberculosis* to switch host reservoirs and expand to broadly infecting mammals as *M. bovis*.

By using validated epitopes from the MTBC from the Immune Epitope Database, the genome sequence of a uniquely attenuated strain of *M. bovis* – strain Ravenel – was analyzed for mutations in epitope-producing regions to identify changes. Such changes were rare among a reference virulent strain and an outbreak strain we also sequenced, but more commonplace

in a reference attenuated strain and strain Ravenel. No changes were predicted *in silico* to destabilize the mutant proteins, indicating the substitutions likely allow similar protein functionality, but may confer differential immune recognition by the host, a process known to be critical to mycobacterial infection and persistence and which could contribute to observed attenuation.

Finally, MTBC members are believed to be some of the few species of mycobacteria that carry "bacterial immune system" of CRISPR/Cas systems, and the MTBC and strains of closely related *M. canettii* are reported to be the only *Mycobacterium* to possess Type III-A systems. By searching for homologous *cas* genes and genetic contexts, previously unreported Cas systems of 4 classified types were found across a range of pathogenic and saprophytic mycobacteria, the closest related species outside the MTBC like *M. lacus, M. shinjukuense* showed only non-Type III-A systems, but the Type III-A system was identified in an unusual environmental *Mycobacterium*. The frequency of Cas systems and the rare nature of the MTBC Cas Type III-A support searching for it as a marker for the evolutionary origins of tuberculosis.

The work herein demonstrates the utility of existing datasets and software for discovery and provides three distinct paths forward for researchers to study the evolution and adaptation of *Mycobacterium* species to new hosts and environments: over 100 genetic loci associated with differentiation of specialist *M. tuberculosis* vs. generalist *M. bovis*, an understudied path for pathogen attenuation by epitope variation in the unique background of the MTBC, and the expansion of the mycobacterial CRISPR/Cas repertoire but continued rarity of the Type III-A system supports its use as a marker to trace the evolutionary history of the MTBC.

Dedicated to the lighthouse keepers in my life.

ACKNOWLEDGEMENTS

A comprehensive list of acknowledgements would exceed the length of this dissertation. An abridged version follows:

Foremost, thank you to Dr. Srinand Sreevatsan, for believing in me through everything, and for many years of mentorship, patience, encouragement, and thoughtful conversations (and funding). Thanks to my committee; Dr. Shannon Manning, Dr. Xuefei Huang, and Dr. Arjun Krishnan, for your support, feedback, and for taking the time from your own lives to guide me. I'm grateful to my many peers, and all the time we shared. A special thanks is essential for Dr. Fernanda Miyagaki Shoyama and Dr. Syeda Anum Hadi, and for the vast, ineffable powers generated by placing tired graduate students in the same rooms together for extended periods.

To my parents, Christine and Alan: I would not be who I am without everything you've done, and I am forever grateful for all of it. To my younger siblings, Lily and Kirby: you made it to adulthood and even appear to be successful in spite of my actions as the older brother, which is laudable and humbling.

Thank you to my friends who have become my family – Ren, Nick, Hunter, Emily, Erica, Chris, Autumn, and Alex. Thank you for helping me see all that life can be. Our histories, present, and futures together keep me going wherever we find ourselves. To promises made, time shared, and everything we can accomplish when we work together. I love you all. Each one of you knows how important you are to me, but now that I've published it, I can self-cite and remind you this way.

To Louie, the cockatiel on my shoulder throughout this PhD: you remain a menace.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION AND BACKGROUND 1
Introduction: 1
Classifying mycobacteria: 3
Mycobacterial cell wall composition: 6
Mycobacterial genomics and evolution: 7
Diagnostics of mycobacteria: 12
Pathogenesis of mycobacteria: 15
***M. tuberculosis* variant *tuberculosis* – Tuberculosis: 16**
***M. tuberculosis* variant *bovis* – Bovine tuberculosis: 19**
***M. tuberculosis* variant *bovis* strain BCG – The tuberculosis vaccine strain: 21**
***M. canettii* – The smooth tubercule cause of human tuberculosis: 22**
Beyond the MTBC – Non-tuberculous mycobacteria and disease: 23
***M. tuberculosis*-associated phylotype species – Tuberculosis-like NTM of humans: 24**
***M. leprae*, *M. lepromatosis*, and *M. lepramurium* – Hansen's disease and murine leprosy: 24**
***M. avium* complex – Diverse diseases of humans and animals: 26**
***M. kansasii* complex – A mix of saprophytes and pathogens: 27**
***M. marinum* and *M. ulcerans* – Evolution of mycolactone-producing mycobacteria: 28**
***Mycobacterium xenopi* – A significant aquatic-origin NTM pathogen: 30**
***Mycobacterium abscessus* complex – Fast-growing, drug-resistant pathogens: 31**
Mycobacterial epidemiology in summary: 32
Vaccine development in mycobacteria: 33
Conclusions: 35

CHAPTER 2: GWAS OF *M. TUBERCULOSIS* COMPLEX ISOLATES REVEALS GENES ASSOCIATED WITH DIVERGENCE INTO *M. BOVIS* 37
Abstract: 37
Introduction: 38
Materials and Methods: 39
Results: 41
Discussion: 44
Tables: 50
Figures: 66

CHAPTER 3: *M. BOVIS* STRAIN RAVENEL SHOWS CHANGES IN EPITOPES THAT MAY CONTRIBUTE TO ATTENUATION 70
Abstract: 70
Introduction: 71
Materials and Methods: 74
Results: 76
Discussion: 79
Tables: 88
Figures: 91

CHAPTER 4: MYCOBACTERIA BROADLY CONTAIN UNANNOTATED CRISPR/CAS SYSTEMS 95
Abstract: 95
Introduction: 96
Materials and Methods: 99
Results: 102
Discussion: 111
Tables: 120
Figures: 124

CHAPTER 5: CONCLUSIONS AND FUTURE RESEARCH 126

BIBLIOGRAPHY 131

CHAPTER 1: INTRODUCTION AND BACKGROUND

Introduction:

Mycobacteria include around 200 bacterial species[1,2]. Members of this shifting taxon share unique attributes of long carbon chain mycolic acids in their cell walls, characteristic acid-fastness, and an unusually high genomic guanine/cytosine content (GC%)[3,4]. Despite their commonalities, the diversity of mycobacteria is enormous. Pereira *et al.* aptly described mycobacteria as "colonizers of the total environment," and they range from environmental saprophytes living in soil, water, air, to opportunistic and obligate pathogens of birds, herptiles, terrestrial and aquatic mammals, invertebrates, and even protozoa[4]. Mycobacteria also frequently show remarkable resistance against acid and oxidative stresses, along with survival through a range of temperatures even including some types of pasteurization[5–8]. These adaptations, and resistance against various disinfectants, contribute to their frequent identification and persistence in the built environment, including water distribution and treatment infrastructure[8], and hospitals[9]. Many mycobacteria are non-pathogenic, but a growing number are understood to be opportunistic pathogens, and some of the deadliest human pathogens across the entirety of civilization are contained within the taxon[4,10,11]. The taxonomic grouping *Mycobacterium tuberculosis* complex (MTBC), for example, holds its infamous namesake, *Mycobacterium tuberculosis* (MTB). This professional pathogen causes tuberculosis (TB) in humans, a progressive and often fatal granulomatous infection of the lungs. The most recent common ancestor to the MTBC is thought to have arisen around 70,000 years ago and throughout recorded history, TB has caused significant mortality in human populations around the world[12]. Evidence of tuberculosis has been found in human remains from nearly

10,000 years ago, and written descriptions of TB were made in ancient India and China. From the 1600s to 1800s, it is estimated 1 in 4 deaths in Europe and North America were caused by TB, and among members of general working society in England, up to a third of this group died of the disease[12]. In 2014, modeling estimated 1.7 billion people worldwide are latently infected with MTB, and more than a million die to the disease annually[13]. A variant within the MTBC called *Mycobacterium tuberculosis* variant *bovis*, or *M. bovis*, infects an enormous range of animal species besides humans, causing tremendous losses not only to people, but to animal agriculture and conservation efforts as well[14]. Outside of the *Mycobacterium tuberculosis* complex are *Mycobacterium leprae* and *M. lepromatosis*, a pair of closely related mycobacteria that cause leprosy/Hansen's disease (HD)[15]. Like tuberculosis, HD has been present throughout human evolution, and molecular studies suggest it has migrated along with humans out of Africa and around the world[16]. While most *M. leprae* or *M. lepromatosis* infections do not result in disease, for those who develop HD, it can lead to permanent nerve, tissue, and cartilage damage and loss, joint deformities, and paralysis[17]. Human skeletal evidence in India shows the scars of HD dated to 2000BCE, and like few other diseases in human history, HD has always carried an immense social stigma[16]. It is worth noting use of the historical name – leprosy – has been discouraged by the World Health Organization due to this historical and modern social stigma and exclusion from society, and Hansen's disease is preferred[16]. Jumping to another broad class of relevant mycobacterial pathogens, non-tuberculous mycobacteria (NTM) are a growing global threat[4,18]. NTM are the category of mycobacteria outside the MTBC and excluding *M. leprae/lepromatosis.* This vast swath of species includes environmental mycobacteria that may be incidental pathogens, as well as pathogens of the

immunocompromised and immunocompetent[4,18]. The presentation of these infections varies greatly. Some cause pulmonary disease clinically similar to *M. tuberculosis,* as in *M. riyadhense*[19], but pulmonary involvement is common across species[20]. Wound and soft tissue infections that do not respond to standard antibacterial therapy are another typical sign, but most presentations are non-specific[18]. For details on the variety of symptoms and diseases, see Tables 1 and 2 in the manuscript by Dr. Pennington *et al.*, 2021[18]. NTM burden appears to be growing globally, though the reasons for this are unclear, and NTM-caused disease can have a significant mortality rate from 7% to nearly 70%, depending on species and normal clinical variables[4,18,20]. Just from this limited introduction, it's evident that mycobacteria are worthy of study. However, many facets of mycobacteria remain unknown despite the dedicated efforts of countless thousands of researchers over eons. Even the taxonomic classification and what is meant by *Mycobacterium* is in flux.

Classifying mycobacteria:

Mycobacterium refers to the gamut of ~200 accepted species, but in the last 5 years, reclassification has been proposed to separate the historical genus *Mycobacterium* into five genera[1]. Gupta, Lo, and Son (2018) note that by comparative phylogenomics, *Mycobacterium* reliably forms five monophyletic clades, and they propose splitting these clades into the genera *Mycolicibacterium, Mycolicibacter*, *Mycolicibacillus*, and *Mycobacteroides*, and *Mycobacterium*[1]. Others have argued that while a robust scientific basis for *Mycobacterium* being split into separate genera exists, splitting this taxon now will only result in confusion, that such a split does not serve any patient-focused purpose, and that reordering the taxonomy of so many species that cause human disease may in fact even lead to patient harm[2,21]. Finally,

Gupta, Lo, and Son do discuss in their manuscript that earlier taxonomic labels remain valid and can be used instead, and they do not propose that older taxonomy be erased entirely[1]. In this document, species will be referred to under *Mycobacterium* by default, and the split genera will be utilized only if they yield better at-a-glance estimation of relatedness relevant to a discussion, but it is important to note that this is not a settled subject and whether this taxonomic reclassification is widely adopted remains to be seen.

Mycobacteria have since the 1960s been sorted into two general groups: slow-growers, and fast-growers[1,22,23]. Slow-growers take more than 7 days to form colonies on solid media, and fast growers fewer than 7[1]. Slow-growing mycobacteria are more often of clinical relevance, and include those in the MTBC, *M. avium* complex (MAC) species, *M. kansasii* (MKC) complex species, *M. xenopi, M. ulcerans*, and *M. marinum*[3,18,23]. Fast-growing mycobacteria are more often clinically isolated incidentally, and are typically environmental saprophytes, but pathogenic fast-growers do exist like *M. abscessus* and *M. fortuitum*[3,18].

Another means of classification is in pigmentation, with pigmented mycobacteria producing yellow carotenoids[3]. The production of these compounds is triggered by light exposure in some species, and produced in the absence of light in others[3,4]. The Runyon classification system, developed in the 1950s, uses speed of growth (fast or slow growers) and pigmentation (pigmented with light, pigmented without light) to bin “atypical mycobacteria” or NTM, but is rarely utilized against more recent and robust methods of classification[3,24,25].

Classification of mycobacterial species shifts regularly as well. Since *Mycobacterium tuberculosis* was studied by Koch in the late 1800s, *Mycobacterium tuberculosis* (MTB) has been

a species in constant flux. Closely related organisms, like *M. tuberculosis* variant *bovis* (MBO) have been argued to represent a lineage of MTB, or a separate species. Further complicating this are other variants in the complex, like *M. tuberculosis* variant *caprae*, which had previously been classified as *M. tuberculosis* subspecies *caprae, M. bovis* variant *caprae,* and eventually *M. caprae*[26]. In 2018, MTBC isolates were studied by next-generation sequencing and digital DNA-DNA hybridization (dDDH) and it was concluded that similarity of all MTBC isolates far surpassed the threshold to consider them of the same species[26]. As such, it was proposed that the members of the MTBC be reclassified as variants of *M. tuberculosis*[26]. This extended even to *M. canettii*, the "smooth tubercule bacillus" believed to be at the evolutionary perimeter of the complex[23,26,27]. Finally, the case of *Mycobacterium kansasii* is one of species expansion. It was previously believed that *M. kansasii* included 7 subtypes, but two separate phylogenomic analyses by Tagini *et al.* and *Jagielski* et al. identified that these subtypes were genetically distinct enough to stand as different species within a newly created *M. kansasii* complex (MKC)[28,29].

The genera, species, and lineages of mycobacteria remain in flux, but they do share some major physiological characteristics. These species are generally non-motile rods, found in aerobic to microaerobic conditions, and while their niches vary, mycobacteria stain acid-fast – that is, acid decolorization is not effective – and this is a characteristic seen in only a handful of genera (*Gordonia*, *Mycobacterium*, *Nocardia*, *Rhodococcus*, and *Tsukamurella*)[3,4]. This acid-fastness is due to the atypical cell wall composition that features 60-90 carbon atom chain mycolic acids[3,4]. While mycolic acids are not exclusive to *Mycobacterium*, the length and prevalence of these long chain compounds are unique[4]. This waxy, unusual mycobacterial cell

wall is one proposed origin for the name *Mycobacterium:* the Greek-derived prefix myco-, for "fungal," signifies the genus occupies in a space between fungi and bacteria in its characteristics[4,25]

Mycobacterial cell wall composition:

Like many topics in mycobacteria, the composition of the cell wall is not well-understood. The arrangement of the mycobacterial cell wall is unique. Mycobacteria are often referred to as "Gram-positive" despite their acid-fastness, as they possess a Gram-positive-like peptidoglycan (PG) cell wall outside their membrane[4,11,25]. While this layer is very similar to the cell wall of Gram-positive bacteria, they also have a pseudo-periplasmic space and an outer, atypical membrane, leaving to an overall structure with similarities to Gram-negative bacteria, yet one evolutionarily completely distinct[11,30–32]. Even the PG itself is modified in some mycobacteria compared to PG found in traditional Gram-positive organisms[31]. Whether then a classification as Gram-positive is scientifically useful is debated[32], though others contend this pathway remains viable as a therapeutic target for antibiotics[31]. In detail, Mycobacteria bear a typical lipid bilayer plasma membrane (PM), a thin, poorly understood granular layer, and a layer of PG nearly as thick as that seen in Gram-positives that is partially linked to a distal layer of arabinogalactan (AG)[32]. The mycobacterial plasma membrane has been found to contain lipoarabinomannan (LAM) and trehalose monomycolate (TMM), as well as over 2,000 membrane proteins in *M. smegmatis*[33]. The critical MmpL complexes and ESAT-6 secretion systems are found here in *M. tuberculosis*[33–35]. Between the PM and PG is a periplasm-like space, or pseudo-periplasm[4,32,33]. Beyond and bound to the AG layer is the thick mycobacterial outer membrane, or the mycomembrane (MM), rich with the previously mentioned long chain

mycolic acids (MAs)[4,33]. The complex of PG-AG-MM is known as mycoloyl-arabinogalactan-peptidoglycan (mAGP)[4,33]. Beyond mAGP is the outermost layer (OL) – referred to as the capsule in pathogenic species – and the composition of this layer varies significantly between mycobacteria[4,32,33]. In HD-causing species, the capsule is made up of phenolic glycolipids (PGLs) and glycopeptidolipids (GPLs)[33,36]. In other mycobacteria including MTB, OL-localized PGLs are rare or absent[33]. For most slow-growers, the capsule is comprised of glucans, other polysaccharides, and polypeptides, while fast-growers exhibit a protein-dominant OL[4,33]. In either case, lipid composition of this OL/capsule is usually minimal[4]. Like in the plasma membrane, this fraction contains LAM, as well as TMM and TDM[32,33]. These molecules are known virulence factors in MTB, with LAM and TDM both known to impair host immune responses[32,37]. TMM is shuttled outwards from the cytoplasm by MmpL3, where it is believed to crucially contribute MA to the assembly of pAGP before processing to TDM through the Antigen 85 (Ag85) complex[32]. The synthesis and upkeep of a critical system like the mycobacterial cell wall necessitates an expansive genetic and metabolic repertoire[38,39].

Mycobacterial genomics and evolution:

M. tuberculosis variant *tuberculosis* strain H37Rv was first genome-sequenced in 1998[39]. It was quickly recognized that MTB showed an atypically high skew towards GC content (%GC), possessed a significant quantity of genes involved in lipid and fatty acid processing, had an unusual, novel class of genes encoding repetitive sequence proteins believed involved in pathogenesis (later termed PE/PPE genes), and was rich in insertion sequence elements[4,11,39,40]. When sequences from an *M. leprae* isolate and *M. tuberculosis* variant *bovis* isolate followed in 2001[41] and 2003[42], respectively, it began to be understood that some of these features were a

canon of mycobacterial life. *M. avium* subspecies *paratuberculosis*, a widespread agricultural pathogen of ruminants that causes Johne's disease, was found to have a GC% of almost 70%, MTBC species around 65%, and *M. leprae* at 57.7%, though a relationship was also noted by Marri, Bannantine, and Golding in 2006 that higher %GC values are associated with higher genomic gene coding content[40]. As many more genomes have been sequenced, high GC content has persisted as a mycobacterial hallmark[4]. Others, such as diversity and frequency of insertion sequences, or presence and type of CRISPR/Cas systems, have been seen to vary species-to-species[27,40,43,44]. The presence of widespread PE/PPE genes is a genetic separation point between slow-growers (which have numerous PE/PPE genes) and fast-growers (which do not appear to have undergone similar duplication events)[45,46].

Horizontal gene transfer (HGT) is another differentiating point for mycobacteria, and although early studies that suggested mycobacteria simply did not participate in HGT and this thinking has since been overturned, it is now understood that the mechanism of HGT in this group of organisms is often distributive conjugal transfer (DCT)[47,48]. This atypical type of conjugative transfer is unlike classical conjugation in that no *oriT* sequence is necessary, and classic transfer genes evaded detection[48]. Plasmids in mycobacteria are rare, and were confirmed not to play a role in this process, though one instance of a separate, conjugative plasmid has been documented in *M. marinum*[48,49]. Instead, mycobacteria in non-planktonic culture (i.e., growing on solid media or cohabiting biofilms) are able to directly transfer chromosomal DNA from a donor to recipient cell and integrate seemingly by homologous recombination, without an *oriT* sequence to mediate integration, and thought to utilize the mycobacterial ESX-1 and ESX-4 systems in unclear capacities[48]. This results in a unique, almost

meiosis-like mosaic genome[48,50]. Evidence has been shown in some isolates of the *M. kansasii* complex where mosaic genomes generated through DCT has led to reduction in pathogenicity [51].

Members of the MTBC are said to be clonal[23,52–54]. That is, in the MTBC, the genetic material of ancestral isolates is passed to progeny as replicated, without horizontal gene transfer or recombination events, and diversity within the complex is largely restricted solely by the base frequency of mutations. This is in contrast to NTM species, which regularly and widely undergo DCT[48]. *M. canettii* – sometimes said to be *M. tuberculosis* variant *canettii*, and other times left just outside the complex with its 98% genomic identity with MTB – shows a great deal of genetic diversity among its isolates, and readily undergoes DCT with other mycobacteria as a part of this process[27,55]. Signatures of DCT have been detected in very rare cases between MTB and *M. canettii* (a non-clonal organism), but it is speculated that due to the strong niche separation, this is a moot point as MTBC isolates simply do not have the opportunity to undergo transfers outside of their own population[48]. In more recent research, Madacki *et al.* strikingly demonstrated MTBC members including *M. tuberculosis* var. *tuberculosis* and *M. tuberculosis* var. *bovis* are actually able to donate chromosomal DNA through DCT, but were unable to ever serve as recipient[55]. The MBO BCG Pasteur strain was even demonstrated to transfer its region of difference 5 deletion to *M. canettii* STB-L, showing such transfers can have more than trivial effects[55]. Furthermore, they showed this mechanism is independent of ESX-1, in contrast with the findings in *M. smegmatis*[47,55]. Despite this, the clonality of MTBC members has remained evident in genomics, and apart from the absence of typical HGT, MTB has very

long generation times (≥20hr *in vitro*) and has been shown to have a modest mutational frequency[56,57].

As discussed briefly earlier, the *Mycobacterium tuberculosis* complex contains the human-adapted *M. tuberculosis* variant *tuberculosis* (MTB), and animal-adapted *M. tuberculosis* variants *bovis* (MBO), *caprae*, *orygis*, *microti, pinnipedii* and others. As a clonal population, even within MTB are evident evolutionary lineages, first comprehensively explored by Gagneux *et al*. in 2006 utilizing large sequence polymorphisms (LSPs) and regions of difference (RDs) to assign six global MTB lineages affecting humans worldwide[58]. It was also suggested such lineages could reflect host adaptation to different human populations worldwide, and a later analysis identified "specialist" and "generalist" lineages of MTB, with the former being geographically constricted, and the latter being prevalent globally[59]. Sublineage-level designations were also made, particularly for generalist lineage 4 of MTB, and it was noted that lineage 4 showed more variation in genomic regions encoding human T-cell epitopes than the strict epitope conservation shown in other lineages[59]. Curiously, it has been shown too that regions in the MTB genome encoding T-cell epitopes are actually hyperconserved – that is, they stand out as more highly conserved than the rest of the already inflexible genome[60]. Moving beyond LSPs and RDs, Coll *et al.* in 2014 published a "SNP barcode" of single nucleotide polymorphisms (SNPs) fixed in lineages and sublineages worldwide and meant to be used as a straightforward means of assessing ancestry of an unknown isolate[61]. Besides MTB, researchers used spoligotypes (based on patterns of presence/absence in genomic spacer sequencers) in placing MBO isolates into "clonal complexes," with an underlying hypothesis that such variation may reflect adaptation to different cattle hosts[62]. Years later, Zimpel *et al.* noted the existence of

global MBO lineages assignable by SNPs in a similar approach as that taken by Coll *et al*.[14]. However one chooses to assign or explain these lineages, their existence supports the clonality of the MTBC. As genomics has advanced as a field, reconstructions into MTBC evolutionary history has revealed a pattern of MTB evolution from an *M. canettii*-like environmental ancestor towards strict human pathogenesis, a gradual division into ancient and modern MTB lineages, and a much more evolutionarily recent split of the animal adapted variants towards pathogenesis outside of the human host[23,26,63].

Stepping further back, it is speculated that other evolutionary intermediaries exist between the *M. canettii*/MTBC clade and NTM, including *M. lacus, M. deciphiens*, *M. shinjukuense*, and *M. riyadhense*[10,64]. These isolates have been shown to fall more distantly but consistently into a monophyletic clade with *M. canettii*/MTBC isolates[10]. Further back still, phylogenomics suggests *M. marinum, M. avium* complex (MAC) species and *M. kansasii* complex species are more distantly related but still phylogenetically adjacent[11]. As distance increases from the clonal MTBC, the evolutionary picture before more muddled. *M. abscessus* (MAB), an NTM and pathogen of the lung commonly seen in patients with cystic fibrosis (CF), is evolutionarily far from the MTBC, and yet in recent years appears to be spreading clonally in both CF and non-CF individuals worldwide[65]. *M. ulcerans,* members of the MAC, and *M. kansasii* complex all cause substantial worldwide human disease and mortality, but their source remains unclear[4]. In all of these cases, while these species are thought to be closer to the prototuberculosis ancestor in terms of their free-living, environmental lifecycle, we are no closer to understanding where they reside, how and why they infect certain hosts when they do, and how best to control them.

Diagnostics of mycobacteria:

As this review started, the simplest diagnostic is by staining. The mycomembrane characteristically retains carbol-fuschin dye even after decolorization by acid-alcohol, rendering them "acid-fast" organisms[18]. A number of stains can used, but the most frequent are the Ziehl-Neelsen and modified Kinyoun stains, staining of MTB patient sputum has been reported to be most sensitive and specific by auramine O staining if fluorescence microscopy is available, and for HD-associated infections, Wade-Fite staining is suggested[18,66,67]. Regardless, one glaring problem was noted long ago – Koch's Paradox[68]. Mycobacteria are acid-fast, until they aren't; scientists discovered non-acid-fast bacilli from TB patients just a year after Koch reported techniques for identifying these TB-causing organisms, but these atypical bacilli were still able to cause the same disease in lab animals, which also led to the bacteria regaining their acid-fast nature[68]. Treatment with certain antibiotics like isoniazid also causes loss of acid-fastness[68]. Staining also cannot differentiate mycobacteria, nor does it definitively confirm the presence of mycobacteria since other taxa of bacteria can also stain acid-fast[3,4].

This brings us to the gold standard for tuberculosis identification: mycobacterial culture[18,69]. As alluded to earlier, culture is necessary to place organisms in Runyon groups, MTB and MBO can be differentiated based on their growth in differential media, and NTM still require culturing to differentiate the many possible species without genotypic investigation[4,18,67]. Culture-based confirmation and differentiation of MTBC isolates is routine, and commercial liquid culture kits are widely used, but NTM require more deliberation and lack standardized workflows for identification[67]. Culture has numerous drawbacks of its own, including that mycobacteria often come from diverse polymicrobial environments and

decontamination of non-*Mycobacterium* species is necessary, culture can take weeks or months for results, and the fastidious nature of many mycobacteria (e.g., pH, atmosphere, temperature, nutrient availability) can make specific isolation a challenge, particularly in resource-poor settings[4].

Another category of test relies on immunological responses generated by the host, including the widely used tuberculin skin test (TST, or intradermal comparative cervical tuberculin skin test in animals)[12,67,70,71]. An obvious drawback to reliance on immunological responses is for immunocompromised patients who may not mount effective or typical responses and therefore show false negatives, for newly infected individuals who may not yet have mounted an immune response to the antigenic cocktail in tuberculin, in late stages of infection where *Mycobacterium* antigen-specific T cell anergy can be observed, and furthermore, earlier vaccination against MTB through the BCG vaccine (discussed later in this chapter) may still produce positive reactions in the absence of infection[67,69,70,72–74]. Enzyme-linked immunosorbent assays (ELISAs) against mycobacterial infections are also widely used, typically focused on the release of interferon-γ (IFN-γ) (interferon gamma release assay, IGRA) released in a sample experimentally exposed to TB antigens, ESAT-6 and CFP-10, to detect prior mycobacterial exposure in a more sensitive and specific manner[75,76]. Commercially available IGRAs include QuantiFERON-TB Gold (QFT) for detection in humans, and BOVIGAM TB detection in ruminants. Despite this, many NTMs possess and express ESAT-6 and CFP-10, and indeed, some research has been done repurposing TB IGRAs for *M. kansasii* infections – put simply, a positive IGRA does not confirm MTBC, and even when this is the case, it does not differentiate

active vs. latent disease[67,77]. Infections or even incidental colonization by NTMs can lead to false positives by TST or IGRA[4,18,72,78].

Beyond these lie molecular assays, such as the GeneXpert MTB/RIF kit that has the recommendation of the WHO for use at the primary initial MTB diagnostic test[67,73]. This test is capable of detection of MTBC genomic DNA within 2 hours, and can give indications on potential for drug resistance against Rifampin as well, though a clear drawback to this is that such technologies only detect active TB, when patients are already infectious[67]. Diagnosis of non-MTB mycobacteria has fewer options, and the dated standard outside of culture remains pulsed field gel electrophoresis (PFGE)[4]. Typing techniques can also rely on the presence of insertion sequences in different species for multiplex PCR or restriction fragment length polymorphism (RFLP) analysis, including in *M. avium*[4,18,79]. Matrix-assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS) has been described as an excellent means of distinguishing many NTM, but it does require technical expertise, equipment, sufficiently dense and pure biomass for testing, and a spectroscopic profile for each species, not all of which currently exist[4,18,80].

Genome sequences have long been used for differentiation and diagnostics of mycobacteria, from spoligotyping in the 90s to separate MTB isolates into different groups based on the numbers of CRISPR repeats[81], utilization of lineage-fixed SNPs and regions of difference for RT-PCR melt-curve analysis[82], the aforementioned pattern and variety of insertion sequences[79,83], and 16s rRNA-based classification[4,18]. The availability of whole genome sequencing and the ability to interrogate the entire genome of an isolate or pangenome of taxa holds great potential for developing new diagnostic targets, whether they be species-specific

DNA for PCR amplification, SNPs for typing organisms or classifying drug resistance genotypes, or new biomarkers for detection[61,75,84–86], but does have the drawback that WGS technology routinely requires a culture step for enrichment of sufficient DNA for sequencing. A promising and developing approach combines pathogen-origin biomarker discovery by high resolution proteomics with the global wealth of whole genome sequencing data to identify pathogen-specific targets that yield high specificity and without the need for a host immune response[87–89]. In any case, mycobacterial diagnostics is a field with great potential for growth, particularly for distinguishing and differentiating the growing plethora of non-tuberculous mycobacteria of clinical concern[4,18,90].

Another area of increased attention for bacterial genome sequences is in the utilization of genome wide association studies (GWAS). This technique, leveraging the expansion genome sequence data from the falling cost and rising availability of Illumina sequencing, allows researchers to statistically associate phenotypes like drug resistance with genotypes, whether by the presence or absence of certain genes or pathways, or the existence of variants like SNPs in specific genes, at the scale of hundreds to thousands of genomes at a time[84,91,92]. While this technique has successfully been applied to detect drug resistance-associated alleles in MTB, application to other problems and mycobacterial species has been limited.

Pathogenesis of mycobacteria:

Continuing the trend, our understanding of mycobacterial pathogenesis is incomplete, but we know a variety of mechanisms of different mycobacterial species from different hosts and environments. Most work has been done in MTB, and this section will start with the basics from this model.

***M. tuberculosis* variant *tuberculosis* – Tuberculosis:**

M. tuberculosis var. *tuberculosis* is an obligate, intracellular pathogen with no known reservoir outside humans, where it can cause a chronic, granulomatous disease typically of the lung[86]. Entry of infectious MTB into in humans can have a variety of outcomes and clinical presentations. In brief, viable bacilli enter an individual, often through the airway, where they are either neutralized upon reaching the lung epithelium, or replicate and translocate to the interstitial layers of the lung, where they can induce the formation of a granuloma[86,93]. In this non-infectious phase, the bacteria are held at bay by the immune system and persist in a state of dormancy[86,93,94]. Individuals who test TST positive and/or IGRA positive, do not show symptoms of disease, and with clean chest X-rays are considered latent cases[94]. Nearly 25% of the human population is estimated to carry a latent tuberculosis infection (LTBI)[13]. Most people with LTBI live healthy lives without further complications, but for 5-15% of people, MTB re-emerges from latency months or years later and resumes causing active disease[75,93,94]. Treatment is through a multi-antibiotic regimen, typically lasting at least 3 months for LTBI, and 6 months for active TB[94,95]. Left untreated, tuberculosis can lead to severe tissue damage and inflammation, spread to multiple organ systems, and cause death in half or more of cases[96]. Drug (and multi-drug resistance) is commonplace in human TB and remains a growing problem[75,95,96].

At the tissue level, the entire process is more complex, nuanced, and only partially understood. The minimum infectious dose is believed to be only single-digit numbers of bacilli[97]. When inhaled, MTB encounters alveolar epithelial cells and their secreted alveolar lining fluid (ALF) that enhances phagocytosis[86]. ALF is reportedly a critical step in initial MTB

infection control[98]. Infection of alveolar macrophages either leads to prompt pathogen-killing and clearance of the disease, or mycobacterial persistence and replication inside the macrophage[75,86,93]. If the latter takes place, replication eventually leads to macrophage death, spilling infectious bacilli into the local environment and signaling for recruitment of additional macrophages, which start the cycle anew[86]. An inability to clear the infection leads to infected alveolar macrophages migrating into the lung interstitium[69,86,99]. In this environment, MTB infiltrates dendritic cells (DCs) which travel to lymph nodes where subsequent T-cell priming occurs, before recruitment to the site of infection[86,93,99]. Effective cell-mediated immunity develops 2-6 weeks after infection, and the adaptive immune response from $CD4^+$ T cells, plus the innate and adaptive response from macrophages, leads to both a slowing of mycobacterial replication and the formation of a granuloma[69,86,93,99,100]. The granuloma is initially an aggregation of macrophages and DCs, and incorporates B and T cells later[100]. At this stage, within the hypoxic confines of the inner granuloma, MTB enters a state of dormancy where a host is subclinically infected, and the immune system and MTB are at a long-term standoff[86,93]. Through factors that are not fully understood, MTB can become activated again, emerging from the granuloma, replicating, and overcoming an ineffectual host immune response and causing clinical disease[69,100]. This reactivation causes host infectiousness, and widespread damage from the host immune response's attempt to contain the pathogen[69,75,100].

At the bacterial level, the infection again becomes more complex. *M. tuberculosis* displays a sophisticated repertoire mechanisms for host immune subversion, starting with its initial ingestion by macrophages primarily through non-opsonic phagocytosis through recognition of pathogen-associated molecular patterns (PAMPs) like LAM and mannosylated

LAM (Man-LAM), with Man-LAM specifically known to impair phagosome-lysosome fusion[101,102]. Under normal circumstances, ingested bacteria in the phagosome are carried along as it fuses with endosomes and lysosomes, leading to an increase in acidification-inducing V-ATPase proton pumps, as well as exposure to hydrolases and oxidative stresses, ultimately causing bacterial death[101]. MTB subverts this process at multiple steps. Secretion of MTB PtpA, a tyrosine phosphatase, leads to dephosphorylation of host protein VPS33B involved in vesicle trafficking, as well as disruption of V-ATPase proton pump assembly[101]. Acidification is also directly countered through production of phagolysosome disruptor 1-tuberculosinyladenosine (1-TbAd), which neutralizes low pH and cause deformation of lysosome structure[103]. Saha *et al*. recently demonstrated that infected macrophages have heightened intracellular cAMP levels, known to cause depolymerization of actin networks around the cell and preventing effective transit towards lysosomes[104]. Exposure to LAM and TDM are also shown to induce NF-κB-driven expression of mIR-33, which results in macrophages accumulating lipids, transitioning towards the TB-friendly phenotype of lipid droplet-laden foamy macrophages[105–107]. At the same time, MTB PknG impairs phagosome maturation and disrupts proper NF-κB function. MTB ESX-1 plays numerous roles, secreting both immunodominant antigens ESAT-6 and CFP-10, and the former has been shown to play a role with PDIM in destruction of and escape from the phagosome[108]. MTB EsxH, secreted through ESX-3, impairs the macrophage's phagosome ability to repair damage caused to the membrane. MTB NdkA, a GTP-ase activating protein, has also been demonstrated to hinder fusion, and to interfere with the macrophage antimicrobial oxidative burst by interfering with the assembly and function of the NADPH oxidase complex along with MTB CpsA and PPE2[86,101,109]. Infected macrophages also fail to properly express MHC II, and this

phenotype is reproducible *in vitro* even if macrophages are exposed to killed MTB[101]. Research has identified MTB lipoproteins LprA, LprG, and LpqH behind this suppression, and separate work showed MTB serine hydrolase Hip1 actively hinders MHC II expression[101,110]. Even as macrophages become dysregulated and seek the failsafe of apoptosis, MTB proteins SecA2, SodA, NuoG, NdkA, Eis, PknE, SigH, PtpA, ESX-1, ESX-5, MPT64, and Rv3654c have all been shown to play roles in restriction of macrophage apoptosis or promotion of necrotic cell death, allowing MTB to replicate freely until necrosis, which produces inflammation and provides easier access to new host cells[69,101,108]. This is not a comprehensive list, but instead an illustration of a subset of mycobacterial genes involved in manipulating and subverting the host immune system. This fine level of control may come from MTB evolving alongside humans as an obligate pathogen, and Comas *et al.*'s findings of T cell epitopes in MTB showing hyperconservation make sense in the context of how the pathogen often benefits from host recognition in specific ways[60,108,111].

***M. tuberculosis* variant *bovis* – Bovine tuberculosis:**

While genetically very similar to MTB (99.95%+ nucleotide identity), *M. tuberculosis* variant *bovis* (MBO) surprises with its vast host range in comparison to MTB[14]. This organisms causes bovine tuberculosis (bTB), and in primary hosts like cattle, it manifests as a chronic, progressive, and granulomatous disease[14,85]. It often affects the respiratory system and can be transmitted through the air, though a clear understanding of transmission remains surprisingly absent for MBO[85]. It has been argued by Behr and Waters (2014) that tuberculosis in general could be primarily a lymphatic disease, with lung involvement important but secondary to lymphatic involvement[112]. In MBO infections, evidence has been built that indirect and oral

transmission may be a significant means of infection[112,113]. In the UK, where badgers (*Meles meles*) are a highly susceptible reservoir species, GPS tracking has revealed badgers and cattle rarely interact directly but still transmit disease between them, implying an environmental and likely oral route of transmission is in play[113]; badger-to-badger transmission through bites has also been described as a significant secondary route, with disease occurring in a higher proportion of those exposed by bites than the latent infection seen in aerosol/oral transmission, and faster and more robust disease development manifested by lesions throughout the body[114]. In deer, a reservoir species in parts of North America with frequent spillover to domesticated cattle, bTB produces lesions primarily in the retropharyngeal lymph nodes, though lung and other lymphatic lesions do occur[114,115]. Like MTB, MBO is found worldwide, and different lineages have been described[14]. Analysis by Loiseau and Menardo *et al.* (2020) estimated MBO originated in domesticated cattle in East Africa or potentially the Near East before accompanying humans in settling around the world[116]. Human disease with bTB does occur, though its frequency depends on country and population-specific risk factors and is still uncertain overall[117]. Human disease with bTB is more often associated with extra-pulmonary disease, but pulmonary involvement does occur[117]. Like with many mycobacterial diseases, clinical symptoms are not distinctive and include fever and weight loss[76]. Indeed, the primary source of infection in humans is through the consumption of contaminated animal products[76]. MBO is intrinsically resistant to pyrazinamide treatment, so differentiation of this disease from that caused by MTB can have critical clinical implications[118]. It has also been noted that in some cases, the causative agent in humans is not MBO, but actually *M. tuberculosis* variant *orygis* (MOR), including a case of transmission from human to a dairy cow in New

Zealand[119], and multiple studies showing human MOR cases linked to parts of India and South Asia[120–122]. This spillover to and from humans and animals is complicated, but reflects the reality of MTBC evolution, transmission, and distribution.

