bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

A New Phylogenetic Framework for the Animal-adapted *Mycobacterium tuberculosis* Complex

Running title

Evolutionary history of animal MTBC

Daniela Brites[1,2*], Chloe Loiseau[1,2], Fabrizio Menardo[1,2], Sonia Borrell[1,2], Maria Beatrice Boniotti[3], Robin Warren[4], Anzaan Dippenaar[4], Sven David Charles Parsons[4], Christian Beisel[5], Marcel A. Behr[6], Janet A Fyfe[7], Mireia Coscolla[8†], Sebastien Gagneux[1,2†]

*Correspondence: d.brites@swisstph.ch

[†]equal contribution

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland

[2] University of Basel, Basel, Switzerland

[3] Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia-Romagna: Centro Nazionale di Referenza per la Tubercolosi Bovina, Brescia, Italy

[4] SAMRC Centre for TB Research; DST/NRF Centre of Excellence for Biomedical Tuberculosis Research; Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, South Africa.

[5] Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland.

[6] McGill International TB Centre; Infectious Diseases and Immunity in Global Health, McGill University Health Centre Research Institute. Montréal, Canada

[7] Mycobacterium Reference Laboratory, Victoria Infectious Diseases Reference Laboratory, Peter Doherty Institute, Melbourne, Victoria, Australia

[8] Institute for Integrative Systems Biology (I2SysBio), University of Valencia-CSIC, Valencia, Spain

Keywords: host-pathogen interactions, specificity, host range, genetic diversity, whole-genome sequencing

bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Abstract

Tuberculosis (TB) affects humans and other animals and is caused by bacteria from the *Mycobacterium tuberculosis* complex (MTBC). Previous studies have shown that there are at least nine members of the MTBC infecting animals other than humans; these have also been referred to as ecotypes. However, the ecology and the evolution of these animal-adapted MTBC ecotypes are poorly understood. Here we screened 12,886 publicly available MTBC genomes and newly sequenced 17 animal-adapted MTBC strains, gathering a total of 529 genomes of animal-adapted MTBC strains. Phylogenomic and comparative analyses confirm that the animal-adapted MTBC members are paraphyletic with some members more closely related to the human-adapted *Mycobacterium africanum* Lineage 6 than to other animal-adapted strains. Furthermore, we identified four main animal-adapted MTBC clades that might correspond to four main host shifts; two of these clades are proposed to reflect independent cattle domestication events. Contrary to what would be expected from an obligate pathogen, MTBC nucleotide diversity was not positively correlated with host phylogenetic distances, suggesting that host tropism in the animal-adapted MTBC seems to be driven more by contact rates and demographic aspects of the host population rather than host relatedness. By combining phylogenomics with ecological data, we propose an evolutionary scenario in which the ancestor of Lineage 6 and all animal-adapted MTBC ecotypes was a generalist pathogen that subsequently adapted to different host species. This study provides a new phylogenetic framework to better understand the evolution of the different ecotypes of the MTBC and guide future work aimed at elucidating the molecular mechanisms underlying host specificity.

bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Introduction

Tuberculosis (TB) remains a major concern both from a global health and economic point of view. With an estimated 10.4 million new human cases and 1.7 million fatalities every year, TB kills more people than any other infectious disease (WHO 2014). Moreover, bovine TB is responsible for an estimated US$3 billion annual economic loss in livestock production globally (Waters et al. 2012) and represents an ongoing threat for zoonotic TB in humans (Olea-Popelka et al. 2017). The causative agents of TB in humans and animals are a group of closely related acid-fast bacilli collectively known as the *Mycobacterium tuberculosis* complex (MTBC) (Brites et al. 2017, Malone et al. 2017). The human-adapted MTBC comprises five main phylogenetic lineages generally referred to as *Mycobacterium tuberculosis* sensu stricto (i.e. MTBC lineages 1-4 and lineage 7) and two lineages traditionally known as *Mycobacterium africanum* (i.e. MTBC lineages 5 and 6) (de Jong et al. 2010, Brites et al. 2017, Yeboah-Manu et al. 2017). Among the animal-adapted members of the MTBC, some primarily infect wild mammal species (Malone et al. 2017). These include *Mycobacterium microti* (a pathogen of voles) (Brodin et al. 2002), *Mycobacterium pinnipedii* (seals and sea lions) (Cousins et al. 2003), *Mycobacterium orygis* (antelopes) (van Ingen et al. 2012) and the "dassie bacillus" (rock hyrax) (Mostowy et al. 2004), which have been known for a long time, as well as the more recently discovered *Mycobacterium mungi* (mongooses) (Alexander et al. 2010), *Mycobacterium suricattae* (meerkats) (Parsons et al. 2013) and the "chimpanzee bacillus" (chimpanzees) (Coscolla et al. 2013). *Mycobacterium bovis* and *Mycobacterium caprae* on the other hand are mainly found in domesticated cattle and goats, but do also frequently spill over into many different wild animal species (Malone et al. 2017). *Mycobacterium canettii* is also considered part of the MTBC based on nucleotide identity; however *M. canettii* is likely an environmental microbe only occasionally causing opportunistic infections in humans (Koeck et al. 2010, Supply et al. 2013). We therefore use the term MTBC to refer to all the above mentioned members except *M. canettii*. Many of the names of the animal-adapted MTBC species were originally coined based on the animal they were first isolated from. For example, *M. orygis* was first identified in a captive oryx (van Soolingen et al. 1994) but has since then been isolated from many different host species including humans (van Ingen et al. 2012). Thus, the actual host range of *M. orygis* remains ill-defined (Malone et al. 2017). Similarly, for many of the animal-adapted members of the MTBC, only a few representatives have been isolated so far (e.g. only one in the case of the chimpanzee bacillus), limiting inferences with respect to the host range of these microbes. Moreover when studying host tropism, it is important to differentiate between maintenance hosts, in which the corresponding MTBC members traverse their full life cycle, including the transmission to secondary hosts, and spillover hosts, in which the infection leads to a dead end with no onward transmission (Malone et al. 2017). For example, *M. tuberculosis* sensu stricto is well adapted to transmit from human to human (Brites et al. 2015) and is occasionally isolated from cattle (Ameni et al. 2010). However *M. tuberculosis* sensu stricto is avirulent in cattle (Whelan et al. 2010, Villarreal-Ramos et al. 2018) and transmission from an animal back to humans is extremely rare (Murphree et al. 2011). Conversely, *M. bovis* is well adapted to transmit among cattle and does occasionally infect humans, mainly through the consumption of raw milk (Muller et al. 2013) or close contact with infected cattle, but transmission of *M. bovis* among immuno-competent humans is similarly uncommon (Blazquez et al. 1997).

The different members and phylogenetic lineages of the MTBC share a high nucleotide identity (>99.9%), and it has recently been suggested that they should be regarded as part of the same bacterial species (Riojas et al. 2018). The fact that these lineages also occupy different ecological niches, which is reflected in their host-specific tropism, supports a

bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

distinction into separate ecotypes (Smith et al. 2005). Yet, the host range of many of these animal-adapted MTBC members remain poorly defined, with respect to both maintenance and spillover hosts (Malone et al. 2017). In this study, we present and discuss a new phylogenetic framework based on whole genome sequences covering all known MTBC ecotypes. Based on this novel framework, we challenge previous assumptions regarding the evolutionary history of the MTBC as a whole, and point to new research directions for uncovering the molecular basis of host tropism in one of the most important bacterial pathogens.

bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Methods

MTBC genome dataset

We downloaded 12,886 genomes previously published and accessible from the sequence read archive (SRA) repository by December 2017 (Menardo et al. 2018). After mapping and calling of variants (see below), phylogenetic SNPs as in (Steiner et al. 2014) were used to classify genomes into human-adapted MTBC if they belonged to lineages 1 to 7 and if not, into non-human (hereafter referred to as "animal") MTBC. All genomes determined as animal MTBC, as well as those classified as L5 or L6, were used for downstream analysis. We have furthermore newly sequenced four *M. orygis* genomes, two dassie bacillus genomes, eight *M. microti*, two *M. bovis* and one *M. caprae* (Supplementary Table 1). For downstream analysis, we selected the genomes published in (Comas et al. 2013) as representatives of other human MTBC, giving a total 851 genomes used in the downstream analysis (Supplementary Table 1).

Bacterial culture, DNA extraction and whole-genome sequencing

The MTBC isolates were grown in 7H9-Tween 0.05% medium (BD) +/- 40mM sodium pyruvate. We extracted genomic DNA after harvesting the bacterial cultures in the late exponential phase of growth using the CTAB method (Belisle et al. 1998). Sequencing libraries were prepared using NEXTERA XT DNA Preparation Kit (Illumina, San Diego, USA). Multiplexed libraries were paired-end sequenced on an Illumina HiSeq2500 instrument (Illumina, San Diego, USA) with 151 or 101 cycles at the Genomics Facility of the University of Basel. In the case of the *M. microti* isolates, DNA was obtained using the QIAamp DNA mini kit (Qiagen, Hilden, Germany) and libraries also prepared with the NEXTERA XT DNA Preparation Kit, were sequenced on an Illumina MiSeq using the Miseq Reagent Kit v2, 250-cycle paired-end run (Illumina, San Diego, USA).

Bioinformatics analysis:

Mapping and variant calling of Illumina reads

The obtained FASTQ files were processed with Trimmomatic v 0.33 (SLIDINGWINDOW: 5:20) (Bolger et al. 2014) to clip Illumina adaptors and trim low quality reads. Reads shorter than 20 bp were excluded from the downstream analysis. Overlapping paired-end reads were merged with SeqPrep v 1.2 (overlap size = 15) (https://github.com/jstjohn/SeqPrep). We used BWA v0.7.13 (mem algorithm) (Li et al. 2010) to align the reads to the reconstructed ancestral sequence of MTBC obtained as reported (Comas et al. 2010). Duplicated reads were marked by the Mark Duplicates module of Picard v 2.9.1 (https://github.com/broadinstitute/picard) and excluded. To avoid false positive calls, Pysam v 0.9.0 (https://github.com/pysam-developers/pysam) was used to exclude reads with alignment score lower than (0.93*read_length)-(read_length*4*0.07)), corresponding to more than 7 miss-matches per 100 bp. SNPs were called with Samtools v 1.2 mpileup (Li 2011) and VarScan v 2.4.1 (Koboldt et al. 2012) using the following thresholds: minimum mapping quality of 20, minimum base quality at a position of 20, minimum read depth at a position of 7x and without strand bias. Only SNPs considered to have reached fixation within an isolate were considered (at a within-host frequency of ≥90%). Conversely, when the SNP within-isolate frequency was ≤10% the ancestor state was called. Mixed infections or contaminations were discarded by excluding genomes with more than 1000 variable positions

bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

with within-host frequencies between 90% and 10% and genomes for which the number of within-host SNPs was higher than the number of fixed SNPs. Additionally, we excluded genomes with average coverage lower than 15x (after all the referred filtering steps). All SNPs were annotated using snpEff v4.11 (Cingolani et al. 2012), in accordance with the *M. tuberculosis* H37Rv reference annotation (NC_000962.3). SNPs falling in regions such as PPE and PE-PGRS, phages, insertion sequences and in regions with at least 50 bp identities to other regions in the genome were excluded from the analysis (Stucki et al. 2016). SNPs known to confer drug resistance as used in (Steiner et al. 2014) were also excluded from the analysis. Customized scripts were used to calculate mean coverageper gene corrected by the size of the gene. Gene deletions were determined as regions with no coverage to the reference genome. To identify deletions of regions and genes absent from the chromosome of H37Rv (e.g. RD900) unmapped reads resultant from the previous mapping procedure were mapped with reference to *M. canettii* (SRX002429) annotated using as reference NC_015848.1, following the same steps described above.