***M. tuberculosis* variant *bovis* strain BCG – The tuberculosis vaccine strain:**

Discussion of MBO in the context of greater mycobacterial disease requires mention of the BCG vaccine. As covered in Luca and Mihaescu's review, development of a vaccine against tuberculosis began in 1900 by scientists Albert Calmette and Camille Guérin[123]. From 1908, the pair grew a virulent isolate of MBO on a bile, glycerine, and potato medium that they had noted was capable of attenuating isolates grown in earlier years, and they passaged this strain forward 230 times until it failed to cause progressive tuberculosis in lab animals[123]. This was the origin of *M. tuberculosis* variant *bovis* strain Bacillus Calmette-Guérin, *M. bovis* BCG, or simply BCG[123]. At a molecular level, BCG has long been known to differ from fully virulent MTBC members by regions of difference (RDs)[75,85,124]. Chief among them is RD1, which encodes the ESAT-6 and CFP-10 proteins essential for full virulence in MTB and MBO[85]. Different BCG strains exist worldwide, and each has its own pattern of RDs and other modifications[124]. In humans, BCG is shown to confer limited protective benefits, though the exact measure of this protection varies based on factors including the age of the individual, which environmental mycobacteria the population has been exposed to, and the geographic region both in terms of latitude and climate[125]. While many strains of BCG do exist and each vaccine is unique in its genetic background, research has not shown a clear difference between these strains in protective efficacy[124,125]. What is clear is that the type of protection conferred by BCG in humans is inadequate – it is efficacious at preventing disseminated tuberculosis in children, but ineffectual

at stopping pulmonary tuberculosis and thus transmission[125]. BCG vaccination is also rarely associated with disease, either BCGitis (local) or BCGosis (disseminated)[126]. BCG is also given as a vaccine in animals, while the success varies based on the animal in question[85,127]. Previously, it has been found that vaccination of cattle with BCG may reduce the severity and transmission of tuberculosis, but that the improvements are modest enough that the current standard of "test and slaughter" – that is, testing a herd for TB, and culling the herd if a positive is found – is more effective than trying to vaccinate in many cases[127]. Another complicating factor is the ability to differentiate infected from vaccinated animals (DIVA). Animals vaccinated with BCG react to tuberculin skin testing as if they were tuberculosis-positive, meaning more specific and expensive testing is necessary to assess whether a herd is merely vaccinated or harboring a tuberculosis outbreak[127,128]. Nevertheless, a meta-analysis by Srinivasan *et al.* (2021) reports a ~25% protective efficacy of BCG vaccination in cattle, and suggests by modeling that this is enough of a benefit to justify its inclusion in bovine tuberculosis control when used in combination with other measures[129]. Vaccination of other animals has shown more limited benefits, with little change in susceptibility of disease and instead only a delay in onset or decrease in severity[127].

***M. canettii* – The smooth tubercule cause of human tuberculosis:**

M. canettii is the sometimes-member of the MTBC at its periphery – it is the most diverse in any comparison between MTBC members, and diversity even within a small number of geographically constrained MCAN strains is greater than that observed between variants within the MTBC worldwide[27]. The digital DNA-DNA hybridization (dDDH) approach by Riojas *et al.* that provided clear evidence that MTBC members were variants of the same species

reported a mean dDDH value of 89.8% to MTB H37Rv, compared to values of ~97.1–98.6% for other MTBC variants[26]. While clearly distinct, this is far above the 70% cutoff for species-delineation for standard DDH, and the 80% cutoff used by the authors for dDDH, indicating MCAN can be classified as *M. tuberculosis* variant *canettii*[26]. That said, MCAN genomes do not show MTBC-like clonality, readily undergo horizontal gene transfer, cause phenotypically varied tuberculosis in humans, do not transmit human-to-human and are believed to come from an undiscovered environmental reservoir, and show striking morphological differences on solid media[26,27,50,130]. Most cases of *M. canettii* have been identified in the Republic of Djibouti on the Horn of Africa, and of these cases, most appear to be part of an epidemic clone since the 1980s[130]. In a study by Blouin *et al.*, 8.7% of TB cases across 3 years at the Bouffard Military Hospital were found to be caused by MCAN[130]. Unlike others in the MTBC, MCAN genomes show several different types of CRISPR systems, believed to be very rare among Mycobacteria[43,50,130]. MCAN isolates are intrinsically more resistant to certain anti-TB drugs, and diagnosing an infection as MCAN rather than MTB is thus clinically important[131].

Beyond the MTBC – Non-tuberculous mycobacteria and disease:

Outside the *M. tuberculosis* complex of species are numerous mycobacteria capable of causing human disease, ranging from opportunistic and limited infections to progressive, fatal illnesses. Mycobacteria in this group – those that cause Hansen's disease (HD) and those collectively referred to as NTMs, have caused substantial harm worldwide, and disease caused by NTMs has been growing for unclear reasons for decades[4,9,20]. Describing all diseases, niches, and knowledge on NTMs is beyond the scope of this review, but some of the key pathogens and groups will be discussed.

***M. tuberculosis*-associated phylotype species – Tuberculosis-like NTM of humans:**

Similarly to *M. canettii*, other NTM species exist which have been proposed to serve as a sort of evolutionary representative of the transition between environmental mycobacteria and professional human pathogen capable of causing tuberculosis. These include species like *M. riyadhense*, identified in Saudi Arabia and initially misdiagnosed as MTB[64]; *M. lacus,* discovered from a case of bursitis in Canada[132]; *M. shinjukense*, isolated from multiple immunocompetent patients across Japan[133]; or *M. decipiens,* found originally in cases from the United States and Italy of extrapulmonary granulomatous growths[134]. These globally distributed species have all been found to show genetic similarities to the MTBC and cause human disease, leading to the creation of a new categorization system – the *M. tuberculosis*-associated phylotype, or MTBAP[10]. This group, specifically containing the species above, fall into a monophyletic group based on a 107 conserved gene phylogeny by Sapriel and Brosch[10]. These infections are not widespread, but do represent the closest identified relatives to *M. canettii* and the MTBC, with *M. decipiens* the closest of the set[10,134].

M. leprae*, *M. lepromatosis*, and *M. lepramurium* – Hansen's disease and murine leprosy*:

Across human history, few diseases have been as feared or stigmatized as leprosy, or more appropriately, Hansen's disease[16]. Affected individuals can suffer scarring, nerve damage, loss of digits, and skeletal deformities, as well as some neurological conditions that have been assigned to either a possible mechanism of the disease, or more likely, the result of people with HD historically experiencing systemic, permanent exclusion from society[16,17,135]. The causative agent of this illness is either *M. leprae* or *M. lepromatosis*, closely related mycobacteria both believed to primarily spill over from animal reservoirs[15]. *M. leprae* was identified in the 1800s

as the cause of HD, and the contribution by *M. lepromatosis* only recognized in 2008[15]. The vast majority of HD cases are caused by *M. leprae*, but the exact amount caused by *M. lepromatosis* remains unclear[15]. *M. leprae* is a difficult organism to work with in a laboratory setting, and the only way to culture it is in the reservoir host of armadillos[15,136]. Transmission from human-to-human does occur, and while many routes of transmission have been postulated, none have conclusively been demonstrated as responsible for disease transmission[136,137]. Both species have been shown to have undergone massive genome reduction and gene decay, with an abundance of pseudogenes and very small genome sizes compared to other mycobacteria[41,138,139].

M. lepraemurium is, like *M. leprae*, an organism with a genome that has undergone extensive genome reduction, with nearly 1/3 of its genes as pseudogenes[140]. It was first discovered in rats, and is classically known as the cause of murine leprosy, though it can cause disease in other hosts as well[140,141]. At the time of its discovery, it was thought to be the causative agent of Hansen's disease, but since then, researchers have come to appreciate that *M. leprae/lepromatosis* and *M. lepraemurium* have many differences[141]. While both cause similar diseases in their hosts, are non-cultivable, show immunological cross-reactivity, and are structurally similar, they have very different mechanisms of action within a host macrophage, with the former showing a TB-like impairment of phagolysosomal maturation and escape from the phagosome, and the latter preferring to remain inside the phaogosome and takes advantage of lysosomal enzymes in its lifecycle[141]. Murine leprosy does not show a preference for peripheral nervous tissue either, a key distinction from the species causing HD[140]. *M. lepraemurium* enters the phagosome without triggering the normal oxidative burst, bypassing

the need to subvert or neutralize it as in other mycobacteria[141]. Curiously, modern genome sequencing has found *M. lepraemurium* is not particularly related to the *M. leprae/lepromatosis* cluster, instead falling within the *M. avium* complex[140].

***M. avium* complex – Diverse diseases of humans and animals:**

M. avium complex (MAC) members include *M. avium* and its subspecies (*paratuberculosis, avium/silvaticum, hominissuis*) and *M. intracellulare*, as well as the previously mentioned *M. lepraemurium*[79,140,142]. *M. avium* subspecies *hominissuis* is the dominant human clinical isolate from the complex worldwide, causing disease of the pulmonary and lymphatic systems, as well as wound and soft tissue infections, with an increased burden of disease in immunocompromised individuals[79]. In developed countries, disease from the MTBC variants is less frequent than disease from NTM, and MAC subspecies are a dominant subset of these infections[4,20,24,143]. Treatment of human MAC infections requires long antibiotic treatment regimens, and culture conversion is sometimes difficult to achieve even after 12 months[24,142].

M. avium ssp. *avium*, sometimes referred to as ssp. *silvaticum*, primarily causes disease in birds, and disease manifestation is primarily non-pulmonary, affecting the liver, intestines, spleen, as well as the bone marrow[144]. While this disease does pose a threat to birds broadly, it is not commonplace in the poultry industry, where sanitation and production practices limit its potential for its primarily oral (contaminated soil and water) spread[144,145]. Affected flocks are culled, and treatment not indicated due to concerns of encouraging resistance to human anti-tuberculosis drugs[145].

M. avium ssp. *paratuberculosis* (MAP), on the other hand, has major impacts on the cattle industry around the world, causing a chronic, progressive, granulomatous enteric pathology and subsequent wasting called Johne's disease (JD)[74,146,147]. The disease presents very slowly, with infections often occurring through mother-calf transmission shortly after birth and clinical signs not developing until years later[74,147]. JD is untreatable, and immediate isolation and/or culling affected animals is the recommended approach for control[147]. Unfortunately, MAP infection of dairy cattle is extremely common in many countries including the United States, where prevalence estimates suggest that up to 70% of dairy herds contain infected animals, as well as 5-10% of beef herds[147]. JD is associated with chronic wasting and weight loss, as well as a reduction in milk production[74,147]. MAP is both present in the milk of infected animals, and is seemingly able to survive commercial pasteurization, leading to its isolation from products like powdered infant formula, milk, and cheese[7,148,149].

***M. kansasii* complex – A mix of saprophytes and pathogens:**

The *Mycobacterium kansasii* complex (MKC) is a recent development, a reclassification and elevation of seven *M. kansasii* subtypes into independent species within a complex[28]. Some of these species, like *M. attenuatum* or *M. innocens* are considered most likely either non-pathogenic and incidental colonizers, or only pathogenic under exceptional circumstances, while *M. kansasii* is a frequent human pathogen with substantial morbidity and mortality in certain populations[20,28,29,51]. The sources of these mycobacteria are uncertain, but believed to be environmental[4]. *M. kansasii* has been found rarely in soil and natural water sources, but is more commonly isolated in tap and hot water systems, as well as an instance of a housecat in Japan[4,77,150]. Prevalence appears higher in more urban, industrialized environments, and

mycobacteria are known to be frequent and resilient in built water systems, but isolates from patients appear genetically clonal and distinct from identified environmental isolates thus far, complicating the epidemiology[4,77]. What has been noted is a gradual decline in prevalence as opposed to the generally increasing trend for NTM infections as a whole[77]. In humans, *M. kansasii* infections are pulmonary in 90%+ of cases, and present similarly to tuberculosis, and disseminated disease is rare except in cases of immunosuppression[77]. *M. persicum* is a less frequent cause of disease, and one that predominantly infects immunocompromised/HIV-positive individuals[29,151,152]. Fairly little is known about the lifestyle or reservoir for most members of this complex despite the clinical importance of former subtypes I and II (*kansasii* and *persicum*).

***M. marinum* and *M. ulcerans* – Evolution of mycolactone-producing mycobacteria:**

There are multiple mycobacterial species that are capable of mycolactone production, which are grouped into a loose taxonomic group of "mycolactone-producing mycobacteria" (MPM)[153–155]. Mycolactones are large polyketide compounds with potent cytotoxic and immunosuppressive effects, capable of disruption of normal cell migration and adhesion, induction of necrosis, as well as analgesic and immunomodulatory effect [153–156]. MPM are associated with human disease, but an interesting study by Hammoudi *et al*. suggests the function of mycolactone may actually be as a chemoattractant for fungi in polymicrobial environmental communities, such that *M. ulcerans* can exploit the nutrient-rich degradation pathways of fungi for its own nutrient uptake[154,157].

Buruli ulcer is a neglected tropical disease characterized by large, severe, necrotizing soft tissue infections that are paradoxically nearly painless, identified most commonly in Ghana,

Benin, and Côte d'Ivoire, southeastern Australia, as well as parts of Central and South America and the rest of the Western Pacific[156,158]. It is caused by *M. ulcerans*, whose pathogenesis is caused by mycolactone synthesized from a large plasmid pMUM001 found in this species[156,158]. *M. ulcerans* is predominantly identified in human hosts, but no evidence exists to support transmission between people[158]. Instead, it has often been isolated in water, and associated organisms like fish, frogs, plants, fungi, and invertebrates, including in a robust environmental survey by Garchitorena *et al*. of 32 aquatic communities in Cameroon sampled across a year[158,159]. It is believed that either exposure to contaminated water sources, or possibly vector-borne transmission through organisms like water bugs which have previously been shown to carry *M. ulcerans*, or similarly by mosquitoes in Australia[159]. Buruli ulcer has also been described in koalas, horses, and possums, with one case of a possum *M. ulcerans* isolate's genome sequencing returning only 2 SNPs against a human clinical isolate in the region, supporting a shared exposure to an environmental source[158].

M. marinum is another pathogen with an aquatic environmental source, is genetically very similar to *M. ulcerans*, and is believed to perhaps represent an earlier lineage before *M. ulcerans* split off evolutionarily as the latter acquired the pMUM001 plasmid and underwent genome reduction as it specialized[155,160]. *M. marinum* is technically a fast-growing mycobacterium by growth rate, but groups cleanly with slow-growers by 16s rRNA[160]. Indeed, *M. marinum* groups proximally with the MTBC clade in phylogenies, and it is used as a fast-growing proxy of *M. tuberculosis* infection in some fish models, rapidly developing into disease with associated MTB-like systemic and granulomatous infections in said models[160,161]. Further, *M. marinum* contains ESX-1, ESAT-6, and CFP-10 critical to MTB virulence, and even returns

false positives by TB tuberculin skin testing and interferon gamma release assays in human cases, highlighting more similarities[160,161]. *M. marinum* grows at lower temperatures, and was first identified as a disease in humans by skin infections, usually at the extremities like the hands[160,161]. These infections were traced to aquatic sources like swimming pools before chlorination became commonplace, and many human infections in the present day are instead due to contamination from hobbyist aquarium exposures[160]. Some *M. marinum* isolates from infected fish from the Red Sea region have since been shown to also produce mycolactones, though human disease by mycolactone-producing *M. marinum* has yet to be identified[160].

Other rare disease-causing mycobacterial isolates, like *M. liflandii* and *M. pseudoshotsii* have also been identified as very close to *M. marinum* and *M. ulcerans* (98%+ genetic identity) and capable of different types of mycolactone production[154–156]. Strains of *M. marinum, M. liflandii*, and *M. ulcerans* have all been identified as associated with or disease-causing in amphibians as well[4,155].

***Mycobacterium xenopi* – A significant aquatic-origin NTM pathogen:**

Another major NTM burden on human health is from *M. xenopi*, originally identified from lesions on *Xenopus laevis*, the African clawed frog[162]. Since its discovery, *M. xenopi* infections have become an increasingly common and deadly infection of humans (51%-69% 5 year mortality rate), especially across Europe[20,162,163]. Infections can be divided into pulmonary cavitary, nodular, or infiltrate forms[20,163]. As with many mycobacteria, tap and hot water systems have been implicated as a source, and rare infections have been observed in other species, such as domestic cats, ferrets, birds, and swine[4,164]. A genetically related pathogen, *M.*

heckeshornense, named for the Heckeshorn Lung Clinic in Germany where it was originally isolated, is also known to cause disease in humans[165–167].

***Mycobacterium abscessus* complex – Fast-growing, drug-resistant pathogens:**

The last specific group to be discussed in this review are the species of the *M. abscessus* complex (MABC), a group of related organisms that are often drug-resistant, increasing in prevalence, and characterized by severe pulmonary infection in humans[168]. Like the MTBC and MKC groups, taxonomic classification for the MABC has shifted since its initial discovery, when *M. abscessus* was shortly after discovery suggested to be a subspecies of *M. chelonae*[168,169]*.* Since, it has been recognized both that *M. chelonae* and *M. abscessus* are distinct enough by DDH to be different species, and that *M. massiliense, M. bolletii* and *M. abscessus* can be considered subspecies of a newly defined *M. abscessus* complex[168,169]. Presently, *M. abscessus* is said to contain 3 genetically similar subspecies with divergent clinical presentations – *M. abscessus* ssp. *abscessus*, *M. abscessus* ssp. *bolletii*, and *M. abscessus* ssp. *massiliense*[168].

The complex is the primary cause of pulmonary illness caused by fast-growing mycobacteria, and is unusual among NTM in that WGS-based epidemiological studies have concluded that MABC isolates can indeed be transmitted human-to-human rather than solely acquired from contaminated environments[169]. Regardless, like many NTM, MABC is known to be a frequent isolate from water and built water infrastructure, including in showerheads and faucets[169]. MABC infections are intrinsically resistant to all anti-tuberculosis drugs, and *in vitro* susceptibility to other compounds does not reliably translate to clinical efficacy, posing a significant threat to health and leading to a relatively frequent misdiagnosis in some countries of MABC infection as multi-drug resistant TB, leading to ineffectual treatment for months[168,169].

While cystic fibrosis patients are at enhanced risk of disease by MABC members, infection prevalence is increasing overall[168,169].

Mycobacterial epidemiology in summary:

The range of mycobacterial species that are of interest to clinicians is overwhelming, and the previously listed organisms are only a subset. Broadly, pathogenic mycobacteria can be separated a few ways.

First, the *Mycobacterium tuberculosis* complex (MTBC): a closely related group of obligate, intracellular pathogenic organisms that infects humans and animals worldwide. MTB primarily infects humans, and inflicts the greatest burden of human disease, particularly in developing countries and resource-poor settings, and even more so in areas with high HIV prevalence. MBO infects a wide range of mammals, including humans, but transmission to humans is usually through exposure to infected animals and contaminated products like unpasteurized milk from infected animals. MBO makes up a small fraction of most diagnosed human TB cases currently, but historical prevalence was high, estimated at 66% of TB infections in New York in one study from 1912[115]. MBO is still a relevant pathogen in the complex, but it is more restricted in its distribution, especially in wealthier countries[170].

As infections in these wealthier countries have decreased over the years, another has greatly risen in these same places: non-tuberculous mycobacteria (NTM). This broad group contains many mycobacteria with many presentations of disease across many groups, and can be subdivided further into whether the individual species cause disease in immunocompromised individuals only, or are capable of infection in immunocompetent people

as well. Species in the MTB-associated phylotype mentioned previously (such as *M. shinjukuense* or *M. riyadhense*) have been observed to cause severe pulmonary and extrapulmonary disease in immunocompetent individuals, but other mycobacteria outside this group can also cause varied diseases in humans. NTM infections of immunocompromised individuals are often associated with HIV infection, and the identification and rise of *M. avium* infections since the 80s coincided with the emergence of AIDS[4,171]. The epidemiology of NTM infections is very complex, as each species can differ substantially[4]. The most commonly identified NTM infections differ region-to-region, with *M. avium* and *M. abscessus* dominant in parts of East Asia and the United States, *M. ulcerans* widespread in West and Central Africa, *M. kansasii* and *M. malmoense* are commonplace across Europe[172]. In any case, differential diagnosis depends heavily not just on patient history and geography, but also on the presentation of disease, from soft tissue and wound infections, to pulmonary symptoms, to lymphatic manifestations[4]. To illustrate further, the British Thoracic Society's guidelines for diagnosis only of NTM presenting pulmonary disease comprises 57 pages of material[173]. A recommended overall review of NTMs by Pereira *et al.* (2020) is used throughout this writeup[4]. For information more focused on clinical perspectives, risk factors, and disease distribution, the review by Sharma and Upadhyay (2020) is suggested[174]. NTM are a complicated group, and prevalence and our understanding of it shifts year-to-year.

Vaccine development in mycobacteria:

Vaccinology is a dense topic with a scope large enough to be its own review, but will be discussed briefly here. Its importance in the context of mycobacterial diseases is obvious, but like the BCG vaccine discussed earlier in this section, vaccination for mycobacteria is limited and

fraught with problems. For MTB, the correlates of protection are not even understood, and vaccine development proceeds with only an incomplete picture of what protection might look like[175]. It is believed but not certain that a Th1-driven response is best for protection, led by recognition of T cell epitopes by $CD4^+$ T cell, although in recent years attention has been drawn to the role of $CD8^+$ T cells in combination[69]. Contributions by B cells and antibodies are thought to be more minor, in comparison[69]. A degree of protection has been observed in some models against certain NTM by administration of the BCG vaccine, but the mechanisms by which this occurs are unclear, and conversely, prior exposure to NTMs seems to reduce the protective effects of BCG[4,78]. Overall, very little research is ongoing in vaccine development for NTMs[78]. Multiple vaccines have been developed against *M. avium* ssp. *paratuberculosis* to control Johne's disease in cattle, but they have all fallen far short of the goal – they do not prevent infection, do not eliminate clinical disease, and do not stop transmission, instead only reducing or slowing all three[176]. Many attempts have been made to develop improved vaccines against tuberculosis, with particular attention focused on the design of subunit vaccines such as the ID93 vaccine long in development, but whether recently shown immunogenicity in humans is protective or not remains unclear[177,178]. As of 2020, 16 candidate vaccines were undergoing clinical development or testing. Many attempts have been made to develop improved vaccines against tuberculosis, with particular attention focused on recent subunit vaccines such as the ID93 vaccine or the M72/$AS01_E$ vaccines, both of which are based on recombinant fusion proteins with a combination of adjuvants[177–180]. The former has recently shown immunogenicity and a good safety profile in humans, including in a thermostable configuration that does not require refrigeration, but whether it is protective or not remains to be seen[177,178,181]. The

M72/AS01$_E$ vaccine is designed to prevent LTBI from progressing to active disease, and a 3 year clinical trial in South Africa, Kenya, and Zambia showed 54% protection from LTBI progressing to clinical disease[179]. Advances in vaccinology against TB have been made, especially in a growing recognition that the most immunogenic antigens may be counterproductive targets, that PE/PPE family proteins are both significantly involved with virulence but that their antigenic variability and overlap in sequence between many different proteins may help obfuscate effective targeting, that previously underappreciated B cells play important roles in memory T cell responses essential to a protection though an overreliance on this cell type biases towards an unproductive Th2 response, and the importance of boosting the response to any selected vaccine targets by inclusion of appropriate adjuvants[175,177,180,182–184]. Alternative delivery platforms, such as nanoparticle or endospore vehicles to administer antigens to mucosal surfaces, or newly popularized mRNA vaccine technologies may also have roles to play[185–188]. Vaccine development in the face of all that is unknown about mycobacteria is daunting, and mycobacterial researchers should consider their own work in the context of how basic science can advance this topic and yield desperately needed translational benefits.

Conclusions:

Mycobacteria are ubiquitous. Pathogenic mycobacteria range from infamous plagues like tuberculosis to rare opportunistic infections[4,12]. Affected hosts span a wide range, from humans indeed most mammals, to birds, fish, herptiles, and even free-living amoebae[4,144,155,189]. Many act through unique mechanisms of immune infiltration and subversion, and a common theme is that they remain poorly understood for the impacts they cause. Mycobacterial evolution has been a convoluted subject of great debate for as long as mycobacterial existence

has been recognized, but with the genomics era well underway and improved availability of whole genome sequencing, researchers are finally able to drill down to differences at the individual nucleotide level but applied across entire genomes, pangenomes, and taxonomic groups. With this remarkable technique comes an overwhelming amount of data, and scientists must be responsible both in generating additional data carefully, in ways that are clear, reproducible, and complete, and in taking advantage of the vast amounts of data waiting in databases worldwide to begin making larger discoveries that help us interpret the past, present, and future of mycobacteria.

CHAPTER 2: GWAS OF *M. TUBERCULOSIS* COMPLEX ISOLATES REVEALS GENES ASSOCIATED WITH DIVERGENCE INTO *M. BOVIS*

Abstract:

While *Mycobacterium tuberculosis* complex (MTBC) variants are clonal, variant *tuberculosis* is a human-adapted pathogen, and variant *bovis* infects many hosts. Markers of adaptation into variants were sought by bacterial genome-wide association study (bGWAS) of single nucleotide polymorphisms (SNPs) extracted from 6,360 MTBC members from varied hosts and countries. bGWAS concordantly identified 120 loci associated with variant classification and certain hosts. Among this group are multiple changes in cholesterol and fatty acid metabolism, pathways previously proposed to be important for host adaptation, including Mce4F (part of the fundamental cholesterol intake Mce4 pathway), 4 FadD and FadE genes (playing roles in cholesterol and fatty acid utilization), and other targets like Rv3548c and PTPB, genes shown essential for growth on cholesterol by transposon studies. These findings could support that adaptation to new hosts involves adjustments in uptake and catabolism of cholesterol and fatty acids, similar to the proposed specialization to different populations in MTBC variants and MTB lineages as determined by alterations to their lipid composition. Future studies are required to elucidate how the associations between cholesterol profiles and pathogen utilization differences between hosts and MTBC variants, as well as the investigation of uncharacterized genes discovered in this study. This information will likely provide an understanding on the diversification of MBO away from humans and specialization towards a broad host range.

Introduction:

The *Mycobacterium tuberculosis* complex (MTBC) has afflicted human and animal health since the dawn of civilization. This ancient pathogen, typified by *M. tuberculosis* variant *tuberculosis* (MTB), infects humans primarily and is considered specialized for this niche[59,190]. Its subversion of host immune responses, dormancy in granulomas for years or decades, and transmissibility suggest fine adaptation to humans, potentially to per-lineage adaptation to different human populations[52,58]. MTB infects non-human hosts, including primates, and other animals (such as cattle) more infrequently[191,192]. On the other hand, *M. tuberculosis* variant *bovis* (MBO) is a generalist pathogens – its host range includes foxes, seals, cattle, cervids, lions, dogs, mustelids, badgers, and others[190,193–196]. The eponymous bovine reservoir is one of several for MBO[196], including white-tailed deer, elk, or bison in the US and Canada[115,195,197], red deer and wild boar populations in Spain[193], European badgers in the UK and Ireland[114,127], and possums in New Zealand[127]. Despite host range differences, MTBC variants show a rigid population structure[14]. From the initial whole genome sequencing, researchers were surprised to find MTB and MBO shared 99.95% nucleotide identity, excluding genomic deletions in MBO[42]. Within a few years of the first MBO genome being sequenced, research began on what might drive these differences, including gene expression[198], omics analysis[194], and metabolism[199,200], among others. Meaningful variations have been reported, but it remains uncertain how MBO has evolved towards a generalist lifestyle away from a presumed MTB-like specialist ancestor.

To help address this gap in knowledge, WGS datasets were collected for MTBC variants from diverse hosts and countries. Paired-end read SRA datasets with metadata including

country and host of isolation, and MTBC variant (n=6,360 taxa, plus reference and outgroup) were used to create a set of 9,755 SNPs for a bacterial genome-wide association study (bGWAS). This sought to detect loci associated with classification as MTB or MBO, as well as any detectable host-specific markers (e.g., SNPs associated with isolation from cervids). Using RAxML-NG[201], prewas[202], and TreeWAS[203], bGWAS was performed with isolates classified by MBO (1) or not (0). A set of 120 loci was identified, which were also identified by bGWAS of phenotypes classified as a host of *Bovidae* (1) or not (0), and *Homo sapiens* (1) or not (0). Adaptation to specific hosts was not detectable with this approach, the analysis of which could be improved through routine sequencing of isolates from non-standard host types around the world, which are currently rare and geographically biased. The 120 SNPs identified provide a trove of genes and pathways implicated in adaptation towards a generalist lifecycle, including loci across cholesterol and fatty acid uptake, catabolism, and downstream processing pathways, important for central metabolism in MTBC organisms and critical for pathogenesis[204–208]. These findings support closer investigation into how MTB and MBO utilize these pathways in different animal models, and how host cholesterol and lipid profiles could contribute to pathogen host preferences.

Materials and Methods:

Existing datasets were collected, including those where bGWAS was performed to answer other questions. Dong *et al*. (2022) genome-sequenced 74 Chinese cattle MBO isolates, and performed bGWAS analysis on a set of 3,227 MBO isolates from around the world[209]. Additionally, sequences used by Coll *et al.* in designing the MTBC SNP barcode are a validated set of primarily human MTB isolates[61]. Both datasets were included in this analysis, along with

many smaller sets. FASTQ download URLs were acquired through SRA-Explorer[210], formatted for Globus-CLI[211,212] and downloaded to MSU's High Performance Computing Center (HPCC) for processing. After transfer, single-end read data were excluded, and Snippy[213] run with default parameters, paired-end read input, and MTB H37Rv as the reference genome (AL123456.3). A complete list of accession IDs for all sequences used are provided in Supplemental File S1-1. Rare cases of genomes with unusually high numbers of SNPs (over 3,000) were excluded, as were genomes with alignment coverage <90% of the reference length (cutoff value < 3.96mbp). Additionally, metadata were compiled for all isolates (Supplemental File S1-1), and only isolates with host, MTBC variant, and country of isolation were included. Taxonomic classification by Kraken 2[214] revealed several isolates primarily contained plant or insect genomes instead, and were also excluded. Remaining paired-end read sets were selected (n=6,360) to build the final "snippy-core" core SNP alignment, along with the H37Rv reference and *M. canettii* (GCF_000253375.1) as an outgroup, while masking PE/PPE genes using the H37Rv-specific .bed file of coordinates provided by default in Snippy. Core SNPs were used for phylogenetic tree generation in RAxML-NG[201] (substitution model GTR+G selected by ModelTest-NG[215]); bootstrap analysis: seed=774900118, bootstrap trees=300; tree search analysis: seed=4949250770, 50 parsimony-based and 50 random-based starting trees for tree search; applying bootstrap support to best ML tree: --consense MRE). On a Windows 10 desktop PC, RStudio (2022.07.2+576)[216], R (v4.0.5)[217], and the R package vcfR (v1.12.0)[218] were used to generate a vcfR object for import with prewas[202]. In prewas (v1.1.1), the VCF object containing variant calls was processed with an input tree generated from RAxML-NG, the H37Rv GFF3 file, and with ancestral reconstruction flag set to TRUE. On an HPCC cluster, a Conda[219] environment

was created containing GCC (v11.2.0)[220], OpenMPI (v4.1.1)[221], and R (v4.1.2). The R package devtools (v2.4.5)[222] was installed, and used to install prewas (v1.1.1)[223] and treeWAS (v1.1)[224] from Github. The .RData object containing prewas output from the desktop PC was uploaded to HPCC and used for ancestral reconstruction state, binary variant matrix, and phylogenetic tree inputs, along with binary metadata phenotype matrices. All other parameters were left at their defaults. For MTBC lineage determination, the Coll *et al.* SNP barcode and SNP-IT tool were used[61,225]. SnpEff (v4.2) was used to annotate variants separately[226].TreeWAS generated default Manhattan plots and distribution graphics, and text output was collected in .csv files.

All sequence data used in this project are publicly available through NCBI and ENA. Accession IDs for all data are recorded in the table Supplemental File S1-1, with BioProjects in Column A, and corresponding SRA identifiers in Column B per sequence.

Results:

Core SNP extraction was successful for 6,360 isolates, reference, and *M. canettii* outgroup. Isolates were from 27 countries (Fig. 1-1A); included 2,096 MTB (including reference), 4,105 MBO, 152 variant *caprae*, and 8 variant *orygis* (Fig. 1-1B); and across 30 hosts (Fig. 1-1C). The core SNP set is provided as Supplemental File S1-2 and may be informative for other scientific inquiries beyond this GWAS. SnpEff-annotated VCFs for all extractions are also provided (Supplemental File S1-3), as well as a SNP set without PE/PPE masking (Supplemental File S1-4). The masked core SNP alignment was for phylogenetic tree generation by RAxML-NG (Supplemental File S1-5), which shows splits based on MTBC variant, but is only used for GWAS

and not intended for visualization due to its scale. After prewas and ancestral reconstruction, a final set of 7,524 variants over the 6,362 taxa was used for treeWAS input.

TreeWAS runs three tests for statistical significance – the terminal, simultaneous, and subsequent tests. The terminal test identifies broad associations between genotype and phenotype looking only at terminal nodes in the tree; the simultaneous test more stringently identifies deterministic relationships of genotype and phenotype, without necessitating the relationship occur at all branches; the subsequent test utilizes the terminal test, but adds ancestral state reconstruction to analyze all nodes of the tree[203]. A thorough explanation of these tests is provided by Dr. Collins on the treeWAS GitHub page[227]. The simultaneous and subsequent tests were used initially, as an ancestral reconstruction was available. When analyzing by phenotype of MBO (1) and Not MBO (0), treeWAS produced significant loci for both simultaneous and subsequent statistical scoring metrics (Figure 1-2). Analysis by phenotype of *Bovidae* also produced the same 120 significant loci by the subsequent test but did not produce loci for the simultaneous test (Figure 1-3). Another search by phenotype of *Homo sapiens* produced the same 120 loci again by subsequent test, and no loci by simultaneous test (Figure 1-4). A phenotype of "non-standard hosts" (where *Homo sapiens* is the standard host for MTB, *Bovidae* for MBO, *Capra* for MCP, and *Oryx* for MOR) yielded no significant hits (Supplemental File S1-6). Analysis by phenotype *Cervidae* returned no significance (Supplemental File S1-7), but phenotype *Meles meles* (European badger) showed an unusual pattern by subsequent statistical test where nearly all SNPs clustered just around the significance cutoff, yielding hundreds of loci technically exceeding the cutoff yet are tightly clustered with those under it (Supplemental File S1-8A). The simultaneous score did not show

similar hits (Supplemental File S1-8B. A similar pattern was seen for *Sus* (Supplemental File S1-9). To investigate if these associations reflected geographic effects for *Meles meles,* as all badger samples were from the UK, a GWAS analysis was performed by phenotype of UK origin, which yielded no significance by the subsequent test, and a single locus by simultaneous test (Supplemental File S1-10). This hit, for a variant present in only two isolates, is a spurious result. For *Meles meles* and *Sus,* it would appear the composition of the dataset makes it difficult to detect significance against a background already tightly associated with specific genotypes. MCP produced no significant hits (Supplemental File S1-11), and MOR was not attempted due to low representation of *M. orygis* samples in the dataset (n=8).

The 32 loci identified by the simultaneous test for *Bovidae* are listed in Table 1-1. As mentioned for the spurious hit for UK samples, the simultaneous test can report loci as associated even if a SNP is only present in a few isolates. These false positives are included in the data tables, but are shaded in gray and should not be considered meaningful. Subtracting these spurious hits, the simultaneous test identifies 22 loci, all of which are also identified by the subsequent test. The 120 loci concordantly identified as associated by *Bovidae*, MBO, and *Homo sapiens* by the subsequent test are listed in Table 1-2. The subset of loci called by both tests in MBO (n=22) are presented in Table 1-3.

After GWAS, several apparent genotypic edge cases arose. FadD11, for example, was highlighted as significantly associated by a single non-synonymous variant, FadD11 L286S, which appeared fixed in MBO and MCP, while MTB and MOR showed WT nearly exclusively. Of 4,105 MBO isolates and 152 MCP isolates, only 1 MBO isolate showed WT at this position, suggesting a reversion in this genome. Likewise, of 2,087 MTB and 8 MOR isolates, only 9 MTB

isolates showed L286S. These genotypic exceptions were checked further: a Chinese cattle isolate SRR16278270 for the 1 MBO outlier, and 9 UK MTB isolates from humans for the MTB outliers (Table 1-4). These 10 isolates' VCF files were checked against the SNP barcode[61], with lineage-determining positions searched per VCF through Unix command "grep" and the SNP coordinate. The MBO isolate bore no MBO lineage-determining SNPs, and instead was cleanly typable as MTB lineage 2.2.1 (Table 1-5). Evidently, this isolate is a case of bovine MTB being incorrectly identified as MBO when uploaded to NCBI. Likewise, of the 9 human MTB cases that stood out, all bore the 3 lineage-determining SNPs for MBO (Table 1-4), and no MTB lineage-determining markers. In these instances, 9 cases of human MBO were misclassified as MTB. These results were validated using SNP-IT software[225], which also typed ERR387001 as a BCG strain, suggesting a case of BCG-osis misdiagnosed or mislabeled as TB. After correcting these calls, there is a perfect divide between MTB/MOR and MBO/MCP, with 100% of MTB/MOR isolates showing WT, and 100% of MBO/MCP isolates bearing the SNP. This discrepancy may have affected robustness of GWAS based on phenotype of "MBO variant." However, it is noted that comparisons for "host *Bovidae"* and "host *Homo sapiens"* are unaffected by this, and all 3 analyses produced perfect concordance of their 120 associated loci by subsequent tests, suggesting this mislabeling had minimal influence.