Phylogenetic reconstructionAll 851 selected genomes were used to produce an alignment containing only polymorphic sites. The alignment was used infer a Maximum likelihood phylogenetic tree using the MPI parallel version of RaxML (Stamatakis 2006). The model GTR implemented in RAxML was used, and 1,000 rapid bootstrap inferences followed by a thorough maximum-likelihood search (Stamatakis 2006) was performed in CIPRES (Miller et al. 2010).The best-scoring Maximum Likelihood topology is shown. The phylogeny was rooted using *M. canettii*. The topology was annotated using the package ggtree (Guangchuang et al. 2017) from R (Team 2018) and Adobe Illustrator CC. Taxa images were obtained from http://phylopic.org/. To remove redundancy and obtain a more even representation of the different MTBC groups for analysis of population structure and genetic diversity, we applied Treemer (Menardo et al. 2018) with the stop option *-RTL* 0.95, i.e. keeping 95% of the original tree length. The resulting reduced dataset was used for further analysis.

Population structure and genetic diversity

Population structure was evaluated using Principal Component Analysis (PCA) on SNP differences using the R package *adegent* (Jombart 2008). Genetic diversity was measured as raw pair-wise SNP differences for each MTBC lineage and ecotype if there were more than four genomes from a different geographic location, and as mean nucleotide diversity per site π using the R package *ape* (Paradis et al. 2004). π was calculated as the mean number of pair-wise mismatches among a set of sequences divided by the total length of queried genome base pairs which comprise the total length of the genome after excluding repetitive regions (see above) (Hartl et al. 2006). Confidence intervals for π were obtained by bootstrapping (1000 replicates) by re-sampling with replacement the nucleotide sites of the original alignments of polymorphic positions using the function *sample* in R (Team 2018). Lower and upper levels of confidence were obtained by calculating the 2.5th and the 97.5th quantiles of the π distribution obtained by bootstrapping.

bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Results and Discussion

Genome-based phylogeny reveals multiple animal-adapted clades

We assembled a total of 851 whole-genome sequences covering all known MTBC lineages and ecotypes. These included 834 genomes published previously, as well as four *M. orygis* genomes, two dassie bacillus genomes, eight *M. microti*, two *M. bovis* and one *M. caprae* newly sequenced here (Supplementary Table 1). We used a total of 56,195 variable single nucleotide positions extracted from these genome sequences to construct a phylogenetic tree rooted with *M.canettii*, the phylogenetically closest relative of the MTBC (Supply et al. 2013) (Figure 1). Our findings support the classification of the human-adapted MTBC into seven main phylogenetic lineages as previously reported (Gagneux et al. 2006, Gagneux et al. 2007, Firdessa et al. 2013). Classical genotyping studies and genomic deletion analyses indicated a single monophyletic clade for all the animal-adapted MTBC defined by clade-specific deletions in the Regions of Difference (RD) 7, 8, 9 and 10 (Brosch et al. 2002, Mostowy et al. 2002), and our new genome-based analysis confirms that all known animal-adapted members of the MTBC share a common ancestor at the branching point which is characterized by these deletions. Of note, the human-adapted MTBC Lineage 6 also shares this common ancestor, which has led to the hypothesis that Lineage 6 might have an unknown animal reservoir (Smith et al. 2006); however no such reservoir has yet been identified (Yeboah-Manu et al. 2017). Due to the limitations of standard genotyping (Comas et al. 2009) and the limited phylogenetic resolution of RDs in the MTBC (Hershberg et al. 2008), previous classifications have considered all animal-adapted ecotypes as part of one phylogenetic clade, recently referred to as MTBC "Lineage 8" (Gonzalo-Asensio et al. 2014). However, our new genome-based data revealed that these animal-adapted ecotypes form separate animal-adapted clades, some of which are paraphyletic. For the purpose of this study, we discuss four of these animal-adapted clades which we named Clade A1 to A4.

The animal-adapted MTBC Clade A1

One important finding from our phylogenomic analysis was that *M. mungi*, *M. suricattae*, the dassie bacillus and the chimpanzee bacillus form a separate Clade A1, which clusters with the human-adapted MTBC Lineage 6 (Figure 2). Based on limited previous genotyping data (Huard et al. 2006), it was hypothesized that the dassie bacillus shared a common ancestor with *M. africanum* (i.e. MTBC Lineage 6) (Huard et al. 2006, Brites et al. 2015). Our new whole genome data now confirms this hypothesis, and at the same time, highlight the fact that Clade A1 is more closely related to the human-adapted Lineage 6 of the MTBC than to the other animal-adapted ecotypes. This observation has important implications for our understanding of the original emergence of the animal-adapted strains and the evolutionary history of the MTBC as a whole. Specifically, considering that Lineage 5 is human-adapted and basal to the RD7-10 defined lineages, according to the most parsimonious evolutionary scenario, the common ancestor defined by the deletions in RD7-10 was a human-adapted pathogen (Brosch et al. 2002, Mostowy et al. 2002), and given that MTBC Lineage 6 is human-adapted (de Jong et al. 2010, Yeboah-Manu et al. 2017), the jump into animal hosts had to occur at least twice. Alternatively, if this common ancestor was already animal-adapted, it had to jump back into humans during the emergence of Lineage 6. A slight modification of this latter scenario would see the RD7-10 common ancestor as a generalist capable of infecting and causing disease in multiple host species, which was followed by a host-specialization of the different ecotypes. The generalist notion could be further extended to the evolution of the MTBC as a whole. According to the most common view, the MTBC

bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

emerged as a human pathogen (Brosch et al. 2002, Mostowy et al. 2002, Smith et al. 2009). This notion is supported by the fact that except for the lineage defined by deletions in RD7-10, all other MTBC lineages are human-adapted (Brites et al. 2015). Moreover, all known *M. canettii* isolates have been obtained from human TB patients (Supply et al. 2013), suggesting that the common ancestor of all the MTBC was able to cause infections in humans. However according the latest available epidemiological data (Koeck et al. 2010), *M. canettii* and the other so-called smooth TB bacilli are most likely opportunistic pathogens with a reservoir in the environment (Supply et al. 2017). Hence, the ancestor of the MTBC could also have been a generalist initially, which then adapted to the various host species over time (Smith et al. 2009).

Another important characteristic of clade A1 is the absence of the region encoded by RD1 in *M. mungi*, *M. suricattae*, the dassie bacillus (Supplementary Table S2). RD1 encodes proteins that are essential virulence factors for MTBC in humans (further discussed below). Our data confirm that *M. mungi*, *M. suricattae*, the dassie bacillus all have deleted the region corresponding to RD1. This deletion is not present in the chimp bacillus suggesting that RD1 might be essential for virulence in primates as proposed previously (Dippenaar et al. 2015).

The animal-adapted MTBC Clade A2

Similar to Clade A1 that comprises pathogens adapted to wild animals, Clade A2 consists of two ecotypes mainly affecting wild animals, namely *M. microti* and *M. pinnipedii*. In addition, Clade A2 also includes MTBC genomes isolated from pre-Columbian human remains published previously (Bos et al. 2014). These ancient genomes are most closely related to *M. pinnipedii*, suggesting possible cases of zoonotic TB transmission resulting from the handling and consumption of seal or sea lion meat at the time. Contemporary *M. pinnipedii* is known to infect humans occasionally (e.g. zoo keepers or seal trainers), but no human-to-human transmission has been documented to date. *M. microti* was originally isolated from voles in the 1930s (Wells 1937), but has since then been found in cats, pigs, llamas and immune-compromised humans (Brodin et al. 2002, Frota et al. 2004, Smith et al. 2009). Here we report 8 new *M. microti* genomes isolated from wild boar. Based on the 15 *M. microti* genomes included in this analysis, some host-specificity of particular sub-groups with this ecotype might be suggested, but analysis of a larger sample is needed to explore this possibility further. To our knowledge, *M. microti* has not been reported outside Europe, as infections in llamas pertain to captive animals in Europe (Oevermann et al. 2004) and represent probable spillovers from other hosts. Furthermore, the *M. microti*–like strain isolated from a rock hyrax has been likely misclassified (Clarke et al. 2016). Many of the animals species infected by *M. microti* occur across Eurasia which might therefore also correspond to the geographic range of *M. microti*. One of the important characteristics of all *M. microti* strains is the deletion of RD1 (Brodin et al. 2002), which is independent of the one described for the some of the members of Clade A1, and which is the most important virulence attenuating mutation in the *M. bovis* BCG vaccine (Pym et al. 2002). In support of the low virulence of *M. microti* in humans, and in contrast to *M. bovis* and *M. orygis* (see below), we detected only one infection with *M. microti* (from an immune-compromised patient (van Soolingen et al. 1998) among all the human isolates queried in the public domain (see methods).

The animal-adapted MTBC Clade A3

bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

In contrast to the animal Clades A1 and A2 that include multiple MTBC ecotypes infecting various wild animal host species, A3 comprises only genomes belonging to *M. orygis*. Even though *M. orygis* has been isolated from many different wild and domestic animals (Dawson et al. 2012, Gey van Pittius et al. 2012, van Ingen et al. 2012, Thapa et al. 2015, Thapa et al. 2016, Rahim et al. 2017), a large proportion of isolates reported to date are actually from human TB patients. One of the first detailed studies reporting on the genotypic properties of *M. orygis* strains included a total of 22 isolates, 11 of which originated from humans (van Ingen et al. 2012). The majority of the remaining isolates came from various zoo animals from the Netherlands and South Africa, which included three waterbucks, two antelopes, one deer and one oryx. A recent study from New York reported whole genome data from eight *M. orygis* isolates from human patients (Marcos et al. 2017). Another recent report from Birmingham, UK identified 24 *M. orygis* among 3,128 routinely collected human MTBC isolates (Lipworth et al. 2017). Similarly, eight *M. orygis* isolates were reported among 1,763 human TB cases from Victoria, Australia (Lavender et al. 2013), the genomes of four of which are newly reported here (Figure 1 and Figure 2). Importantly, all human *M. orygis* isolates, for which the relevant information was reported, were from patients born in India, Pakistan, Nepal or "South Asia", except for one with a reported origin in "South East Asia" (Dawson et al. 2012, van Ingen et al. 2012, Lavender et al. 2013, Marcos et al. 2017). This also includes one patient who immigrated from India to New Zealand and infected a dairy cow there (Dawson et al. 2012). One recent study reported 18 *M. orygis* isolates from dairy cattle in Bangladesh (Rahim et al. 2017). These isolates grouped into three distinct MIRU-VNTR clusters, with the largest cluster including two additional *M. orygis* isolates from captive monkeys. The authors propose that *M. orygis* is endemic among wild and domestic animals across South Asia and thus of relevant One Health significance. Based on the available evidence summarized above, and given that *M. orygis* shares a common ancestor with Clade 4 (Figure 1), which comprises *M. bovis* and *M. caprae* primarily adapted to domestic animals (further discussed below), we extend this notion and hypothesize that *M. orygis* is primarily a pathogen of cattle in South Asia, leading to zoonotic TB in humans through e.g. the consumption of raw milk. This scenario is the most parsimonious explanation for why *M. orygis* has repeatedly been isolated from South Asian migrants living in low TB-endemic countries in Europe, USA and Australia (Dawson et al. 2012, van Ingen et al. 2012, Lavender et al. 2013, Marcos et al. 2017). The genetic distance among the *M. orygis* identified in this study also supports this scenario, as the genomes of these isolates differ on average by 231 SNPs, suggesting independent infections in their countries of origin (Figure 3). Broader in-depth molecular analyses of cattle TB in South Asia, for which little data currently exist despite it representing a major public health threat (Rahim et al. 2017, Srinivasan et al. 2018) are needed to verify our hypothesis. Regarding *M. orygis* reported in animals other than cattle, our hypothesis would suggest that these likely represent spillovers from infected cattle, similar to the situation seen in *M. bovis* (Malone et al. 2017). In support of this view, except for one case isolated from a free-ranging rhinoceros in Nepal (Thapa et al. 2016), all *M. orygis* reported in un-domesticated animals were associated with zoos, farms or other forms of captivity where these wild animals might have come into contact with *M. orygis* infected cattle or humans (Gey van Pittius et al. 2012, van Ingen et al. 2012, Thapa et al. 2015, Rahim et al. 2017).

The animal-adapted MTBC Clade A4

Clade A4 includes the classical members of the animal-adapted MTBC, i.e. *M. bovis*, *M. caprae* and all the *M. bovis* BCG vaccine strains (Figure 2). Much work has been done on the genetic characterization of these MTBC members (Mostowy et al. 2005, Huard et al. 2006,

bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Smith et al. 2006, Muller et al. 2009, Copin et al. 2014, Malone et al. 2017), and thus we will not discuss these in any further details here. One exception is the deletion RD900, which has been described as a region specific to L6 and for which, presence and absence in *M. bovis* has been disputed (Bentley et al. 2012, Malone et al. 2017). The results of mapping with respect to *M. canettii* reads which remained unmapped to the chromosome of H37Rv, revealed that RD900 is polymorphic within *M. bovis*, within BCG strains and within *M. caprae*. In contrast, the region encoded by RD900 is deleted in all *M. orygis* genomes analyzed (Figure 2).

We end this section by speculating that if our hypothesis regarding the host range of *M. orygis* is true, Clade A3 and Clade A4 might reflect the two independent cattle domestication events known to have occurred in the Fertile Crescent and Indus Valley, respectively (Loftus et al. 1994). The corresponding domesticated forms emerging form the ancestral aurochs (*Bos primigenious*) are the sub-species *Bos taurus* and *Bos indicus*. Hence, *M. bovis* might have adapted to *B. taurus* whereas *M. orygis* might be better adapted to *B. indicus*. While highly speculative at this stage, this hypothesis could be tested experimentally (Villarreal-Ramos et al. 2018).

MTBC genetic diversity and host specificity

From an ecological perspective, pathogen diversity is generally positively correlated with host diversity especially in the case of obligate pathogens (Kamiya et al. 2014). Given the broad MTBC range of hosts, we explored how the genetic diversity is partitioned within the MTBC and if the genetic diversity of the animal-adapted MTBC was higher than that of the human-adapted MTBC. To obtain a more balanced representation of the different MTBC groups and remove redundancy caused by an over-representation of very closely related isolates which tell us little about macro-evolutionary processes, we used Treemer (Menardo et al. 2018) and reduced our dataset from 851 to 367 genomes while keeping 95% of the original total tree length. We performed principal component analysis (PCA) on the matrix of SNP distances correspondent to the non-redundant data set (n=367) (Figure 4). The resultant groups correspond largely to the results obtained with the phylogenenetic approach. The first principal component (PC1) explains 20.5% of the variation in genetic differences and highlights the contrast between "modern" human MTBC lineages (Lineages 2, 3 and 4) and Lineages 1, 5 and 7, which on their own formed very distinct groups. Lineage 6 appears closer to the animal MTBC but separated from clade A1. Interestingly, despite a clear separation between the human-adapted and animal-adapted MTBC (except for Lineage 6), PC1 contrasts more prominently the different human-adapted lineages than the different animal-adapted ecotypes (Figure 4).

As a measure of genetic diversity, we estimated the mean nucleotide diversity per site (π) of human versus animals isolates. The estimates indicate that two randomly picked human isolates differ on average by 0.0345% nucleotide differences (95% CI: 0.0337%-0.0352%) whereas animal isolates differ on average by 0.0313% (95% CI: 0.0305%-0.0321%). Despite non-overlapping confidence intervals, the difference between both π estimates is very small (0.003%) indicating that the genetic diversity which has emerged within animal and human MTBC is of similar magnitude. The estimates of π reflect both the diversity within each lineage/ecotype, as well as the diversity between lineages/ecotypes, resulting from older evolutionary events leading to the emergence of the latter. Whereas our sampling of the human MTBC reflects both pre- and post-lineage diversification reasonably well, the animal MTBC samples are most likely a poor representation of the genetic diversity resulting from

bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

diversification processes within each ecotype, with the possible exception of *M. bovis* (Figure 3). We thus compared the raw SNP differences among one random representative of each human and animal-adapted MTBC lineage and ecotype (Figure 5). The SNP differences accumulated in the different human-adapted lineages can be as high, or even higher than the genetic differences that separate MTBC strains infecting a broad taxonomic range of mammal species other than humans. Thus host-specificity in the MTBC cannot be easily explained only by quantitative genetic differences among the different animal-adapted MTBC ecotypes. In the light of the fact that in the MTBC, as in other bacteria, genomic variants caused by large deletions are pervasive (Bolotin et al. 2015) and genomes evolve towards a reduction of gene content as no horizontal gene transfer has been found in extant populations of the MTBC, it is also unlikely that the acquisition of new genes underlies host specificity. In support of this, after mapping reads using *M. canettii* as a reference, we found no regions that would be present in all representatives of each the different animal ecotype genomes and absent from human-adapted MTBC genomes. Several genomic deletions have been described in the genomes of animal-adapted MTBC members which we could also confirm here (Supplementary Table 2). Some of those deletions, e.g. RD1 and RD5, have been shown to impact virulence in different ways (Lewis et al. 2003, Dippenaar et al. 2015, Ates et al. 2018). In the case of RD1 and RD5, the deletion events seem to have occurred independently in different animal MTBC ecotypes (Figure 2) suggesting that the former have provided a fitness gain and were involved in the adaptation to new hosts (Brodin et al. 2002, Dippenaar et al. 2015, Ates et al. 2018). However, RD5 has also independently evolved and shown to impact virulence in the human adapted L2 Beijing sub-lineage (Ates et al. 2018).Taken together, this suggests that MTBC genomes are extremely robust in terms of host adaptation, and that interactions between different genes in the different ecotypes will be key determinants of host specificity in the MTBC.

Evolutionary scenarios for the evolution of the animal-adapted MTBC

The different MTBC members have adapted to infect a broad range of mammalian species, ranging from micro-mammals with short life-spans to humans, indicating that host shifts to distantly related hosts have occurred throughout the evolution of the MTBC. However, these host shifts have not emerged from any random phylogenetic branch of the MTBC as most of the human-adapted MTBC lineages are monophyletic and possibly locally adapted to different human populations (Fenner et al. 2013, Gagneux 2018). Host range expansion seems to have occurred after the split between Lineage 5 and the ancestor of Lineage 6 and all the animal ecotypes. The most parsimonious explanation is thus that the ancestor pathogen of the extant animal-adapted MTBC and Lineage 6 was a generalist with the ability to cause infections in many different kinds of hosts. A series of genetic events have been put forward by (Gonzalo-Asensio et al. 2014) to explain the decreased virulence of *M. africanum* L5 and L6 and the animal MTBC members compared to *M. tuberculosis* sensu stricto. A nonsynonymous mutation on the codon 71 in the *phoR* gene (Figure 2) which has emerged in the common ancestor of *M. africanum* L5 and L6 and of the animal-adapted strains, if transferred to a *M. tuberculosis* sensu stricto background leads to decreased virulence in mice and primary macrophages (Gonzalo-Asensio et al. 2014). This decrease in virulence is mediated by a decrease in the secretion of ESAT-6 which among other virulence factors is regulated by *phoPR* genes. The work of (Gonzalo-Asensio et al. 2014) shows that in L6, the loss of virulence was compensated by the RD8 deletion which restored the secretion of ESAT-6 independently of *PhoPR*. RD8 is common to L6 and all the animal ecotypes (but not L5, Figure 2), thus how the effects of *PhoPR* are restored in L5 remains unknown. This and related events could be at the origin of a putative generalist pathogen with compromised

bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

virulence in its original human host, and for which infecting other hosts represented fitness gains leading to the host range expansion we see today.

Based on the known geographic ranges of the animal-adapted MTBC ecotypes, we suggest two main divisions after the emergence of the ancestor of L6 and the animal ecotypes (Anc_{L6-A}, Figure 6); A series of specialization events which have occurred within Africa leading to the emergence of L6 in humans and clade A1 in several wild mammal species. With the exception of the chimp bacillus, these ecotypes have all been sampled in Southern Africa (Clarke et al. 2016). However, the extant geographic distributions of the hosts are not restricted to Southern Africa (except for Meerkats), additionally they have several overlapping areas and as a whole, form a continuum ranging from West-Africa to Southern-Africa (see http://www.iucnredlist.org/). Another series of specialization events might have happened outside Africa as suggested by the extant distribution of *M. orygis* and *M. microti* (Figure 6). Given that the maintenance hosts of strains that comprise A3 and A4 are domesticated species, one possible scenario is that the ancestor of $Anc_{A2-A3-A4}$ was carried by human populations as they migrated from Africa to the rest of the world (Figure 6). This ancestor could have been transferred posteriorly to different cattle and other livestock species which were domesticated outside Africa and independently in different parts of world as suggested in the discussion of clade A3 above, and become extinct in human populations. The example of the three human Peruvian mummies circa 1000 years old, which were infected with what is known today as *M. pinnipedii* (Bos et al. 2014), despite the difficulties in defining if humans were maintenance or spill-over hosts, illustrate the plausibility of such a scenario. Alternatively, $Anc_{A2-A3-A4}$ might have been brought outside Africa by another migratory species with close contact to livestock. The jump from the ancestor $Anc_{A2-A3-A4}$ to clade A2, which comprises such different host species, is not easily explained without invoking an environmental reservoir. This cannot be excluded as *M. bovis* and *M. microti* can possibly survive in the environment (Courtenay et al. 2006, Kipar et al. 2014).