Discussion:

Despite remarkable similarity between MTB and MBO, evidence of clear divides was present SNPs highlighted as associated by treeWAS analysis.

The Fad family of proteins are important in MTBC, with MTB known to carry 36 FadD and FadE loci[228,229]. GWAS identified SNPs in FadD11, FadE5, FadE27, and FadE32 associated with differentiation into MBO. These genes are involved in fatty acid and cholesterol handling inside the environment of the host[208]. Mycobacterial reliance on cholesterol is known to be critical for pathogenesis, and MTB features around 80 genes involved with cholesterol balance and metabolism[230]. Disruption of cholesterol import is severely disruptive to infection and persistence[230,231].

Cholesterol intake in MTBC requires a functional Mce4 system[231]. A missense mutation (A734G, Asp254Gly) in Mce4F was seen to be fixed in all MBO/MCP isolates, and only a single MTB isolate from Russia (ERR108427) which bore a unique SNP signature: G3836739A (lineage 4.8), and G1759252T (lineage 4.9). No other SNPs for lineage 4 or any other lineage were identified. A literature search turned up Congo type MTB that can present lineage 4.8 and 4.9 SNPs, but only in combination with 4.7 SNPs that were absent in this sample[232]. Except this atypical MTB specimen, another clear split by this *mce4f* SNP separated MTB/MOR and MBO/MCP.

Catabolized cholesterol products fuel core acyl-CoA metabolism pathway, as well as polyketide synthesis, a pathway already known to differ between MBO and MTB[40]. GWAS identified two separate missense mutations in *ppsD*, and synonymous changes in *ppsB* and *pks15*, all polyketide synthase genes. Genes annotated in roles of cholesterol and fatty acid metabolism or in pathways downstream of these processes among all loci identified by the MBO subsequent tests are shown in Table 1-6. These associated SNPs are scattered across lipid

metabolic pathways and include members whose exact function is unclear. Ten out of fourteen of these SNPs are non-synonymous.

Other identified loci with functions separate from cholesterol and lipid metabolism include AccD1, involved in leucine degradation[233], which bore a fixed SNP of Phe343Leu. SNP 1739294 in the essential isoleucyl-tRNA synthetase IleS causes a Pro926Ala substitution, SNP 3152421 in the essential prolyl-tRNA synthetase ProS yields His177Arg, SNP 3371365 in the essential glutamyl-tRNA amidotransferase subunit GatA causes Ala24Thr, and a synonymous change is seen at 1260537 for methionine synthesis gene MetE. Related to translational machinery, SNP 3198332 causes Thr259Met in essential elongation factor Tsf. SNP 1129160 impacts both RpfB (resuscitation promoting factor B) and KsgA (a dimethyladenosine transferase) genes, resulting in RpfB Ala357Val and a synonymous mutation in KsgA. RpfB is thought to be involved in the transition from dormancy to active replication, and is co-transcribed with *ksgA* and *ispE*, genes involved in ribosome maturation and cell wall synthesis, respectively[234]. After accounting for aforementioned mislabeling cases in deposited isolates, these SNPs are all fixed and exclusive in this dataset either in MBO and MCP, or MBO alone. MTBC physiology and function remain uncertain, and 48/120 loci identified are in genes annotated only by locus identifier and generically, like “conserved protein” or “possible oxidoreductase” even after PE/PPE gene filtering, removing a largely uncharacterized family comprising ~10% of MTBC genes. Even among genes with fuller annotations, nearly all include “probable” in their descriptions. The genes presented in Table 6 are not comprehensive and given the uncertainty in function across many loci, other important genes both inside and outside lipid metabolism almost certainly exist in the bGWAS output of Table 2. Loci associated

with adaptation towards new hosts and lifestyles are useful then to highlight for characterization, as it narrows the still vast pool of MTBC genes with uncertain functions towards a subset of genes with fixed changes in some variants.

Any associated loci may signify adaptive roles in differentiation from a specialist infection by MTB and a pathogen with a much broader host range, like MBO. It is well-reported that members of the MTBC are clonal, and not only is horizontal gene transfer vanishingly rare, mutation rates in members of this complex are low (~$2x10^{-10}$ mutations per cell division)[56]. Given a 99.95% genetic identity between MTB and MBO, <2000 polymorphisms differentiate divergent variants (disregarding RDs/large sequence polymorphisms), and fixed changes are an even smaller subset. While this research cannot draw conclusions about how specific changes might alter metabolism or virulence to better reflect new host environs, it does highlight multiple SNPs across multiple genes in MTBC metabolic pathways. Metabolic differences are known to exist between variants and even between lineages of MTB, including in lipid profile[207,235]. While these data are only associations, they may support findings by Griffin *et al.* in 2013 reporting cholesterol utilization in MTB is key to host adaptation[230]. MTB drives macrophages to import lipids for utilization as an energy and carbon source[106,204,205,236]. The human cholesterol profile is LDL-dominant[237], as is the guinea pig[237], a model of tuberculosis that better recapitulates human disease[238] vs. the mouse model[191,238,239], an animal model with an HDL-dominant cholesterol profile and a lower overall cholesterol load[240,241]. The MBO bovine host is HDL-dominant, for comparison[242]. Others have reported that MTB infections are influenced differently by HDL vs. LDL cholesterol[243]. MTB is known to exploit lipid-rich "foamy macrophages," and research has shown MTB trehalose dimycolate and other factors are

associated with lipid droplet and foamy macrophage formation[204,205,236]. Foamy macrophage formation is associated with higher levels and intake of circulating LDL cholesterol, but recent research found MTB-infected macrophages have a different lipid profile from foamy macrophages characterized in atherosclerosis and other diseases[240], indicating disease-specific responses lead to buildup of certain lipids[106]. Finally, higher HDL levels are known to counteract foamy macrophage formation through classical LDL intake, HDL suppresses TNFα production in MTB-infected macrophages, and mice are more resistant to foamy macrophage formation, compared to humans, and guinea pig, rabbit, or primate models[237,240,243]. Variation in use of cholesterol and fatty acids is known to exist between MTBC variants and lineages. Finally, though research is more limited in this area, studies have demonstrated mice are more susceptible to disease and death by MBO infection than by MTB[244,245]. Biological reality is undoubtedly far more complex, but from existing literature, host lipid profiles differ, lipid availability and sequestration are key to MTB virulence, animal models with a lipid profile closer to humans better reproduce "classic" granulomatous tuberculosis by MTB as seen in humans, and MTB lineages and MTBC variants utilize lipids differentially. Cholesterol/fatty acid metabolic pathways associated by bGWAS showing variant-specific changes between MTB and MBO are suggestive of a potential contributor towards host adaptation.

It is also important to note that this dataset is necessarily only a small sampling of publicly available genomes, which are themselves an infinitesimal fraction of the true numbers of MTBC infections. With ~6,000 genomes represented and 1.6 million TB deaths in 2022, this dataset represents less than half a percent even of fatal human TB cases in a single year, without even touching on the billions of latently infected humans and an untold number of

animal infections both presently and over history. Furthermore, sampling from human and animal hosts is biased by limits of availability for whole genome sequencing technology, and the necessary willingness to sequence isolates, especially from animals. As such, all findings must be remembered in their context. This GWAS takes a step towards answering what may play roles in host adaptation and variant diversification, but more work remains.

In summary, many SNPs differentiate MTBC variants, and importantly, they may inform research into genes that differ between variants, narrowing the pool of uncharacterized proteins to study in the MTBC. Future work in MTBC host adaptation should investigate from host and pathogen sides how available lipid and cholesterol pools in bovine, murine, and other non-human hosts may modulate pathogenesis.

Tables:

SNP	Locus	Protein	Description	Essentiality Notes from Mycobrowser
147873			Intergenic, upstream of elongation factor G FusA2 (Rv0120c)	n/a
184727	Rv0156	PntAb	Probable NAD(P) transhydrogenase (subunit alpha) PntAb [second part; integral membrane protein] (pyridine nucleotide transhydrogenase subunit alpha) (nicotinamide nucleotide transhydrogenase subunit alpha)	n/a
268277	Rv0224c		Possible methyltransferase (methylase)	In vitro essential per multiple studies (Minato 2019; DeJesus 2017; Sassetti 2003; Griffin 2011)
277862			Intergenic, downstream of FadE4 (Rv0231) and upstream of probable transcriptional regulatory protein (probably TetR/AcrR-family) (Rv0232)	n/a
438069	Rv0359		Probable conserved integral membrane protein	n/a
1234657	Rv1108c	XseA	Probable exodeoxyribonuclease VII (large subunit) XseA (exonuclease VII large subunit)	n/a
1390284	Rv1248c		Multifunctional alpha-ketoglutarate metabolic enzyme	In vitro essential per multiple studies (Minato 2019; Carvalho 2010; Sassetti 2003; Griffin 2011)
1478312	Rv1317c	AlkA	Probable bifunctional regulatory protein and DNA repair enzyme AlkA (regulatory protein of adaptative response) (methylphosphotriester-DNA—protein-cysteine S-methyltransferase)	n/a
1499291	Rv1330c	PncB1	Nicotinic acid phosphoribosyltransferase PncB1	n/a

Table 1-1: SNPs significantly associated with classification as MBO by Simultaneous statistical test. GWAS results by treeWAS showing single nucleotide polymorphisms (coordinate relative to MTB H37Rv in SNP column) associated with classification of MTBC isolates as M. *tuberculosis* variant *bovis*. For SNPs within genes or ORFs, the classification and putative function is listed, as well as select information about essentiality from Mycobrowser [https://mycobrowser.epfl.ch/]. Gray shading indicates false positive hits.

Table 1-1 (cont'd)

1586961	Rv1410c		Aminoglycosides/tetracycline-transport integral membrane protein	Essential in murine macrophages (Rengarajan 2005) and murine spleen (Sassetti and Rubin 2003)
1739294	Rv1536	IleS	Isoleucyl-tRNA synthetase IleS	In vitro essential per multiple studies (Minato 2019; DeJesus 2017; Lamichhane 2003; Griffin 2011)
1763524	Rv1559	IlvA	Probable threonine dehydratase IlvA	In vitro essential (DeJesus 2017; Griffin 2011), non-essential in rich media (Minato 2019)
1830295	Rv1628c		Conserved protein	n/a
2314425	Rv2056c	RpsN2	30S ribosomal protein S14 RpsN2	Disruption provides growth advantage (DeJesus 2017)
2475888	Rv2210c	IlvE	Branched-chain amino acid transaminase IlvE	In vitro essential (DeJesus 2017; Sassetti 2003; Griffin 2011), non-essential in rich media (Minato 2019)
2528773	Rv2254c		Probable integral membrane protein	n/a
2658676	Rv2379c	MbtF	Peptide synthetase MbtF (peptide synthase)	n/a
2682593	Rv2388c	HemN	Probable oxygen-independent coproporphyrinogen III oxidase HemN (coproporphyrinogenase) (coprogen oxidase)	Essential in murine spleen (Sassetti and Rubin, 2003)
2912516	Rv2585c		Possible conserved lipoprotein	n/a
2927291	Rv2598		Conserved hypothetical protein	n/a
3140153	Rv2833c	UgpB	Probable Sn-glycerol-3-phosphate-binding lipoprotein UgpB	Disruption provides growth advantage (DeJesus 2017)
3143890	Rv2837c		Conserved protein	n/a
3235485	Rv2922c	Smc	Probable chromosome partition protein Smc	n/a
3371365	Rv3011c	GatA	Probable glutamyl-tRNA(GLN) amidotransferase (subunit A) GatA (Glu-ADT subunit A)	In vitro essential per multiple studies (Minato 2019; DeJesus 2017; Sassetti 2003; Griffin 2011)
3534980	Rv3166c		Conserved hypothetical protein	n/a
3773023	Rv3361c		Conserved protein	n/a

Table 1-1 (cont'd)

3877256	Rv3456c	RplQ	50S ribosomal protein L17 RplQ	In vitro essential (Minato 2019; Griffin 2011), or mutant shows growth defect (DeJesus 2017)
3904490	Rv3484	CpsA	Possible conserved protein CpsA	Essential in murine spleen (Sassetti and Rubin, 2003)
3922919	Rv3504	fadE26	Probable acyl-CoA dehydrogenase FadE26	n/a
4157578	Rv3712		Possible ligase	In vitro essential per multiple studies (Minato 2019; DeJesus 2017; Sassetti 2003; Griffin 2011)
4171113	Rv3725		Possible oxidoreductase	Disruption provides growth advantage (DeJesus 2017)
4281133	Rv3816c		Possible acyltransferase	n/a

SNP	**Locus**	**Protein**	**Description**	**Essentiality Notes from Mycobrowser**
22264	Rv0018c	PstP	Involved in regulation (using dephosphorylation of a specific phosphorylated substrate)	Required for survival in murine macrophages (Rengarajan 2005)
23714	Rv0019c	FhaB	Conserved protein with FHA domain, FhaB	Required for survival in murine macrophages (Rengarajan 2005)
147873			Intergenic, upstream of elongation factor G FusA2 (Rv0120c)	n/a
184727	Rv0156	PntAb	Probable NAD(P) transhydrogenase (subunit alpha) PntAb [second part; integral membrane protein] (pyridine nucleotide transhydrogenase subunit alpha) (nicotinamide nucleotide transhydrogenase subunit alpha)	n/a

Table 1-2: SNPs significantly associated with classification as MBO by Subsequent statistical test. GWAS results by treeWAS showing single nucleotide polymorphisms (coordinate relative to MTB H37Rv in SNP column) associated with classification of MTBC isolates as M. *tuberculosis* variant *bovis* (MBO). For SNPs within genes or ORFs, the classification and putative function is listed, as well as select information about essentiality from Mycobrowser [https://mycobrowser.epfl.ch/].

Table 1-2 (cont'd)

212254			Intergenic, upstream of transmembrane protein (Rv0180)	n/a
217863	Rv0186	BglS	Possibly involved in degradation [catalytic activity: hydrolysis of terminal, non-reducing beta-D-glucose residues with release of beta-D-glucose]	n/a
262160	Rv0218		Probable conserved transmembrane protein	Essential in murine spleen (Sassetti and Rubin, 2003)
268277	Rv0224c		Possible methyltransferase (methylase)	In vitro essential per multiple studies (Minato 2019; DeJesus 2017; Sassetti 2003; Griffin 2011)
277862			Intergenic, downstream of FadE4 (Rv0231) and upstream of probable transcriptional regulatory protein (probably TetR/AcrR-family) (Rv0232)	n/a
294198	Rv0244c	FadE5	Probable acyl-CoA dehydrogenase FadE5	Required for growth on cholesterol (Griffin 2011)
386060			Intergenic, upstream of glpQ2 (Rv0317c)	n/a
397386			Intergenic, downstream of putative dehydrogenase/reductase (Rv0331) and upstream of hypothetical protein (Rv0332)	n/a
398034	Rv0332		Conserved protein	n/a
411100	Rv0342	IniA	Isoniazid inductible gene protein IniA	n/a
1027445	Rv0921		Possible resolvase for IS1535	n/a
1029936	Rv0923c		Conserved hypothetical protein	n/a
1125316	Rv1006		Unknown protein	Disruption provides growth advantage (DeJesus 2017)
1129160	Rv1010	KsgA	Probable dimethyladenosine transferase KsgA (S-adenosylmethionine-6-N', N'-adenosyl(rRNA) dimethyltransferase) (16S rRNA 53emethylase) (high level kasugamycin resistance protein KsgA) (kasugamycin dimethyltransferase)	n/a

Table 1-2 (cont'd)

1129160	Rv1009	RpfB	Probable resuscitation-promoting factor RpfB	n/a
1234657	Rv1108c	XseA	Probable exodeoxyribonuclease VII (large subunit) XseA (exonuclease VII large subunit)	n/a
1260537	Rv1133c	MetE	Probable 5-methyltetrahydropteroyltriglutamate—homocysteine methyltransferase MetE (methionine synthase, vitamin-B12 independent isozyme)	In vitro essential (DeJesus 2017; Sassetti 2003; Griffin 2011), non-essential in rich media (Minato 2019)
1307958	Rv1175c	FadH	Probable NADPH dependent 2,4-dienoyl-CoA reductase FadH (2,4-dienoyl coenzyme A reductase) (4-enoyl-CoA reductase)	n/a
1377140	Rv1234		Probable transmembrane protein	n/a
1393003	Rv1248c		Multifunctional alpha-ketoglutarate metabolic enzyme	In vitro essential per multiple studies (Minato 2019; Sassetti 2003; Griffin 2011; Carvalho 2010)
1425641	Rv1276c		Conserved hypothetical protein	n/a
1458076	Rv1301		Conserved protein	In vitro essential (Sassetti 2003; Griffin 2011), non-essential in rich media (Minato 2019)
1478312	Rv1317c	AlkA	Probable bifunctional regulatory protein and DNA repair enzyme AlkA (regulatory protein of adaptative response) (methylphosphotriester-DNA—protein-cysteine S-methyltransferase)	n/a
1496289	Rv1328	GlgP	Probable glycogen phosphorylase GlgP	n/a
1499291	Rv1330c	PncB1	Nicotinic acid phosphoribosyltransferase PncB1	n/a
1562049	Rv1387	PPE20	PPE family protein PPE20	n/a
1609445	Rv1431		Conserved membrane protein	n/a
1671658	Rv1481		Probable membrane protein	In vitro essential per multiple studies (Minato 2019; DeJesus 2017; Griffin 2011)

Table 1-2 (cont'd)

1681928	Rv1491c		Conserved membrane protein	n/a
1684979	Rv1493	MutB	Probable methylmalonyl-CoA mutase large subunit MutB (MCM)	n/a
1739294	Rv1536	IleS	Isoleucyl-tRNA synthetase IleS	In vitro essential per multiple studies (Minato 2019; DeJesus 2017; Griffin 2011; Lamichhane 2003)
1754572	Rv1550	FadD11	Probable fatty-acid-CoA ligase FadD11 (fatty-acid-CoA synthetase) (fatty-acid-CoA synthase)	n/a
1766620	Rv1562c	TreZ	Maltooligosyltrehalose trehalohydrolase TreZ	n/a
1794234	Rv1593c		Conserved protein	n/a
1804248	Rv1604	ImpA	Probable inositol-monophosphatase ImpA (imp)	n/a
1804315	Rv1604	ImpA	Probable inositol-monophosphatase ImpA (imp)	n/a
1830295	Rv1628c		Conserved protein	n/a
1834859	Rv1630	RpsA	30S ribosomal protein S1 RpsA	In vitro essential per multiple studies (Minato 2019; DeJesus 2017; Griffin 2011; Sassetti 2003)
1971029	Rv1744c		Probable membrane protein	n/a
2013589	Rv1779c		Possible integral membrane protein	n/a
2082865	Rv1836c		Conserved protein	n/a
2092688	Rv1843c	GuaB1	Probable inosine-5'-monophosphate dehydrogenase GuaB1(imp dehydrogenase) (IMPDH) (IMPD)	Disruption provides growth advantage (DeJesus 2017)
2104270	Rv1856c		Possible oxidoreductase	Disruption provides growth advantage (DeJesus 2017)
2280081	Rv2032	Acg	Conserved protein Acg	n/a
2475116	Rv2210c	IlvE	Branched-chain amino acid transaminase IlvE	In vitro essential (Sassetti 2003; Griffin 2011; DeJesus 2017), non-essential in rich media (Minato 2019)

Table 1-2 (cont'd)

2475888	Rv2210c	IlvE	Branched-chain amino acid transaminase IlvE	In vitro essential (Sassetti 2003; Griffin 2011; DeJesus 2017), non-essential in rich media (Minato 2019)
2502757	Rv2229c		Conserved protein	n/a
2528773	Rv2254c		Probable integral membrane protein	n/a
2529798	Rv2256c		Conserved hypothetical protein	n/a
2606813	Rv2333c	Stp	Integral membrane drug efflux protein Stp	n/a
2646542	Rv2364c	Era	Probable GTP-binding protein Era	In vitro essential (Sassetti 2003; Griffin 2011), non-essential in rich media (Minato 2019)
2658676	Rv2379c	MbtF	Peptide synthetase MbtF (peptide synthase)	n/a
2659542	Rv2379c	MbtF	Peptide synthetase MbtF (peptide synthase)	n/a
2682593	Rv2388c	HemN	Probable oxygen-independent coproporphyrinogen III oxidase HemN (coproporphyrinogenase) (coprogen oxidase)	Essential in murine spleen (Sassetti and Rubin, 2003)
2692875	Rv2396	AprC	Acid and phagosome regulated protein C, PE-PGRS family protein PE_PGRS41	n/a
2760147	Rv2458	MmuM	Probable homocysteine S-methyltransferase MmuM (S-methylmethionine:homocysteine methyltransferase) (cysteine methyltransferase)	Disruption provides growth advantage (DeJesus 2017)
2809318	Rv2495c	BkdC	Probable branched-chain keto acid dehydrogenase E2 component BkdC	n/a
2812742	Rv2498c	CitE	Probable citrate (pro-3S)-lyase (beta subunit) CitE (citrase) (citratase) (citritase) (citridesmolase) (citrase aldolase)	n/a
2817446	Rv2502c	AccD1	Probable acetyl-/propionyl-CoA carboxylase (beta subunit) AccD1	Essential in murine spleen (Sassetti and Rubin, 2003)
2912516	Rv2585c		Possible conserved lipoprotein	n/a
2927291	Rv2598		Conserved hypothetical protein	Disruption provides growth advantage (DeJesus 2017)

Table 1-2 (cont'd)

2932890	Rv2605c	TesB2	Probable acyl-CoA thioesterase II TesB2 (TEII)	n/a
2997325	Rv2681		Conserved hypothetical alanine rich protein	Required for growth on cholesterol (Griffin 2011)
3032137	Rv2720	LexA	Repressor LexA	n/a
3041679	Rv2729c		Probable conserved integral membrane alanine valine and leucine rich protein	n/a
3042353	Rv2729c		Probable conserved integral membrane alanine valine and leucine rich protein	n/a
3055922	Rv2742c		Conserved hypothetical arginine rich protein	n/a
3140153	Rv2833c	UgpB	Probable Sn-glycerol-3-phosphate-binding lipoprotein UgpB	n/a
3142580	Rv2836c	DinF	Possible DNA-damage-inducible protein F DinF	n/a
3152421	Rv2845c	ProS	Probable prolyl-tRNA synthetase ProS (proline—tRNA ligase) (PRORS) (global RNA synthesis factor) (proline translase)	Essential in vitro (Minato 2019; DeJesus 2017; Griffin 2011; Sassetti 2003) and in murine spleen (Sassetti and Rubin 2003)
3157785	Rv2849c	CobO	Probable cob(I)alamin adenosyltransferase CobO (corrinoid adenosyltransferase) (corrinoid adotransferase activity)	n/a
3158719	Rv2850c		Possible magnesium chelatase	n/a
3159237	Rv2850c		Possible magnesium chelatase	n/a
3174591	Rv2862c		Conserved hypothetical protein	n/a
3189664	Rv2879c		Conserved hypothetical protein	n/a
3198332	Rv2889c	Tsf	Probable elongation factor Tsf (EF-ts)	In vitro essential (Sassetti 2003; Griffin 2011; DeJesus 2017; Minato 2019)
3213089	Rv2903c	LepB	Probable signal peptidase I LepB (SPASE I) (leader peptidase I)	In vitro essential (Sassetti 2003; Griffin 2011; DeJesus 2017; Minato 2019)

Table 1-2 (cont'd)

3223303	Rv2914c	PknI	Probable transmembrane serine/threonine-protein kinase I PknI (protein kinase I) (STPK I) (phosphorylase B kinase kinase) (hydroxyalkyl-protein kinase)	Required for growth on cholesterol (Griffin 2011), mutant shows increased growth in THP-1 cells, SCID mice show faster mortality with mutant (Gopalaswamy 2009)
3235715	Rv2922c	Smc	Probable chromosome partition protein Smc	n/a
3254695	Rv2932	PpsB	Phenolpthiocerol synthesis type-I polyketide synthase PpsB	In vitro essential in CDC1551 (Lamichhane 2003), not in H37Rv (Griffin 2011; DeJesus 2017; Minato 2019)
3262628	Rv2934	PpsD	Phenolpthiocerol synthesis type-I polyketide synthase PpsD	n/a
3267715	Rv2934	PpsD	Phenolpthiocerol synthesis type-I polyketide synthase PpsD	n/a
3282079	Rv2940c	Mas	Probable multifunctional mycocerosic acid synthase membrane-associated Mas	n/a
3320554	Rv2947c	Pks15	Probable polyketide synthase Pks15, involved in the biosynthesis of phenolphthiocerol glycolipids.	n/a
3355417	Rv2997		Possible alanine rich dehydrogenase	n/a
3371365	Rv3011c	GatA	Probable glutamyl-tRNA(GLN) amidotransferase (subunit A) GatA (Glu-ADT subunit A)	In vitro essential (Sassetti 2003; Griffin 2011; DeJesus 2017; Minato 2019)
3388682	Rv3029c	FixA	Probable electron transfer flavoprotein (beta-subunit) FixA (beta-ETF) (electron transfer flavoprotein small subunit) (ETFSS)	In vitro essential (Sassetti 2003; Griffin 2011), non-essential in rich media (Minato 2019)
3517567	Rv3151	NuoG	Probable NADH dehydrogenase I (chain G) NuoG (NADH-ubiquinone oxidoreductase chain G)	n/a
3534980	Rv3166c		Conserved hypothetical protein	n/a
3540144	Rv3171c	Hpx	Possible non-heme haloperoxidase Hpx	n/a
3565449	Rv3195		Conserved hypothetical protein	n/a
3594851	Rv3218		Conserved protein	n/a

Table 1-2 (cont'd)

3595427	Rv3218		Conserved protein	n/a
3624710	Rv3244c	LpqB	Probable conserved lipoprotein LpqB	In vitro essential (Sassetti 2003; Griffin 2011; DeJesus 2017; Minato 2019)
3664615	Rv3282		Conserved hypothetical protein	n/a
3678929	Rv3296	Lhr	Probable ATP-dependent helicase Lhr (large helicase-related protein)	n/a
3690854	Rv3303c	LpdA	NAD(P)H quinone reductase LpdA	n/a
3770588	Rv3356c	FolD	Probable bifunctional protein FolD: methylenetetrahydrofolate dehydrogenase + methenyltetrahydrofolate cyclohydrolase	In vitro essential (Sassetti 2003; Griffin 2011; DeJesus 2017; Minato 2019)
3857161	Rv3437		Possible conserved transmembrane protein	Disruption provides growth advantage (DeJesus 2017)
3877256	Rv3456c	RplQ	50S ribosomal protein L17 RplQ	In vitro essential (Minato 2019; Griffin 2011)
3904490	Rv3484	CpsA	Possible conserved protein CpsA	Essential in murine spleen (Sassetti and Rubin, 2003)
3907958	Rv3488		Conserved hypothetical protein	n/a
3912636	Rv3494c	Mce4F	Mce-family protein Mce4F	Required for growth on cholesterol (Griffin 2011)
3924350	Rv3505	FadE27	Probable acyl-CoA dehydrogenase FadE27	n/a
3977910	Rv3538		Probable dehydrogenase. Possible 2-enoyl acyl-CoA hydratase.	n/a
3987645	Rv3548c		Probable short-chain type dehydrogenase/reductase	Required for growth on cholesterol (Griffin 2011)
4004604	Rv3563	FadE32	Probable acyl-CoA dehydrogenase FadE32	Required for growth on cholesterol (Griffin 2011), essential in murine spleen (Sassetti and Rubin, 2003)

Table 1-2 (cont'd)

4034908	Rv3593	LpqF	Probable conserved lipoprotein LpqF	In vitro essential (Sassetti 2003; Griffin 2011; Minato 2019)
4047039	Rv3604c		Probable conserved transmembrane protein rich in alanine and arginine and proline	In vitro essential (Sassetti 2003; Griffin 2011; Minato 2019)
4083511	Rv3645		Probable conserved transmembrane protein	In vitro essential (DeJesus 2017; Griffin 2011)
4090661	Rv3649		Probable helicase	Essential in murine spleen (Sassetti and Rubin, 2003)
4157578	Rv3712		Possible ligase	In vitro essential (Sassetti 2003; Griffin 2011; DeJesus 2017; Minato 2019)
4171113	Rv3725		Possible oxidoreductase	Disruption provides growth advantage (DeJesus 2017)
4242970	Rv3793	EmbC	Integral membrane indolylacetylinositol arabinosyltransferase EmbC (arabinosylindolylacetylinositol synthase)	In vitro essential (Sassetti 2003; Goude 2008; Griffin 2011; DeJesus 2017; Minato 2019)
4278968	Rv3813c		Conserved protein	n/a

SNP	**Locus**	**Protein**	**Description**	**Essentiality Notes from Mycobrowser**
147873			Intergenic, upstream of elongation factor G FusA2 (Rv0120c)	n/a

Table 1-3: SNPs concordantly significantly associated with classification as MBO by Subsequent and Simultaneous statistical tests. GWAS results by treeWAS showing single nucleotide polymorphisms (coordinate relative to MTB H37Rv in SNP column) associated with classification of MTBC isolates as M. *tuberculosis* variant *bovis* (MBO). For SNPs within genes or ORFs, the classification and putative function is listed, as well as select information about essentiality by transposon mutagenesis studies from Mycobrowser [https://mycobrowser.epfl.ch/]. This list is a subset of only variants called in both Table 1-1 and Table 1-2.

Table 1-3 (cont'd)

184727	Rv0156	PntAb	Probable NAD(P) transhydrogenase (subunit alpha) PntAb [second part; integral membrane protein] (pyridine nucleotide transhydrogenase subunit alpha) (nicotinamide nucleotide transhydrogenase subunit alpha)	n/a
268277	Rv0224c		Possible methyltransferase (methylase)	In vitro essential per multiple studies (Minato 2019; DeJesus 2017; Sassetti 2003; Griffin 2011)
277862			Intergenic, downstream of FadE4 (Rv0231) and upstream of probable transcriptional regulatory protein (probably TetR/AcrR-family) (Rv0232)	n/a
1234657	Rv1108c	XseA	Probable exodeoxyribonuclease VII (large subunit) XseA (exonuclease VII large subunit)	n/a
1478312	Rv1317c	AlkA	Probable bifunctional regulatory protein and DNA repair enzyme AlkA (regulatory protein of adaptative response) (methylphosphotriester-DNA—protein-cysteine S-methyltransferase)	n/a
1499291	Rv1330c	PncB1	Nicotinic acid phosphoribosyltransferase PncB1	n/a
1739294	Rv1536	IleS	Isoleucyl-tRNA synthetase IleS	In vitro essential per multiple studies (Minato 2019; DeJesus 2017; Lamichhane 2003; Griffin 2011)
1830295	Rv1628c		Conserved protein	n/a
2475888	Rv2210c	IlvE	Branched-chain amino acid transaminase IlvE	In vitro essential (DeJesus 2017; Sassetti 2003; Griffin 2011), non-essential in rich media (Minato 2019)
2528773	Rv2254c		Probable integral membrane protein	n/a
2658676	Rv2379c	MbtF	Peptide synthetase MbtF (peptide synthase)	n/a
2682593	Rv2388c	HemN	Probable oxygen-independent coproporphyrinogen III oxidase HemN (coproporphyrinogenase) (coprogen oxidase)	Essential in murine spleen (Sassetti and Rubin, 2003)

Table 1-3 (cont'd)

2912516	Rv2585c		Possible conserved lipoprotein	n/a
2927291	Rv2598		Conserved hypothetical protein	n/a
3140153	Rv2833c	UgpB	Probable Sn-glycerol-3-phosphate-binding lipoprotein UgpB	Disruption provides growth advantage (DeJesus 2017)
3371365	Rv3011c	GatA	Probable glutamyl-tRNA(GLN) amidotransferase (subunit A) GatA (Glu-ADT subunit A)	In vitro essential per multiple studies (Minato 2019; DeJesus 2017; Sassetti 2003; Griffin 2011)
3534980	Rv3166c		Conserved hypothetical protein	n/a
3877256	Rv3456c	RplQ	50S ribosomal protein L17 RplQ	In vitro essential (Minato 2019; Griffin 2011), or mutant shows growth defect (DeJesus 2017)
3904490	Rv3484	CpsA	Possible conserved protein CpsA	Essential in murine spleen (Sassetti and Rubin, 2003)
4157578	Rv3712		Possible ligase	In vitro essential per multiple studies (Minato 2019; DeJesus 2017; Sassetti 2003; Griffin 2011)
4171113	Rv3725		Possible oxidoreductase	Disruption provides growth advantage (DeJesus 2017)

Lineage Marker	**Bovis**		
Accession ID	**C1427476T**	**A2831482G**	**C3624593T**
ERR017796	✓	✓	✓
ERR026636	✓	✓	✓

Table 1-4: SNP typing improperly labeled MTBC isolates by MBO-lineage markers. Nine pathogen isolates from humans (underlined) were classified when deposited into NCBI as MTB, but their genotypes by GWAS did not align with other MTB isolates. These isolates were checked for three MBO lineage-determining SNPs as reported by Coll *et al.* (2014), and all 9 were found to possess these SNPs, indicating a misclassification in the database. This was confirmed by SNP-IT (Lipworth et al., 2019), which also reported one isolate was a BCG strain (starred). Conversely, one isolate (SRR16278270) was classified as MBO, but was shown not to possess any MBO lineage-determining SNPs, supporting a misclassification of an MTB isolate as MBO.

Table 1-4 (cont'd)

ERR046747	✓	✓	✓
ERR046748	✓	✓	✓
ERR046749	✓	✓	✓
ERR046954	✓	✓	✓
ERR046961	✓	✓	✓
ERR046989	✓	✓	✓
ERR387001*	✓	✓	✓
SRR16278270	X	X	X
*Flagged by SNP-IT software as BCG strain			

Lineage Marker	**Lineage 2**	**Lineage 2**	**Lineage 2**	**Lineage 2**	**Lineage 2.2**	**Lineage 2.2**	**Lineage 2.2**	**Lineage 2.2.1**	**Lineage 2.2.1**
Accession ID	G497491A	C811753T	A1834177C	T2543395C	C1849051T	G2505085A	C2775361T	C797736T	C3498198T
SRR16278270	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1-5: SNP typing improperly labeled MBO isolate by MTB-lineage markers. Isolate SRR16278270 was deposited in NCBI as an MBO isolate from cattle, but its genotype by GWAS did not align with other MBO isolates. This isolate was checked for MTB lineage-determining SNPs as reported by Coll *et al.* (2014), and was found to contain all SNPs for MTB lineage 2.2.1 and none for MBO (Table 1-4), indicating a misclassification in the database. This was confirmed by SNP-IT (Lipworth et al., 2019).

SNP	G1P1	G0P0	G1P0	G0P1	Locus	Protein	Change	Description	Notes
181672	4100	2100	161	1	Rv0153c	PtbB	Asp105Gly	Phosphotyrosine protein phosphatase PTPB (protein-tyrosine-phosphatase) (PTPase)	Required for growth on cholesterol (Griffin 2011)
294198	4100	2250	9	8	Rv0244c	FadE5	Glu479Ala	Probable acyl-CoA dehydrogenase FadE5	Required for growth on cholesterol (Griffin 2011)
1684979	4100	2080	172	1	Rv1493	MutB	Synonymous	Probable methylmalonyl-CoA mutase large subunit MutB (MCM)	Downstream in cholesterol to propionyl-CoA metabolic pathways (Wilburn 2018)
1754572	4100	2100	161	1	Rv1550	FadD11	Leu286Ser	Probable fatty-acid-CoA ligase FadD11 (fatty-acid-CoA synthetase) (fatty-acid-CoA synthase)	n/a
2997325	4100	2100	161	1	Rv2681		Ala196Val	Conserved hypothetical alanine rich protein	Required for growth on cholesterol (Griffin 2011)
3223303	4100	2250	9	8	Rv2914c	PknI	Synonymous	Probable transmembrane serine/threonine-protein kinase I PknI (protein kinase I) (STPK I) (phosphorylase B kinase kinase) (hydroxyalkyl-protein kinase)	Required for growth on cholesterol (Griffin 2011), mutant shows increased growth in THP-1 cells, SCID mice show faster mortality with mutant (Gopalaswamy 2009)
3912636	4100	2100	162	1	Rv3494c	Mce4F	Asp245Gly	Mce-family protein Mce4F	Required for growth on cholesterol (Griffin 2011)

Table 1-6: Genes identified by GWAS associated with fatty acid and cholesterol metabolism. A subset of genes in the pathways of lipid and cholesterol intake, metabolism, and utilization were identified with SNPs by GWAS, with a split roughly between one genotype in MTB ± MOR and a divergent genotype in MBO ± MCP. Columns 2-5 indicate presence of genotype **G** (SNP = 1, WT = 0) and phenotype **P** (MBO classification = 1, non-MBO classification = 0). The misclassification of 9 MBO isolates as MTB, and 1 MTB isolate as MBO (Tables 1-4, 1-5) are evident in the G1P0 and G0P1 columns for many variants.