The biology of pathogen jumps into new hosts involves three main steps (Woolhouse et al. 2005); i) exposure of the pathogen to a new environment, i.e. contact rates between hosts or between hosts and an environmental reservoir, ii) the ability to infect the new host, which most commonly decreases with the genetic distance from the ancestral host, and iii) transmissibility within the new host population. Generally, when the complete host ranges and the known geographic distributions are taken into account in the animal-adapted MTBC ecotypes, geographic proximity between hosts and therefore contact rates seemed to have played a more important role in determining host range and specialization than the phylogenetic distance among hosts. A corollary of these considerations and given that one important contributor to ii) is the ability to avoid or supress the host immune system, is that the immune repertoire of the host may have played a less important role in determining the host range of the different animal MTBC ecotypes compared to i) or iii) as long as the hosts were mammalian species. There are exceptions to this, e.g within Clade A1 moongoses and meerkats belong to the same taxonomic family (Clarke et al. 2016). However, in this case, host geographic range, ecology and phylogenetic distances are not independent, blurring conclusions. One important characteristic common to all host species in which the different MTBC members cause sustainable infections is that they attain high population densities, even if predominantly seasonally as in the case of pinnipeds (Cassini 1999). This characteristic might have been one of the most important determinants in the evolution of the different MTBC ecotypes and in particular, of their mode of transmission. Whereas the ability to cause pulmonary infections is essential for transmission among humans, in other animals, routes of infection other than aerosol transmission seem to play an important role, e.g. grazing contaminated pasture leads probably to a significant proportion of infections by

bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

M. bovis in cattle (Phillips et al. 2003), *M. mungi* can transmit directly through abrasions resultant from foraging activity of banded mongoose (Alexander et al. 2010, Malone et al. 2017), and transmission through skin lesions in *M. microti* has also been suggested (Kipar et al. 2014).

Concluding remarks

There are several reports about animal-adapted members of the MTBC infecting humans, wild and domestic animals, but an overarching analysis of all information available is required. In this study, we have combined all available information about animal-adapted MTBC strains and expanded it by sequencing more animal-adapted MTBC strains gathering the most comprehensive whole genome dataset of animal-adapted MTBC to date. We have used genomic analysis to elucidate the evolutionary history of the animal-adapted MTBC and have confirmed that the former are paraphyletic and that at least four different main clades can be defined. The phylogeny presented together with the known host range would be compatible with two realistic scenarios during the evolutionary history of the non-human MTBC, both involving more than one host jump. One scenario would present the ancestor of the group including L6 and all animal-adapted clades as a generalist capable of infecting a wide group of mammals, and different host adaptations would have occurred thereafter. An alternative scenario proposes that the ancestor of L6 and animal-adapted MTBC was adapted to humans, and subsequent host jumps lead to the host specificity of the four clades.

We found no correlation between genetic diversity of the pathogen and the phylogenetic distance of the host, as animal-adapted MTBC strains are not more diverse in average than human-adapted strains. Based on the current known host-ranges and geography of the animal-adapted MTBC, we propose that host expansion has been driven to a great extent by host geographical proximity, i.e. by contact rates among different species of mammals, and by high host population densities rather than by host genetic relatedness.

bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Author contributions

DB, CL, MC and FM have analysed the data. DB, SG and MC wrote the manuscript. BB, RW, AD, SP, MB, CB, SB, and JF contributed reagents and performed the experiments.

Funding

This work was supported by by the Swiss National Science Foundation (grants 310030_166687, IZRJZ3_164171, IZLSZ3_170834 and CRSII5_177163), the European Research Council (309540-EVODRTB) and SystemsX.ch.

Acknowledgements

Calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing core facility at University of Basel. Library preparation and sequencing was carried out in the Genomics Facility Basel.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Supplementary Material

SupplementaryTable1.xls

SupplementaryTable2.xls

Figure legends

Figure 1.
Maximum Likelihood topology of 322 human-adapted and 529 animal-adapted MTBC members. Branch lengths are proportional to nucleotide substitutions and the topology is rooted with *Mycobacterium canettii*. Support values correspond to bootstrap values. Main large deletions defining the animal-adapted MTBC are indicated by red arrows.

Figure 2.
Topology showed in Figure 1 after collapsing all human-adapted branches. Branch lengths are proportional to nucleotide substitutions and the topology is rooted with *Mycobacterium canettii*. Main large deletions discussed in the text are indicated by red arrows. Those deletions which are polymorphic in terms of presence or absent within main clades are indicated in italics.

bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 3. Pair-wise SNP distances within lineage and ecotype from human and animal-adapted MTBC, respectively. Each box corresponds to the 25% and 75% quantiles, the black line represents the median and the whiskers extend to 1.5 times the interquartile range.

Figure 4.
Principal Component Analysis (PCA) derived from whole-genome SNPs. The two first principal components are shown.

Figure 5.
Pair-wise SNP distances between one randomly chosen representative of each human adapted MTBC lineage A) and animal-adapted MTBC B).

Figure 6.
Schematic illustration of the putative evolutionary history of the animal-adapted MTBC. The length of the branch is not proportional to genetic distances.

bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

596 **References**
597
598 Alexander, K. A., Laver, P. N., Michel, A. L., Williams, M., van Helden, P. D., Warren, R.
599 M., et al. (2010). Novel Mycobacterium tuberculosis Complex Pathogen, M. mungi. Emerg
600 Infect Dis 16:1296-1299.

601 Ameni, G., Vordermeier, M., Firdessa, R., Aseffa, A., Hewinson, G., Gordon, S. V., et al.
602 (2010). Mycobacterium tuberculosis infection in grazing cattle in central Ethiopia. Vet
603 J.doi:10.1016/j.tvjl.2010.05.005.

604 Ates, L. S., Dippenaar, A., Ummels, R., Piersma, S. R., van der Woude, A. D., van der Kuij,
605 K., et al. (2018). Mutations in ppe38 block PE_PGRS secretion and increase virulence of
606 Mycobacterium tuberculosis. Nat Microbiol 3:181-188.doi:10.1038/s41564-017-0090-6.

607 Ates, L. S., Sayes, F., Frigui, W., Ummels, R., Damen, M. P. M., Bottai, D., et al. (2018).
608 RD5-mediated lack of PE_PGRS and PPE-MPTR export in BCG vaccine strains results in
609 strong reduction of antigenic repertoire but little impact on protection. PLoS Pathog
610 14:e1007139.doi:10.1371/journal.ppat.1007139.

611 Belisle, J. T. and Sonnenberg, M. G. (1998). Isolation of genomic DNA from mycobacteria.
612 Methods Mol Biol 101:31-44.doi:10.1385/0-89603-471-2:31.

613 Bentley, S. D., Comas, I., Bryant, J. M., Walker, D., Smith, N. H., Harris, S. R., et al. (2012).
614 The genome of Mycobacterium africanum West African 2 reveals a lineage-specific locus
615 and genome erosion common to the M. tuberculosis complex. PLoS Negl Trop Dis
616 6:e1552.doi:10.1371/journal.pntd.0001552.

617 Blazquez, J., Espinosa de Los Monteros, L. E., Samper, S., Martin, C., Guerrero, A., Cobo, J.,
618 et al. (1997). Genetic characterization of multidrug-resistant Mycobacterium bovis strains
619 from a hospital outbreak involving human immunodeficiency virus-positive patients. J Clin
620 Microbiol 35:1390-1393.

621 Bolger, A. M., Lohse, M. and Usadel, B. (2014). Trimmomatic: a flexible trimmer for
622 Illumina sequence data. Bioinformatics 30:2114-2120.doi:10.1093/bioinformatics/btu170.

623 Bolotin, E. and Hershberg, R. (2015). Gene Loss Dominates As a Source of Genetic
624 Variation within Clonal Pathogenic Bacterial Species. Genome Biol Evol 7:2173-
625 2187.doi:10.1093/gbe/evv135.

626 Bos, K. I., Harkins, K. M., Herbig, A., Coscolla, M., Weber, N., Comas, I., et al. (2014). Pre-
627 Columbian mycobacterial genomes reveal seals as a source of New World human
628 tuberculosis. Nature 514:494-497.doi:10.1038/nature13591.

629 Brites, D. and Gagneux, S. (2015). Co-evolution of Mycobacterium tuberculosis and Homo
630 sapiens. Immunol Rev 264:6-24.doi:10.1111/imr.12264.

631 Brites, D. and Gagneux, S. (2017). The Nature and Evolution of Genomic Diversity in the
632 Mycobacterium tuberculosis Complex. Adv Exp Med Biol 1019:1-26.doi:10.1007/978-3-
633 319-64371-7_1.

bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Brodin, P., Eiglmeier, K., Marmiesse, M., Billault, A., Garnier, T., Niemann, S., et al. (2002). Bacterial artificial chromosome-based comparative genomic analysis identifies Mycobacterium microti as a natural ESAT-6 deletion mutant. Infect Immun 70:5568-5578.

Brosch, R., Gordon, S. V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., et al. (2002). A new evolutionary scenario for the Mycobacterium tuberculosis complex. Proc Natl Acad Sci U S A 99:3684-3689.

Cassini, M. H. (1999). The evolution of reproductive systems in pinnipeds. Behavioral Ecology 10:612-616.doi:DOI 10.1093/beheco/10.5.612.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6:80-92.doi:10.4161/fly.19695.

Clarke, C., Van Helden, P., Miller, M. and Parsons, S. (2016). Animal-adapted members of the Mycobacterium tuberculosis complex endemic to the southern African subregion. J S Afr Vet Assoc 87:1322.doi:10.4102/jsava.v87i1.1322.

Comas, I., Chakravartti, J., Small, P. M., Galagan, J., Niemann, S., Kremer, K., et al. (2010). Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. Nat Genet 42:498-503.doi:10.1038/ng.590.

Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K. E., Kato-Maeda, M., et al. (2013). Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat Genet 45:1176-1182.doi:10.1038/ng.2744.

Comas, I., Homolka, S., Niemann, S. and Gagneux, S. (2009). Genotyping of genetically monomorphic bacteria: DNA sequencing in mycobacterium tuberculosis highlights the limitations of current methodologies. PLoS ONE 4:e7815.doi:10.1371/journal.pone.0007815.

Copin, R., Coscolla, M., Efstathiadis, E., Gagneux, S. and Ernst, J. D. (2014). Impact of in vitro evolution on antigenic diversity of Mycobacterium bovis bacillus Calmette-Guerin (BCG). Vaccine 32:5998-6004.doi:10.1016/j.vaccine.2014.07.113.

Coscolla, M., Lewin, A., Metzger, S., Maetz-Rennsing, K., Calvignac-Spencer, S., Nitsche, A., et al. (2013). Novel Mycobacterium tuberculosis complex isolate from a wild chimpanzee. Emerg Infect Dis 19:969-976.doi:10.3201/eid1906.121012.

Courtenay, O., Reilly, L. A., Sweeney, F. P., Hibberd, V., Bryan, S., Ul-Hassan, A., et al. (2006). Is Mycobacterium bovis in the environment important for the persistence of bovine tuberculosis? Biol Lett 2:460-462.doi:10.1098/rsbl.2006.0468.

Cousins, D. V., Bastida, R., Cataldi, A., Quse, V., Redrobe, S., Dow, S., et al. (2003). Tuberculosis in seals caused by a novel member of the Mycobacterium tuberculosis complex: Mycobacterium pinnipedii sp. nov. Int J Syst Evol Microbiol 53:1305-1314.

Dawson, K. L., Bell, A., Kawakami, R. P., Coley, K., Yates, G. and Collins, D. M. (2012). 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.doi:10.1128/JCM.01652-12.

bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

de Jong, B. C., Antonio, M. and Gagneux, S. (2010). Mycobacterium africanum--review of an important cause of human tuberculosis in West Africa. PLoS Negl Trop Dis 4:e744.doi:10.1371/journal.pntd.0000744.

Dippenaar, A., Parsons, S. D., Sampson, S. L., van der Merwe, R. G., Drewe, J. A., Abdallah, A. M., et al. (2015). Whole genome sequence analysis of Mycobacterium suricattae. Tuberculosis (Edinb) 95:682-688.doi:10.1016/j.tube.2015.10.001.

Fenner, L., Egger, M., Bodmer, T., Furrer, H., Ballif, M., Battegay, M., et al. (2013). HIV infection disrupts the sympatric host-pathogen relationship in human tuberculosis. PLoS Genet 9:e1003318.doi:10.1371/journal.pgen.1003318.

Firdessa, R., Berg, S., Hailu, E., Schelling, E., Gumi, B., Erenso, G., et al. (2013). Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia. Emerg Infect Dis 19:460-463.

Frota, C. C., Hunt, D. M., Buxton, R. S., Rickman, L., Hinds, J., Kremer, K., et al. (2004). Genome structure in the vole bacillus, Mycobacterium microti, a member of the Mycobacterium tuberculosis complex with a low virulence for humans. Microbiology 150:1519-1527.

Gagneux, S. (2018). Ecology and evolution of Mycobacterium tuberculosis. Nat Rev Microbiol 16:202-213.doi:10.1038/nrmicro.2018.8.

Gagneux, S., Deriemer, K., Van, T., Kato-Maeda, M., de Jong, B. C., Narayanan, S., et al. (2006). Variable host-pathogen compatibility in Mycobacterium tuberculosis. Proc Natl Acad Sci U S A 103:2869-2873.

Gagneux, S. and Small, P. M. (2007). Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. Lancet Infect Dis 7:328-337.

Gey van Pittius, N. C., Perrett, K. D., Michel, A. L., Keet, D. F., Hlokwe, T., Streicher, E. M., et al. (2012). Infection of African buffalo (Syncerus caffer) by oryx bacillus, a rare member of the antelope clade of the Mycobacterium tuberculosis complex. J Wildl Dis 48:849-857.doi:10.7589/2010-07-178.

Gonzalo-Asensio, J., Malaga, W., Pawlik, A., Astarie-Dequeker, C., Passemar, C., Moreau, F., et al. (2014). Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. Proc Natl Acad Sci U S A.doi:10.1073/pnas.1406693111.

Guangchuang, Y., K., S. D., Huachen, Z., Yi, G. and Tsan-Yuk, L. T. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution 8:28-36.doi:10.1111/2041-210X.12628.

Hartl, D. L. and Clarck, A. G. (2006). Principles of population genetics. Sunderland, MA, Sinauer Associates, Inc.

Hershberg, R., Lipatov, M., Small, P. M., Sheffer, H., Niemann, S., Homolka, S., et al. (2008). High Functional Diversity in Mycobacterium tuberculosis Driven by Genetic Drift and Human Demography. PLoS Biol 6:e311.

bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

712 Huard, R. C., Fabre, M., de Haas, P., Claudio Oliveira Lazzarini, L., van Soolingen, D.,
713 Cousins, D., et al. (2006). Novel Genetic Polymorphisms That Further Delineate the
714 Phylogeny of the Mycobacterium tuberculosis Complex. J Bacteriol 188:4271-4287.

715 Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers.
716 Bioinformatics 24:1403-1405.doi:10.1093/bioinformatics/btn129.

717 Kamiya, T., O'Dwyer, K., Nakagawa, S. and Poulin, R. (2014). Host diversity drives parasite
718 diversity: meta-analytical insights into patterns and causal mechanisms. Ecography 37:689-
719 697.doi:10.1111/j.1600-0587.2013.00571.x.

720 Kipar, A., Burthe, S. J., Hetzel, U., Rokia, M. A., Telfer, S., Lambin, X., et al. (2014).
721 Mycobacterium microti tuberculosis in its maintenance host, the field vole (Microtus
722 agrestis): characterization of the disease and possible routes of transmission. Vet Pathol
723 51:903-914.doi:10.1177/0300985813513040.

724 Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., et al. (2012).
725 VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome
726 sequencing. Genome Res 22:568-576.doi:10.1101/gr.129684.111.

727 Koeck, J. L., Fabre, M., Simon, F., Daffe, M., Garnotel, E., Matan, A. B., et al. (2010).
728 Clinical characteristics of the smooth tubercle bacilli "Mycobacterium canettii" infection
729 suggest the existence of an environmental reservoir. Clin Microbiol
730 Infect.doi:10.1111/j.1469-0691.2010.03347.x.

731 Lavender, C. J., Globan, M., Kelly, H., Brown, L. K., Sievers, A., Fyfe, J. A., et al. (2013).
732 Epidemiology and control of tuberculosis in Victoria, a low-burden state in south-eastern
733 Australia, 2005-2010. Int J Tuberc Lung Dis 17:752-758.doi:10.5588/ijtld.12.0791.

734 Lewis, K. N., Liao, R., Guinn, K. M., Hickey, M. J., Smith, S., Behr, M. A., et al. (2003).
735 Deletion of RD1 from Mycobacterium tuberculosis mimics bacille Calmette-Guerin
736 attenuation. J Infect Dis 187:117-123.

737 Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association
738 mapping and population genetical parameter estimation from sequencing data. Bioinformatics
739 27:2987-2993.doi:10.1093/bioinformatics/btr509.

740 Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler
741 transform. Bioinformatics 26:589-595.doi:10.1093/bioinformatics/btp698.

742 Lipworth, S. I. W., Jajou, R., de Neeling, H., Bradley, P., van der Hoek, W., Iqbal, Z., et al.
743 (2017). A novel multi SNP based method for the identification of subspecies and associated
744 lineages and sub-lineages of the Mycobacterium tuberculosis complex by whole genome
745 sequencing. bioRxiv.doi:10.1101/213850.

746 Loftus, R. T., MacHugh, D. E., Bradley, D. G., Sharp, P. M. and Cunningham, P. (1994).
747 Evidence for two independent domestications of cattle. Proc Natl Acad Sci U S A 91:2757-
748 2761.

749 Malone, K. M., Farrell, D., Stuber, T. P., Schubert, O. T., Aebersold, R., Robbe-Austerman,
750 S., et al. (2017). Updated Reference Genome Sequence and Annotation of Mycobacterium
751 bovis AF2122/97. Genome Announc 5.doi:10.1128/genomeA.00157-17.

bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

752 Malone, K. M. and Gordon, S. V. (2017). Mycobacterium tuberculosis Complex Members
753 Adapted to Wild and Domestic Animals. Adv Exp Med Biol 1019:135-154.doi:10.1007/978-
754 3-319-64371-7_7.

755 Marcos, L. A., Spitzer, E. D., Mahapatra, R., Ma, Y., Halse, T. A., Shea, J., et al. (2017).
756 Mycobacterium orygis Lymphadenitis in New York, USA. Emerg Infect Dis 23:1749-
757 1751.doi:10.3201/eid2310.170490.

758 Menardo, F., Loiseau, C., Brites, D., Coscolla, M., Gygli, S. M., Rutaihwa, L. K., et al.
759 (2018). Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of
760 diversity. BMC Bioinformatics 19:164.doi:10.1186/s12859-018-2164-8.

761 Miller, M. A., Pfeiffer, W. and Schwartz, T. (2010). Creating the CIPRES Science Gateway
762 for inference of large phylogenetic trees. Proceedings of the Gateway Computing
763 Environments Workshop (GCE), New Orleans, LA.

764 Mostowy, S., Cousins, D. and Behr, M. A. (2004). Genomic interrogation of the dassie
765 bacillus reveals it as a unique RD1 mutant within the Mycobacterium tuberculosis complex. J
766 Bacteriol 186:104-109.

767 Mostowy, S., Cousins, D., Brinkman, J., Aranaz, A. and Behr, M. A. (2002). Genomic
768 deletions suggest a phylogeny for the Mycobacterium tuberculosis complex. J Infect Dis
769 186:74-80.

770 Mostowy, S., Inwald, J., Gordon, S., Martin, C., Warren, R., Kremer, K., et al. (2005).
771 Revisiting the evolution of Mycobacterium bovis. J Bacteriol 187:6386-6395.

772 Muller, B., Durr, S., Alonso, S., Hattendorf, J., Laisse, C. J., Parsons, S. D., et al. (2013).
773 Zoonotic Mycobacterium bovis-induced tuberculosis in humans. Emerg Infect Dis 19:899-
774 908.doi:10.3201/eid1906.120543.

775 Muller, B., Hilty, M., Berg, S., Garcia-Pelayo, M. C., Dale, J., Boschiroli, M. L., et al.
776 (2009). African 1; An Epidemiologically Important Clonal Complex of Mycobacterium bovis
777 Dominant in Mali, Nigeria, Cameroon and Chad. J Bacteriol 191: 1951-
778 1960.doi:10.1128/JB.01590-08.

779 Murphree, R., Warkentin, J. V., Dunn, J. R., Schaffner, W. and Jones, T. F. (2011). Elephant-
780 to-human transmission of tuberculosis, 2009. Emerg Infect Dis 17:366-
781 371.doi:10.3201/eid1703.101668.

782 Oevermann, A., Pfyffer, G. E., Zanolari, P., Meylan, M. and Robert, N. (2004). Generalized
783 tuberculosis in llamas (Lama glama) due to Mycobacterium microti. J Clin Microbiol
784 42:1818-1821.