Table 1-6 (cont'd)

3924350	4100	2100	161	1	Rv3505	FadE27	Ala218Val	Probable acyl-CoA dehydrogenase FadE27	n/a
3987645	4100	2250	9	8	Rv3548c		Met218Val	Probable short-chain type dehydrogenase/reductase	Required for growth on cholesterol (Griffin 2011)
4004604	4100	2250	9	8	Rv3563	FadE32	Gln105Arg	Probable acyl-CoA dehydrogenase FadE32	Required for growth on cholesterol (Griffin 2011), essential in murine spleen (Sassetti and Rubin, 2003)
3254695	4100	2100	161	1	Rv2932	PpsB	Synonymous	Phenolpthiocerol synthesis type-I polyketide synthase PpsB	In vitro essential in CDC1551 (Lamichhane 2003), not in H37Rv (Griffin 2011; DeJesus 2017; Minato 2019)
3262628	4100	2100	161	1	Rv2934	PpsD	Met127Ile	Phenolpthiocerol synthesis type-I polyketide synthase PpsD	n/a
3267715	4100	2100	161	1	Rv2934	PpsD	Glu1823Ala	Phenolpthiocerol synthesis type-I polyketide synthase PpsD	n/a
3320554	4100	2100	161	1	Rv2947c	Pks15	Synonymous	Probable polyketide synthase Pks15, involved in the biosynthesis of phenolphthiocerol glycolipids.	n/a

Figures:

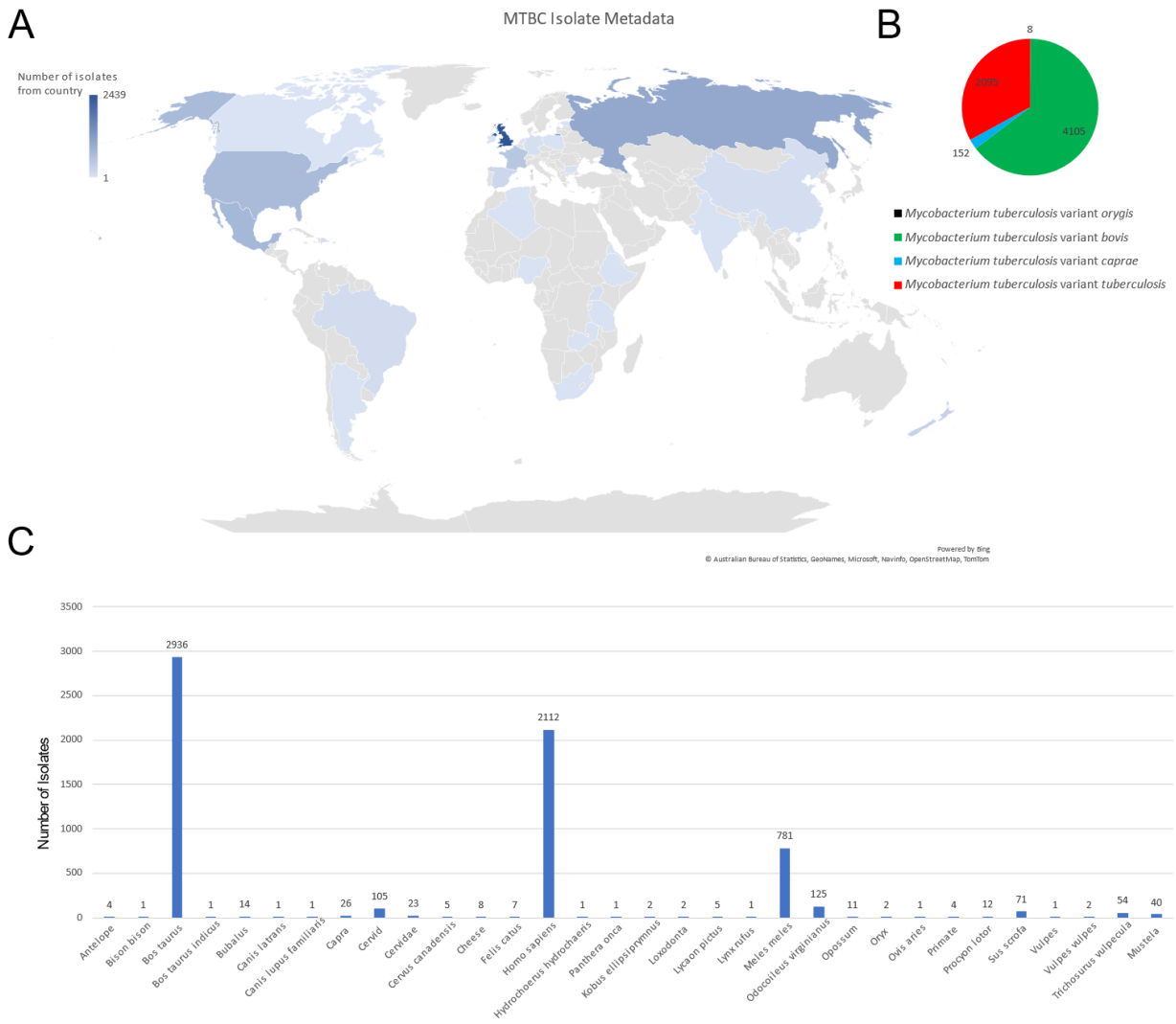

Figure 1-1: MTBC isolate collection metadata. **A)** Geographical distribution of isolates worldwide, where darker colors represent more isolates from that country. **B)** MTBC variant makeup of the collection. Values are based on NCBI/ENA SRA designations. **C)** Host origin for the collection. *Bos taurus* is the dominant host type, followed by *Homo sapiens* and *Meles meles*.

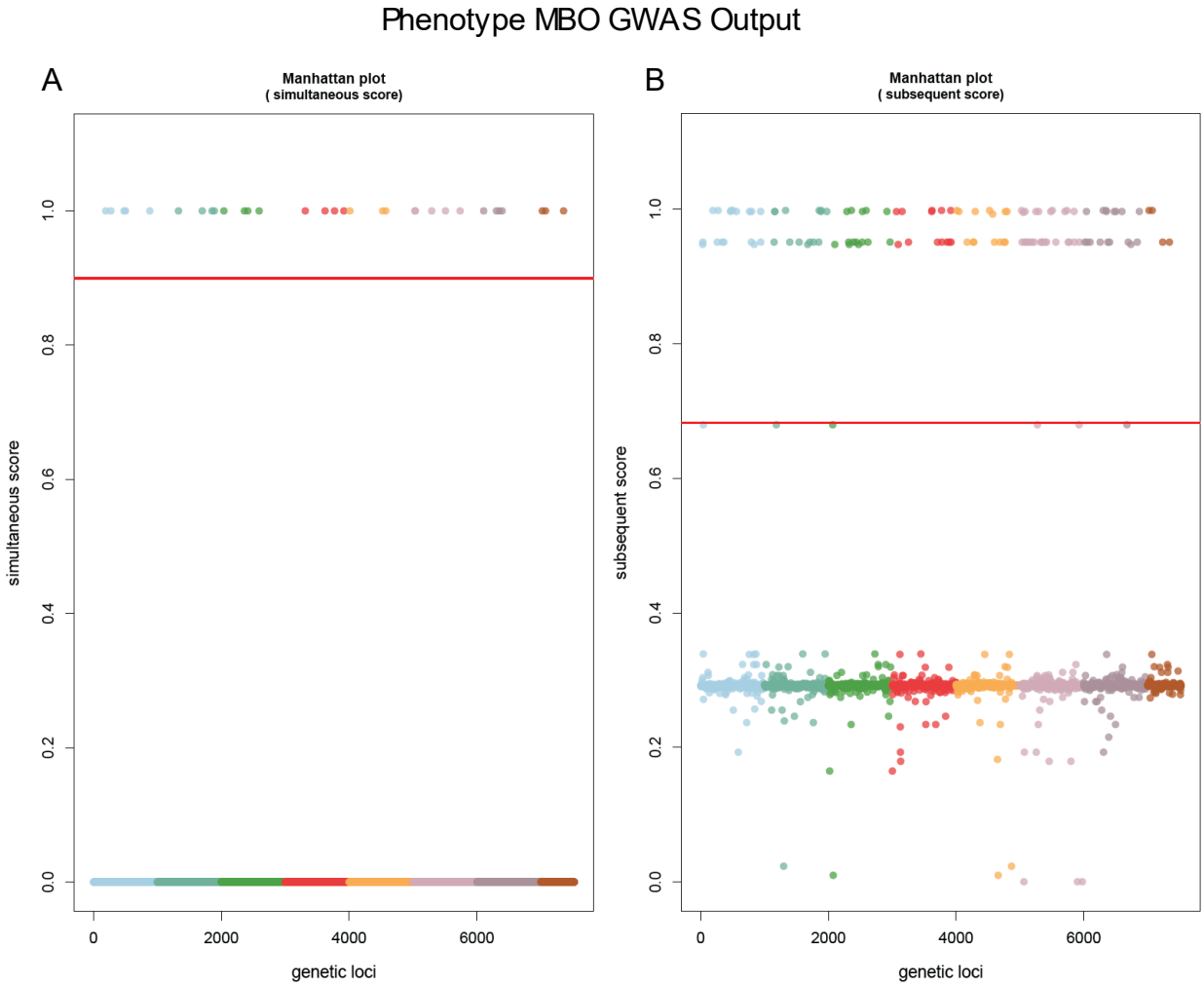

Figure 1-2: Manhattan plots of bGWAS results for phenotype "*M. tuberculosis* variant *bovis.*" X-axes represent SNPs in the order presented in the input binary genotype matrix. Y-axes represent arbitrary units for significance of association, where higher values (approaching 1) indicate a greater association between a genotype and a phenotype on the tree, vs. no association (0). **A)** Simultaneous test of association, showing 32 loci ranked to be significant, of which 10 are of dubious quality (Table 1-1). **B)** Subsequent test of association, showing 120 loci are ranked to be significant. Details for each locus are available in Tables 1-1 (simultaneous) and 1-2 (subsequent).

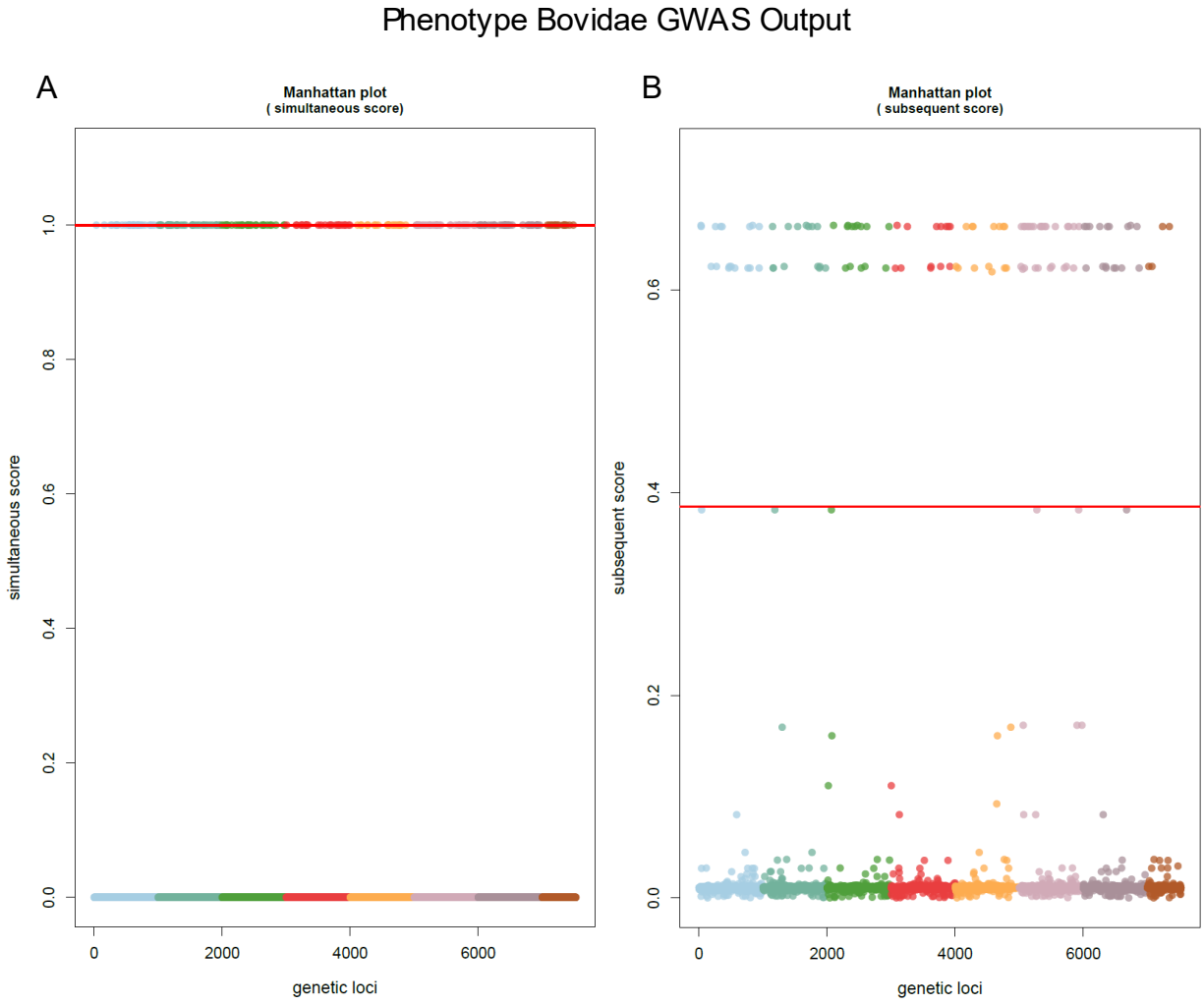

Figure 1-3: Manhattan plots of bGWAS results for phenotype "*Bovidae*" host. **A)** Simultaneous test of association, showing no significantly ranked loci. **B)** Subsequent test of association, showing 120 loci are ranked to be significant. The 120 loci identified here are identical to those seen in Table 1-2 and Figure 1-2.

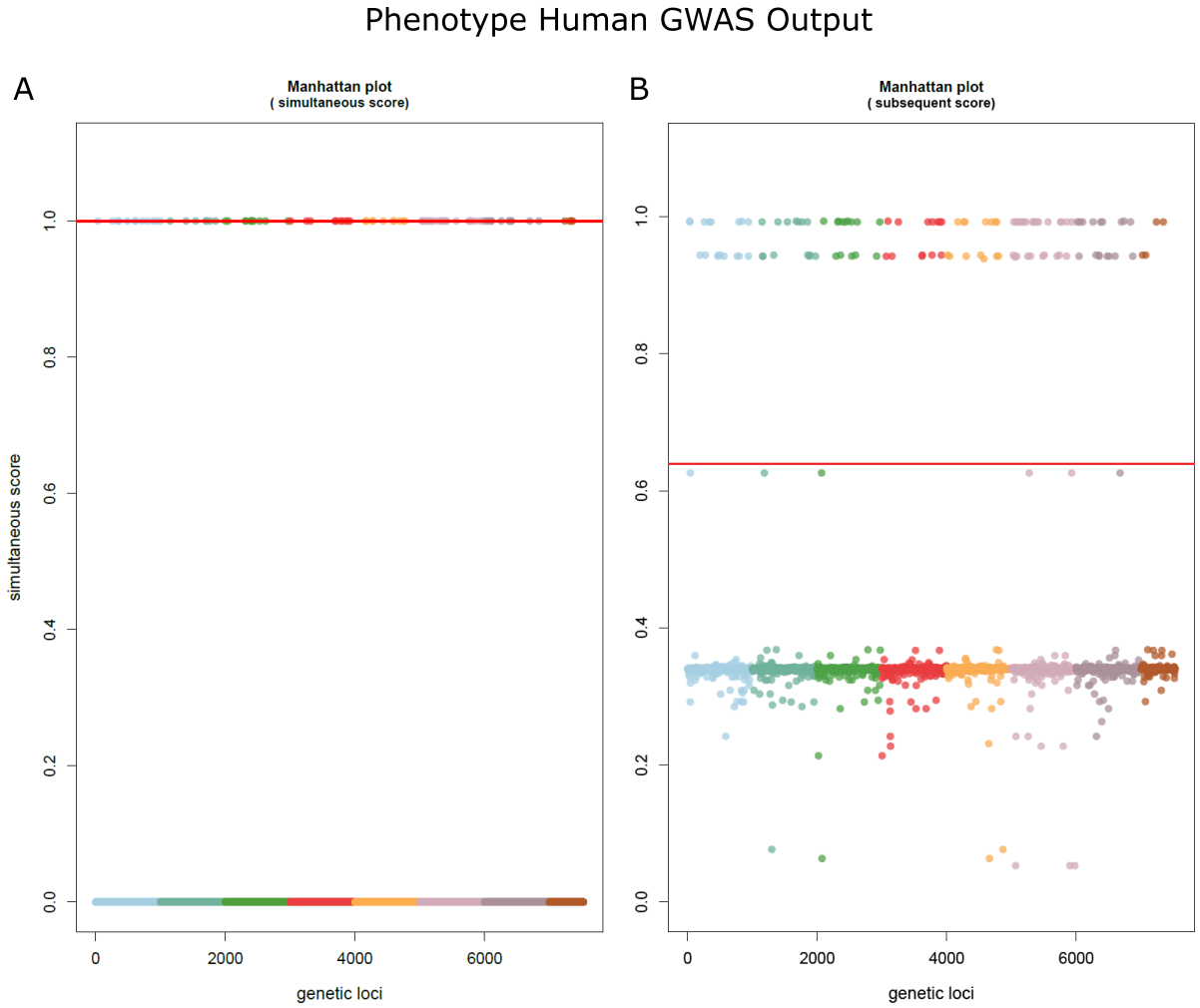

Figure 1-4: Manhattan plots of bGWAS results for phenotype "*Homo sapiens*" host**. A)** Simultaneous test of association, showing no significantly ranked loci. **B)** Subsequent test of association, showing 120 loci are ranked to be significant. The 120 loci identified here are identical to those seen in Table 1-2 and Figure 1-2 and Figure 1-3.

CHAPTER 3: *M. BOVIS* STRAIN RAVENEL SHOWS CHANGES IN EPITOPES THAT MAY CONTRIBUTE TO ATTENUATION

Abstract:

Tuberculosis, caused by *Mycobacterium tuberculosis* complex (MTBC) organisms, affects a range of humans and animals globally. Mycobacterial pathogenesis involves manipulation of the host immune system, partially through antigen presentation. Epitope sequences across the MTBC are evolutionarily hyperconserved, suggesting their recognition is advantageous for the bacterium. *Mycobacterium tuberculosis* var. *bovis* (MBO) strain Ravenel is an isolate known to provoke a robust immune response in cattle, but typically fails to produce lesions and persist. Unlike attenuated MBO BCG strains that lack the critical RD1 genomic region, Ravenel is classic-type MBO genetically, suggesting genetic variation is responsible for defective pathogenesis. This work explores variation in epitope sequences in MBO Ravenel by whole genome sequencing, and contrasts such variation against a fully virulent clinical isolate, MBO strain 10-7428. Validated MTBC epitopes (n=4,818) from the Immune Epitope Database were compared to their sequences in MBO Ravenel and MBO 10-7428. Ravenel yielded 3 modified T cell epitopes, in genes *rpfB*, *argC*, and *rpoA*, with changes not *in silico* predicted to significantly affect protein stability. In contrast, no T cells epitopes were changed in 10-7428. Considering T cell epitope hyperconservation across MTBC variants, these altered MBO Ravenel epitopes support a contribution to its attenuation. The affected genes may provide clues on basic pathogenesis, and if so, be feasible targets for reverse vaccinology.

Introduction:

Over 100 years since vaccination by *M. tuberculosis* variant *bovis* (MBO) strain BCG began, tuberculosis remains the deadliest single infectious agent in the world for humans. Our understanding of *Mycobacterium tuberculosis* complex (MTBC) pathogenesis remains lacking, both in humans and in the myriad non-human hosts MTBC variants attack. An underlying mechanism behind this is the subversion and misdirection of the host immune response, achieved both directly by targeted alteration of host kinase signaling cascades, and indirectly by intentional presentation of conserved antigens that drive specific immune feedback[69,75,100,108]. In a background of mycobacterial antigenic hyperconservation, this research asked what level of epitope variation can be detected in each an attenuated and a virulent MBO strain, and if variation may signal loss of virulence through dysregulation of host immune control.

At a simplistic level, a host's immune response to infection – including recognition of pathogen markers, response by innate and adaptive immunity, and ultimate pathogen clearance – depends on the receipt of signals differentiating self and foreign molecules. Antigens here generically refer to such markers that a host can react to in mounting a pathogen-specific immune response. Under the classic model, after initial infection, pathogen antigens recognized by the immune system are under selective pressure to change, and the host in return faces selective pressure to maintain recognition of changing targets in a process long referred to as an evolutionary arms race[46,60,246]. Over the course of the infection, immunity develops, targeting of specific antigens arises, the infection is suppressed, and re-infection by the same pathogen later hindered. On the other hand, immune responses to mycobacterial pathogens like *M. tuberculosis* are well-known to be more nuanced and skewed towards a cell-

mediated immune response[100,247,248]. Early research suggested a limited role of B cells or antibodies in protection, and although these subsets do yield benefits[69,175,249], it is known that a successful immune response against MTB infection absolutely requires $CD4^+$ and $CD8^+$ T cell activity[69,75,250]. Subsequent work has focused heavily on a process of initial engulfment of MTB by macrophages, subsequent intracellular MTB replication and cell signaling that leads to a Type IV hypersensitivity response wherein T cells contribute towards granuloma formation to seal in the antigen provoking the response[69,75,100]. In most infections, this successful response drives the granulomatous encasement of MTB, not its sterilization[183,247]. While most patients with latent TB infection (LTBI) do not progress again to active disease, MTB is able to survive indefinitely within the confines of the granuloma, and in reactivation, MTB drives caseation of the granuloma core and subsequent spillage of bacilli from the granuloma into the lung, allowing dissemination and respiratory transmission to others[69,75,183,205]. This reactivation is again believed to involve MTB host manipulation by antigen presentation in a hypersensitized host state, and is not associated with an increased bacterial load, indicating some degree of bacterial control by antigen presentation over the host immune response drives disease[183,205]. Amidst this backdrop, it has been observed that T cell epitopes in the MTBC are hyperconserved, equivalent to levels seen in the most essential MTB genes[60,251]. This supports a model wherein T cell recognition of MTB epitopes is essential to bacterial survival[60,100,183,251]. After millennia of coevolution in human hosts, MTB has developed a strategy of immune subversion not rooted in antigenic obfuscation, but in the intentional presentation of conserved targets that manipulate the host immune system into responding in unproductive ways that, at least at a species-scale, confer fitness benefits for the pathogen[183,252]. In short, T cell responses

are necessary to control tuberculosis, but MTB has also evolved to exploit these same responses, provoking specific immune reactions that can ultimately benefit MTB and perpetuate disease[60,175,183,251,253].

Given the specialization of pathogenic mycobacteria into manipulation of the host by intentional presentation of specific, hyperconserved T cell epitopes, this work posits that evidence of attenuation and adaptation can be found in epitope variation. Changes observed in known T cell epitopes in attenuated strains may represent a contributor to such attenuation, and identification of such epitopes would therefore be informative for pathways necessary for mycobacterial subversion of host immune responses. While most vaccine strategies continue to target immunodominant antigens like ESAT-6 or the Antigen 85 complex, the identification of other antigens where variation is associated with dysregulation of host immune manipulation could lead to selection of better, more protective targets.

This work sought to analyze variation of known MTBC epitopes between a subset of virulent and attenuated *M. tuberculosis* variant *bovis* (MBO) strains, with a hypothesis that variation in T cell epitopes will be more frequent in attenuated strains, and that these variant epitopes may contribute to observed attenuation. To explore this topic, polymorphisms were extracted from the recently sequenced genomes of MBO strain Ravenel, a naturally attenuated cattle strain that does not carry the causative genomic lesions of BCG strains, and MBO strain 10-7248, a fully virulent cattle clinical isolate[254–256]. For additional comparison, the MBO BCG-1 (Russia) vaccine strain believed closest to the now-lost ancestral BCG strain[124,257] and the MBO strain AF2122/97 reference[42,258] were analyzed, the former to assess epitope variation in a truly dysfunctional MBO strain, and the latter to exclude variation fixed in the MBO genetic

background for any epitopes characterized only in non-MBO complex members. . Validated MTBC epitopes collected from the Immune Epitope Database (IEDB) were compared against SNPs in the genomes of the four selected strains, with a focus on 1-3 amino acid changes to represent variation through point mutations, and to limit both nonspecific hits on other epitopes and the analysis of more extensive variations known to exist in PE/PPE genes that buck the trend of epitope conservation[46]. The analysis is represented by the schematic in Figure 2-1.

As expected, very few changes were identified overall. Variation was seen in multiple T cell epitopes for attenuated strains and none in virulent 10-7428, which only showed a B cell epitope altered in a PE/PPE gene already known to be variable. These affected T cell epitopes may play a role in the attenuation seen in MBO Ravenel, which does not produce persistent disease or lesions in cattle but does produce a robust cell-mediated immune response[256]. Further investigation into whether these and other variant T cell epitopes can contribute to a loss in MTBC host immune manipulation is warranted.

Materials and Methods:

The US National Institutes of Health and the Department of Health and Human Services jointly support and operate the Immune Epitope and Analysis Database (IEDB), a comprehensive database of over 1.5 million experimentally validated epitopes from a range of species and diseases[259]. The IEDB (accessed September 2022) was used to collect all epitopes validated in the *Mycobacterium tuberculosis* complex (ID:77643, n=4,894), primarily from *Mycobacterium tuberculosis* var. *tuberculosis* (n=4,111) and *M. tuberculosis* var. *bovis* (n=783), including 356 MBO BCG strains[259]. These sequences include non-peptide targets like LAM which were removed (n=46), along with B cell-specific discontinuous peptide sequences (n=2), and

peptides that had undergone post-translational modifications (n=28), both because they are unable to be processed through tblastn and because the presence or absence of these modifications could definitionally not be verified by screening against whole genome sequences. The remaining 4,818 sequences were converted to FASTA format and uploaded to the High Performance Computing Center at Michigan State University (Supplemental File S2-1). TBLASTN v2.10.0+ was utilized with custom databases built from four genome assemblies[260,261]. Two of these strains of *M. tuberculosis* var. *bovis* were sequenced and analyzed by our lab: attenuated Ravenel (GCF_018305025.1) and virulent 10-7428 (GCF_018305045.1)[255]. From existing databases, the attenuated reference BCG-1 (Russia) (GCF_001483905.1), and virulent reference AF2122/97 (GCF_000195835.3) were selected. The BLAST+ package (v2.11.0)[261], preinstalled on HPCC, was loaded and the makeblastdb command used with default parameters to generate local databases based on each of the four genomes mentioned. . All TBLASTN searches were with default BLAST parameters, with the filtered list of 4,818 epitope sequences searched against each genome database[261]. Some highly repetitive epitopes – particularly those from PE/PPE genes – were caught by filtering, Karlin-Altschul parameters were not calculated by the BLAST algorithm, and they were subsequently excluded from analysis. The TBLASTN output was produced twice, once in pairwise-alignment format, and once in tabular format. Tables were viewed in Excel, sorted by number of mismatches, sequences with gaps were excluded, and the subsets of homology hits with 0, 1, 2, or 3 mismatches separated into their own respective sheets. The XLOOKUP function was used to find whether any mismatched epitopes had a 100% match against any variant epitope in the Match sheet, and epitopes that appeared only to vary from the IEDB listed sequences – that is, epitopes that truly do not

perfectly match known epitope sequences in IEDB – were moved into their own sheets. Finally, for strains Ravenel, 10-7428, and BCG-1, the list of epitopes that varied from IEDB sequences was queried against the same list of TBLASTN searches from MBO reference strain AF2122/97 in order to identify and dismiss candidate epitopes that were identical to the MBO reference and may simply represent the MBO genetic background. The pairwise BLAST results list was then analyzed for each remaining mismatch, epitope IDs queried on IEDB for details, and results recorded. A schematic representation of this workflow is presented in Figure 2-1. Unique epitopes in Ravenel, 10-7428, and BCG-1 Russia were queried through UniProt and NCBI, and effects on protein stability examined by three high-performing software tools benchmarked by Pancotti *et al.* (2022) – DDGun3D, PremPS, and INPS – as well as an additional recently released tool, DynaMut2, using *in vitro* crystal structures preferentially and DeepMind's AlphaFold *in silico* predictions when *in vitro* structures were not available[263–267]. Finally, BLAST searches were also performed with the full-length proteins against their H37Rv and AF2122/97 counterparts to explore potential compensatory mutations.

Results:

For Ravenel, 4,130 raw epitope hits were recorded against the IEDB dataset (Table 2-1: Raw Hits). Of these, 2,488 were a perfect match to a characterized epitope, which are considered uninformative as this work seeks epitope variation (Table 2-1: Matches). In Ravenel, filtering the remaining subset to epitopes showing a single amino acid mismatch vs. a characterized epitope, 100% coverage of the Ravenel hit vs. the IEDB query, and with no gaps vs. the IEDB query returned 303 matches (Table 2-1: Degenerate Mismatches). In most cases in the epitope dataset, sequences are *M. Tuberculosis* var. *tuberculosis*-derived, and additionally,

one epitope can have multiple known variants in the database – thus, degenerate mismatches – and so each mismatch was then interrogated for a perfect match against other variant epitope sequences, as well as against the AF2122/97 epitope hits to reduce the chance variants were simply fixed in the MBO background. Of the 303 Ravenel single amino acid hits against the epitope dataset, 188 mismatches were found to have a 100% match against a different variant epitope in the dataset and were therefore excluded, leaving 115 epitopes in Ravenel that differed by 1 amino acid from the IEDB sequences. Comparing these 115 epitopes to the output for the epitope workflow for AF2122/97 yielded 100% matches for 110, leaving just 5 epitopes in Ravenel that were mismatches from known sequences, and different from the same epitope sequence in the MBO reference strain AF2122/97 (Table 2-1: Unique Mismatches). Expanding to 2 amino acid mismatches yielded 293 degenerate epitopes, and 0 unique epitopes. Finally, 3 mismatches returned 289 degenerate epitopes, and filtered to 1 unique epitope. This process was repeated for 10-7428 and BCG-1. It is important to note that for reference AF2122/97, the process does not include subtracting epitopes that may be the wild-type in MBO, yielding an artificially high count of 121 1AA changes, 28 2AA changes, and 13 3AA changes. Table 2-1 reports the initial results for each of the 4 strains.

Most mismatches in Ravenel, 10-7428, and BCG-1 mapped perfectly to an existing variation in AF2122/97, leaving only a handful of changes not observed elsewhere (Table 2-1: Unique Mismatches). In Ravenel, five potentially impactful single amino acid mismatches were initially recorded. Epitope ID 229352, a 15aa epitope and one of only two known antigenic regions in RpfB, showed Glu263Gly. Epitope 595988 in ArgC presented Tyr20His. Epitope 597585 in RpoA showed Glu75Asp. These substitutions are provided as Supplemental File S2-2.

The remaining two mismatches (in epitopes 163642 and 163423) both affected the same gene, *esxJ*, but this region showed an unusual pattern in the Ravenel assembly with the gene broken into 3 ORFs each containing partial starts and stops, with an identical broken pattern seen in the 10-7428 assembly suggesting this may be a systemic assembly error, and so changes in this gene were marked as unreliable. The three amino acid mismatch in Ravenel (epitope 100593) was also unreliable, mapping to an unknown PE/PPE family gene on an unplaced contig. These excluded sequences are listed in Supplemental File S2-3. In contrast, virulent strain 10-7428 showed a single 1 amino acid substitution, PPE42 Asn232Asp in B cell epitope 10022 (Supplemental File S2-4). Strain 10-7428 showed a 3 amino acid mismatch beyond this, again in partial *esxJ* hits could be an assembly error (Supplemental File S2-5) which led to subsequent exclusion of this epitope.

As expected, BCG-1 Russia showed the most variation in this analysis, a finding that likely arises both from its historical age relative to the others in the dataset as well as its impaired functionality. Antigen 85B showed a substitution Phe140Leu that impacted four overlapping epitopes; PPE genes featured multiple variants, including three epitopes inconclusively from the highly homologous PPE18/PPE19 cluster of genes known to regularly contain alterations to known T-cell epitopes[46] and one in PPE25; the virulence-associated serine protease MarP showed one; and one change was observed in an Mce family protein (Supplemental File S2-6).

The values for AF2122/97 are relative to the data from IEDB with is predominated by MTB epitopes, and so its unique mismatches values appear much greater than the other

strains. A more objective comparison would be analyzing degenerate mismatches (Table 2-1: Degenerate Mismatches), where strains appear similar.

Predictions of effects on protein stability by ΔΔG values were performed and recorded for observed single mutations in Ravenel and 10-7428 (Table 2-2)[263–267]. Full length protein sequences for RpfB, ArgC, and RpoA in Ravenel, and PPE42 in 10-7428 were searched by BLASTP against the AF2122/97 sequences for potential compensatory mutations, but only the originally detected single amino acid substitution was seen across the entire protein for each of these. For RpoA, given its presence in the large multi-subunit RNA polymerase complex, additional searches were performed for Ravenel's RpoB and RpoC sequences, but these were unchanged relative to the AF2122/97 reference.

Discussion:

It is known that, with the exception of some members of the PE/PPE gene family, T cell epitopes are hyperconserved across the MTBC[46,60,251]. This is thought to reflect an evolutionary strategy of intentional host immune manipulation by presentation of T cell epitopes to drive specific immune responses that the pathogen can leverage to its advantage[60,183,251,253]. In this work, epitopes were studied to assess whether variation in T cell epitopes might be associated with attenuation by the pathogen's loss of immunomodulatory potential. An initial investigation was performed using recently sequenced MBO Ravenel, an attenuated strain that provokes a robust immune response in cattle but does not cause lesions or persistent disease. In comparison, fully virulent cattle isolate MBO strain 10-7428 was also analyzed, along with MBO BCG-1 and MBO AF2122/97. Overall, the number of changes observed across the three strains

compared to AF2122/97 was small, but was expected based on existing knowledge about T cell epitope conservation[60,251]. Analysis shows 3 changes to T cell epitopes across Ravenel, impacting dormancy, arginine biosynthesis, and the alpha chain of the DNA-dependent RNA polymerase. In contrast, no T cell epitope changes were seen in 10-7428, which bore only a single change in a B cell epitope of PPE42, an immunogenic gene already known to exhibit antigenic variation[46,182]. These results are compatible with the hypothesis that T cell variation could impair pathogenesis, though further investigation and confirmatory testing of these specific epitopes are required.

Ravenel changes:

RpfB, or resuscitation promoting factor B, is one of five *rpf* genes known to be crucial in the transition from mycobacterial dormancy back to growth and infection dissemination. It is known that deletion of any one *rpf* gene still allows normal *in vitro* or *in vivo* growth but significantly impaired reactivation in a murine model of MTB infection, and furthermore that deleterious effects in infection and persistence are dramatic in a double-knockout background[268]. MBO Ravenel is capable of cattle infection, but after provoking a strong immune response it fails to produce lesions in most experimentally infected animals, unlike virulent strains[256,269]. The change (Glu263Gly) observed in epitope 229352 was earlier reported as a SNP[256]. The large, polar, charge-bearing glutamic acid to a small, flexible, non-polar glycine is a major alteration, and across homologues, position 263 is almost always either a glutamic acid, or an aspartic acid (the COG3583 consensus sequence residue). RpfB in Ravenel is unique in NCBI's Identical Protein Groups database, and TBLASTN searches of the NCBI NR database and WGS database filtered by *Mycobacterium tuberculosis* complex (taxid: 77643) showed no hits

for this substitution outside Ravenel. Per a partial crystal structure by Ruggiero *et al.*, residue 263 (Figure 2-2, green highlight) is in a linking region connecting two different three-strand stretches of beta-sheets, and the enhanced flexibility caused by replacement of this large, charged residue with a glycine could destabilize this region[270]. However, *in silico* ΔΔG predictions yielded inconsistent results ranging from weakly stabilizing to weakly destabilizing (Table 2-2: RpfB) and what effects this may have on the larger protein structure and function are uncertain. Since the catalytic residue and binding pocket is more than 30 amino acids upstream, it is unlikely that catalysis is directly compromised by this change, but the influence on tertiary structure and function is an open question that needs to be further evaluated. Regardless, this substitution is considerable at an epitope level, so even if protein functionality is maintained, altered immune recognition of the RpfB protein may still affect virulence. Epitope 229352 is T-cell epitope shown previously to elicit a $CD8^+$ dominant response and release of IFN-γ and TNF-α[271]. As such, RpfB remains an interesting candidate for modulation of virulence in *M. tuberculosis* var. *bovis* Ravenel and beyond.

ArgC, or N-acetyl-gamma-glutamyl-phosphate reductase, is an oxidoreductase found in the L-arginine biosynthesis pathway[272] and well-conserved across Actinobacteria. Transposon mutagenesis in MTB has identified all Arg members as essential *in vitro*[273], and *argB* and *argF* knockouts are efficiently sterilized from murine infection models in both C57BL/6 and immunocompromised SCID mice, with Δ*argB* infections of the latter group resulting in 100% survival to 300 days and complete clearance even at 10^8CFU/mL doses[274]. Less research has been performed with ArgC by comparison, but a modest IFN-γ response against the specific epitope 595988 has been demonstrated by ELISA in *Bos taurus*[275]. Investigation by Gupta *et al.*

of MTB ArgC places Tyr20 near a structural center of ArgC where it forms a hydrogen bond with Glu203[272]. Upon binding substrates, the conformation of ArgC shifts around the region of this mutation[272], which could allow modest functional impacts despite the similar structures of tyrosine and histidine. This Tyr20His alteration (Figure 2-3: black stars) in epitope 595988 (Figure 2-3: green) is observed in a total of 6 MTBC isolates on NCBI: Ravenel, two human MTB isolates (2926STDY5723476, 01-R1463) a cattle MBO isolate (2008/0665), an MBO type strain (ATCC 19210), and an MBO lab isolate (strain 30). Like for RpfB, the impact of this change on a short, linear T cell epitope seems more likely significant than a tyrosine to histidine substitution for overall protein function. While this change is rare, its presence in a small number of human and bovine clinical isolates makes it unlikely to have a major impact on pathogenesis alone. It may instead be a contributor in overall attenuation through cumulative changes in T cell recognition.

RpoA, or the DNA-directed RNA polymerase subunit alpha, is a core subunit of the RNA polymerase (RNAP) complex conserved across bacteria. It is the target of the antibiotic rifampicin, and while this critical enzyme is known to be intensely conserved, in MTB in particular, changes in RpoA-RpoC are known to be associated with rifampicin resistance, especially with changes in RpoB[276,277]. It has also been found in *Mycobacterium* as well as *Salmonella* that compensatory mutations across the genes in the RNAP complex are necessary to offset the fitness deficits of resistance-conferring mutation in these essential genes[276,278,279]. In MBO Ravenel, a change was observed in epitope 597585, leading to RpoA Glu75Asp (Figure 2-4). No other substitutions are seen in RpoA, RpoB, or RpoC. Curiously, RpoA Glu75Asp is seen in all the same strains that carry ArgC Tyr20His change, as well as MBO strains M1009 and

M1010, both clinical isolates from slaughter of two cattle (*Bos taurus*) in Paraguay. Like epitope 595988, RpoA's T cell epitope 597585 has been validated in a bovine interferon gamma release assay[275]. The substitution is not one believed associated with drug resistance, and while the physiochemical differences between aspartic and glutamic acids are minimal, this change may influence immune recognition, or even still incur minor fitness costs given the fundamental role RpoA plays. Regardless, it seems unlikely this epitope would be a major driving factor of observed attenuation in Ravenel and might instead be a contributor among a constellation of changes.