785 Olea-Popelka, F., Muwonge, A., Perera, A., Dean, A. S., Mumford, E., Erlacher-Vindel, E.,
786 et al. (2017). Zoonotic tuberculosis in human beings caused by Mycobacterium bovis-a call
787 for action. Lancet Infect Dis 17:e21-e25.doi:10.1016/S1473-3099(16)30139-6.

788 Paradis, E., Claude, J. and Strimmer, K. (2004). APE: Analyses of Phylogenetics and
789 Evolution in R language. Bioinformatics 20:289-290.

bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

790 Parsons, S. D., Drewe, J. A., Gey van Pittius, N. C., Warren, R. M. and van Helden, P. D.
791 (2013). Novel cause of tuberculosis in meerkats, South Africa. Emerg Infect Dis 19:2004-
792 2007.doi:10.3201/eid1912.130268.

793 Phillips, C. J., Foster, C. R., Morris, P. A. and Teverson, R. (2003). The transmission of
794 Mycobacterium bovis infection to cattle. Res Vet Sci 74:1-15.

795 Pym, A. S., Brodin, P., Brosch, R., Huerre, M. and Cole, S. T. (2002). Loss of RD1
796 contributed to the attenuation of the live tuberculosis vaccines Mycobacterium bovis BCG
797 and Mycobacterium microti. Mol Microbiol 46:709-717.

798 Rahim, Z., Thapa, J., Fukushima, Y., van der Zanden, A. G. M., Gordon, S. V., Suzuki, Y., et
799 al. (2017). Tuberculosis Caused by Mycobacterium orygis in Dairy Cattle and Captured
800 Monkeys in Bangladesh: a New Scenario of Tuberculosis in South Asia. Transbound Emerg
801 Dis 64:1965-1969.doi:10.1111/tbed.12596.

802 Riojas, M. A., McGough, K. J., Rider-Riojas, C. J., Rastogi, N. and Hazbon, M. H. (2018).
803 Phylogenomic analysis of the species of the Mycobacterium tuberculosis complex
804 demonstrates that Mycobacterium africanum, Mycobacterium bovis, Mycobacterium caprae,
805 Mycobacterium microti and Mycobacterium pinnipedii are later heterotypic synonyms of
806 Mycobacterium tuberculosis. Int J Syst Evol Microbiol 68:324-
807 332.doi:10.1099/ijsem.0.002507.

808 Smith, N. H., Crawshaw, T., Parry, J. and Birtles, R. J. (2009). Mycobacterium microti; more
809 diverse than previously thought. J Clin Microbiol 47:2551-2559.doi:10.1128/JCM.00638-09.

810 Smith, N. H., Gordon, S. V., de la Rua-Domenech, R., Clifton-Hadley, R. S. and Hewinson,
811 R. G. (2006). Bottlenecks and broomsticks: the molecular evolution of Mycobacterium bovis.
812 Nat Rev Microbiol 4:670-681.

813 Smith, N. H., Hewinson, R. G., Kremer, K., Brosch, R. and Gordon, S. V. (2009). Myths and
814 misconceptions: the origin and evolution of Mycobacterium tuberculosis. Nat Rev Microbiol
815 7:pages 537–544.doi:10.1038/nrmicro2165.

816 Smith, N. H., Kremer, K., Inwald, J., Dale, J., Driscoll, J. R., Gordon, S. V., et al. (2005).
817 Ecotypes of the Mycobacterium tuberculosis complex. J Theor Biol

818 Srinivasan, S., Easterling, L., Rimal, B., Niu, X. M., Conlan, A. J. K., Dudas, P., et al.
819 (2018). Prevalence of Bovine Tuberculosis in India: A systematic review and meta-analysis.
820 Transbound Emerg Dis.doi:10.1111/tbed.12915.

821 Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses
822 with thousands of taxa and mixed models. Bioinformatics 22:2688-
823 2690.doi:10.1093/bioinformatics/btl446.

824 Steiner, A., Stucki, D., Coscolla, M., Borrell, S. and Gagneux, S. (2014). KvarQ: targeted and
825 direct variant calling from fastq reads of bacterial genomes. BMC Genomics
826 15:881.doi:10.1186/1471-2164-15-881.

827 Stucki, D., Brites, D., Jeljeli, L., Coscolla, M., Liu, Q., Trauner, A., et al. (2016).
828 Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically
829 restricted sublineages. Nat Genet 48:1535-1543.doi:10.1038/ng.3704.

bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Supply, P. and Brosch, R. (2017). The Biology and Epidemiology of Mycobacterium canettii. Adv Exp Med Biol 1019:27-41.doi:10.1007/978-3-319-64371-7_2.

Supply, P., Marceau, M., Mangenot, S., Roche, D., Rouanet, C., Khanna, V., et al. (2013). Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of Mycobacterium tuberculosis. Nat Genet 45:172-179.doi:10.1038/ng.2517.

Team, R. C. (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria.

Thapa, J., Nakajima, C., Maharjan, B., Poudell, A. and Suzuki, Y. (2015). Molecular characterization of Mycobacterium orygis isolates from wild animals of Nepal. Jpn J Vet Res 63:151-158.

Thapa, J., Paudel, S., Sadaula, A., Shah, Y., Maharjan, B., Kaufman, G. E., et al. (2016). Mycobacterium orygis-Associated Tuberculosis in Free-Ranging Rhinoceros, Nepal, 2015. Emerg Infect Dis 22:570-572.doi:10.3201/eid2203.151929.

van Ingen, J., Rahim, Z., Mulder, A., Boeree, M. J., Simeone, R., Brosch, R., et al. (2012). Characterization of Mycobacterium orygis as M. tuberculosis complex subspecies. Emerg Infect Dis 18:653-655.doi:10.3201/eid1804.110888.

van Soolingen, D., de Haas, P. E., Haagsma, J., Eger, T., Hermans, P. W., Ritacco, V., et al. (1994). Use of various genetic markers in differentiation of Mycobacterium bovis strains from animals and humans and for studying epidemiology of bovine tuberculosis. J Clin Microbiol 32:2425-2433.

van Soolingen, D., van der Zanden, A. G., de Haas, P. E., Noordhoek, G. T., Kiers, A., Foudraine, N. A., et al. (1998). Diagnosis of Mycobacterium microti infections among humans by using novel genetic markers. J Clin Microbiol 36:1840-1845.

Villarreal-Ramos, B., Berg, S., Whelan, A., Holbert, S., Carreras, F., Salguero, F. J., et al. (2018). Experimental infection of cattle with Mycobacterium tuberculosis isolates shows the attenuation of the human tubercle bacillus for cattle. Sci Rep 8:894.doi:10.1038/s41598-017-18575-5.

Waters, W. R., Palmer, M. V., Buddle, B. M. and Vordermeier, H. M. (2012). Bovine tuberculosis vaccine research: historical perspectives and recent advances. Vaccine 30:2611-2622.

Wells, A. Q. (1937). Tuberculosis in wild voles. Lancet 1:1221.

Whelan, A. O., Coad, M., Cockle, P. J., Hewinson, G., Vordermeier, M. and Gordon, S. V. (2010). Revisiting host preference in the Mycobacterium tuberculosis complex: experimental infection shows M. tuberculosis H37Rv to be avirulent in cattle. PLoS One 5:e8527.doi:10.1371/journal.pone.0008527.

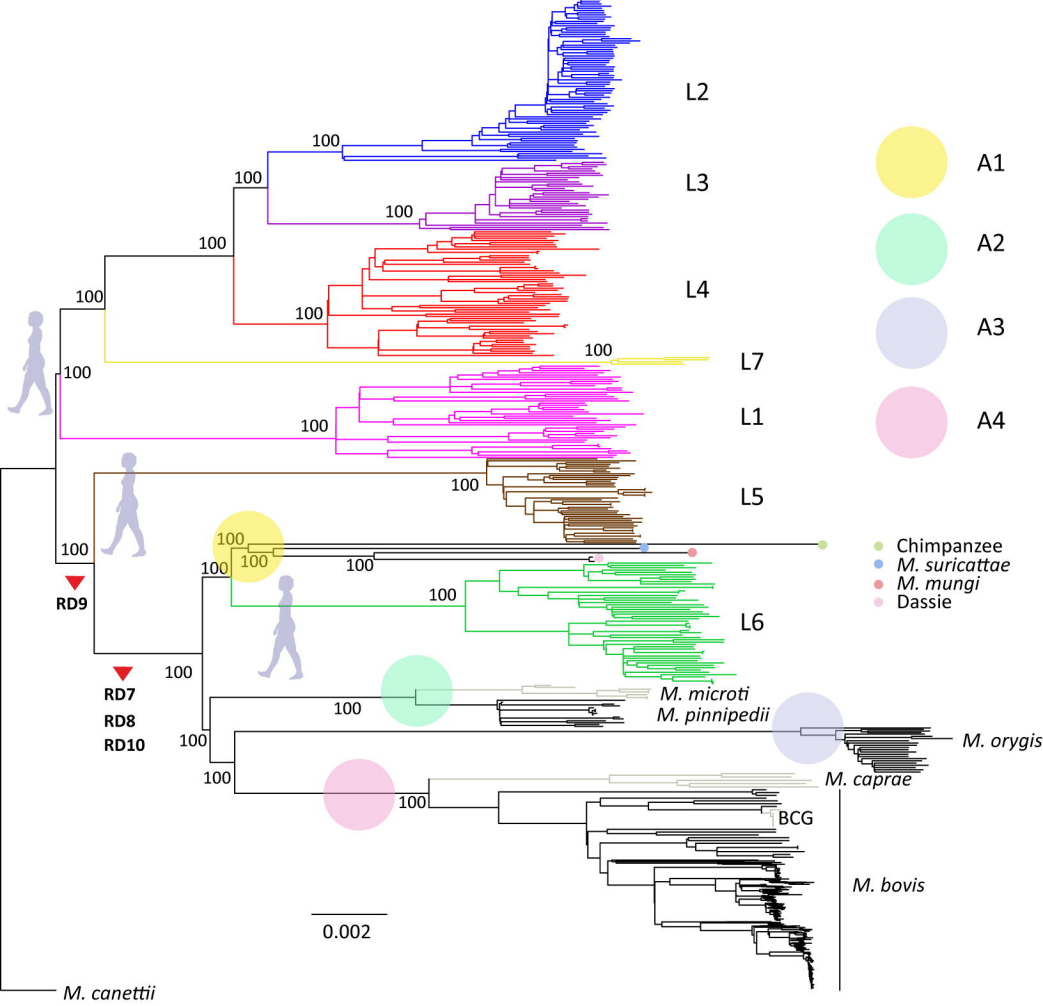
WHO (2014). Global tuberculosis control - surveillance, planning, financing. Geneva, Switzerland, World Health Organization.