A 4th single amino acid mismatch in Ravenel, in epitope 163423, was found to be in a fragmented sequence of EsxJ that appeared to represent assembly error. This epitope mismatch was discarded as unreliable. Subsequently, the last 1aa mismatch (epitope 163642) was found to involve this same gene and was also discarded. Finally, the 3aa mismatch in Ravenel was determined to fall in a partial PPE family protein (WP_152345480.1) that aligns ambiguously with multiple possible PE/PPE genes. Due to the unreliable nature of these genes in sequencing and assembly, this match was disregarded.

Strain 10-7428 change:

Only one unique single epitope change was observed in the virulent strain 10-7428, a modification of PPE42 (Mb2640 in MBO). This protein is known to be highly immunogenic, stimulates a strong humoral response in human patients[280], and is one of the selected antigens in the ID93 subunit vaccine[178,281]. It is suggested that variation in some PE/PPE genes, unlike in most other antigens, may actually be beneficial to MTB in avoiding a productive, Th1-dominant

host immune response[281]. In strain 10-7428, epitope 10022, one of three recognized in PPE42, presented with Asn232Asp. This modification is seen in only one other sequenced isolate. Interestingly, 10022 is a validated B cell epitope unlike the exclusive T cell epitopes changes seen in Ravenel. In contrast with other PE/PPE family genes analyzed in this work, PPE42 is a relatively distinct gene, and sequencing and placement of this substitution is clean in the Ravenel and 10-7428 assemblies. However, as no *in vitro* crystal structure has been obtained for this protein and PE/PPE genes are often poorly understood regardless, any interpretation of this finding must remain limited absent further investigation *in vitro*.

Like in Ravenel, an additional epitope change was reported that also arose from an atypical assembly pattern around *esxJ*, and this epitope was discarded as an artifact.

<u>BCG-1 Russia changes</u>:

The Antigen 85B complex has been known for decades to be a dominant secretion product and immune stimulator[282,283]. Antigen 85B within the MTBC is strongly conserved, with more variation across NTM but noted antibody cross-reactivity between all variants[80]. In BCG-1 Russia, a single substitution was observed, Phe140Leu, that resulted in a change that appears fixed across all BCG strains, with very few non-BCG examples from lab or clinical strains. This single amino acid change modifies four characterized epitopes (149353, 196212, 16926, and 18898) that all overlap this position.

MarP is a serine protease involved in maintenance of intracellular pH during phagocytosis through the processing of peptidoglycan through its interactions with RipA[284]. Loss of MarP or its catalytic residues leave the bacterium severely impaired for pathogenesis *in*

vitro and in a murine model of infection[285,286]. In BCG-1 Russia, epitope 600022 of MarP showed Asn165Thr, a mutation in the flexible linker region between transmembrane anchor and protease domain. Like the Ag85B mutation, this substitution is also fixed across BCG strains and appears again only in the same small number of lab and clinical isolates, raising the possibility these isolates may be evolutionarily similar to BCG, or cases of BCG-osis.

Mce (mammalian cell entry) family proteins are a class of protein involved with entry into host cells and interference with host cell signaling pathways, particularly through the ERK1 and ERK2 MAPK pathway for inducing cytokine production[287]. T cell epitope 20707 presented with Thr46Pro. At a molecular level, the sequence change is another significant structural alteration, though whether this affects the function of this protein or its immune recognition in BCG-1 are beyond the scope of this work.

PE/PPE family changes were numerous in BCG-1, but as before, they are difficult to rely on and are reported only with the caveat that their true sequence and placement in the assembly are unclear.

Predicted Structural Effects:

Finally, *in silico* predictions for ΔΔG were variable and conflicting (Table 2-2), a known problem in the field[288]. The use of these tools may also be complicated by the specific portions of molecules being studied – surface residue substitutions have been shown previously to be more poorly predicted than changes to residues buried within the protein[289], and epitopes herein were exposed residues. Despite this, a message remains in these contradictory findings: no consensus change of a magnitude >0.5kcal/mol in either a stabilizing or destabilizing

direction was measured for any affected protein in Ravenel or 10-7428, and it is therefore less likely that any observed substitutions are sufficient to modify an affected protein enough to ablate functionality. This investigation is premised on the idea that changes to epitope recognition can lead to dysregulation of host manipulation by *Mycobacterium* species. Indeed, if mutant proteins retain their molecular functionality but still alter virulence, modified epitope recognition is a means of describing this contradiction.

Conclusions:

Using epitope data gleaned largely from *M. tuberculosis* var. *tuberculosis*, and after sequencing a uniquely attenuated *M. tuberculosis* variant *bovis* genome in strain Ravenel, the haystack of 4,130 epitope alignments yielded 3 needles of epitopes with non-synonymous changes in characterized epitopes. Two of these epitopes have been validated in bovine IFN-γ release assays, and one in a murine model showing IFN-γ and TNF-α release. These affected genes are of particular importance for bacterial survival: with arginine biosynthesis genes are so critical that even SCID mice unable to develop B and T cell responses are still able to sterilize Arg pathway MTB mutants; DNA-dependent RNA polymerase subunits are fundamental to gene expression and mutants of these proteins are known to show significant fitness deficits; and the resuscitation promoting factor (Rpf) genes are believed important for MTB to reemerge from dormancy in the granuloma. As recent work has demonstrated, MBO Ravenel provokes potent cell-mediated immune responses in the bovine host, but fails to produce granulomatous lesions in nearly all cases, and the process of an infection leading to granuloma formation is known to be tightly associated with the precise presentation of specific T cell epitopes at specific times[183,256]. In contrast to these, an isolate from a dairy cattle outbreak – MBO 10-7428 – was

found to contain no T cell epitope variation, and instead the only change fell in a B cell epitope in a gene where variation is associated with increased virulence. These results are not conclusive, but support a closer inspection of whether changes in T cell epitopes may be an indicator of attenuation in MBO Ravenel and other strains, an investigation of how affected genes are involved in pathogenesis, and whether any of these epitopes may prove useful as subunits for vaccine development, as has already been proposed for RpfB[271]. The analytic process described here is simple, and with limited development could be readily incorporated into standard comparative genomics analysis for MTBC organisms, increasing the amount of information researchers can extract from each experiment.

Tables:

	Ravenel			10-7428			BCG-1 Russia			AF2122/97		
Raw Hits	4130			4539			4303			4635		
Matches	2488			2734			2466			2773		
Degenerate	**1aa**	**2aa**	**3aa**	**1aa**	**2aa**	**3aa**	**1aa**	**2aa**	**3aa**	**1aa**	**2aa**	**3aa**
Mismatches	303	293	289	335	321	314	358	321	320	343	327	328
Unique	**1aa**	**2aa**	**3aa**	**1aa**	**2aa**	**3aa**	**1aa**	**2aa**	**3aa**	**1aa**	**2aa**	**3aa**
Mismatches	5	0	1	2	0	0	8	0	2	121*	28*	13*

Table 2-1: Epitope homology results for four genomes. *M. tuberculosis* var. *bovis* strains AF2122/97 (virulent, reference), BCG-1 Russia (attenuated, reference), 10-7428 (virulent, newly sequenced), and Ravenel (attenuated, newly sequenced). Epitopes (n=4,894) were selected from IEDB.org and filtered to 4,818 to query against the 4 genomic .fna files indexed for TBLASTN. The raw hits row indicates the number of epitopes aligning anywhere in the genome designated per column. Matches indicates 100% similarity to at least one epitope variant in the IEDB dataset. Mismatches indicates alignment but with #aa differences (indicated per column, 1, 2, or 3 mismatches). Finally, Unique Mismatches indicates the mismatched sequence does not also map perfectly to any other variant epitope sequences from IEDB or, with the exception of blue-shaded AF2122/97 cells, that these changes are not observed in AF2122/97 epitopes either. Sequences with gaps or with more than 3 mismatches are not included in the analysis, so values do not sum to the original Raw Hits value. *AF2122/97 values are not subtracted from what may be MBO wild-type variation like other strains, and thus its values should not be compared directly.

Protein	Input	Mutation (chain)	Software	ΔΔG Prediction
RpfB	AF A0A2Z3DFM7 *In silico*	Glu263Gly (A)	DDGun3D	-0.4 kcal/mol
RpfB	AF A0A2Z3DFM7 *In silico*	Glu263Gly (A)	PremPS	0.1 kcal/mol
RpfB	AF A0A2Z3DFM7 *In silico*	Glu263Gly (A)	DynaMut2	-0.58 kcal/mol
RpfB	Sequence, FASTA	Glu263Gly	INPS	-1.02 kcal/mol
RpfB	PDB 3EO5 Partial, *in vitro*	Glu263Gly	DDGun3D	-0.5 kcal/mol
RpfB	PDB 3EO5 Partial, *in vitro*	Glu263Gly	PremPS	0.47 kcal/mol
RpfB	PDB 3EO5 Partial, *in vitro*	Glu263Gly	DynaMut2	-0.39 kcal/mol
ArgC	PDB 7NNI Xray, complex, *in vitro*	Tyr20His (A) Tyr20His (B)	DDGun3D	-1.1 kcal/mol
ArgC	PDB 7NNI Xray, complex, *in vitro*	Tyr20His (A) Tyr20His (B)	PremPS	1.88 kcal/mol 1.89 kcal/mol

Table 2-2: ΔΔG predictions on protein structure of a subset of identified mutations. Software packages DDGun3D, PremPS, and DynaMut2 utilize input of .pdb structures, though PremPS only accepts X-ray crystallography and not Cryo-EM structures. Software INPS uses sequence information exclusively.

Table 2-2 (cont'd)

ArgC	PDB 7NNI Xray, complex, *in vitro*	Tyr20His (A) Tyr20His (B)	DynaMut2	-0.11 kcal/mol -0.08 kcal/mol
ArgC	Sequence, FASTA	Tyr20His	INPS	-1.17 kcal/mol
RpoA	PDB 7Q59 CryoEM, dimer, *in vitro*	Glu75Asp (A) Glu75Asp (B)	DDGun3D	-0.2 kcal/mol -0.2 kcal/mol
RpoA	PDB 7Q59 CryoEM, dimer, *in vitro*	Glu75Asp (A) Glu75Asp (B)	DynaMut2	-1.03 kcal/mol -1.05 kcal/mol
RpoA	PDB 5UHA Xray, complex, *in vitro*	Glu75Asp (A) Glu75Asp (B)	DDGun3D	-0.1 kcal/mol -0.1 kcal/mol
RpoA	PDB 5UHA Xray, complex, *in vitro*	Glu75Asp (A) Glu75Asp (B)	PremPS	0.05 kcal/mol
RpoA	PDB 5UHA Xray, complex, *in vitro*	Glu75Asp (A) Glu75Asp (B)	DynaMut2	-1.03 kcal/mol
RpoA	Sequence, FASTA	E75D	INPS	-0.30 kcal/mol
PPE42	AF P9WHZ5 *In silico*	N232D	DDGun3D	-0.2 kcal/mol
PPE42	AF P9WHZ5 *In silico*	N232D	PremPS	1.26 kcal/mol
PPE42	AF P9WHZ5 *In silico*	N232D	DynaMut2	-1.1 kcal/mol
PPE42	Sequence, FASTA	N232D	INPS	-0.21 kcal/mol

Figures:

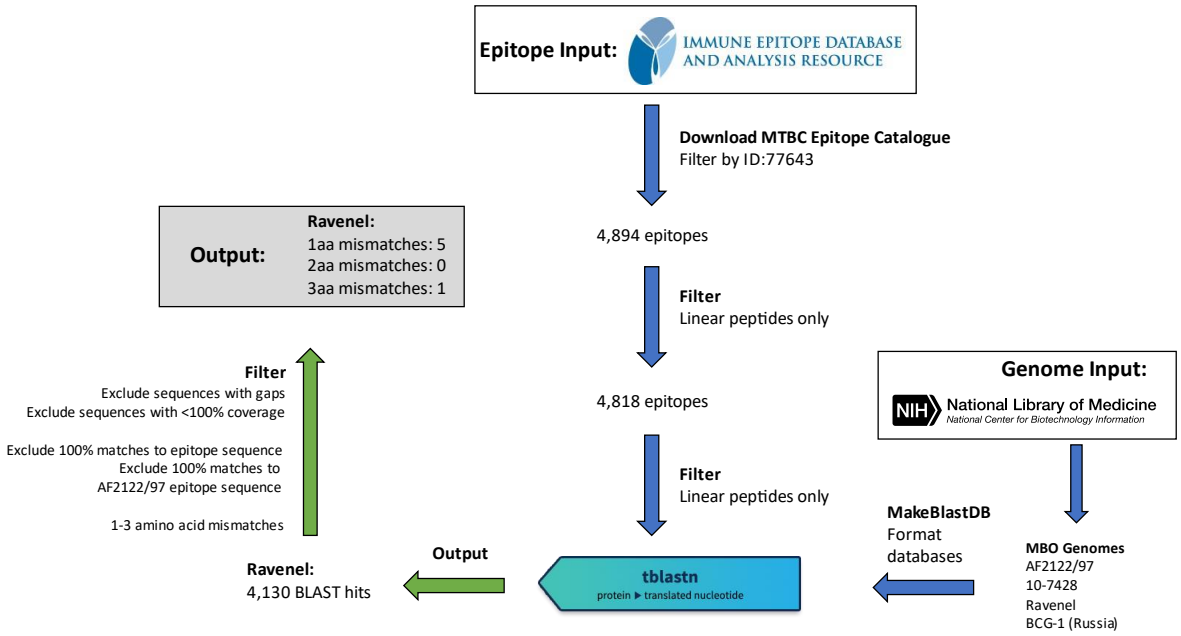

Figure 2-1: Schematic of epitope extraction workflow for MBO strain Ravenel. **Epitope Input:** All classified epitope sequences from any *Mycobacterium tuberculosis* complex members are downloaded and filtered to include linear peptide sequences. **Genome Input:** The MBO reference genome AF2122/97, as well as reference BCG-1 (Russia) genomes were downloaded in .fna format from NCBI RefSeq. Strains Ravenel and 10-7428 were downloaded as draft genomes. Genomes were processed by BLAST+ command makeblastdb to generate searchable BLAST databases. **TBLASTN:** BLAST searches were performed with default parameters, searching 4,818 epitopes against each genome database to match IEDB epitopes to their counterparts encoded in the genome. **Output and Filtering:** BLAST hits were filtered as described in Materials and Methods to identify epitopes with amino acid substitutions vs. the IEDB epitope and that are not seen in reference AF2122/97.

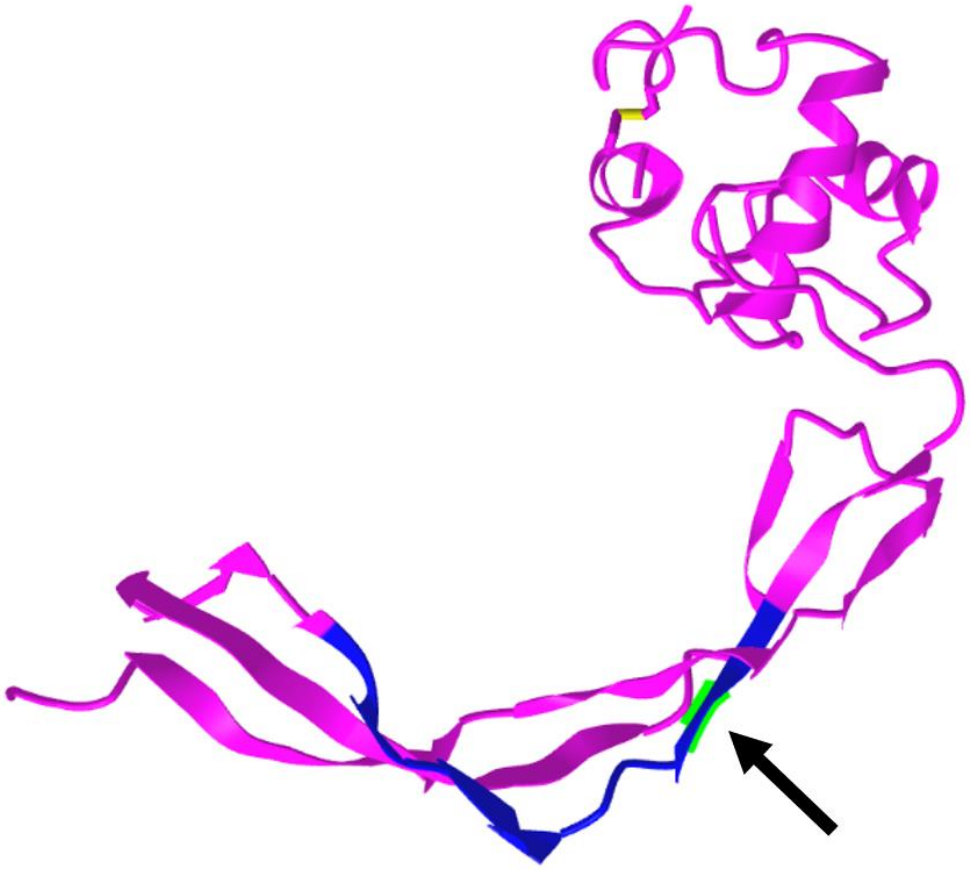

Figure 2-2: RpfB 3D structure (PDB 3EO5). Crystal structure generated by Ruggiero *et al.* (2009) by X-ray diffraction, representing RpfB residues 194-362 making up the G5 domain. Flat arrow shapes represent beta sheets, corkscrews represent alpha helices, and other parts are unstructured. Blue backbone: epitope 229352. Black arrow pointing to green highlight: Residue 263 (E263G in Ravenel).

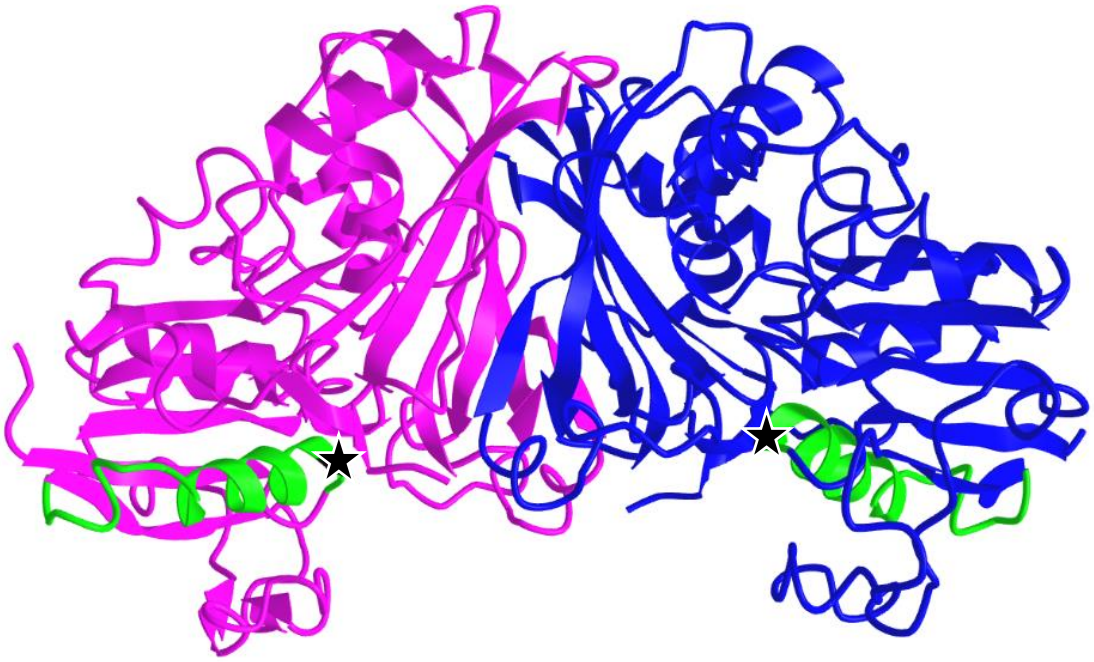

Figure 2-3: ArgC 3D structure (PDB 2NQT). Crystal structure generated by Cherney *et al.* (2006) by X-ray diffraction of full-length protein. Flat arrow shapes represent beta sheets, corkscrews represent alpha helices, and other parts are unstructured. Magenta and blue: ArgC chains A and B, respectively. Green: epitope 595988. Black stars at end of green epitopes: Residue 20 (Y20H in Ravenel). Partial complex of 2NQT shown over full complex of 7NNI used in DDG predictions for visual clarity.

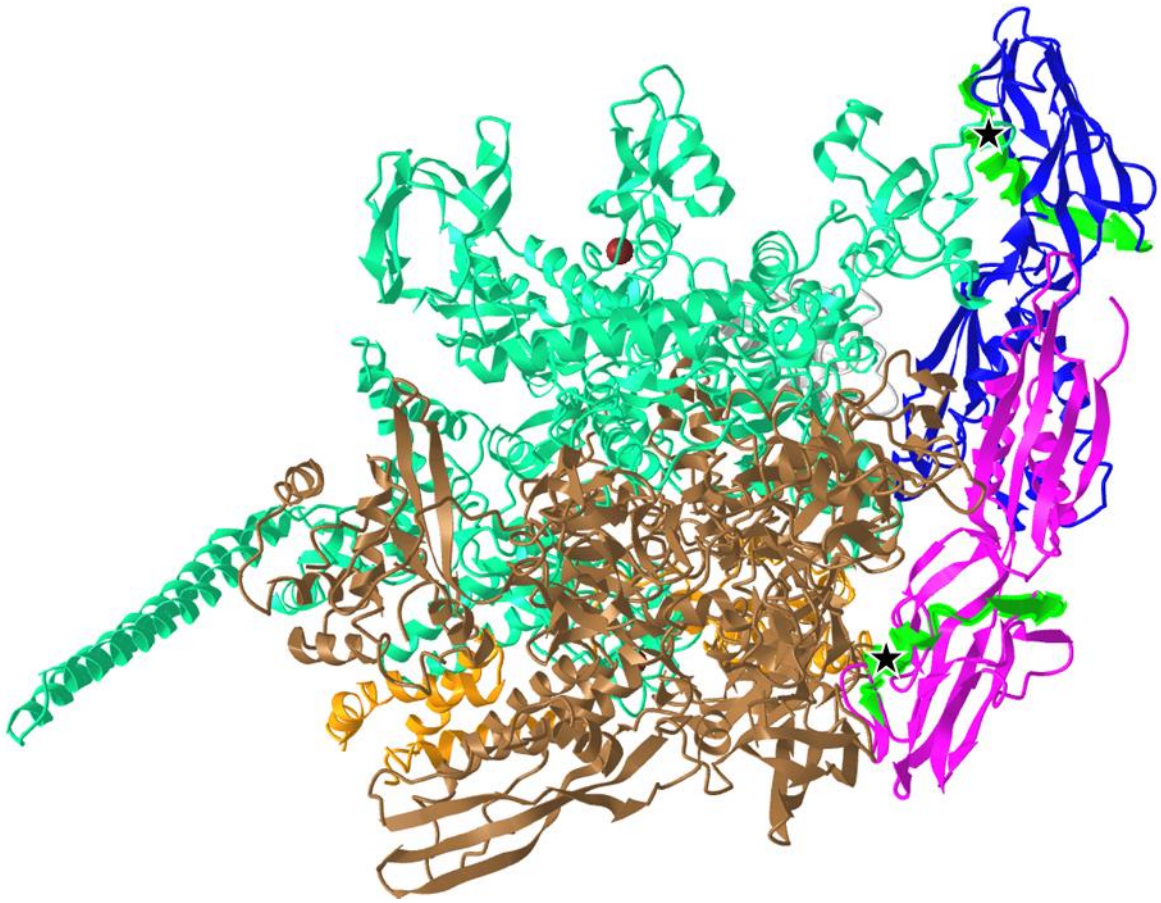

Figure 2-4: RNA polymerase complex 3D structure (PDB 5ZX3). Crystal structure generated by Li and Zhang (2019) by X-ray diffraction, representing the full RNAP complex in association with sigma factor H. Flat arrow shapes represent beta sheets, corkscrews represent alpha helices, and other parts are unstructured. Magenta and blue (right): RpoA chains A and B. Green backbone in RpoA: epitope 597585. Black stars at ends of green epitopes: residue 75 (E75D in Ravenel).

CHAPTER 4: MYCOBACTERIA BROADLY CONTAIN UNANNOTATED CRISPR/CAS SYSTEMS

Abstract:

Bacterial CRISPR/Cas systems target foreign genetic elements like phage, and regulate gene expression, even of the host by some pathogens. The system is a marker for evolutionary history, used for inferences in *Mycobacterium tuberculosis* for 30 years. However, knowledge about mycobacterial CRISPR/Cas systems remains limited. It is believed that Type III-A Cas systems are exclusive to *M. canettii* and the *M. tuberculosis* complex (MTBC) of organisms, and that very few of the >200 diverse species of non-tuberculous mycobacteria (NTM) possess any CRISPR/Cas system. This work sought unreported CRISPR/Cas loci across NTM to better understand mycobacterial evolution, particularly in species phylogenetically near the MTBC. Analysis of available mycobacterial genomes revealed Cas systems are widespread across *Mycobacteriaceae*, and that some species contain multiple types. Phylogeny of Cas loci shows scattered presence in many NTM, with variation even within-species, suggesting gains/losses of these loci occur frequently. Cas Type III-A systems were identified in pathogenic *Mycobacterium heckeshornense* and geological environmental isolate *Mycobacterium* SM1. In summary, mycobacterial CRISPR/Cas systems are numerous, Type III-A systems are unreliable as a marker for MTBC evolution, and mycobacterial horizontal gene transfer appears to be a frequent source of genetic variation.

Introduction:

The impacts of mycobacteria on human and animal health are difficult to overstate. Apart from the well-known effects by members of the *M. tuberculosis* complex (MTBC), *M. avium* complex members cause widespread losses to animal agriculture, numerous non-tuberculous mycobacteria cause disease in humans, mycobacterial biofilms cause difficulties for water treatment and hospital infection control, and some species are merely innocuous saprophytes. With such diversity, it is perhaps a surprise that their genomes share much in common. Between *M. smegmatis*, a non-pathogenic saprophyte, and *M. tuberculosis*, *M. avium*, *M. leprae*, and *M. abscessus* – four highly divergent pathogenic species in the *Mycobacterium* genus – there are over 1,000 genes with more than 50% amino acid identity[290]. While this is less than a quarter of genes in *M. tuberculosis*, *M. leprae* has undergone extensive genome reduction and has only ~1600 genes in total[138], meaning this conserved core gene set accounts for ~70% of the *leprae* genome. The closeness of the genetic background between mycobacteria has been discussed since before Illumina sequencing became prevalent and whole genome sequences were rare[40], with one argument for this closeness being an apparent dearth of horizontal gene transfer, at least as seen in MTBC organisms[47,83]. In the last 15 years, advances in mycobacterial genomics have revealed horizontal gene transfer occurring through unusual "distributive conjugative transfer" (DCT) mechanisms in *M. smegmatis*[47,48]. While the extent of this process' effects on mycobacterial diversity remains debated, it is now generally accepted that mycobacterial species undergo an atypical chromosomal genetic transfer that results in meiosis-like genome mosaics[11,47,50,55,291]. It is contested how much – if any – transfer occurs in the strict *M. tuberculosis* complex organisms, but DCT has been demonstrated to

occur at low frequencies in *Mycobacterium canettii*, the "smooth tubercule" species at the periphery of the MTBC and thought to be the most like the proposed *Mycobacterium prototuberculosis* ancestor of modern MTB[11,43,50]. The development of some mycobacteria into some of the most resilient and deadly human pathogens has followed an unclear evolutionary trajectory, and our understanding of genes and pathways in these organisms, how they function, and what roles they may or may not play in pathogenesis or host adaptation is limited. Determining relatedness of organisms has long been crucial in outbreak investigation, and in the 1990s, tuberculosis outbreaks began being differentiated through restriction fragment length polymorphism analysis of mycobacterial insertion sequences[292] and later through "spoligotyping," a technique that exploited unusual loci of spacers and repeats in the MTB genome[81,293]. Later, this pattern of repeats and spacers identified across different bacterial and archaeal genomes became the basis of a genetics revolution as the CRISPR/Cas system began to be understood and exploited for eukaryotic genome engineering[294,295]. While a tremendous amount of research has gone into the study of CRISPR/Cas systems across prokaryotes, study of the system in mycobacteria has remained fairly limited. It has been speculated that the MTBC CRISPR/Cas system may in fact be non-functional[44] and that CRISPR/Cas loci exist exclusively in the MTBC and *M. canettii*[27], but research by Wei *et al.* in 2018 and Grüschow *et al.* in 2019 reported that the Type III-A system found in the MTBC is unique from other Type III-A systems, actively expressed, and targets foreign genetic elements through an unusual cyclic hexa-adenylate signaling pathway instead of the expected nuclease-driven DNA cleavage[296,297]. More attention has since been paid to *M. canettii* and its diversity, including the multiple CRISPR/Cas system types contained across its strains – Type III-A in STB-A

and STB-D, Type I-C in STB-K, STB-L, and STB-J, type I-E in STB-G and STB-I, and a variant Type I-C in STB-E and STB-H[27]. The diversity in these systems and their presumed exclusive presence in these species have spurred hypotheses that *M. canettii* acquired these systems from other non-mycobacterial environmental species, and that an evolutionary history of the MTBC might be able to be derived through the tracing of this uptake[43]. In 2021, Singh *et al.* performed an exploration of 141 *Mycobacteriaceae* genomes to determine if CRISPR/Cas systems existed outside the *M. canettii*/MTBC cluster[43]. They reported confirmation that Type III-A systems were exclusive to *M. canettii*/MTBC group, but also identified the rare presence of alternative systems in a handful of other *Mycobacteriaceae.* In this work, searches for CRISPR/Cas loci outside the MTBC were performed, and have revealed a number of previously unannotated CRISPR/Cas systems across *Mycobacteriaceae.* Surprisingly, these systems show very little conservation, with variation in the presence and types of systems even between different isolates of the same species. The distribution of systems suggests that CRISPR/Cas loci regularly undergo exchange through DCT or other means of horizontal gene transfer. Further, with identification of a Type III-A Cas system in the environmental isolate *Mycobacterium* sp. *SM1*, and in a clinically relevant, pathogenic NTM, *M. heckeshornense*, this genetic locus' believed exclusivity to the MTBC clade is lost. Additionally, species proposed to represent evolutionary intermediates between environmental mycobacteria and professional pathogens – *M. riyadhense, M. shinjukuense,* and *M. canettii*[10,64] – have CRISPR/Cas systems distinct from those found in the *MTBC*, which may suggest that the clonal state of the MTBC is the reason for Type III-A complexes' persistence, and that such systems may have already been gained and lost in more mutable mycobacteria. This means non-essential marker loci are too unreliable to trace

evolution from within to outside the complex due to the comparative fluidity of NTM genomes. A lack of conservation between spacer sequences across the complexes and species analyzed support that these systems are actively incorporating new spacers and in diverse environments, and that they are not inactive relics of an ancestral *Mycobacterium*. Finally, this work expands on the recent findings by Singh *et al.* from the MTBC[43], including by analysis of proximal Cas gene clusters, but also provides a note of caution in inferring ancestry through the Type III-A system and indeed any Cas systems in *Mycobacterium*; the discovery of many seemingly complete and diverse CRISPR/Cas loci across mycobacteria does away with the notion of its scarcity; and frequent variation in isolates even within a species supports that mycobacterial horizontal gene transfer plays a large role in ongoing diversification of these organisms.

Materials and Methods:

To collect loci for initial analysis, a database search through CRISPROne and CRISPR-Cas++ was conducted[298,299]. These databases include analysis of >32,000 genomes for CRISPROne (2018 release), and >36,000 bacterial strains and >500 archaeal strains for CRISPR-Cas++ (2022 release). These were explored for putative hits in *Mycobacterium*. Most hits in both databases are for *M. tuberculosis* and its variants (e.g., 560 entries out of 786 for CRISPR-Cas++). Questionable matches in each database come with scoring metrics, with most "Evidence Level 1" hits through CRISPR-Cas++'s CRISPRCasFinder being false positives. Potential loci of interest were collected from both databases, though CRISPROne's inclusion of draft genome sequences yielded far more data. Searches were performed with default parameters except where otherwise noted. Secondary searches of these databases utilized input of contigs

or chromosomes downloaded from NCBI's repository to screen for homologues not included in the CRISPRone and CRISPRCasFinder pre-built databases[300].

To determine conservation or these sequences across *Mycobacterium* species and to assess any homology hits within the larger context of a locus necessary for actual CRISPR/Cas function, any putative hits in *Mycobacterium* were searched through TBLASTN and BLASTN of mycobacterial genomes, including the WGS database for *Mycobacterium* (taxid: 1763), and subsequently *Mycobacteriales*[260]. Expanded searches were also conducted through the Bacterial and Viral Bioinformatics Resource Center website (BV-BRC.org)[301]. Genomic data uploaded to BV-BRC is annotated through a different pipeline (RASTtk) than NCBI (PGAP), and both were queried for completeness[302,303].

After putative detection of two different classes of CRISPR/Cas complex in some members of the *M. kansasii* complex, the complexes were split and searched individually using a similar strategy. TBLASTN queries were performed using the MK13 Locus I-C Cas8c (a multirole protein in type I-C complexes) and Cas5 (Cas I-C), and the MK13 Locus I-G Cas1 (containing a characteristic Cas4 fusion in I-G complexes) sequence against the WGS dataset filtered by *Mycobacteriales* (taxid: 85007) with default parameters except increasing allowed return sequences to 500. Hits were selected from this list for investigation to determine the likelihood they were real positives and to type any putative CRISPR system(s) identified. Searches were also performed using PLfams models through BV-BRC to identify matches in other genomes, and assess genomic context of homologues in other genomes.

CRISPR repeat and spacer sequences were downloaded from CRISPRCasFinder. Because of intermittent availability of the CRISPR-Cas++ webserver's CRISPR Repeats and Spacers BLAST tool, the dataset was downloaded from the website and custom BLAST databases built on MSU'S HPCC to query spacers against for conservation using "makeblastdb" in the BLAST+ package (v2.7.1)[261]. Using these data, "blastn" searches were performed with default parameters in all cases.

To assess relatedness and divergence of complexes across *Mycobacterium*, a subset of identified Cas protein sequences from each annotated or putative complex were acquired from NCBI and BV-BRC, using CRISPRCasFinder, CRISPROne, and NCBI's ORF Finder to identify genes that had escaped annotation in several cases. The 49 sequences, concatenated end-to-end starting from upstream and proceeding towards Cas2 but not including the CRISPR region (Supplemental File S3-1), were aligned by MUSCLE (v3.8.31, option -diags, all other parameters default) which yielded an alignment (Supplemental File S3-2) of 3,943 positions (maximum sequence length = 2,984aa, including gaps introduced post-alignment)[304]. ModelTest-NG was used to calculate the best amino acid substitution model, which recommended VT+G4+F (Supplemental File S3-3)[305]. A maximum likelihood tree was then constructed using RAxML-NG (v1.0.1, with options --all, substitution model VT+G4+F, seed=212667, --bs-tree autoMRE to automatically determine bootstrap value convergence, 50 random starting trees, 50 parsimony starting trees)[201]. The output tree had bootstrap tree support values appended to nodes using raxml-ng --support with the output files from tree generation. The tree was visualized in FigTree v1.4.4, exported as an SVG, and labeled and colored in Inkscape[306,307]. Raw tree files in .nwk format are provided as Supplemental File S3-4.

For translation of open reading frames and predicted *cas* gene sequences by CRISPRCasFinder, Expasy's Translate tool was used[308]. For comparison of protein sequence percent identity, the SIM Alignment Tool was used[308].

Locus maps were created through BV-BRC's Compare Region Viewer. Exported SVGs were edited in Inkscape software to include only the CRISPR/Cas regions of interest unless other genes were a part of the locus (e.g., transposase insertions) and apply text labels. Locus maps are only representative and not to scale between loci, but each locus was visualized with BV-BRC Compare Region Viewer option set to 20,000nt.

Results:

"*cas3*"/TatC:

As previously published, very few hits were initially returned by either database query for *Mycobacterium* species outside the *M. tuberculosis* complex already known to host a system. One broadly conserved element with similarity to type-I systems' Cas3 protein was returned but found to be very similar to the non-Cas RNA helicase TatC and subsequently discounted as a false positive.

Orphan *M. avium* 104 locus:

M. avium strain 104 presented a high-confidence, orphaned CRISPR locus, which has been previously identified by others[43,44]. Direct repeat (DR) sequences and spacer sequences in this locus were separately searched by BLASTN against CRISPR-Cas++ datasets (custom BLAST databases built on direct_repeat/spacer_seqName.fsa and direct_repeat/spacer_taxon.fsa) but

only returned hits in *M. avium* 104[298]. A BLASTN search of a concatenated *M. avium* 104 repeat and spacer sequence (TGCTCCCCGCGTAAGCGGGGATGAACCGGTCGGTCACTGCGGTGGTGTCCTGTGCATGCTCC) in the WGS database filtered by *Mycobacteriales* does return a match for the full 13 repeat CRISPR locus in *M. avium* Chester/M. avium subsp. hominissuis str. ATCC 700898, an isolate ATCC identifies as coming from a human host in 1983[309]. This timeline corresponds to the isolation of the *M. avium* str. 104 in 1983 from a patient in California, USA[310], so this match may be spurious. Regardless, no *cas* loci are observed in either strain and this does appear to truly be orphaned as a CRISPR array. A BLASTN search of this *M. avium* 104 CRISPR sequence plus several thousand flanking bases against the WGS database for Mycobacteriales (taxid 85007) yielded hits with near-complete coverage only in *M. avium* subsp. *hominissuis* strains 101, ATCC 700898, and GM44. Other hits in *M. fortuitum* and *abscessus* returned only a fragment of this query (7%-~30%), and a partial alignment of *M. abscessus* subspecies *massiliense* strain 618 (contig FVWY01000006.1) appears to show a different 9 repeat orphan CRISPR locus per CRISPRone and CRISPRCasFinder with a partial hit to the locus in few *M. avium* strains. A comparison of the DR sequences and spacer sequences shows little shared homology and it is unlikely these hits are related.

M. innocens MK13 loci:

Two higher-confidence CRISPR loci and upstream arrays of *cas*-like genes were initially identified in the draft genome of *M. kansasii* complex member strain *M. innocens* MK13 by CRISPRone. The first complex, comprising 10,974bp of the 15,118bp contig UPHQ01000292.1, contained 7 *cas*-associated gene homologues in sequence, with a CRISPR locus immediately

downstream containing 46 predicted spacers and 47 repeats. As a draft sequence, this was not searchable directly through CRISPRCas++'s CRISPRCasFinder, so the contig UPHQ01000292.1 was downloaded from NCBI and uploaded for annotation separately. CRISPRCasFinder predicts 7 generic *cas*-associated genes and, alternatively, 4 *cas* Type IC-associated genes, along with the same major CRISPR locus of 46 spacers and 47 repeats, showing 0% spacer conservation and 92.61% repeat conservation (Figure 3-1). While functional labeling is putative without *in vitro* confirmation of activity, the locus will be referred to as locus "I-C" for convenience. The second complex, making up 10,580bp of 37,563bp contig UPHQ01000073.1, was reported by CRISPRone to contain 7 *cas*-associated gene homologues in sequence, with another CRISPR locus downstream containing 15 spacers and 14 repeats. CRISPRCasFinder reports this contig contains 5 *cas*-like genes, annotates 4 as "Type I-U," and separately annotates 1 gene immediately downstream as Type I-D. The nomenclature for Cas type I-U complexes – standing for "type I, Unknown" – has been updated to Type I-G after initial characterization in recent years[311]. As such, this complex will be referred to as locus or type "I-G" going forward. In type I-G systems, the *cas1* gene – containing a unique fusion of Cas1 and Cas4 – is a distinguishing feature[311,312], and this fusion is seen here (Figure 3-2). CRISPRCasFinder also reports 14 spacers and 15 repeats, with 0% spacer conservation and 93.13% repeat conservation. CRISPRCasFinder results are presented in Table 3-1, and CRISPROne results in Table 3-2. Tables 3-3 and 3-4 list the predicted functional roles of the proteins that make up the I-C and I-G loci, respectively.

M. innocens MK13 CRISPR Sequences for Discovery:

To explore whether related species have also acquired or maintained these CRISPR loci, a BLAST search of the more stringent NR database with DR sequences from M. *innocens* MK13

was performed but yielded no matches. A subsequent search of the WGS database filtered by organisms in the taxon *Mycobacterium* (taxid: 1763) does return hits (Table 3-5). Before proceeding, the *M. persicum* MK4 contig UPHM01000048.1 returned as containing a 100% hit to DR sequences was uploaded to CRISPR-Cas++ and searched by CRISPRCasFinder, returning a homologous set of *cas*-like genes and a 50 spacer CRISPR locus containing the DR identified by BLAST. Despite very high conservation of DR sequences between these two members of the *M. kansasii* complex, spacer sequences were unique for both. A BLAST search of the WGS database filtered by *Mycobacterium* (taxid: 1763) with the locus I-C DR consensus yielded 22 hits, and the same strategy with locus I-G returned 24 hits (Table 3-5). Some of these hits matched multiple contigs from the same sequencing run (e.g., table 3-5, *M. persicum* AFPC-000227 with 4 separate contigs). Hits for I-C were seen in *M. persicum* (10/15), *M. innocens* (4/15), and 1 from *M. attenuatum* (1/15). Expanding the search to *Actinobacteria* (taxid: 201174) only added one additional hit in *Mycobacterium* sp. SM1, an environmental isolate from a mud volcano in Italy[313]. Contigs for this organism were downloaded and searched by CRISPRCasFinder, which returned the aforementioned I-G locus, as well as an unexpected second locus on a separate contig. Other hits for I-G included *M. canettii* (8/20), *M. riyadhense* (4/20), *M. innocens* (4/20), *M. ostraviense* (2/20), and *M. gastri* (2/20). Expansion of this DR to *Actinobacteria* yielded no additional hits. The secondary locus identified on *Mycobacterium* sp. SM1 contigs was annotated by CRISPROne and CRISPRCasFinder as a type III-A CRISPR/Cas system.

A BLASTN megablast search of *M. canettii* (taxid: 78331) with low-complexity filter turned off of the 7 DR sequences from locus I-C yielded modest-confidence hits in *M. canettii* strain STB-K on multiple contigs for 5/7 DR sequences, and the contig with 54 matches proximal

to each other (JAHVHL010000002.1) was run through CRISPRCasFinder along with the other two major contigs, identifying a homologous I-C CRISPR/Cas region with 53 spacers. However, in this genome, a second CRISPR locus of 28 spacers with 99.4% DR conservation was located ~240,000bp away from the nearest predicted *cas* genes, directly at the end of the contig (coordinates: 970,599-972,654, contig length 972,672). The repeat consensus was unique from those located proximal to the *cas* locus, and an identical match to this alternate DR sequence was found at the start of another contig in the assembly (NZ_JAHVHL010000003.1, coordinates: 27-851bp). Searching the contigs again with the CRISPRCasFinder option "Unordered" ticked to allow for *cas* gene hits that do not include a complete ordered locus identified genes starting at 1023bp (untyped cas2), a type ID *cas1*, and continuing through a "type IU" locus of 4 genes. A literature search indicated a resequencing effort by Blouin *et al.* in 2014 identified the presence of two systems in STB-K, though they classified one as Type I-C and one as a Type I-C variant as type I-G systems had not been characterized at the time of publication[130]. With evidence of a variety of existing CRISPR/Cas systems across the genus, the search expanded to incorporate Cas protein sequences for system discovery rather than relying on the identification of CRISPR repeats, which can be difficult to sequence and assemble and may not be reliably identified.

M. innocens MK13 Cas Sequences for Discovery:

Using BV-BRC's PLfams browser to search the motif of the Cas2-like gene in *M. innocens* MK13, and searching genus-specific families for hits in *Mycobacterium*, two reference or representative quality strains have loci showing structural homology to the putative MK13 CRISPR/Cas locus: *Mycobacterium gastri* DSM 43505, and *Mycobacterium ostraviense* FDA-ARGOS_1613, two species recently reclassified into the *M. kansasii* complex of organisms[29].

Searching the *M. ostraviense* FDA-ARGOS_1613 assembly with CRISPRCasFinder returns a 127 spacer CRISPR array with 97%+ DR conservation, but CRISPRCasFinder does not identify the adjacent conserved locus as a Cas cluster like it does for *M. kansasii* MK13. CRISPROne does identify both the CRISPR array and the 5 gene Cas cluster directly adjacent.

A TBLASTN of the MK13 Type I-G Cas1 protein returned hits in two broad groups – a high similarity group (75%+ AA identity), and a partial similarity group (~26-40% AA identity). High similarity homologues were identified in *Mycobacterium kansasii* complex strains *gastri*, *ostraviense*, and *innocens*, as well as in *Mycobacterium canettii* and *Mycobacterium riyadhense*. Partial similarity homologues were identified again in the aforementioned *Mycobacterium kansasii* complex strains, and a wide group of genera outside *Mycobacterium* (*Corynebacterium, Nocardia, Rhodococcus*) and in *Mycobacterium tuberculosis* variant *tuberculosis*. High similarity homologues were noted as showing 99%+ query coverage against Type I-G Cas1, and partial similarity groups showed roughly 50-70% query coverage. Several hits fall outside these categories, such as 99% coverage/47-60% identity *Mycolicibacterium hassiacum* and *Prescottella subtropica, Gordonia paraffinivorans*, and others. Of this set, *Mycobacterium heckeshornense* stood out as featuring two sets of hits, one with 68% coverage, and one with 31% coverage. For *Mycobacterium heckeshornense* strain DSM 44428, contigs JACKTA010000047.1 and JACKTA010000078.1 were downloaded from NCBI in FASTA format and searched on CrisprCas++, and although contig 47 is short (4,529bp), it was predicted to contain a 36 spacer CRISPR locus, *cas2*, and a partial *cas1*. Contig 78 (86,845bp) also contains a predicted complex close to the end of the locus, but contains some predicted *cas* genes of Type I-U by CRISPRCasFinder. CRISPROne finds the same system for contig 47, but does not find the

contig 78 complex credible enough to call. The second *M. heckeshornense* WGS dataset (strain RLE) has the same two hits, and interestingly, the two contigs bearing these hits are nearly the same lengths (85,024 bp and 4,395bp, respectively). Unfortunately, SRA data for the two *M. heckeshornense* assemblies are unavailable publicly and reconstruction of these possible loci is not possible at this time. However, the complete genome of *M. heckeshornense* JCM 15655 was searched next and a locus – disrupted by an transposase and with the structure of [*csb1-csb2-cas3*-hypothetical protein-<IS1380>-*cas1-cas2*] – was extracted for sequence comparison. A second locus was not evident in this genome. An additional genome – *M. heckeshornense* JMUB5695 – was also identified to contain a high-confidence Cas Type III-A locus, with significant similarity to the complex inside the MTBC and some *M. canettii* strains (Figure 3-3). The same genome also contains a second locus reminiscent of the order seen in *M. heckeshornense* JCM 15655 and *Mycobacterium marinum* VIMS9, as well as one in *M. shinjukuense* JCM 14233, though still containing an unannotated hypothetical protein in-frame.

Investigation of *M. hassiacum* returned two hits in CRISPRCasFinder. This genome was reported to contain a CRISPR array by Singh *et al*.[43], but both *cas* loci presented with an atypical structure of *cas3 – csb2 – csb1 – hypothetical protein – cas1 – cas2*. This hypothetical protein contains no detected conserved domains and fits neatly between *csb1* and *cas1* in both loci. It is not clear whether this system is active or not, but both are included in the sequence alignment (Supplemental File S3-1). The first *cas* locus contains 44 CRISPR spacers downstream, and the other has a severely disrupted CRISPR locus punctuated with four repeats of IS6120 transposases between short CRISPR repeats before an intact 19 spacer array. This same atypical *cas* locus structure with a protein of unknown function in the operon was found in *M.*

longobardus DSM 45394 as well, with a 54 spacer CRISPR locus in place downstream, which may represent a novel Cas-like gene.

M. austroafricanum DSM 44191 presented another interesting arrangement. This genome too was reported to contain a CRISPR array[43], and appears to have both an atypical Cas locus and a homologue in *Mycobacterium* sp. *D3*. A full repertoire of 5 *cse* is found upstream of the CRISPR array, and *cas1* and *cas2* are instead found immediately downstream (Figure 3-4).

Some *M. xenopi* strains may have two systems, but one with a Cas1/4 fusion homologue is split across multiple contigs. The complete system (Cas Type I-E) is upstream of a reported CRISPR repeat region[43], but the putative type I-G system is not recoverable and cannot be reconstructed from the available sequence.

Mycobacterium vanbaalenii strain JOB5 does not appear to have a chromosomal Cas system, but returns an unusual locus from CRISPROne and CRISPRCasFinder (latter with search option "Unordered"), with homologues of *csf1, csf4, csf2, and csf3* appearing 10kb upstream of a small, low-evidence 3 spacer CRISPR locus. Such a finding may very well be noise, but does have similarities to wildly divergent, plasmid-borne Type 4 Cas systems[314]. A TBLASTN search of the NR database for the four concatenated genes returns 95 hits, nearly all of which are on plasmids. Representative sequences were taken from a *Mycobacterium fluoranthenivorans* 2A plasmid (CP059893.1), *Mycobacterium* sp. *YC-RL4* plasmid pMYC1 (CP015597.1), and the chromosomal sequence of *Mycobacterium rhodesiae* NBB3. Although confidence in these sequences making up a complete Type IV system is low, because these systems are known to be extremely divergent even within subtypes and due to the difficulty in identifying them due

to the lack of reliable signature genes, they are still being included with the appropriate caution towards further interpreting the results without *in vitro* confirmation of any function.

M. riyadhense MR-246 appeared both in the Locus IG *cas1* TBLASTN search, as well as the search using the MK13 direct repeat sequence, with both hits independently finding the same region in the chromosome (CAJMWP010000001.1). CRISPRCasFinder and CRISPRone independently identify a 111 spacer CRISPR array with 97% DR conservation and 0% spacer conservation, though only CRISPRone identifies upstream *cas* genes. The subtype in this chromosome is reported as I-U (I-G), but the locus is missing a *cas2* gene call. When re-running CRISPRCasFinder and ticking the "Unordered" option to allow for hits outside a longer locus, it too returned the same hits as CRISPROne, but again missing a *cas2* gene call, which presumably led this software to initially discard the other 5 genes upstream of the CRISPR array. NCBI's ORF Finder output from the full locus (4,227,944-4,244,859, 16,916bp) identified a 95aa open reading frame directly downstream of *cas1* and upstream of the CRISPR array, and a BLASTP search of this protein against the UniProtKB database identified 17 moderate confidence alignments (E value range = 6^{-11} to 0.003) against Cas2 proteins. No other hits were reported. NCBI's CDD Search reports a hit on a Cas2 domain in this sequence (from position 5-70, E-value = 5.54^{-17}). Given the divergence of the *cas2*-like ORF in *M. riyadhense* led to CRISPRCasFinder missing the entire locus, and CRISPROne identified 5/6 genes plus the CRISPR array but did not flag the *cas2* homologue, the entire nucleotide sequence of the locus from ORF Finder was also used for an unfiltered BLASTN search of the nt database, and a broad scan of the WGS database filtered by hits in Actinomycetota. The NR database returned a hit with 16,914 of coverage and only a single non-identical nucleotide from *M. riyadhense* strain NTM, as well as high

confidence hits (E value = 0.0) in *M. ostraviense* str. FDAARGOS 1613, and *M. canettii* strains CIPT 140070010, 140070008, and 140070017.

Mycobacterium lacus JCM 15657 presented with a striking 146 repeat CRISPR locus, and subsequent investigation showed unannotated upstream *cas2* and *cas1* genes, but the rest of the locus appeared unusual with GenBank annotation noting frameshifts present. A BLASTP search of the sequence upstream of Cas1 hit numerous IS3 family transposases scattered throughout *Mycobacterium lacus, kansasii,* and *persicum* strains, as well as in other genera. It appears this may have been a functional system broken by the insertion of this transposase in the past, and a scan of similar disruptions finds an IS3 family transposase where Cas1 and Cas2 would be found in the CRISPR/Cas locus of *M. tuberculosis* strain 36918. This disruption appears widespread across Beijing-like strains (as can be seen in the lineage 2 isolate in Figure 3-3), with a BLASTN search of the locus returning dozens of nearly identical sequences with 100% query coverage containing the disrupted region from other MTB Beijing-like isolates.

The phylogenetic tree of 49 putative concatenated Cas protein complexes generated robust branches, both for different *Cas* complex types and for divergence of these complexes among *Mycobacterium* species (Figure 3-5). Bootstrap analysis reached statistical convergence after 500 replicates. Complexes of similar types grouped together as expected.

Discussion:

Despite longstanding knowledge that CRISPR/Cas systems are present in *Mycobacterium tuberculosis*, a species with very limited horizontal gene transfer compared to other bacteria[315], investigation into the prevalence of these systems across other Mycobacteria has been limited.

In recent years, some research has been performed, such as the work by Singh *et al.* in exploring the extent of CRISPR/Cas systems outside the MTBC, but their reasonably cautious focus on completed genomes and seeking CRISPR arrays rather than complete CRISPR/Cas gene sets limited the scope of potential discovery[43]. Using available software like CRISPROne and CRISPRCasFinder in combination with a more fine-toothed approach of manually investigating many draft genomes and checking for ORFs missed by annotation packages has revealed a substantial and previously unreported spread of multiple Cas types across *Mycobacterium*, spanning the range from free-living organisms to pathogens.

First, any *in silico* predictions must be taken with caution. These data show high-confidence homologues to CRISPR/Cas systems exist outside the MTBC, but they do not prove this without a functional validation. Despite CRISPROne and CrisprCas++ algorithms being designed to minimize false positives, the widespread annotation of a TatC homologue as an orphan Cas3 protein in *Mycobacterium* is a reminder that all output requires careful interpretation.

<u>The *Mycobacterium avium* complex does not contain Cas loci:</u>

The CRISPR locus identified in *M. avium* strain 104 was not returned in any other *M. avium* complex species in the NCBI NR database. As seen in large sequence polymorphisms that contain unique genetic elements not observed in other mycobacteria, like those involved in metal acquisition and regulation in *M. avium* subspecies *paratuberculosis*[316,317], members in the MAC do have the potential to take up foreign genetic elements. Likewise, recent research has begun to elucidate the frequency and importance of HGT in *Mycobacterium* despite its limits in

some species[315]. The identification of homologous sequence containing this orphan CRISPR in other closely related (or possibly identical, in the case of *M. avium* ATCC 700898) *M. avium* isolates at least minimizes the likelihood that it is a sequencing error, but the closest match outside these immediate relative *M. avium* isolates appears to be a different orphan CRISPR in *M. abscessus* subspecies *massiliense*. Regardless, the exact origin of these orphan CRISPR loci remains unknown and was not able to be traced with DR or spacer sequences. No further evidence of any Cas or Cas-like genes was identified in any *M. avium* complex species, which looks to be devoid of these loci with the exception of this orphaned CRISPR element in a single strain.

Other phylogenetically diverse mycobacterial taxa contain numerous, diverse CRISPR/Cas systems:

This work demonstrates the prevalence and diversity of CRISPR/Cas systems in the clinically important *Mycobacterium kansasii* complex of organisms. *M. innocens* MK13's locus I-C showed homologues within the *M. kansasii* complex, with most concentrated in *M. persicum*, a recently characterized opportunistic pathogen in the complex originally isolated from sputum in multiple cases of human disease in Iran[151]. Other hits were in species not considered pathogenic – *M. innocens* and *M. attenuatum*[51]. With the phylogenetic comparison of sequences, an incomplete locus in *M. malmoense* – a frequent NTM infection and one of global clinical relevance [318] – grouped near this cluster (Figure 3-5).

In contrast to this grouping within the MKAN complex, MK13's locus I-G returned a surprising number of hits in pathogenic mycobacteria outside the complex – 8 in *M. canettii*,

the smooth tubercule species at the edge of the *Mycobacterium tuberculosis* complex; and 4 in *M. riyadhense*, a human NTM pathogen first identified in Saudi Arabia[319]. The other hits were in *M. innocens*, *M. gastri*, and *M. ostraviense*, and *Mycobacterium* sp. *SM1*, largely non-pathogenic or opportunistic pathogen organisms[28,51], and an environmental isolate in the case of *SM1*[320]. *M. innocens*, it seems, possesses two potential CRISPR/Cas systems that bifurcate either within the MKAN complex, or outside of it.

Existing literature can tell us much about the composition of complexes identified here. Cas5 is involved in pre-crRNA processing into crRNA units[321–323]. In Cas type I systems, Cas6 typically functions with Cas5 and Cas7 in crRNA processing, but Type I-C systems are unique in that Cas6 is absent, Cas5 functions on its own, and Cas8c is a signature of these types of loci[321–324], streamlining the Cascade complex and allowing full functionality with only Cas5, Cas7c, and Cas8c, all of which are intact in MK13 locus I-C. This suggests that, at a minimum, *M. innocens* MK13 possesses intact ability to process pre-crRNA into its target-binding form for a Cas complex. Similarly, type I-G systems have a fusion of Cas1 and Cas4[325], which is also observed in MK13's locus I-G *cas1* gene, showing an N-terminal Cas4 domain and a C-terminal Cas1 domain per NCBI's CDD search. This provides a complete complex of Cas3, Cas8, Cas7, Csb2, Cas1 (with the Cas4 fusion), and Cas2.

The *M. riyadhense* MR-246 CRISPR/Cas locus identified by both CRISPR repeat sequence and type I-G Cas1 BLAST searches is an interesting example where the stringency of software to eliminate false positives is likely obscuring identification of new features. The comparatively low percent identity values of this ORF (28%-49%) combined with the short sequence length of Cas2 proteins[326] likely contribute to both the relatively modest E values and the exclusion of this

divergent Cas2 protein from identification by both software tools used. The nearly identical loci (1 mismatch out of 16,914bp) between *M. riyadhense* MR-246 and *M. riyadhense* NTM raise questions, and BioSample records (SAMEA7003857, SAMN12495011) indicate the samples were taken in 2016 and 2018, respectively, from human disease cases in Saudi Arabia. An additional 2,000nt on each side of the *M. riyadhense* NTM locus aligned to MR-246 was extracted and a BLAST search run against other *riyadhense* isolates to assess how much variation occurs between other clinical isolates, as well as to assess by CRISPROne if additional CRISPR spacers had been acquired in the newer 2018 NTM isolate. The same number of spacers were identified, and differing numbers were seen in other *riyadhense* isolates. This tight similarity may represent that *M. riyadhense* isolates MR-246 and NTM originated from a common point of exposure, but this is beyond the scope of this investigation. *M. riyadhense* has been proposed as being an intermediate of sorts between environmental and professionally pathogenic mycobacteria in the MTBC[64]. It is tempting to speculate about following the Type I-G breadcrumbs down evolutionary path from environmental organisms like *M. innocens*, through *riyadhense,* towards the *prototuberculosis*-analogue of *canettii* and finally into *M. tuberculosis*, and it has been reported that the Type I-G-bearing *M. canettii* STB-K is one of the most distant from other *canettii* members[27]. That said, a Type I-C complex was observed in STB-K and not in the MTBC nor in other proposed evolutionary stepping stone species like *M. shinjukuense* or *M. riyadhense*[10,64], so tracing any such path is complicated.

<u>The Type III-A Cas system is not exclusive to the M. tuberculosis complex:</u>

In searching homologues for type I-G complexes, the curious case of *Mycobacterium* sp. *SM1* arose, which was unexpectedly found to contain a disrupted form of the Cas Type III-A

system previously reported to exist exclusively in the MTBC and some isolates of *M. canettii*[43]. An analysis of the operon arrangement suggests that SM1's locus is missing *csm3* and *csm2*, and its *csm1* gene is truncated, with the area disrupted by an IS3 family-like transposase. TBLASTN searching this transposable element returns a long range of matches to *M. lacus*, *M. riyadhense*, and numerous *M. tuberculosis* complex strains. In MTB H37Rv, its homologue is annotated as a possible IS1557 transposase [mycobrowser.epfl.ch/genes/Rv1313c]. Little is currently known about *M.* sp. *SM1*, but it is annotated as having been discovered in an Italian mud volcano[313], and an abstract from the International Meeting of the Microbiological Society of Korea in October of 2022 by Awala *et al*. describe the species as an "extremely acidophilic" bacterium, growing in an optimal pH of 2-4[320]. How this organism or its ancestors could have either received or donated this Cas system to other mycobacteria is another fascinating open question, but one beyond the scope of this work. However, in exploring the homology of this complex to other Type III-A systems in *Mycobacterium,* another homologous but intact Cas Type III-A locus was found in *M. heckeshornense* JMUB5695, with an arrangement nearly identical to that observed in the MTBC except with spacer variation in its two downstream CRISPR arrays. This organism is known to cause severe pulmonary disease in immunocompetent adults worldwide[166,167,327].

<u>CRISPROne reports homology of conserved, plasmid-borne genes in some mycobacteria and unusual Type IV Cas systems:</u>

Lastly, and with a caveat of caution, it is possible that some mycobacteria possess Type IV Cas systems. The abundance of four clustered genes highlighted by CRISPROne as potential homologues to Type IV proteins across mycobacterial plasmid sequences from many different

species offer that mycobacterial plasmids may themselves be involved in targeting other plasmids[314]. However, these systems are notoriously difficult to predict, and while this merits further follow-up, it is not sufficient to definitively assign these loci as Type IV Cas systems purely by computational analysis.

Conclusions:

In summary, this work has uncovered numerous mycobacterial CRISPR/Cas systems. Rather than being an oddity or an indicator of the genetic rigidity of the *Mycobacterium tuberculosis* complex, these systems are frequent across environmental species, opportunistic pathogens, and professional pathogens. Stranger still is the absence of such systems in genetically close organisms; in the *M. kansasii* complex, *M. kansasii* and *M. pseudokansasii* were not observed to contain *cas* genes, despite the fact that the former has even been reported to be a DCT recipient from both *M. persicum* and *M. attenuatum*, CRISPR/Cas bearing species within the complex[51] – or even the variation in the presence or type of systems within the same species, as was previously observed in *M. canettii* species but has now been expanded to across *Mycobacterium.* The presence or absence of one or more systems in one organism seems to have little bearing on whether all members of that species or its closest relatives will have the same loci. However, one pattern did emerge – some broad groups of mycobacteria simply do not appear to have CRISPR/Cas systems. Barring an orphan CRISPR locus in *M. avium* 104, also called *M. avium* ssp. *hominissuis* (MAH), no evidence of any Cas proteins, other CRISPR arrays, or even variation in the MAH CRISPR locus was found across any member of the *Mycobacterium avium* complex. Through BV-BRC, a search of 493 *M. avium* genomes for Cas domains returned only a single partial hit for *Cas2* on a 521bp contig from a study

characterizing hypervariable genomic islands from German MAH strains[328]. While it is now understood that many mycobacteria undergo horizontal gene transfer by DCT, this form of exchange remains poorly understood. Because systems exist sporadically within even tight phyletic groups, it may be that CRISPR/Cas systems are not ancestral to mycobacteria, at least in recent history. Instead, horizontal gene transfer could be a primary mechanism through which CRISPR/Cas systems are acquired in mycobacteria. It is known that plasmids carry CRISPR/Cas systems[329], and that both the systems on plasmids as well as genomic systems often target mobile genetic elements like plasmids[314,330]. The rarity of mycobacterial plasmids has been noted[48], and plasmids that do transfer between species are also compartmentalized, as reported by Ummels *et al*., who reported the discovery of a conjugative plasmid capable of exchange between slow-growing mycobacteria, but this unique plasmid was unable to be transmitted to any fast-growing mycobacteria[331]. While plasmids are still believed to be important and do play key roles in mycobacteria[332], the mechanism of DCT, where chromosomal DNA is transferred between mycobacterial cells, appears to be the predominant means of horizontal gene transfer[48]. This unusual meiotic-like means of genetic exchange, the high frequency of transposable elements across mycobacteria, and a low frequency of other mechanisms of gene transfer could explain the inconsistent but still widespread prevalence of a variety of CRISPR/Cas systems across mycobacterium. Additionally, both the disruption of the Cas system in hypervirulent *M. tuberculosis* lineage 2 isolates, and the numerous instances of mobile genetic elements having already disrupted newly discovered loci herein, CRISPR/Cas systems do not appear essential to most mycobacterial lifestyles.

While their historical use in tracing clonality and evolution of *M. tuberculosis* preceded even an understanding of their function, it would seem CRISPR/Cas systems are not suited to direct evolutionary extrapolations. Instead, their presence and absence appears – like mycobacteria themselves – enigmatic, implying both a frequent flow of genetic material in and out of most mycobacteria, and that the NTM pangenome may substantially expand as more isolates are sequenced. Future work should confirm functionality of these newly reported systems, including whether several phylogenetically diverse systems including hypothetical proteins may represent novel Cas types or functions.

Tables:

Coordinates	CRISPR/Cas Type	Number of Elements	Genes	Orientation
UPHQ01000292.1 3713-11095	CAS (generic)	7	*cas2*, *cas4*, *cas3*, *cas3*, *cas5c*, *cas7c*, *cas8c*	Forward
UPHQ01000292.1 5878-10803 (Overlapping)	CAS Type I-C	4	*cas1*, *cas5c*, *cas7c*, *cas8c*	Forward
UPHQ01000292.1 11226-14630	CRISPR Locus	46 spacers	n/a	n/a
UPHQ01000073.1 15821-22597	CAS Type I-U	4	*cas3*, *csb1*, *csb2*, *csx17*	Forward
UPHQ01000073.1 22601-24241	CAS Type I-D	1	*cas1*	Forward
UPHQ01000073.1 24696-25749	CRISPR Locus	14 spacers	n/a	n/a

Table 3-1: CRISPRCasFinder output for *M. innocens* MK13 contigs UPHQ01000292.1 and UPHQ01000073.1.

Coordinates	CRISPR/Cas Type	Number of Elements	Genes	Orientation
UPHQ01000292.1 3656-11095	CAS Type I/I-C	7	*cas3*, *cas5*, *cas8c*, *cas7b*, *cas4*, *cas1*, *cas2*	Forward
UPHQ01000292.1 11226-14630	CRISPR Locus	46 spacers	n/a	n/a
UPHQ01000073.1 15169-24525	CAS Type I/I-U	7	*csm3gr7*, *cas3*, *cas8u1*, *cas7*, *csb2gr5*, *cas1*, *cas2*	Forward
UPHQ01000073.1 24696-25749	CRISPR Locus	14 spacers	n/a	n/a

Table 3-2: CRISPROne output for *M. innocens* MK13 contigs UPHQ01000292.1 and UPHQ01000073.1.

Coordinates	Annotation	Activity	Association
UPHQ01000292.1 3656-5881	Cas3 HD	Helicase activity[323] Endonuclease activity[333]	CAS Complex
UPHQ01000292.1 5878-6516	Cas5	crRNA Processing[321–323]	Cascade Complex
UPHQ01000292.1 6516-8264	Cas8c	crRNA Processing[321] CAS Recruitment[321]	Cascade Complex
UPHQ01000292.1 8266-9171	Cas7c	crRNA Processing[321]	Cascade Complex
UPHQ01000292.1 9164-9778	Cas4	Exonuclease activity[334]	Spacer Acquisition Complex
UPHQ01000292.1 9769-10803	Cas1	Endonuclease activity[323] Spacer integration[323]	Spacer Acquisition Complex
UPHQ01000292.1 10805-11095	Cas2	Structural[323] Spacer integration[323]	Spacer Acquisition Complex

Table 3-3: Components of *M. innocens* MK13's Locus I-C Cas complex. The makeup of this complex classifies it as a complete member of the CAS Type I-C group.

Coordinates	Annotation	Functional Role	Association
UPHQ01000073.1 15169-15681	Csm3Gr7	crRNA Processing[323,335] (Cas7-like)	Type III, Interference Complex
UPHQ01000073.1 16001-18160	Cas3 HD	Helicase activity[323] Endonuclease activity[333]	Type I, CAS Complex
UPHQ01000073.1 18160-20256	Cas8u/ Cas8g*	crRNA Processing[323,325] CAS Recruitment[325]	Type I, Cascade Complex Type I, Interaction between Cascade and CAS Complexes
UPHQ01000073.1 20253-21212	Cas7	crRNA Processing[325]	Type I, Cascade Complex
UPHQ01000073.1 21215-22597	Csb2gr5	crRNA Processing by Cas5/Cas6 Fusion Activity[325]	Type I, Cascade Complex

Table 3-4: Components of *M. innocens* MK13's Locus I-G CAS complex. In 2019, the CAS Type I-U (for "unknown") complex was renamed to Type I-G after it was found that Cas1 in this class is a fusion of Cas4 and Cas1 domains, but the exact roles of proteins in this complex remain poorly understood. The makeup of this complex classifies it as a complete member of the CAS Type I-G group, but Csm3 does not play a known role in this system and its presence here is atypical. CRISPRCasFinder does not predict this ORF, and subsequent searches suggest this may be a false positive.

Table 3-4 (cont'd)

UPHQ01000073.1 22601-24241	Cas1*	Endonuclease activity[323] Spacer integration[323] Cas4 Nuclease Activity[311]	Type I, Spacer Acquisition Complex
UPHQ01000073.1 24238-24525	Cas2	Structural[323] Spacer integration[323]	Type I, Spacer Acquisition Complex

DR Origin	**Strain**	**Query Coverage (Identity)**	**Contig Accession Number**
MK13 locus I-C	*M. innocens* MK13 (source)	100% (100%)	UPHQ01000292.1
MK13 locus I-C	*M. persicum* MK4	100% (100%)	UPHM01000048.1
MK13 locus I-C	*M. persicum* MK42	100% (100%)	UPHL01000057.1
MK13 locus I-C	*M. persicum* MK15	100% (100%)	UPHK01000053.1
MK13 locus I-C	*M. innocens* 49-11	100% (100%)	NKRC01000001.1
MK13 locus I-C	*M. persicum* 12MK	100% (100%)	MWQA01000001.1
MK13 locus I-C	*M. persicum* 7MK	100% (100%)	MWKZ01000001.1 MWKZ01000001.1
MK13 locus I-C	*M. persicum* 3MK	100% (100%)	MWKX01000001.1
MK13 locus I-C	*M. persicum* 8MK	100% (100%)	MWKV01000001.1
MK13 locus I-C	*M. persicum* AFPC-000227	100% (100%)	MVIF01000358.1 MVIF01000183.1 MVIF01000100.1 MVIF01000083.1
MK13 locus I-C	*M. persicum* 1010001469	100% (100%)	LWCM01000086.1
MK13 locus I-C	*M. innocens* 1010001493	100% (100%)	LWCK01000074.1
MK13 locus I-C	*M. innocens* 1010001454	100% (100%)	LWCH01000340.1
MK13 locus I-C	*M. persicum* CSURQ1465	100% (100%)	CADEAW010000279.1 CADEAW010000263.1 CADEAW010000143.1 CADEAW010000058.1
MK13 locus I-C	*M. attenuatum* MK41	100% (97.30%)	UPHT01000066.1
MK13 locus I-G	*M. innocens* MK13 (source)	100% (100%)	UPHQ01000073.1
MK13 locus I-G	*M. ostraviense* 241/15	100% (100%)	NKRE01000001.1
MK13 locus I-G	*M. innocens* 49/11	100% (100%)	NKRC01000001.1
MK13 locus I-G	*M. innocens* 1010001493	100% (100%)	LWCK01000025.1

Table 3-5: BLASTN results for *M. innocens* MK13 direct repeat (DR) consensus sequences from locus on UPHQ01000292.1 ("locus I-C") and locus on UPHQ01000073.1 ("locus I-G") identified by CRISPRCasFinder searched against WGS database filtered to include species in *MycobacterI-Gm* (taxid: 1763).

Table 3-5 (cont'd)

MK13 locus I-G	*M. ostraviense* 1010001458	100% (100%)	LWCI01000115.1
MK13 locus I-G	*M. innocens* 1010001454	100% (100%)	LWCH01000288.1
MK13 locus I-G	*M. gastri* DSM 43505	100% (100%)	LQOX01000131.1
MK13 locus I-G	*M. riyadhense* MR-246	100% (100%)	CAJMWP010000001.1
MK13 locus I-G	*M. riyadhense* MR-244	100% (100%)	CAJMWO010000001.1
MK13 locus I-G	*M. riyadhense* MR-206	100% (100%)	CAJMWK010000001.1
MK13 locus I-G	*M. riyadhense* MR-1023	100% (100%)	CAJMWI010000001.1
MK13 locus I-G	*M. gastri* 'Wayne'	100% (100%) 100% (100%) 100% (100%)	AZYN01000299.1 AZYN01000281.1 AZYN01000120.1
MK13 locus I-G	*M. canettii* STB-K	94% (97.06%) 94% (97.06%)	JAHVHL010000003.1 JAHVHL010000002.1
MK13 locus I-G	*M. canettii* NLA000701671	94% (97.06%)	JACTAP010000068.1
MK13 locus I-G	*M. canettii* CPIT 140070013 (2013)	94% (97.06%)	CAON01000366.1
MK13 locus I-G	*M. canettii* CPIT 140070002	94% (97.06%)	CAOL01000412.1
MK13 locus I-G	*M. canettii* CPIT 140070013 (2022)	94% (97.06%)	CAMJXS010000071.1
MK13 locus I-G	*M. canettii* Percy1101	94% (97.06%)	CAKKKT010000093.1
MK13 locus I-G	*M. canettii* Percy258	94% (97.06%)	CAKKKM010000107.1
MK13 locus I-G	*M. canettii* Percy525	94% (97.06%)	CAKKKA010000091.1
MK13 locus I-G	*M. canettii* CIPT 140070002	94% (97.06%)	CAJJDU010000059.1

Figures:

Figure 3-1: Locus organization of *Mycobacterium innocens* strain MK13 Cas Type I-C system and CRISPR region. The first seemingly complete complex identified in this work in the *M. kansasii* complex was identified in *Mycobacterium innocens* strain MK13, and includes the full expected repertoire of Cas proteins in the usual arrangement for Cas Type I-C systems. Coloring of CRISPR region indicates direct repeat conservation, with each color representing a unique repeat. Visualization from BV-BRC Compare Region Viewer and modified in Inkscape.

Figure 3-2: Locus organization of *Mycobacterium innocens* strain MK13 Cas Type "I-U"/I-G system and CRISPR region. A second seemingly complete CRISPR/Cas locus was identified in MK13, but of Cas Type I-G. This locus contains all expected proteins in the usual arrangement for a functional Type I-G system, including the fusion of the Cas4 domain to the N-terminal region of Cas1. Coloring of CRISPR region indicates direct repeat conservation, with each color representing a unique repeat. Visualization from BV-BRC Compare Region Viewer and modified in Inkscape.

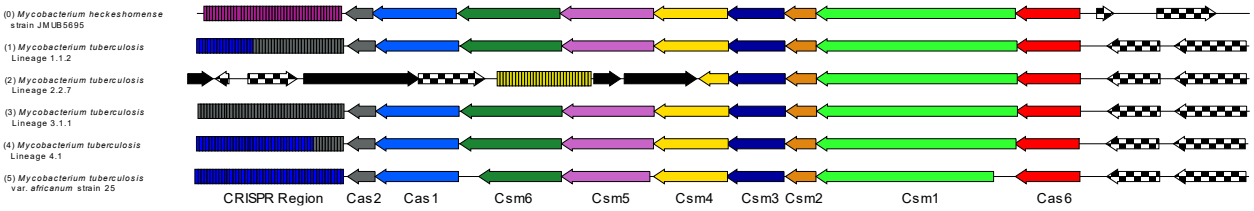

Figure 3-3: Comparison of Cas Type III-A system in *M. heckeshornense* strain JMUB5695 vs. MTBC members. The discovery of a Cas Type III-A system outside the *M. tuberculosis* complex prompted investigation into similarity between it and examples inside the complex. Genes with checkerboard coloring are unrelated to the complex, and genes colored solid black are transposase sequences. A comparison of Cas10 (Csm1, bright green) sequences between *M. heckeshornense* and the MTBC isolates above showed 72-73% identity by Expasy's SIM tool. When comparing *M. heckeshornense* to the insertion element-truncated Cas10 in *Mycobacterium* sp. *SM1*, overlapping regions shared 78% identity. Finally, Comparing *SM1* Cas10 to *MTB* Cas10 yielded 70-71%.