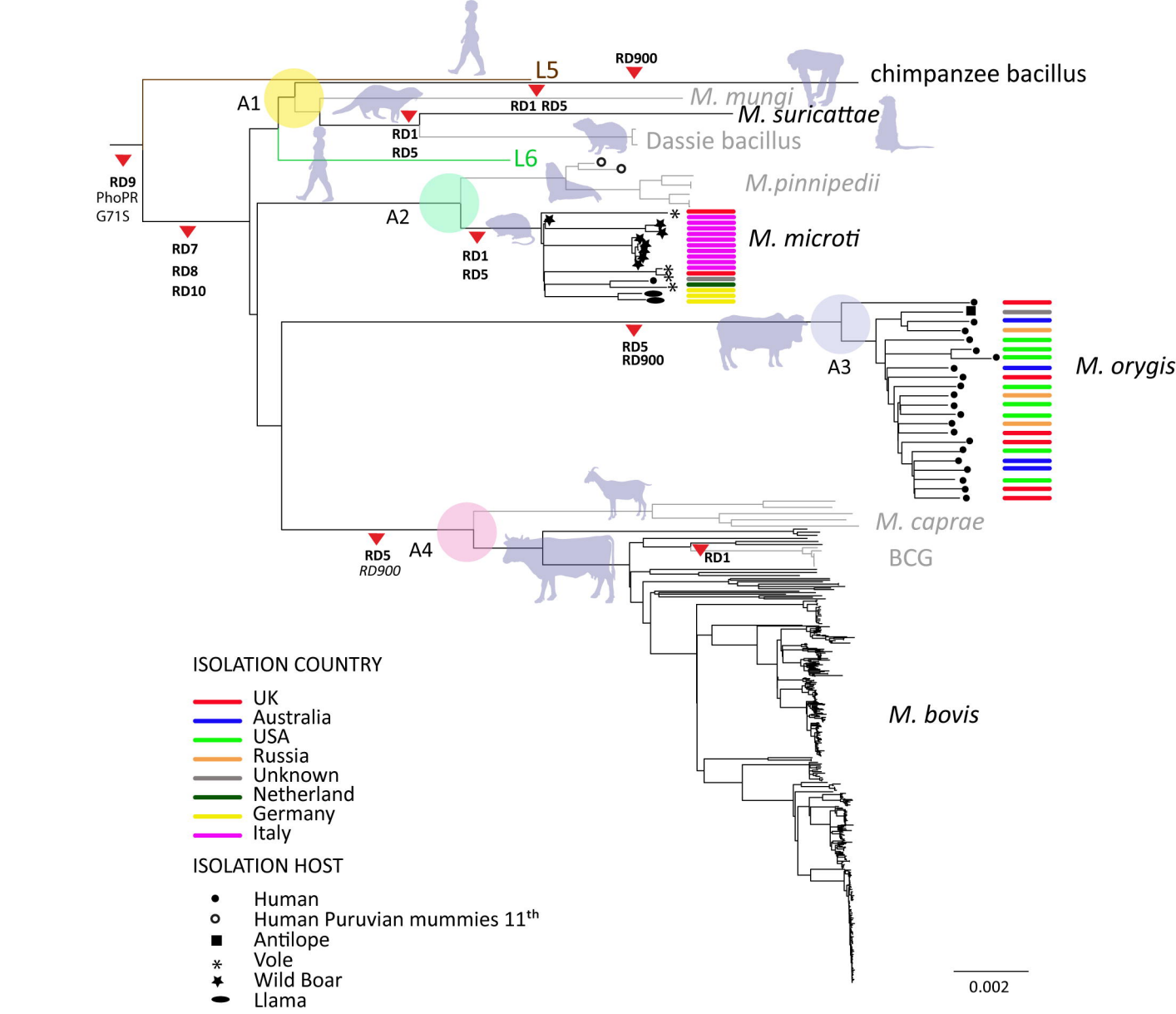
Woolhouse, M. E. J., Haydon, D. T. and Antia, R. (2005). Emerging pathogens: the epidemiology and evolution of species jumps. Trends in ecology & evolution 20:238-244.

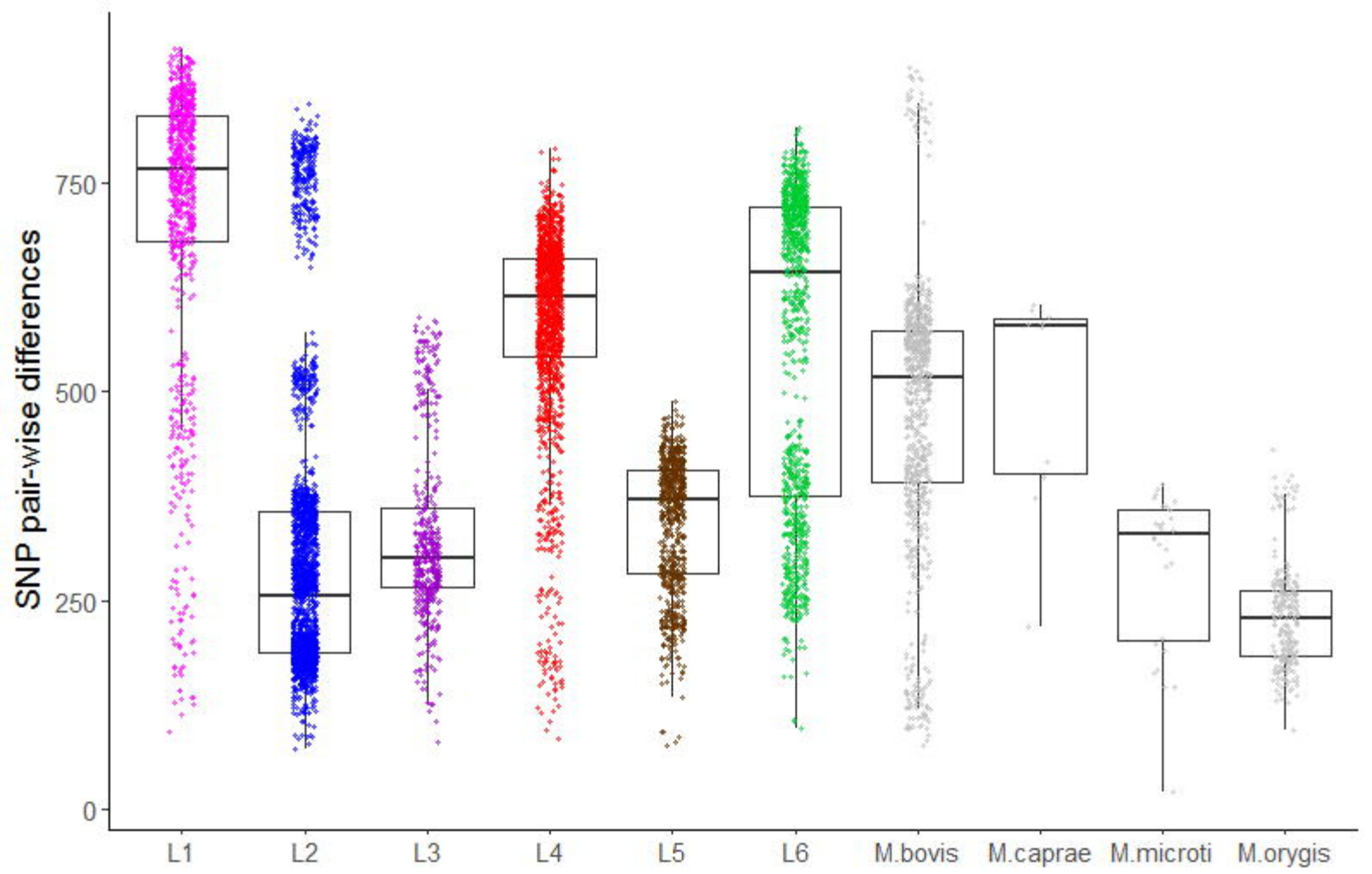
bioRxiv preprint doi: https://doi.org/10.1101/384263; this version posted August 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

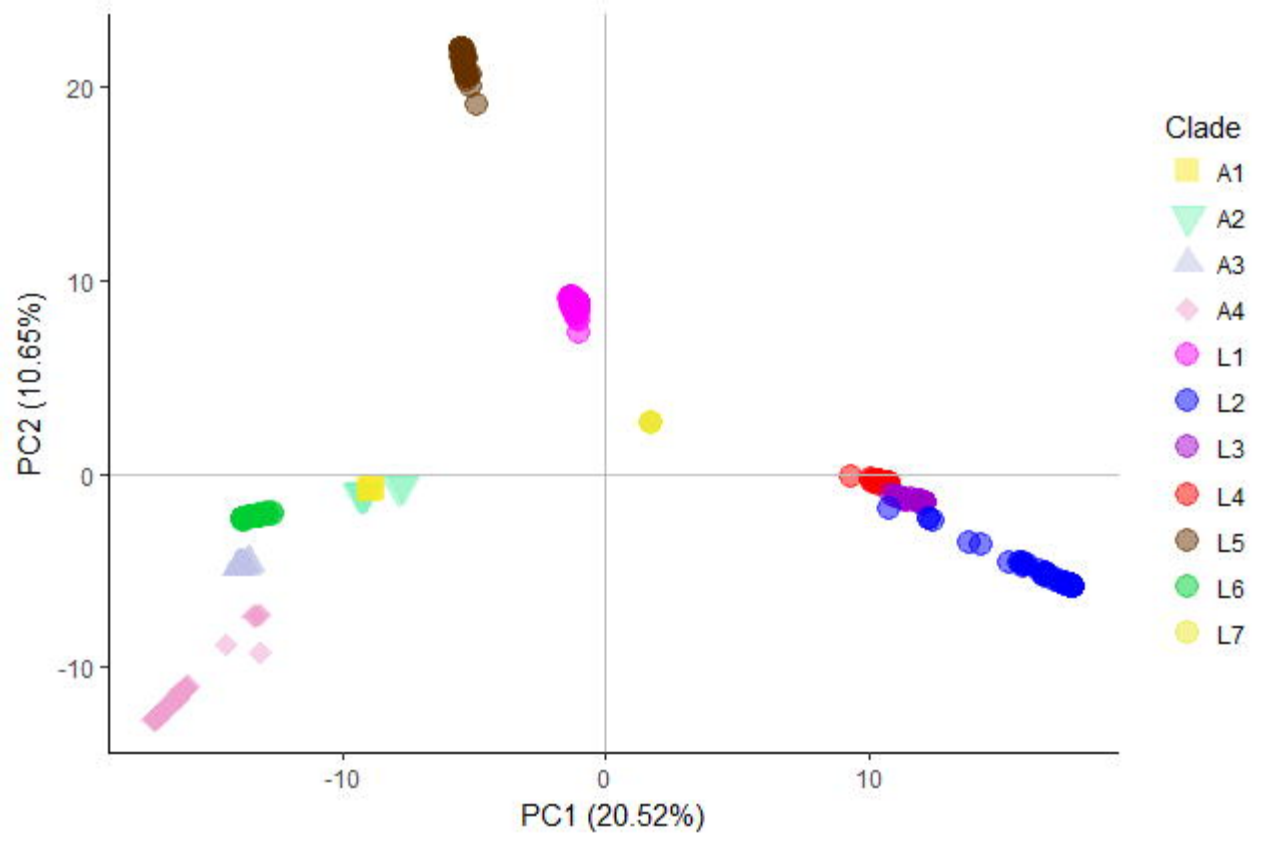
869 Yeboah-Manu, D., de Jong, B. C. and Gehre, F. (2017). The Biology and Epidemiology of
870 Mycobacterium africanum. Adv Exp Med Biol 1019:117-133.doi:10.1007/978-3-319-64371-
871 7_6.