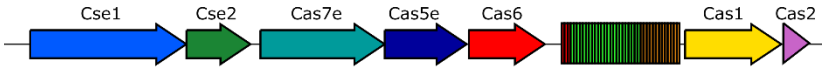

Figure 3-4: Structure of *M. austroafricanum* Cas locus. Species *austroafricanum* presented with an arrangement of *cse* genes upstream of the CRISPR locus, and the Cas1-Cas2 cluster downstream.

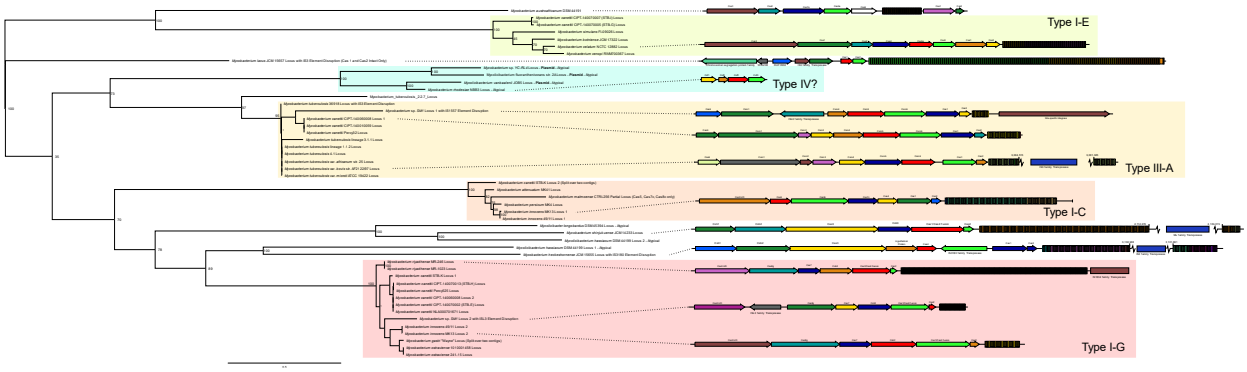

Figure 3-5: Phylogeny and organization of putative Cas systems across *Mycobacteriales*. **Left: Phylogenetic tree of mycobacterial Cas complexes.** A maximum likelihood tree was generated out of 49 sequences comprised of putative Cas complex protein sequences (detected either by existing annotation, CRISPROne/CRISPRCasFinder annotation, and/or NCBI Conserved Domain Search annotation. Proteins are sequentially concatenated from upstream towards *cas2* in all cases except the low-confidence Type IV loci (topmost clade) which do not normally contain *cas2*. Sequences were aligned by MUSCLE and a phylogenetic tree generated by RAxML-NG with bootstrapping. Branches are drawn to scale, with the scale bar representing average amino acid substitutions per site. Bootstrap values (percentage concordant, out of 500 replicates) are shown by the nodes, except where noted by *****, indicating labels have been removed for clarity of visualization but that at least value falls under 70%. Tree visualized and rooted at the midpoint by FigTree. **Right:** Representative locus maps of identified Cas systems. Organization of types of Cas complexes identified in this work. Locus maps created through BV-BRC Compare Regions Viewer. Dashed lines from tree indicate the genetic origin of each locus shown, and genes are labeled as described above. Complexes are grouped into their existing types when possible (e.g., Type I-G). CRISPR repeat sequences are indicated by colored rectangles, with each box indicating a spacer and repeat, and differently colored boxes representing a specific CRISPR repeat sequence within the locus. In most cases, flanking genes are removed from the visualization for clarity, but transposases/integrases/insertion sequence elements are labeled and included. Both tree and locus visualizations are not exhaustive or comprehensive in coverage, and other loci and arrangements across *Mycobacteriales* exist. Figure created in Inkscape.

CHAPTER 5: CONCLUSIONS AND FUTURE RESEARCH

Mycobacterial diseases impact humans and animals worldwide, causing tremendous morbidity, mortality, and economic impacts. Despite their importance, the scientific knowledge of mycobacteria is limited. Mycobacteria are often difficult to isolate, grow, and characterize, and the study of their mechanisms of survival and adaptation against environmental and within-host stresses is incomplete. With the onset of the genomics era, researchers can now analyze isolates at the fundamental level of nucleotides across an entire genome and compare through catalogues of genomes. Advances in computational power and the field of bioinformatics has led to the development of software and pipelines capable of processing millions of reads of sequences in minutes, robust statistical assessment of differences across terabytes of data, and discovery of genomic patterns and features previously overlooked. These advances are tools with previously unimaginable capacity, but like any tools, they are useless unless a user has a purpose for them. If we are to take advantage of the power these tools offer the world, scientific discovery using big data requires clarity of purpose. In this work, data and software freely available to researchers around the world were utilized to answer questions about mycobacterial adaptation and evolution, demonstrating that advances in knowledge await researchers using just the techniques and software available to us now. Investigation of markers for host adaptation of the *Mycobacterium tuberculosis* complex species was conducted by genome-wide association study using over 6,000 genomes from different hosts and countries around the world. Thousands of validated epitope sequences for MTBC members were leveraged to explore whether unique attenuation in *Mycobacterium tuberculosis* variant *bovis* may be associated with changes in host immune recognition as detected by SNPs from

whole genome sequencing. Finally, the evolution of mycobacteria was examined by investigating the conservation and diversity of CRISPR/Cas systems across the family. Each of these techniques uses existing data in new ways, yields avenues for future research and discovery, and advances our understanding of the enigmatic group of mycobacteria.

The results of the GWAS of MTBC isolates yielded a plethora of results with 120 loci concordantly identified as associated by the subsequent statistical test against phenotype of *Bovidae* host, MBO classification, and MTB classification. Among these loci were numerous SNPs fixed in MBO or MBO/MCP showing missense mutations in fatty acid and cholesterol metabolism genes, including genes shown to be essential for growth on cholesterol by transposon mutagenesis studies[230]. Cholesterol intake in particular and fatty acid utilization in general is fundamental to MTBC pathogenesis, as intracellular mycobacteria feed off host lipids for central metabolism and to maintain their mycolic acid-rich cell walls, constituents of which also play virulence roles inside the host[107]. Additionally, several genes involved in translational machinery were affected, as were 4 lipoproteins, various reductases, numerous membrane proteins, 10 transferases, and others. This work cannot assign significance to these changes without *in vitro* and *in vivo* validation, but it does provide a focused list of changes between variants MTB and MBO in the complex to start such investigations. Searches for markers of adaptation to other hosts, like cervids or badgers, yielded inconclusive results. It is possible that a GWAS performed only on the subset of MBO isolates may identify meaningful changes for MBO that enters non-bovine reservoir hosts like white-tailed deer, ringtail possums, or European badgers, but in any case, the results will undoubtedly be better supported with additional sequencing of isolates from these relatively underrepresented species. Furthermore,

a means of reliably assessing changes in PE/PPE genes – comprising a substantial ~10% of coding sequences in the MTBC – would be a boon for analysis of host adaptation[46]. These genes are difficult to accurately sequence and place by Illumina technologies, but are known for their importance in pathogenicity, as well as their sequence variability relative to other genes across the genome, and they are likely to play a role in the complete explanation for MTBC pathogenesis and host adaptation[46,60].

An exploration of epitope variation in attenuated strain Ravenel and virulent strain 10-7428 of *M. tuberculosis* var. *bovis* yielded a handful of modified T cell epitopes in Ravenel, and no such changes in 10-7428. The cell-mediated immune response in the MTBC is fundamental both to successful control of the disease, but is also understood to be necessary to MTB pathogenesis through immune response subversion[69,183]. MTBC T cell epitopes are hyperconserved, reflective of their role in an evolutionary arms race where recognition of specific epitopes at specific times can drive unproductive or disease-promoting reactions in a host[60,251]. In 10-7428, the only recognized modification was to a B cell epitope in a PE/PPE gene PPE42, a family of proteins where antigenic variation is associated with increased virulence and a specific protein selected as a subunit in the ID93 TB vaccine[46,182]. The defective MBO strain BCG-1 (Russia), with impaired pathogenesis due to the deletion of the RD1 region, showed numerous changes to epitopes, which could be explained by a loss of selective pressure on epitope conservation in the background of defective pathogenesis by gene losses. These findings are suggestive that variation in T cell epitopes may be associated with attenuation, and could provide insights into the specific genes and pathways involved in normal host immune subversion. Future research should build upon this initial investigation to compare more

genomes and assess frequency of T cell epitope variations. Finally, adding this step of epitope variation analysis into standard whole-genome sequencing and variant extraction pipelines should be both straightforward and informative.

While CRISPR/Cas systems have been studied and exploited in MTB for decades for assessing evolutionary similarities between strains, exploration of the existence of these systems outside of the *M. tuberculosis* complex has remained scant[44,81]. Earlier research suggested Cas systems are either exclusive to the MTBC, or that only a few species contain them[43,44]. The Type III-A Cas system found in the MTBC has been reported to be an exclusive marker of the complex across *Mycobacterium*, and has been used by Singh *et al.* to suggest an evolutionary history for the complex that involved ancestral horizontal gene transfer from *Firmicutes*[43]. In contrast, this work finds Cas systems of multiple subtypes exist throughout the family *Mycobacteriaceae*, including what appear to be novel genetic arrangements or incorporating genes of unknown function. Furthermore, a Type III-A Cas complex is observed both in the opportunistic NTM *M. heckeshornense* and in an unusual environmental isolate in *M.* sp. *SM1*, complicating evolutionary inferences that utilize this marker locus. Some species of *Mycobacterium* contain multiple types of Cas loci, and others – such as the entire *M. avium* complex – appear to show no signs of Cas genes at all. Finally, substantial variation exists in Cas system presence even between isolates of the same species, suggesting NTM gene content is in regular flux and our understanding of this flow is a greater knowledge gap than recognized.

To summarize, mycobacteria are a group of bacteria with unparalleled impact on human and animal health worldwide, historically and in the present day. While the MTBC causes 1-2 million human deaths per year, as NTM infections continue increasing worldwide in both

morbidity and mortality, and as multi-drug resistance proliferates both in mycobacteria and beyond, the understanding of what constitutes pathogenic potential and how it is modulated is essential. Breakthroughs in these complex fields are facilitated by traditional *in vitro* microbiology and immunology aided and streamlined by the incorporation of powerful *in silico* analysis. Comparative genomics holds incredible potential to untangle niche adaptation, virulence and attenuation, and evolutionary histories for these species. While much remains to be discovered about mycobacteria, these gaps in our knowledge will be best explored by a research synthesis of immunology, mycobacteriology, and bioinformatics.

BIBLIOGRAPHY

1. Gupta, R. S., Lo, B. & Son, J. Phylogenomics and comparative genomic studies robustly support division of the genus Mycobacterium into an emended genus Mycobacterium and four novel genera. *Front. Microbiol.* **9**, 1–41 (2018).

2. Armstrong, D. T. & Parrish, N. Current Updates on Mycobacterial Taxonomy, 2018 to 2019. *J. Clin. Microbiol.* **59**, e0152820 (2021).

3. Murray, P. R., Rosenthal, K. S. & Pfaller, M. A. Mycobacterium and Related Acid-Fast Bacteria. in *Medical Microbiology* 226–240 (Elsevier, 2020).

4. Pereira, A. C., Ramos, B., Reis, A. C. & Cunha, M. V. Non-tuberculous mycobacteria: Molecular and physiological bases of virulence and adaptation to ecological niches. *Microorganisms* **8**, 1–49 (2020).

5. Grant, I. R. *et al.* Viable Mycobacterium avium ssp. paratuberculosis isolated from calf milk replacer. *J. Dairy Sci.* **100**, 9723–9735 (2017).

6. Nambi, S. *et al.* The Oxidative Stress Network of Mycobacterium tuberculosis Reveals Coordination between Radical Detoxification Systems. *Cell Host Microbe* **17**, 829–837 (2015).

7. Gerrard, Z. E. *et al.* Survival of Mycobacterium avium subspecies paratuberculosis in retail pasteurised milk. *Food Microbiol.* **74**, 57–63 (2018).

8. Le Dantec, C. *et al.* Occurrence of mycobacteria in water treatment lines and in water distribution systems. *Appl. Environ. Microbiol.* **68**, 5318–5325 (2002).

9. Weeks, J. W., Segars, K. & Guha, S. The Research Gap in Non-tuberculous Mycobacterium (NTM) and Reusable Medical Devices. *Front. Public Heal.* **8**, 1–5 (2020).

10. Sapriel, G., Brosch, R. & Bapteste, E. Shared Pathogenomic Patterns Characterize a New Phylotype, Revealing Transition toward Host-Adaptation Long before Speciation of Mycobacterium tuberculosis. *Genome Biol. Evol.* **11**, 2420–2438 (2019).

11. Boritsch, E. C. *et al.* A glimpse into the past and predictions for the future: The molecular evolution of the tuberculosis agent. *Mol. Microbiol.* **93**, 835–852 (2014).

12. Barberis, I., Bragazzi, N. L., Galluzzo, L. & Martini, M. The history of tuberculosis: From the first historical records to the isolation of Koch's bacillus. *J. Prev. Med. Hyg.* **58**, E9–E12 (2017).

13. Houben, R. M. G. J. & Dodd, P. J. The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. *PLoS Med.* **13**, 1–13 (2016).

14. Zimpel, C. K. *et al.* Global Distribution and Evolution of Mycobacterium bovis Lineages. *Front. Microbiol.* **11**, 1–19 (2020).

15. Deps, P. & Collin, S. M. Mycobacterium lepromatosis as a Second Agent of Hansen's Disease. *Front. Microbiol.* **12**, 1–7 (2021).

16. Santacroce, L., Prete, R. Del, Charitos, I. A. & Bottalico, L. Mycobacterium leprae: A historical study on the origins of leprosy and its social stigma. *Infez. Med.* **29**, 623–632 (2021).

17. Makhakhe, L. Leprosy review. *South African Fam. Pract.* **63**, 1–6 (2021).

18. Pennington, K. M. *et al.* Approach to the diagnosis and treatment of non-tuberculous mycobacterial disease. *J. Clin. Tuberc. Other Mycobact. Dis.* **24**, 100244 (2021).

19. Godreuil, S. *et al.* Mycobacterium riyadhense Pulmonary Infection, France and Bahrain. *Emerg. Infect. Dis.* **18**, 176–178 (2012).

20. Adzic-Vukicevic, T. *et al.* Clinical features of infection caused by non-tuberculous mycobacteria: 7 years' experience. *Infection* **46**, 357–363 (2018).

21. Tortoli, E. *et al.* Same meat, different gravy: Ignore the new names of mycobacteria. *Eur. Respir. J.* **54**, 19–21 (2019).

22. Tsukamura, M. Identification of mycobacteria. *Tubercle* **48**, 311–338 (1967).

23. Gagneux, S. Ecology and evolution of Mycobacterium tuberculosis. *Nat. Rev. Microbiol.* **16**, 202–213 (2018).

24. Koh, W.-J., Kwon, O. J. & Lee, K. S. Nontuberculous mycobacterial pulmonary diseases in immunocompetent patients. *Korean J. Radiol.* **3**, 145–57 (2002).

25. Goodfellow, M. & Magee, J. G. Taxonomy of Mycobacteria. in *Mycobacteria: I Basic Aspects* (eds. Gangadharam, P. R. J. & Jenkins, P. A.) 1–71 (Chapman & Hall, 1998).

26. Riojas, M. A., McGough, K. J., Rider-Riojas, C. J., Rastogi, N. & Hazbón, M. H. Phylogenomic analysis of the species of the Mycobacterium tuberculosis complex demonstrates that Mycobacterium africanum, Mycobacterium bovis, Mycobacterium caprae, Mycobacterium microti and Mycobacterium pinnipedii are later heterotypic synonyms of Mycob. *Int. J. Syst. Evol. Microbiol.* **68**, 324–332 (2018).

27. Supply, P., Marceau, M., Mangenot, S. & Roche, D. Genome analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of the etiologic agent of tuberculosis. *Nat. Genet.* **45**, 172–179 (2013).

28. Tagini, F. *et al.* Phylogenomics reveal that mycobacterium kansasii subtypes are species-

level lineages. Description of mycobacterium pseudokansasii sp. nov., mycobacterium innocens sp. nov. and mycobacterium attenuatum sp. nov. *Int. J. Syst. Evol. Microbiol.* **69**, 1696–1704 (2019).

29. Jagielski, T. *et al.* Genomic Insights Into the Mycobacterium kansasii Complex: An Update. *Front. Microbiol.* **10**, (2020).

30. Chiaradia, L. *et al.* Dissecting the mycobacterial cell envelope and defining the composition of the native mycomembrane. *Sci. Rep.* **7**, 1–12 (2017).

31. Maitra, A. *et al.* Cell wall peptidoglycan in Mycobacterium tuberculosis: An Achilles' heel for the TB-causing pathogen. *FEMS Microbiol. Rev.* **43**, 548–575 (2019).

32. Vincent, A. T. *et al.* The mycobacterial cell envelope: A relict from the past or the result of recent evolution? *Front. Microbiol.* **9**, 1–9 (2018).

33. Daffé, M. & Marrakchi, H. Unraveling the structure of the mycobacterial envelope. *Gram-Positive Pathog.* 1087–1095 (2019) doi:10.1128/9781683670131.ch65.

34. Bothra, A. *et al.* Phospholipid homeostasis, membrane tenacity and survival of Mtb in lipid rich conditions is determined by MmpL11 function. *Sci. Rep.* **8**, 1–14 (2018).

35. Osman, M. M. *et al.* The C terminus of the mycobacterium ESX-1 secretion system substrate ESAT-6 is required for phagosomal membrane damage and virulence. *Proc. Natl. Acad. Sci. U. S. A.* **119**, 1–9 (2022).

36. Boddingius, J. & Dijkman, H. Subcellular localization of Mycobacterium leprae-specific phenolic glycolipid (PGL-I) antigen in human leprosy lesions and in M. leprae isolated from armadillo liver. *J. Gen. Microbiol.* **136**, 2001–2012 (1990).

37. Rajni, Rao, N. & Meena, L. S. Biosynthesis and Virulent Behavior of Lipids Produced by Mycobacterium tuberculosis : LAM and Cord Factor: An Overview . *Biotechnol. Res. Int.* **2011**, 1–7 (2011).

38. Marrakchi, H., Lanéelle, M. A. & Daffé, M. Mycolic acids: Structures, biosynthesis, and beyond. *Chem. Biol.* **21**, 67–85 (2014).

39. Cole, S. T. *et al.* Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **393**, 537–544 (1998).

40. Marri, P. R., Bannantine, J. P. & Golding, G. B. Comparative genomics of metabolic pathways in Mycobacterium species: Gene duplication, gene decay and lateral gene transfer. *FEMS Microbiol. Rev.* **30**, 906–925 (2006).

41. Cole, S. T. *et al.* Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007–1011 (2001).

42. Garnier, T. *et al.* The complete genome sequence of Mycobacterium bovis. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 7877–7882 (2003).

43. Singh, A. *et al.* Comparative Genomic Analysis of Mycobacteriaceae Reveals Horizontal Gene Transfer-Mediated Evolution of the CRISPR-Cas System in the Mycobacterium tuberculosis Complex . *mSystems* **6**, (2021).

44. He, L., Fan, X. & Xie, J. Comparative genomic structures of Mycobacterium CRISPR-Cas. *J. Cell. Biochem.* **113**, 2464–2473 (2012).

45. McEvoy, C. R., Van Helden, P. D., Warren, R. M. & Van Pittius, N. C. G. Evidence for a rapid rate of molecular evolution at the hypervariable and immunogenic Mycobacterium tuberculosis PPE38 gene region. *BMC Evol. Biol.* **9**, 1–21 (2009).

46. McEvoy, C. R. E. *et al.* Comparative analysis of mycobacterium tuberculosis pe and ppe genes reveals high sequence variation and an apparent absence of selective constraints. *PLoS One* **7**, (2012).

47. Derbyshire, K. M. & Gray, T. A. Distributive Conjugal Transfer: New Insights into Horizontal Gene Transfer and Genetic Exchange in Mycobacteria. *Microbiol. Spectr.* **2**, 1–32 (2014).

48. Gray, T. & Derbyshire, K. Blending genomes: Distributive Conjugal Transfer in Mycobacteria, a sexier form of HGT. *Mol. Microbiol.* **108**, 601–613 (2018).

49. Ummels, R. *et al.* Identification of a novel conjugative plasmid in mycobacteria that requires both type IV and type VII secretion. *MBio* **5**, (2014).

50. Boritsch, E. C. *et al.* Key experimental evidence of chromosomal DNA transfer among selected tuberculosiscausing mycobacteria. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 9876–9881 (2016).

51. Tagini, F., Pillonel, T., Bertelli, C., Jaton, K. & Greub, G. Pathogenic determinants of the mycobacterium kansasii complex: An unsuspected role for distributive conjugal transfer. *Microorganisms* **9**, 1–22 (2021).

52. Freschi, L. *et al.* Population structure, biogeography and transmissibility of Mycobacterium tuberculosis. *Nat. Commun.* **12**, 6099 (2021).

53. Earle, S. G. *et al.* Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat. Microbiol.* **1**, 1–21 (2016).

54. Galagan, J. E. Genomic insights into tuberculosis. *Nat. Rev. Genet.* **15**, 307–320 (2014).

55. Madacki, J. *et al.* Esx-1-independent horizontal gene transfer by mycobacterium tuberculosis complex strains. *MBio* **12**, (2021).

56. Ford, C. B. *et al.* Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nat. Genet.* **43**, 482–488 (2011).

57. Folkvardsen, D. B. *et al.* Genomic epidemiology of a major Mycobacterium tuberculosis outbreak: Retrospective cohort study in a low-incidence setting using sparse time-series sampling. *J. Infect. Dis.* **216**, 366–374 (2017).

58. Gagneux, S. *et al.* Variable host-pathogen compatibility in Mycobacterium tuberculosis. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 2869–2873 (2006).

59. Stucki, D. *et al.* Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat. Genet.* **48**, 1535–1543 (2016).

60. Comas, Ĩ. *et al.* Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. *Nat. Genet.* **42**, 498–503 (2010).

61. Coll, F. *et al.* A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nat. Commun.* **5**, 4–8 (2014).

62. Smith, N. H. The global distribution and phylogeography of Mycobacterium bovis clonal complexes. *Infect. Genet. Evol.* **12**, 857–865 (2012).

63. Pepperell, C. S. Evolution of Tuberculosis Pathogenesis. *Annu. Rev. Microbiol.* **76**, 661–680 (2022).

64. Guan, Q. *et al.* Insights into the ancestry evolution of the Mycobacterium tuberculosis complex from analysis of Mycobacterium riyadhense. *NAR Genomics Bioinforma.* **3**, 1–16 (2021).

65. Bronson, R. A. *et al.* Global phylogenomic analyses of Mycobacterium abscessus provide context for non cystic fibrosis infections and the evolution of antibiotic resistance. *Nat. Commun.* **12**, 1–10 (2021).

66. Cocito, C. & Delville, J. Biological, chemical, immunological and staining properties of bacteria isolated from tissues of leprosy patients. *Eur. J. Epidemiol.* **1**, (1985).

67. Pai, M., Nicol, M. P. & Boehme, C. C. Tuberculosis Diagnostics: State of the Art and Future Directions. in *Tuberculosis and the Tubercle Bacillus* (eds. Jacobs, W. R., McShane, H., Mizrahi, V. & Orme, I. M.) 363–378 (ASM Press, 2017). doi:10.1128/9781555819569.

68. Vilcheze, C. & Kremer, L. Acid-Fast Positive and Acid-Fast Negative Mycobacterium tuberculosis: The Koch Paradox. in *Tuberculosis and the Tubercle Bacillus* (eds. Jacobs, W. R., McShane, H., Mizrahi, V. & Orme, I. M.) 519–532 (ASM Press, 2017). doi:10.1128/9781555819569.

69. Scriba, T. J., Coussens, A. K. & Fletcher, H. A. Human Immunology of Tuberculosis. in

Tuberculosis and the Tubercle Bacillus (eds. Jacobs, W. R., McShane, H., Mizrahi, V. & Orme, I. M.) 213–237 (ASM Press, 2017). doi:10.1128/9781555819569.

70. Zhou, G. *et al.* Interferon-γ release assays or tuberculin skin test for detection and management of latent tuberculosis infection: a systematic review and meta-analysis. *Lancet Infect. Dis.* **20**, 1457–1469 (2020).

71. de Almeida, C. A. S. *et al.* Intradermal comparative cervical tuberculin test in the diagnosis of caprine tuberculosis. *Brazilian J. Microbiol.* **53**, 421–431 (2022).

72. Fernández-Veiga, L. *et al.* Differences in skin test reactions to official and defined antigens in guinea pigs exposed to non-tuberculous and tuberculous bacteria. *Sci. Rep.* **13**, 2936 (2023).

73. Mechal, Y. *et al.* Evaluation of GeneXpert MTB/RIF system performances in the diagnosis of extrapulmonary tuberculosis. *BMC Infect. Dis.* **19**, 1069 (2019).

74. Koets, A. P., Eda, S. & Sreevatsan, S. The within host dynamics of Mycobacterium avium ssp. paratuberculosis infection in cattle: where time and place matter. *Vet. Res.* **46**, 61 (2015).

75. Kanabalan, R. D. *et al.* Human tuberculosis and Mycobacterium tuberculosis complex: A review on genetic diversity, pathogenesis and omics approaches in host biomarkers discovery. *Microbiol. Res.* **246**, 126674 (2021).

76. Kanipe, C. & Palmer, M. V. Mycobacterium bovis and you: A comprehensive look at the bacteria, its similarities to Mycobacterium tuberculosis, and its relationship with human disease. *Tuberculosis* **125**, 102006 (2020).

77. Johnston, J. C., Chiang, L. & Elwood, K. Mycobacterium kansasii. in *Tuberculosis and Nontuberculous Mycobacterial Infections* 725–734 (ASM Press, 2017). doi:10.1128/9781555819866.ch42.

78. Gopalaswamy, R., Shanmugam, S., Mondal, R. & Subbian, S. Of tuberculosis and non-tuberculous mycobacterial infections – a comparative analysis of epidemiology, diagnosis and treatment. *J. Biomed. Sci.* **27**, 74 (2020).

79. Tran, Q. T. & Han, X. Y. Subspecies Identification and Significance of 257 Clinical Strains of Mycobacterium avium. *J. Clin. Microbiol.* **52**, 1201–1206 (2014).

80. Zhang, W. *et al.* Antigen 85B peptidomic analysis allows species-specific mycobacterial identification. *Clin. Proteomics* **15**, 1–10 (2018).

81. Kamerbeek, J. *et al.* Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. *J. Clin. Microbiol.* **35**, 907–914 (1997).

82. Ereqat, S. *et al.* Rapid differentiation of Mycobacterium tuberculosis and M. bovis by high-resolution melt curve analysis. *J. Clin. Microbiol.* **48**, 4269–4272 (2010).

83. Sreevatsan, S. *et al.* Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 9869–9874 (1997).

84. Coll, F. *et al.* Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. *Nat. Genet.* **50**, 307–316 (2018).

85. Vordermeier, H. M., Jones, G. J., Buddle, B. M., Hewinson, R. G. & Villarreal-Ramos, B. Bovine tuberculosis in cattle: Vaccines, DIVA tests, and host biomarker discovery. *Annu. Rev. Anim. Biosci.* **4**, 87–109 (2016).

86. Chandra, P., Grigsby, S. J. & Philips, J. A. Immune evasion and provocation by Mycobacterium tuberculosis. *Nat. Rev. Microbiol.* **20**, 750–766 (2022).

87. Lamont, E. A. *et al.* Circulating Mycobacterium bovis Peptides and Host Response Proteins as Biomarkers for Unambiguous Detection of Subclinical Infection. *J. Clin. Microbiol.* **52**, 536–543 (2014).

88. Wanzala, S. I. *et al.* Evaluation of pathogen-specific biomarkers for the diagnosis of tuberculosis in white-tailed deer (Odocoileus virginianus). *Am. J. Vet. Res.* **78**, 729–734 (2017).

89. Hadi, S. A., Waters, W. R., Palmer, M., Lyashchenko, K. P. & Sreevatsan, S. Development of a multidimensional proteomic approach to detect circulating immune complexes in cattle experimentally infected with Mycobacterium bovis. *Front. Vet. Sci.* **5**, 1–5 (2018).

90. DS Sarro, Y. *et al.* Simultaneous diagnosis of tuberculous and non-tuberculous mycobacterial diseases: Time for a better patient management. *Clin. Microbiol. Infect. Dis.* **3**, 1–4 (2018).

91. Chen, P. E. & Shapiro, B. J. The advent of genome-wide association studies for bacteria. *Curr. Opin. Microbiol.* **25**, 17–24 (2015).

92. Farhat, M. R. *et al.* GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. *Nat. Commun.* **10**, (2019).

93. Long, R., Divangahi, M. & Schwartzman, K. Chapter 2: Transmission and pathogenesis of tuberculosis. *Can. J. Respir. Crit. Care, Sleep Med.* **6**, 22–32 (2022).

94. Shah, M. & Dorman, S. E. Latent Tuberculosis Infection. *N. Engl. J. Med.* **385**, 2271–2280 (2021).

95. World Health Organization. *Implementing the WHO Stop TB Strategy : a handbook for*

national tuberculosis control programmes. (World Health Organization, 2008).

96. Tiemersma, E. W., van der Werf, M. J., Borgdorff, M. W., Williams, B. G. & Nagelkerke, N. J. D. Natural history of tuberculosis: Duration and fatality of untreated pulmonary tuberculosis in HIV negative patients: A systematic review. *PLoS One* **6**, (2011).

97. Donald, P. R. *et al.* Droplets, dust and Guinea pigs: An historical review of tuberculosis transmission research, 1878-1940. *Int. J. Tuberc. Lung Dis.* **22**, 972–982 (2018).

98. Scordo, J. M. *et al.* The human lung mucosa drives differential Mycobacterium tuberculosis infection outcome in the alveolar epithelium. *Mucosal Immunol.* **12**, 795–804 (2019).

99. Cohen, S. B. *et al.* Alveolar Macrophages Provide an Early Mycobacterium tuberculosis Niche and Initiate Dissemination. *Cell Host Microbe* **24**, 439-446.e4 (2018).

100. Ernst, J. D. The immunological life cycle of tuberculosis. *Nat. Rev. Immunol.* **12**, 581–591 (2012).

101. Hmama, Z., Peña-Díaz, S., Joseph, S. & Av-Gay, Y. Immunoevasion and immunosuppression of the macrophage by Mycobacterium tuberculosis. *Immunol. Rev.* **264**, 220–232 (2015).

102. Kang, P. B. *et al.* The human macrophage mannose receptor directs Mycobacterium tuberculosis lipoarabinomannan-mediated phagosome biogenesis. *J. Exp. Med.* **202**, 987–999 (2005).

103. Buter, J. *et al.* Mycobacterium tuberculosis releases an antacid that remodels phagosomes. *Nat. Chem. Biol.* **15**, 889–899 (2019).

104. Saha, S. *et al.* A Bumpy Ride of Mycobacterial Phagosome Maturation: Roleplay of Coronin1 Through Cofilin1 and cAMP. *Front. Immunol.* **12**, 1–17 (2021).

105. Ouimet, M. *et al.* Mycobacterium tuberculosis induces the miR-33 locus to reprogram autophagy and host lipid metabolism. *Nat. Immunol.* **17**, 677–686 (2016).

106. Guerrini, V. *et al.* Storage lipid studies in tuberculosis reveal that foam cell biogenesis is disease-specific. *PLoS Pathog.* **14**, 1–27 (2018).

107. Wilburn, K. M., Fieweger, R. A. & VanderVen, B. C. Cholesterol and fatty acids grease the wheels of Mycobacterium tuberculosis pathogenesis. *Pathog. Dis.* **76**, 1–14 (2018).

108. Upadhyay, S., Mittal, E. & Philips, J. A. Tuberculosis and the art of macrophage manipulation. *Pathog. Dis.* **76**, 1–12 (2018).

109. Srivastava, S., Battu, M. B., Khan, M. Z., Nandicoori, V. K. & Mukhopadhyay, S.

Mycobacterium tuberculosis PPE2 Protein Interacts with p67phox and Inhibits Reactive Oxygen Species Production. *J. Immunol.* **203**, 1218–1229 (2019).

110. Madan-Lala, R. *et al.* Mycobacterium tuberculosis Impairs Dendritic Cell Functions through the Serine Hydrolase Hip1. *J. Immunol.* **192**, 4263–4272 (2014).

111. Domingo-Gonzalez, R., Prince, O., Cooper, A. & Khader, S. A. Cytokines and chemokines in Mycobacterium tuberculosis infection. *Tuberc. Tuber. Bacillus Second Ed.* 33–72 (2017) doi:10.1128/9781555819569.ch2.

112. Behr, M. A. & Waters, W. R. Is tuberculosis a lymphatic disease with a pulmonary portal? *Lancet Infect. Dis.* **14**, 250–255 (2014).

113. Allen, A. R., Ford, T. & Skuce, R. A. Does Mycobacterium tuberculosis var. bovis Survival in the Environment Confound Bovine Tuberculosis Control and Eradication? A Literature Review. *Vet. Med. Int.* **2021**, 1–19 (2021).

114. Gormley, E. & Corner, L. A. L. Pathogenesis of Mycobacterium bovis Infection: The Badger model as a paradigm for understanding tuberculosis in animals. *Front. Vet. Sci.* **4**, 1–11 (2018).

115. VerCauteren, K. C., Lavelle, M. J. & Campa, H. Persistent Spillback of Bovine Tuberculosis From White-Tailed Deer to Cattle in Michigan, USA: Status, Strategies, and Needs. *Front. Vet. Sci.* **5**, 1–13 (2018).

116. Loiseau, C. *et al.* An African origin for Mycobacterium bovis. *Evol. Med. Public Heal.* **2020**, 49–59 (2020).

117. Taye, H. *et al.* Global prevalence of Mycobacterium bovis infections among human tuberculosis cases: Systematic review and meta-analysis. *Zoonoses Public Health* **68**, 704–718 (2021).

118. de Jong, B. C. *et al.* Does Resistance to Pyrazinamide Accurately Indicate the Presence of Mycobacterium bovis ? *J. Clin. Microbiol.* **43**, 3530–3532 (2005).

119. Dawson, K. L. *et al.* Transmission of Mycobacterium orygis (M. tuberculosis Complex Species) from a Tuberculosis Patient to a Dairy Cow in New Zealand. *J. Clin. Microbiol.* **50**, 3136–3138 (2012).

120. Marcos, L. A. *et al.* Mycobacterium orygis lymphadenitis in New York, USA. *Emerg. Infect. Dis.* **23**, 1749–1751 (2017).

121. Eldholm, V., Rønning, J. O., Mengshoel, A. T. & Arnesen, T. Import and transmission of Mycobacterium orygis and Mycobacterium africanum, Norway. *BMC Infect. Dis.* **21**, 562 (2021).

122. Duffy, S. C. *et al.* Reconsidering Mycobacterium bovis as a proxy for zoonotic tuberculosis: a molecular epidemiological surveillance study. *The Lancet Microbe* **1**, e66–e73 (2020).

123. Luca, S. & Mihaescu, T. History of BCG Vaccine. *Maedica (Buchar).* **8**, 53–8 (2013).

124. Behr, M. A. BCG - Different strains, different vaccines? *Lancet Infect. Dis.* **2**, 86–92 (2002).

125. Kuan, R., Muskat, K., Peters, B. & Lindestam Arlehamn, C. S. Is mapping the BCG vaccine-induced immune responses the key to improving the efficacy against tuberculosis? *J. Intern. Med.* **288**, 651–660 (2020).

126. Ying, W. *et al.* Clinical Characteristics and Immunogenetics of BCGosis/BCGitis in Chinese Children: A 6 Year Follow-Up Study. *PLoS One* **9**, e94485 (2014).

127. Buddle, B. M., Vordermeier, H. M., Chambers, M. A. & de Klerk-Lorist, L. M. Efficacy and safety of BCG vaccine for control of tuberculosis in domestic livestock and wildlife. *Front. Vet. Sci.* **5**, 1–17 (2018).

128. Rathnaiah, G. *et al.* Pathogenesis, Molecular Genetics, and Genomics of Mycobacterium avium subsp. paratuberculosis, the Etiologic Agent of Johne's Disease. *Front. Vet. Sci.* **4**, 1–13 (2017).

129. Srinivasan, S. *et al.* A Meta-Analysis of the Effect of Bacillus Calmette-Guérin Vaccination Against Bovine Tuberculosis: Is Perfect the Enemy of Good? *Front. Vet. Sci.* **8**, (2021).

130. Blouin, Y. *et al.* Progenitor 'Mycobacterium canettii' Clone responsible for lymph node tuberculosis epidemic, Djibouti. *Emerg. Infect. Dis.* **20**, 21–28 (2014).

131. Somoskovi, A. *et al.* 'Mycobacterium canettii' isolated from a human immunodeficiency virus-positive patient: first case recognized in the United States. *J. Clin. Microbiol.* **47**, 255–7 (2009).

132. Turenne, C. Mycobacterium lacus sp. nov., a novel slowly growing, non-chromogenic clinical isolate. *Int. J. Syst. Evol. Microbiol.* **52**, 2135–2140 (2002).

133. Saito, H. *et al.* Mycobacterium shinjukuense sp. nov., a slowly growing, non-chromogenic species isolated from human clinical specimens. *Int. J. Syst. Evol. Microbiol.* **61**, 1927–1932 (2011).

134. Brown-Elliott, B. A. *et al.* Mycobacterium decipiens sp. nov., a new species closely related to the Mycobacterium tuberculosis complex. *Int. J. Syst. Evol. Microbiol.* **68**, 3557–3562 (2018).

135. Becerril-Villanueva, E. *et al.* Chronic infection with Mycobacterium lepraemurium induces alterations in the hippocampus associated with memory loss. *Sci. Rep.* **8**, 1–12 (2018).

136. Oliveira, I. V. P. de M., Deps, P. D. & Antunes, J. M. A. de P. Armadillos and leprosy: from infection to biological model. *Rev. Inst. Med. Trop. Sao Paulo* **61**, 1–7 (2019).

137. Bratschi, M. W., Steinmann, P., Wickenden, A. & Gillis, T. P. Current knowledge on Mycobacterium leprae transmission: a systematic literature review. *Lepr. Rev.* **86**, 142–155 (2015).

138. Vissa, V. D. & Brennan, P. J. The genome of Mycobacterium leprae: A minimal mycobacterial gene set. *Genome Biol.* **2**, 1–8 (2001).

139. Han, X. Y. *et al.* Comparative Sequence Analysis of Mycobacterium leprae and the New Leprosy-Causing Mycobacterium lepromatosis. *J. Bacteriol.* **191**, 6067–6074 (2009).

140. Benjak, A. *et al.* Insights from the Genome Sequence of Mycobacterium lepraemurium : Massive Gene Decay and Reductive Evolution. *MBio* **8**, 1–6 (2017).

141. Rojas-Espinosa, O. Murine Leprosy Revisited. in *Current Topics on the Profiles of Host Immunological Response to Mycobacterial Infections* (ed. Tomioka, H.) 97–140 (Research Signpost, 2009).

142. Kim, B. G., Jhun, B. W., Kim, H. & Kwon, O. J. Treatment outcomes of Mycobacterium avium complex pulmonary disease according to disease severity. *Sci. Rep.* **12**, 1–9 (2022).

143. Loebinger, M. R. Mycobacterium avium complex infection: phenotypes and outcomes. *Eur. Respir. J.* **50**, 1701380 (2017).

144. Dhama, K. *et al.* Tuberculosis in Birds: Insights into the Mycobacterium avium Infections. *Vet. Med. Int.* **2011**, 1–14 (2011).

145. Fulton, R. M. Tuberculosis in Poultry. *Merck Manual* https://www.merckvetmanual.com/poultry/tuberculosis/tuberculosis-in-poultry (2019).

146. Collins, M. Paratuberculosis in Ruminants (Johne's Disease). *Merck Manual* https://www.merckvetmanual.com/digestive-system/intestinal-diseases-in-ruminants/paratuberculosis-in-ruminants? (2021).

147. Sweeney, R. W., Collins, M. T., Koets, A. P., McGuirk, S. M. & Roussel, A. J. Paratuberculosis (Johne's Disease) in Cattle and Other Susceptible Species. *J. Vet. Intern. Med.* **26**, 1239–1250 (2012).

148. Botsaris, G. *et al.* Detection of viable Mycobacterium avium subspecies paratuberculosis in powdered infant formula by phage-PCR and confirmed by culture. *Int. J. Food Microbiol.* **216**, 91–94 (2016).

149. Botsaris, G. *et al.* Rapid detection methods for viable Mycobacterium avium subspecies paratuberculosis in milk and cheese. *Int. J. Food Microbiol.* **141**, S87–S90 (2010).

150. Fukano, H. *et al.* Human pathogenic Mycobacterium kansasii (former subtype I) with zoonotic potential isolated from a diseased indoor pet cat, Japan. *Emerg. Microbes Infect.* **10**, 220–222 (2021).

151. Shahraki, A. H. *et al.* Mycobacterium persicum sp. Nov., a novel species closely related to mycobacterium kansasii and mycobacterium gastri. *Int. J. Syst. Evol. Microbiol.* **67**, 1758–1765 (2017).

152. Taillard, C. *et al.* Clinical Implications of Mycobacterium kansasii Species Heterogeneity: Swiss National Survey. *J. Clin. Microbiol.* **41**, 1240–1244 (2003).

153. A Narh, C. Genotyping Tools for Mycobacterium ulcerans Drawbacks and Future Prospects. *Mycobact. Dis.* **04**, (2014).

154. Hammoudi, N., Saad, J. & Drancourt, M. The diversity of mycolactone-producing mycobacteria. *Microb. Pathog.* **149**, 104362 (2020).

155. Yip, M. J. *et al.* Evolution of Mycobacterium ulcerans and Other Mycolactone-Producing Mycobacteria from a Common Mycobacterium marinum Progenitor. *J. Bacteriol.* **189**, 2021–2029 (2007).

156. Yotsu, R. R. *et al.* Buruli Ulcer: a Review of the Current Knowledge. *Curr. Trop. Med. Reports* **5**, 247–256 (2018).

157. Hammoudi, N. *et al.* Mycobacterium ulcerans mycolactones-fungi crosstalking. *Sci. Rep.* **9**, 1–6 (2019).

158. Muleta, A. J., Lappan, R., Stinear, T. P. & Greening, C. Understanding the transmission of Mycobacterium ulcerans: A step towards controlling Buruli ulcer. *PLoS Negl. Trop. Dis.* **15**, e0009678 (2021).

159. Garchitorena, A. *et al.* Mycobacterium ulcerans Ecological Dynamics and Its Association with Freshwater Ecosystems and Aquatic Communities: Results from a 12-Month Environmental Survey in Cameroon. *PLoS Negl. Trop. Dis.* **8**, (2014).

160. Aubry, A., Mougari, F., Reibel, F. & Cambau, E. Mycobacterium marinum. in *Tuberculosis and Nontuberculous Mycobacterial Infections* 735–752 (ASM Press, 2017). doi:10.1128/9781555819866.ch43.

161. Lai, L.-Y., Lin, T.-L., Chen, Y.-Y., Hsieh, P.-F. & Wang, J.-T. Role of the Mycobacterium marinum ESX-1 Secretion System in Sliding Motility and Biofilm Formation. *Front. Microbiol.* **9**, 1–12 (2018).

162. Jiva, T. M., Jacoby, H. M., Weymouth, L. A., Kaminski, D. A. & Portmore, A. C. Mycobacterium xenopi: innocent bystander or emerging pathogen? *Clin. Infect. Dis.* **24**, 226–32 (1997).

163. Andréjak, C. *et al.* Mycobacterium xenopi pulmonary infections: A multicentric retrospective study of 136 cases in north-east France. *Thorax* **64**, 291–296 (2009).

164. St-Jean, G. *et al.* Mycobacterium xenopi systemic infection in a domestic fiery-shouldered conure bird (Pyrrhura egregia). *JMM Case Reports* **5**, 0–4 (2018).

165. Yoshida, M., Fukano, H., Asakura, T., Suzuki, M. & Hoshino, Y. Complete Genome Sequence of Mycobacterium heckeshornense JCM 15655 T , Closely Related to a Pathogenic Nontuberculous Mycobacterial Species, Mycobacterium xenopi. *Microbiol. Resour. Announc.* **10**, 17–19 (2021).

166. Iitoh, E. *et al.* A case of pulmonary Mycobacterium heckeshornense infection in a healthy Japanese man: A case of pulmonary M. heckeshornense infection. *Respir. Med. Case Reports* **30**, 101093 (2020).

167. Van Hest, R. *et al.* Mycobacterium heckeshornense infection in an immunocompetent patient and identification by 16S rRNA sequence analysis of culture material and a histopathology tissue specimen. *J. Clin. Microbiol.* **42**, 4386–4389 (2004).

168. Victoria, L., Gupta, A., Gómez, J. L. & Robledo, J. Mycobacterium abscessus complex: A Review of Recent Developments in an Emerging Pathogen. *Front. Cell. Infect. Microbiol.* **11**, 1–8 (2021).

169. Lopeman, R., Harrison, J., Desai, M. & Cox, J. Mycobacterium abscessus: Environmental Bacterium Turned Clinical Nightmare. *Microorganisms* **7**, 90 (2019).

170. Ayele, W. Y., Neill, S. D., Zinsstag, J., Weiss, M. G. & Pavlik, I. Bovine tuberculosis: An old disease but a new threat to Africa. *Int. J. Tuberc. Lung Dis.* **8**, 924–937 (2004).

171. Piersimoni, C. & Scarparo, C. Pulmonary infections associated with non-tuberculous mycobacteria in immunocompetent patients. *Lancet Infect. Dis.* **8**, 323–334 (2008).

172. Kwak, N. *et al.* Mycobacterium abscessus pulmonary disease: individual patient data meta-analysis. *Eur. Respir. J.* **54**, 1801991 (2019).

173. Haworth, C. S. *et al.* British Thoracic Society guidelines for the management of non-tuberculous mycobacterial pulmonary disease (NTM-PD). *Thorax* **72**, ii1–ii64 (2017).

174. Sharma, S. & Upadhyay, V. Epidemiology, diagnosis & treatment of non-tuberculous mycobacterial diseases. *Indian J. Med. Res.* **152**, 185 (2020).

175. Larsen, S. E., Williams, B. D., Rais, M., Coler, R. N. & Baldwin, S. L. It Takes a Village: The Multifaceted Immune Response to Mycobacterium tuberculosis Infection and Vaccine-Induced Immunity. *Front. Immunol.* **13**, 1–31 (2022).

176. Bannantine, J. P. *et al.* A rational framework for evaluating the next generation of

vaccines against Mycobacterium avium subspecies paratuberculosis. *Front. Cell. Infect. Microbiol.* **4**, 1–11 (2014).

177. Sagawa, Z. K. *et al.* Safety and immunogenicity of a thermostable ID93 + GLA-SE tuberculosis vaccine candidate in healthy adults. *Nat. Commun.* **14**, 1138 (2023).

178. Ong, E., He, Y. & Yang, Z. Epitope promiscuity and population coverage of Mycobacterium tuberculosis protein antigens in current subunit vaccines under development. *Infect. Genet. Evol.* **80**, 104186 (2020).

179. Tait, D. R. *et al.* Final Analysis of a Trial of M72/AS01 E Vaccine to Prevent Tuberculosis . *N. Engl. J. Med.* **381**, 2429–2439 (2019).

180. Scriba, T. J., Netea, M. G. & Ginsberg, A. M. Key recent advances in TB vaccine development and understanding of protective immune responses against Mycobacterium tuberculosis. *Semin. Immunol.* **50**, 101431 (2020).

181. Day, T. A. *et al.* Safety and immunogenicity of the adjunct therapeutic vaccine ID93 + GLA-SE in adults who have completed treatment for tuberculosis: a randomised, double-blind, placebo-controlled, phase 2a trial. *Lancet Respir. Med.* **9**, 373–386 (2021).

182. Ates, L. S. New insights into the mycobacterial PE and PPE proteins provide a framework for future research. *Mol. Microbiol.* **113**, 4–21 (2020).

183. Martinot, A. J. Microbial Offense vs Host Defense: Who Controls the TB Granuloma? *Vet. Pathol.* **55**, 14–26 (2018).

184. Coler, R. N. *et al.* The TLR-4 agonist adjuvant, GLA-SE, improves magnitude and quality of immune responses elicited by the ID93 tuberculosis vaccine: first-in-human trial. *npj Vaccines* **3**, (2018).

185. Batista, M. T. *et al.* Gut adhesive Bacillus subtilis spores as a platform for mucosal delivery of antigens. *Infect. Immun.* **82**, 1414–1423 (2014).

186. Huang, J. M. *et al.* Mucosal delivery of antigens using adsorption to bacterial spores. *Vaccine* **28**, 1021–1030 (2010).

187. Fan, X.-Y. & Lowrie, D. B. Where are the RNA vaccines for TB? *Emerg. Microbes Infect.* **10**, 1217–1218 (2021).

188. Abdellrazeq, G. S. *et al.* A peptide-based vaccine for Mycobacterium avium subspecies paratuberculosis. *Vaccine* **37**, 2783–2790 (2019).

189. Lamrabet, O., Merhej, V., Pontarotti, P., Raoult, D. & Drancourt, M. The Genealogic Tree of Mycobacteria Reveals a Long-Standing Sympatric Life into Free-Living Protozoa. *PLoS One* **7**, e34754 (2012).

190. Mostowy, S. *et al.* Revisiting the evolution of Mycobacterium bovis. *J. Bacteriol.* **187**, 6386–6395 (2005).

191. Kaushal, D., Mehra, S., Didier, P. J. & Lackner, A. A. The non-human primate model of tuberculosis. *J. Med. Primatol.* **41**, 191–201 (2012).

192. Lombard, J. E. *et al.* Human-to-Cattle Mycobacterium tuberculosis Complex Transmission in the United States. *Front. Vet. Sci.* **8**, 1–11 (2021).

193. Naranjo, V., Gortazar, C., Vicente, J. & de la Fuente, J. Evidence of the role of European wild boar as a reservoir of Mycobacterium tuberculosis complex. *Vet. Microbiol.* **127**, 1–9 (2008).

194. Malone, K. M. *et al.* Comparative 'omics analyses differentiate mycobacterium tuberculosis and mycobacterium bovis and reveal distinct macrophage responses to infection with the human and bovine tubercle bacilli. *Microb. Genomics* **4**, (2018).

195. Wobeser, G. Bovine tuberculosis in Canadian wildlife: an updated history. *Can. Vet. J. = La Rev. Vet. Can.* **50**, 1169–76 (2009).

196. Ayele, W. Y., Neill, S. D., Zinsstag, J., Weiss, M. G. & Pavlik, I. Bovine tuberculosis: an old disease but a new threat to Africa. *Int. J. Tuberc. Lung Dis.* **8**, 924–37 (2004).

197. Sunstrum, J. *et al.* Zoonotic Mycobacterium bovis Disease in Deer Hunters -- Michigan, 2002-2017. *Morb. Mortal. Wkly. Rep.* **68**, 807–808 (2019).

198. Rehren, G., Walters, S., Fontan, P., Smith, I. & Zárraga, A. M. Differential gene expression between Mycobacterium bovis and Mycobacterium tuberculosis. *Tuberculosis* **87**, 347–359 (2007).

199. Sohaskey, C. D. & Modesti, L. Differences in nitrate reduction between mycobacterium tuberculosis and mycobacterium bovis are due to differential expression of both narGHJI and narK2. *FEMS Microbiol. Lett.* **290**, 129–134 (2009).

200. Lofthouse, E. K. *et al.* Systems-based approaches to probing metabolic variation within the Mycobacterium tuberculosis complex. *PLoS One* **8**, 1–14 (2013).

201. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).

202. Saund, K., Lapp, Z., Thiede, S. N., Pirani, A. & Snitkin, E. S. Prewas: Data pre-processing for more informative bacterial gwas. *Microb. Genomics* **6**, 1–8 (2020).

203. Collins, C. & Didelot, X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLOS*

Comput. Biol. **14**, e1005958 (2018).

204. Muñoz, S., Rivas-Santiago, B. & Enciso, J. A. Mycobacterium tuberculosis Entry into Mast Cells Through Cholesterol-rich Membrane Microdomains. *Scand. J. Immunol.* **70**, 256–263 (2009).

205. Kim, M. J. *et al.* Caseation of human tuberculosis granulomas correlates with elevated host lipid metabolism. *EMBO Mol. Med.* **2**, 258–274 (2010).

206. Gatfield, J. & Pieters, J. Essential role for cholesterol in entry of mycobacteria into macrophages. *Science (80-.).* **288**, 1647–1650 (2000).

207. Moopanar, K. & Mvubu, N. E. Lineage-specific differences in lipid metabolism and its impact on clinical strains of Mycobacterium tuberculosis. *Microb. Pathog.* **146**, 104250 (2020).

208. Fieweger, Wilburn & VanderVen. Comparing the Metabolic Capabilities of Bacteria in the Mycobacterium tuberculosis Complex. *Microorganisms* **7**, 177 (2019).

209. Dong, Y. *et al.* Genomic analysis of diversity, biogeography, and drug resistance in Mycobacterium bovis. *Transbound. Emerg. Dis.* **69**, e2769–e2778 (2022).

210. Ewels, P. SRA-Explorer.

211. Foster, I. Globus Online: Accelerating and Democratizing Science through Cloud-Based Services. *IEEE Internet Comput.* **15**, 70–73 (2011).

212. Allen, B. *et al.* Software as a service for data scientists. *Commun. ACM* **55**, 81–88 (2012).

213. Seemann, T. snippy: fast bacterial variant calling from NGS reads. *Https://Github.Com/Tseemann/Snippy* (2015).

214. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *bioRxiv* 1–13 (2019) doi:10.1101/762302.

215. Darriba, Di. *et al.* ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Mol. Biol. Evol.* **37**, 291–294 (2020).

216. Allaire, J. RStudio: integrated development for R. *RStudio Team* (2012).

217. RDC, T. A Language and Environment for Statistical Computing. *Vienna, Austria: R Foundation for Statistical Computing* (2010).

218. Knaus, B. J. & Grünwald, N. J. vcfr: a package to manipulate and visualize variant call format data in R. *Mol. Ecol. Resour.* **17**, 44–53 (2017).

219. Anaconda. Anaconda Software Distribution. *Computer software* Vers. 2-2.4.0. (2016).

220. GCC Team. GCC, the GNU Compiler Collection. (2013).

221. Gabriel, E. *et al.* Open MPI: Goals, Concept, and Design of a Next Generation MPI Implementation. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol. 3241 97–104 (2004).

222. Wickham, H., Hester, J. & Chang, W. Tools to make developing R packages easier - Package 'devtools'. (2021).

223. Saund, K., Lapp, Z., Thiede, S. N., Pirani, A. & Snitkin, E. S. Prewas: Data pre-processing for more informative bacterial gwas. *Microbial Genomics* vol. 6 1–8 (2020).

224. Collins, C. & Didelot, X. treeWAS: A phylogenetic tree-based approach to genome-wide association studies in microbes. (2022).

225. Lipworth, S. *et al.* SNP-IT tool for identifying subspecies and associated lineages of Mycobacterium tuberculosis complex. *Emerg. Infect. Dis.* **25**, 482–488 (2019).

226. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin).* **6**, 80–92 (2012).

227. Collins, C. How treeWAS works: Tests of Association. *GitHub repo for treeWAS* https://github.com/caitiecollins/treeWAS/wiki/1.-How-treeWAS-Works#tests-of-association (2018).

228. Glickman, M. S. & Jacobs, W. R. Microbial pathogenesis of Mycobacterium tuberculosis: Dawn of a discipline. *Cell* **104**, 477–485 (2001).

229. Wipperman, M. F., Yang, M., Thomas, S. T. & Sampson, N. S. Shrinking the fadE proteome of mycobacterium tuberculosis: Insights into cholesterol metabolism through identification of an α2β2 heterotetrameric acyl coenzyme a dehydrogenase family. *J. Bacteriol.* **195**, 4331–4341 (2013).

230. Griffin, J. E. *et al.* High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. *PLoS Pathog.* **7**, 1–9 (2011).

231. Pandey, A. K. & Sassetti, C. M. Mycobacterial persistence requires the utilization of host cholesterol. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 4376–4380 (2008).

232. Malm, S. *et al.* New Mycobacterium tuberculosis Complex Sublineage, Brazzaville, Congo. *Emerg. Infect. Dis.* **23**, 423–429 (2017).

233. Ehebauer, M. T. *et al.* Characterization of the Mycobacterial Acyl-CoA Carboxylase Holo Complexes Reveals Their Functional Expansion into Amino Acid Catabolism. *PLOS Pathog.* **11**, e1004623 (2015).

234. Schwenk, S., Moores, A., Nobeli, I., McHugh, T. D. & Arnvig, K. B. Cell-wall synthesis and ribosome maturation are co-regulated by an RNA switch in Mycobacterium tuberculosis. *Nucleic Acids Res.* **46**, 5837–5849 (2018).

235. Gonzalo-Asensio, J. *et al.* Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 11491–11496 (2014).

236. Moopanar, K. & Mvubu, N. E. Lineage-specific differences in lipid metabolism and its impact on clinical strains of Mycobacterium tuberculosis. *Microb. Pathog.* **146**, (2020).

237. Fernandez, M. L. & Volek, J. S. Guinea pigs: A suitable animal model to study lipoprotein metabolism, atherosclerosis and inflammation. *Nutr. Metab.* **3**, 1–6 (2006).

238. Orme, I. M. & Ordway, D. J. Mouse and Guinea Pig Models of Tuberculosis. in *Tuberculosis and the Tubercle Bacillus* 143–162 (ASM Press, 2017). doi:10.1128/9781555819569.ch7.

239. Cooper, A. M. Mouse model of tuberculosis. *Cold Spring Harb. Perspect. Med.* **5**, 1–8 (2015).

240. Oppi, S., Lüscher, T. F. & Stein, S. Mouse Models for Atherosclerosis Research—Which Is My Line? *Front. Cardiovasc. Med.* **6**, 1–8 (2019).

241. Gordon, S. M. *et al.* A Comparison of the Mouse and Human Lipoproteome: Suitability of the Mouse Model for Studies of Human Lipoproteins. *J. Proteome Res.* **14**, 2686–2695 (2015).

242. Duran, M. J., Kannampuzha-Francis, J., Nydam, D. & Behling-Kelly, E. Characterization of Particle Size Distribution of Plasma Lipoproteins in Dairy Cattle Using High-Resolution Polyacrylamide Electrophoresis. *Front. Anim. Sci.* **2**, 1–10 (2021).

243. Inoue, M. *et al.* High-density lipoprotein suppresses tumor necrosis factor alpha production by mycobacteria-infected human macrophages. *Sci. Rep.* **8**, 1–11 (2018).

244. Dong, H., Lv, Y., Sreevatsan, S., Zhao, D. & Zhou, X. Differences in pathogenicity of three animal isolates of Mycobacterium species in a mouse model. *PLoS One* **12**, 1–17 (2017).

245. Medina, E., Ryan, L., LaCourse, R. & North, R. J. Superior virulence of Mycobacterium bovis over Mycobacterium tuberculosis (Mtb) for Mtb-resistant and Mtb-susceptible mice is manifest as an ability to cause extrapulmonary disease. *Tuberculosis* **86**, 20–27 (2006).

246. Brunham, R. C., Plummer, F. A. & Stephens, R. S. Bacterial antigenic variation, host immune response, and pathogen-host coevolution. *Infect. Immun.* **61**, 2273–2276 (1993).

247. Ottenhoff, T. H. M. The knowns and unknowns of the immunopathogenesis of tuberculosis. *Int. J. Tuberc. Lung Dis.* **16**, 1424–1432 (2012).

248. Ramaiah, A. *et al.* Evidence for highly variable, region-specific patterns of T-cell epitope mutations accumulating in mycobacterium tuberculosis strains. *Front. Immunol.* **10**, 1–18 (2019).

249. Chan, J. *et al.* The role of B cells and humoral immunity in Mycobacterium tuberculosis infection. *Semin. Immunol.* **26**, 588–600 (2014).

250. Jasenosky, L. D., Scriba, T. J., Hanekom, W. A. & Goldfeld, A. E. T cells and adaptive immunity to Mycobacterium tuberculosis in humans. *Immunol. Rev.* **264**, 74–87 (2015).

251. Coscolla, M. *et al.* M. tuberculosis T Cell Epitope Analysis Reveals Paucity of Antigenic Variation and Identifies Rare Variable TB Antigens. *Cell Host Microbe* **18**, 538–548 (2015).

252. Baena, A. & Porcelli, S. A. Evasion and subversion of antigen presentation by Mycobacterium tuberculosis: REVIEW ARTICLE. *Tissue Antigens* **74**, 189–204 (2009).

253. Orme, I. M. Development of new vaccines and drugs for TB: limitations and potential strategic errors. *Future Microbiol.* **6**, 161–77 (2011).

254. Waters, W. R. *et al.* Virulence of two strains of mycobacterium bovis in cattle following aerosol infection. *J. Comp. Pathol.* **151**, 410–419 (2014).

255. Brenner, E. P. *et al.* Genome Sequences of Mycobacterium tuberculosis Biovar bovis Strains Ravenel and 10-7428. *Microbiol. Resour. Announc.* **10**, 11–12 (2021).

256. Brenner, E. P. *et al.* Mycobacterium bovis Strain Ravenel Is Attenuated in Cattle. *Pathogens* **11**, 1330 (2022).

257. Narvskaya, O. *et al.* First insight into the whole-genome sequence variations in Mycobacterium bovis BCG-1 (Russia) vaccine seed lots and their progeny clinical isolates from children with BCG-induced adverse events. *BMC Genomics* **21**, 1–12 (2020).

258. Farrell, D., Crispell, J. & Gordon, S. V. Updated functional annotation of the Mycobacterium bovis AF2122/97 reference genome. *Access Microbiol.* **2**, 2–4 (2020).

259. Vita, R. *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343 (2019).

260. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

261. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 1–9 (2009).

262. Pancotti, C. *et al.* Predicting protein stability changes upon single-point mutation: A thorough comparison of the available tools on a new dataset. *Brief. Bioinform.* **23**, 1–12 (2022).

263. Montanucci, L., Capriotti, E., Frank, Y., Ben-Tal, N. & Fariselli, P. DDGun: An untrained method for the prediction of protein stability changes upon single and multiple point variations. *BMC Bioinformatics* **20**, 1–10 (2019).

264. Chen, Y. *et al.* PremPS: Predicting the impact of missense mutations on protein stability. *PLoS Comput. Biol.* **16**, 1–22 (2020).

265. Fariselli, P., Martelli, P. L., Savojardo, C. & Casadio, R. INPS: Predicting the impact of non-synonymous variations on protein stability from sequence. *Bioinformatics* **31**, 2816–2821 (2015).

266. Rodrigues, C. H. M., Pires, D. E. V. & Ascher, D. B. DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Sci.* **30**, 60–69 (2021).

267. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).

268. Russell-Goldman, E., Xu, J., Wang, X., Chan, J. & Tufariello, J. A. M. A Mycobacterium tuberculosis Rpf double-knockout strain exhibits profound defects in reactivation from chronic tuberculosis and innate immunity phenotypes. *Infect. Immun.* **76**, 4269–4281 (2008).

269. Khare, S., Hondalus, M. K., Nunes, J., Bloom, B. R. & Garry Adams, L. Mycobacterium bovis ΔleuD auxotroph-induced protective immunity against tissue colonization, burden and distribution in cattle intranasally challenged with Mycobacterium bovis Ravenel S. *Vaccine* **25**, 1743–1755 (2007).

270. Ruggiero, A. *et al.* Crystal Structure of the Resuscitation-Promoting Factor ΔDUFRpfB from M. tuberculosis. *J. Mol. Biol.* **385**, 153–162 (2009).

271. Lee, J., Kim, J., Lee, J., Shin, S. J. & Shin, E.-C. DNA immunization of Mycobacterium tuberculosis resuscitation-promoting factor B elicits polyfunctional CD8 + T cell responses. *Clin. Exp. Vaccine Res.* **3**, 235 (2014).

272. Gupta, P. *et al.* A fragment-based approach to assess the ligandability of ArgB, ArgC, ArgD and ArgF in the L-arginine biosynthetic pathway of Mycobacterium tuberculosis. *Comput. Struct. Biotechnol. J.* **19**, 3491–3506 (2021).

273. Dejesus, M. A. *et al.* Comprehensive essentiality analysis of the Mycobacterium tuberculosis genome via saturating transposon mutagenesis. *MBio* **8**, (2017).

274. Tiwari, S. *et al.* Arginine-deprivation-induced oxidative damage sterilizes Mycobacterium tuberculosis. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 9779–9784 (2018).

275. Farrell, D. *et al.* Integrated computational prediction and experimental validation identifies promiscuous T cell epitopes in the proteome of Mycobacterium bovis. *Microb. genomics* (2016) doi:10.1099/mgen.0.000071.

276. Li, Q. J. *et al.* Compensatory mutations of rifampin resistance are associated with transmission of multidrug-resistant Mycobacterium tuberculosis Beijing genotype strains in China. *Antimicrob. Agents Chemother.* **60**, 2807–2812 (2016).

277. National Center for Biotechnology Information. PubChem Compound Summary for CID 135398735 (Rifampicin). https://pubchem.ncbi.nlm.nih.gov/compound/Rifampicin (2022).

278. Brandis, G., Wrande, M., Liljas, L. & Hughes, D. Fitness-compensatory mutations in rifampicin-resistant RNA polymerase. *Mol. Microbiol.* **85**, 142–151 (2012).

279. Comas, I. *et al.* Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nat. Genet.* **44**, 106–110 (2012).

280. Chakhaiyar, P. *et al.* Regions of high antigenicity within the hypothetical PPE major polymorphic tandem repeat open-reading frame, Rv2608, show a differential humoral response and a low T cell response in various categories of patients with tuberculosis. *J. Infect. Dis.* **190**, 1237–1244 (2004).

281. Bertholet, S. *et al.* Identification of Human T Cell Antigens for the Development of Vaccines against Mycobacterium tuberculosis . *J. Immunol.* **181**, 7948–7957 (2008).

282. Bentley-Hibbert, S. I., Quan, X., Newman, T., Huygen, K. & Godfrey, H. P. Pathophysiology of antigen 85 in patients with active tuberculosis: Antigen 85 circulates as complexes with fibronectin and immunoglobulin G (Infection and Immunity 67:2 (581-588)). *Infect. Immun.* **67**, 2050 (1999).

283. Wiker, H. G. & Harboe, M. The antigen 85 complex: A major secretion product of Mycobacterium tuberculosis. *Microbiol. Rev.* **56**, 648–661 (1992).

284. Botella, H. *et al.* Mycobacterium tuberculosis protease MarP activates a peptidoglycan hydrolase during acid stress . *EMBO J.* **36**, 536–548 (2017).

285. Vandal, O. H., Pierini, L. M., Schnappinger, D., Nathan, C. F. & Ehrt, S. A membrane protein preserves intrabacterial pH in intraphagosomal Mycobacterium tuberculosis. *Nat. Med.* **14**, 849–854 (2008).

286. Biswas, T. *et al.* Structural insight into serine protease Rv3671c that Protects M.

tuberculosis from oxidative and acidic stress. *Structure* **18**, 1353–63 (2010).

287. Fenn, K., Wong, C. T. & Darbari, V. C. Mycobacterium tuberculosis Uses Mce Proteins to Interfere With Host Cell Signaling. *Front. Mol. Biosci.* **6**, 1–6 (2020).

288. Marabotti, A., Del Prete, E., Scafuri, B. & Facchiano, A. Performance of Web tools for predicting changes in protein stability caused by mutations. *BMC Bioinformatics* **22**, 1–19 (2021).

289. Kellogg, E. H., Leaver-Fay, A. & Baker, D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins Struct. Funct. Bioinforma.* **79**, 830–838 (2011).

290. Judd, J. A. *et al.* A Mycobacterial Systems Resource for the Research Community. *MBio* **12**, 1–15 (2021).

291. Veyrier, F., Pletzer, D., Turenne, C. & Behr, M. A. Phylogenetic detection of horizontal gene transfer during the step-wise genesis of Mycobacterium tuberculosis. *BMC Evol. Biol.* **9**, 1–14 (2009).

292. Otal, I., Martin, C., Vincent-Levy-Frebault, V., Thierry, D. & Gicquel, B. Restriction fragment length polymorphism analysis using IS6110 as an epidemiological marker in tuberculosis. *J. Clin. Microbiol.* **29**, 1252–1254 (1991).

293. Groenen, P. M. A., Bunschoten, A. E., Soolingen, D. van & Errtbden, J. D. A. va. Nature of DNA polymorphism in the direct repeat cluster of Mycobacterium tuberculosis; application for strain differentiation by a novel typing method. *Mol. Microbiol.* **10**, 1057–1065 (1993).

294. Barrangou, R. *et al.* Against Viruses in Prokaryotes. *Science (80-.).* **315**, 1709–1712 (2007).

295. Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67–71 (2010).

296. Grüschow, S., Athukoralage, J. S., Graham, S., Hoogeboom, T. & White, M. F. Cyclic oligoadenylate signalling mediates Mycobacterium tuberculosis CRISPR defence. *Nucleic Acids Res.* **47**, 9259–9270 (2019).

297. Wei, W. *et al.* Mycobacterium tuberculosis type III-A CRISPR/Cas system crRNA and its maturation have atypical features. *FASEB J.* **33**, 1496–1509 (2019).

298. Couvin, D. *et al.* CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.* **46**, W246–W251 (2018).

299. Zhang, Q. & Ye, Y. Not all predicted CRISPR-Cas systems are equal: Isolated cas genes and classes of CRISPR like elements. *BMC Bioinformatics* **18**, 1–12 (2017).

300. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**, D20–D26 (2022).

301. Olson, R. D. *et al.* Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Res.* **51**, 678–689 (2022).

302. Brettin, T. *et al.* RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci. Rep.* **5**, (2015).

303. Tatusova, T. *et al.* NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **44**, 6614–6624 (2016).

304. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

305. Darriba, Di. *et al.* ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Mol. Biol. Evol.* **37**, 291–294 (2020).

306. Inkscape. Inkscape Project open-source graphics editor. https://inkscape.org/.

307. Andrew Rambaut. FigTree. http://tree.bio.ed.ac.uk/software/figtree/.

308. Gasteiger, E. *et al.* ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **31**, 3784–3788 (2003).

309. ATCC Mycobacterium avium strain Chester. https://www.atcc.org/products/700898.

310. Horan, K. L. *et al.* Isolation of the genome sequence strain Mycobacterium avium 104 from multiple patients over a 17-year period. *J. Clin. Microbiol.* **44**, 783–789 (2006).

311. Almendros, C., Nobrega, F. L., McKenzie, R. E. & Brouns, S. J. J. Cas4-Cas1 fusions drive efficient PAM selection and control CRISPR adaptation. *Nucleic Acids Res.* **47**, 5223–5230 (2019).

312. Makarova, K. S. *et al.* Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **18**, 67–83 (2020).

313. Korea Institute of Geoscience and Mineral Resources. Mycobacterium sp. SM1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/genome/GCA_018361265.1/ (2021).

314. Pinilla-Redondo, R. *et al.* Type IV CRISPR-Cas systems are highly diverse and involved in competition between plasmids. *Nucleic Acids Res.* **48**, 2000–2012 (2020).

315. Panda, A., Drancourt, M., Tuller, T. & Pontarotti, P. Genome-wide analysis of horizontally acquired genes in the genus Mycobacterium. *Sci. Rep.* **8**, 1–13 (2018).

316. Semret, M. *et al.* Extensive genomic polymorphism within Mycobacterium avium. *J. Bacteriol.* **186**, 6332–6334 (2004).

317. Bannantine, J. P. *et al.* Genome sequencing of ovine isolates of Mycobacterium avium subspecies paratuberculosis offers insights into host association. *BMC Genomics* **13**, (2012).

318. Posso-Osorio, I. *et al.* Mycobacterium malmoense: an unusual pathogen causing endocarditis, a case report and literature review. *IDCases* **22**, 38–41 (2020).

319. van Ingen, J. *et al.* Mycobacterium riyadhense sp. nov., a non-tuberculous species identified as Mycobacterium tuberculosis complex by a commercial line-probe assay. *Int. J. Syst. Evol. Microbiol.* **59**, 1049–1053 (2009).

320. Samuel Imisi Awala, Joo-Han Gwak, Chanmee Seo, Lorraine A. Bellosillo, Yong-Man Kim, Ok-Ja Si, S.-K. & Rhee. Short-chain Alkanes Consumption in Mycobacterium sp. SM1 Isolated from an Acidic Geothermal Environmen. in *MSK 2022: Intertional Meeting of the Microbiological Society of Korea* B091 (2022).

321. Hochstrasser, M. L., Taylor, D. W., Kornfeld, J. E., Nogales, E. & Doudna, J. A. DNA Targeting by a Minimal CRISPR RNA-Guided Cascade. *Mol. Cell* **63**, 840–851 (2016).

322. Brendel, J. *et al.* A complex of cas proteins 5, 6, and 7 is required for the biogenesis and stability of clustered regularly interspaced short palindromic repeats (CRISPR)-derived RNAs (crRNAs) in haloferax volcanii. *J. Biol. Chem.* **289**, 7164–7177 (2014).

323. Lundgren, M., Charpentier, E. & Fineran, P. C. CRISPR: Methods and protocols. *Cris. Methods Protoc.* 1–366 (2015) doi:10.1007/978-1-4939-2687-9.

324. Garside, E. L. *et al.* Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. *Rna* **18**, 2020–2028 (2012).

325. Shangguan, Q., Graham, S., Sundaramoorthy, R. & White, M. F. Structure and mechanism of the type I-G CRISPR effector. *Nucleic Acids Res.* **50**, 11214–11228 (2022).

326. Samai, P., Smith, P. & Shuman, S. Structure of a CRISPR-associated protein Cas2 from Desulfovibrio vulgaris. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **66**, 1552–1556 (2010).

327. Coitinho, C. *et al.* First case of Mycobacterium heckeshornense cavitary lung disease in the Latin America and Caribbean region. *New Microbes New Infect.* **9**, 63–65 (2016).

328. Sanchini, A. *et al.* A hypervariable genomic island identified in clinical and environmental

Mycobacterium avium subsp. hominissuis isolates from Germany. *Int. J. Med. Microbiol.* **306**, 495–503 (2016).

329. Pinilla-Redondo, R. *et al.* CRISPR-Cas systems are widespread accessory elements across bacterial and archaeal plasmids. *Nucleic Acids Res.* **50**, 4315–4328 (2022).

330. Wheatley, R. M. & MacLean, R. C. CRISPR-Cas systems restrict horizontal gene transfer in Pseudomonas aeruginosa. *ISME J.* **15**, 1420–1433 (2021).

331. Ummels, R. *et al.* Identification of a novel conjugative plasmid in mycobacteria that requires both type IV and type VII secretion. *MBio* **5**, 1–8 (2014).

332. Dumas, E. *et al.* Mycobacterial pan-genome analysis suggests important role of plasmids in the radiation of type VII secretion systems. *Genome Biol. Evol.* **8**, 387–402 (2016).

333. Beloglazova, N. *et al.* Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference. *EMBO J.* **30**, 4616–4627 (2011).

334. Zhang, J., Kasciukovic, T. & White, M. F. The CRISPR Associated Protein Cas4 Is a 5′ to 3′ DNA Exonuclease with an Iron-Sulfur Cluster. *PLoS One* **7**, (2012).

335. Hrle, A. *et al.* Structure and RNA-binding properties of the type III-A CRISPR-associated protein Csm3. *RNA Biol.* **10**, 1670–1678 (2013).