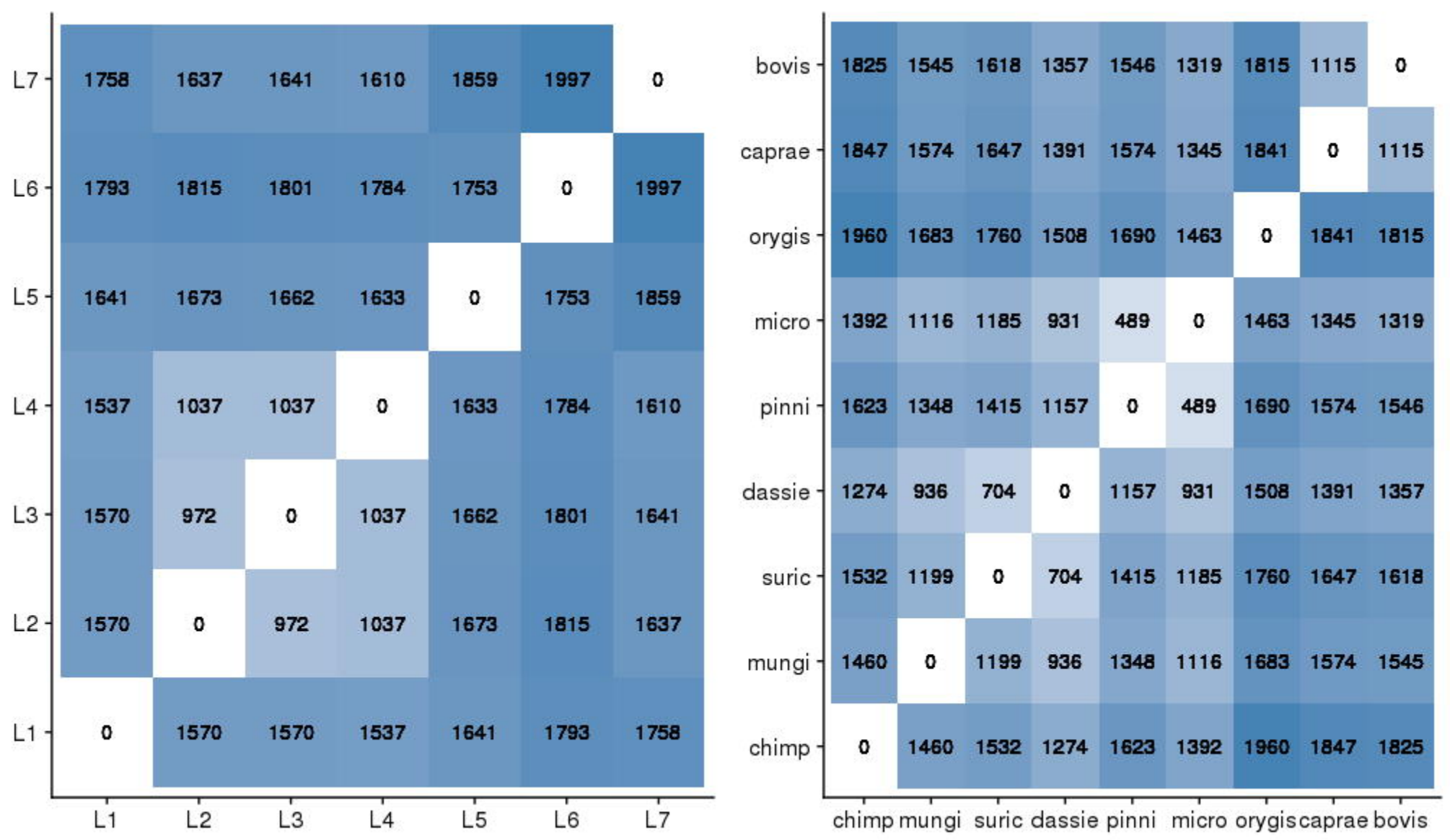
872
873

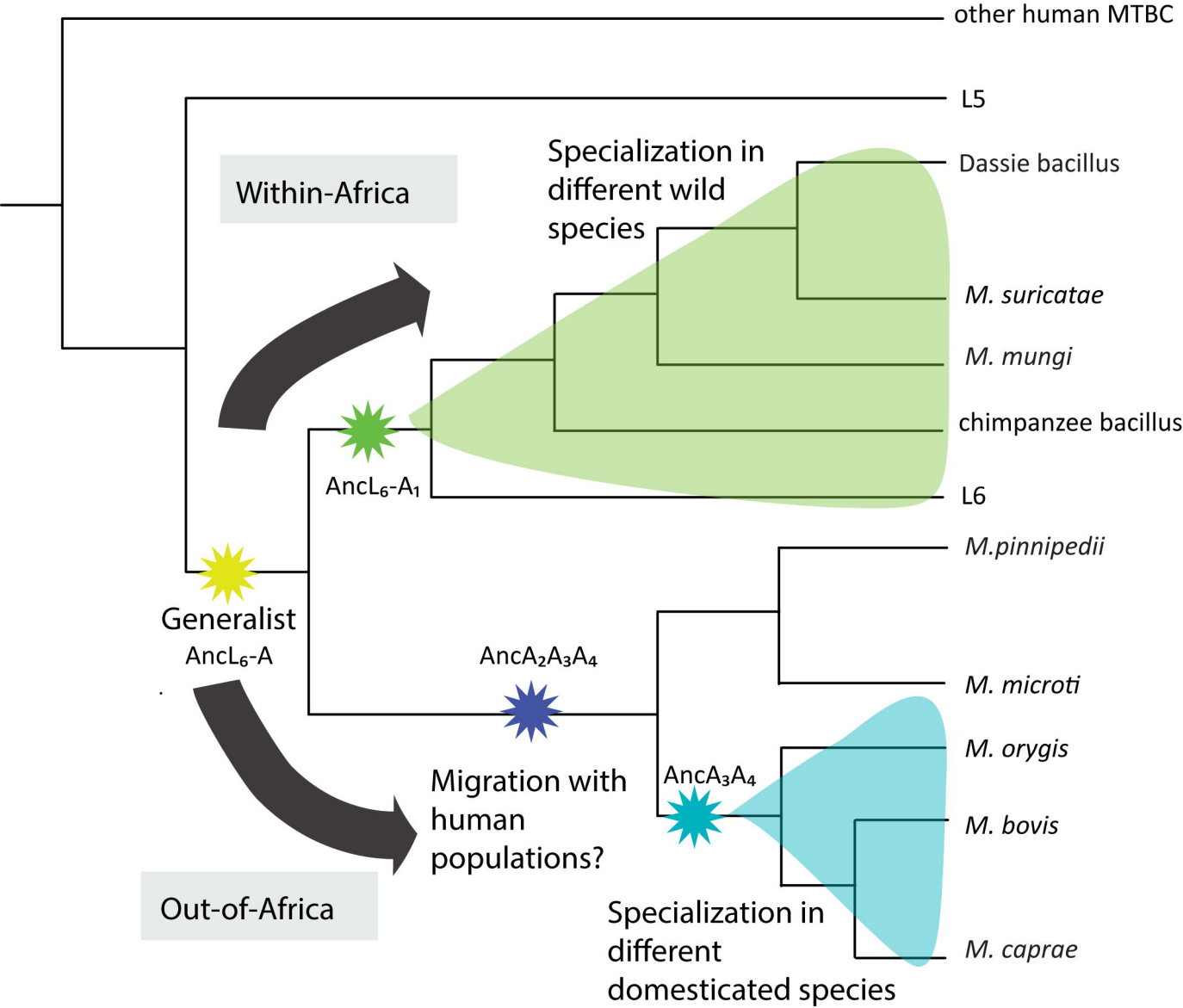

L2
L3
L4
L7
L1
L5
L6
A1
A2
A3
A4
Chimpanzee
M. suricattae
M. mungi
Dassie
RD9
RD7
RD8
RD10
M. microti
M. pinnipedii
M. orygis
M. caprae
BCG
M. bovis
M. canettii
0.002
100


L5
RD900
chimpanzee bacillus
A1
RD1 RD5
M. mungi
M. suricattae
RD1
RD5
Dassie bacillus
L6
M.pinnipedii
RD9
PhoPR
G71S
A2
RD1
RD5
M. microti
RD7
RD8
RD10
RD5
RD900
A3
M. orygis
M. caprae
RD5
RD900
A4
RD1
BCG
M. bovis
ISOLATION COUNTRY
UK
Australia
USA
Russia
Unknown
Netherland
Germany
Italy
ISOLATION HOST
Human
Human Puruvian mummies 11th
Antilope
Vole
Wild Boar
Llama
0.002


SNP pair-wise differences
0
250
500
750
L1
L2
L3
L4
L5
L6
M.bovis
M.caprae
M.microti
M.orygis


20
10
0
-10
PC2 (10.65%)
-10
0
10
PC1 (20.52%)
Clade
A1
A2
A3
A4
L1
L2
L3
L4
L5
L6
L7


L7 1758 1637 1641 1610 1859 1997 0
L6 1793 1815 1801 1784 1753 0 1997
L5 1641 1673 1662 1633 0 1753 1859
L4 1537 1037 1037 0 1633 1784 1610
L3 1570 972 0 1037 1662 1801 1641
L2 1570 0 972 1037 1673 1815 1637
L1 0 1570 1570 1537 1641 1793 1758
L1 L2 L3 L4 L5 L6 L7
bovis 1825 1545 1618 1357 1546 1319 1815 1115 0
caprae 1847 1574 1647 1391 1574 1345 1841 0 1115
orygis 1960 1683 1760 1508 1690 1463 0 1841 1815
micro 1392 1116 1185 931 489 0 1463 1345 1319
pinni 1623 1348 1415 1157 0 489 1690 1574 1546
dassie 1274 936 704 0 1157 931 1508 1391 1357
suric 1532 1199 0 704 1415 1185 1760 1647 1618
mungi 1460 0 1199 936 1348 1116 1683 1574 1545
chimp 0 1460 1532 1274 1623 1392 1960 1847 1825
chimp mungi suric dassie pinni micro orygis caprae bovis


other human MTBC
L5
Within-Africa
Specialization in different wild species
Dassie bacillus
M. suricatae
M. mungi
chimpanzee bacillus
AncL6-A1
L6
M.pinnipedii
Generalist
AncL6-A
AncA2A3A4
M. microti
M. orygis
AncA3A4
Migration with human populations?
M. bovis
Out-of-Africa
Specialization in different domesticated species
M. caprae