Check for updates

RESEARCH ARTICLE

REVISED A new nomenclature for the livestock-associated *Mycobacterium tuberculosis* complex based on phylogenomics

[version 2; peer review: 2 approved]

Michaela Zwyer[1,2], Cengiz Çavusoglu[3], Giovanni Ghielmetti[4], Maria Lodovica Pacciarini[5], Erika Scaltriti[6], Dick Van Soolingen[7,8], Anna Dötsch[1,2], Miriam Reinhard[1,2], Sebastien Gagneux[1,2], Daniela Brites[1,2]

[1]Swiss Tropical and Public Health Institute, Basel, Switzerland
[2]University of Basel, Basel, Switzerland
[3]Department of Medical Microbiology, Ege University Faculty of Medicine, Izmir, Turkey
[4]Institute for Food Safety and Hygiene, Section of Veterinary Bacteriology, University of Zurich, Zurich, Switzerland
[5]National Reference Centre for Bovine Tuberculosis, Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna, Brescia, Italy
[6]Risk Analysis and Genomic Epidemiology Unit, Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia-Romagna, Parma, Italy
[7]National Institute for Public Health and the Environment (RIVM), Bilthoven, Netherlands Antilles
[8]Department of Medical Microbiology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

v2 **First published:** 25 Aug 2021, **1**:100
https://doi.org/10.12688/openreseurope.14029.1
Latest published: 01 Dec 2021, **1**:100
https://doi.org/10.12688/openreseurope.14029.2

Abstract

Background: The bacteria that compose the *Mycobacterium tuberculosis* complex (MTBC) cause tuberculosis (TB) in humans and in different animals, including livestock. Much progress has been made in understanding the population structure of the human-adapted members of the MTBC by combining phylogenetics with genomics. Accompanying the discovery of new genetic diversity, a body of operational nomenclature has evolved to assist comparative and molecular epidemiological studies of human TB. By contrast, for the livestock-associated MTBC members, *Mycobacterium bovis*, *M. caprae* and *M. orygis*, there has been a lack of comprehensive nomenclature to accommodate new genetic diversity uncovered by emerging phylogenomic studies. We propose to fill this gap by putting forward a new nomenclature covering the main phylogenetic groups within *M. bovis*, *M. caprae* and *M. orygis*.
Methods: We gathered a total of 8,736 whole-genome sequences (WGS) from public sources and 39 newly sequenced strains, and selected a subset of 829 WGS, representative of the worldwide diversity of *M. bovis*, *M. caprae* and *M. orygis*. We used phylogenetics and genetic diversity patterns inferred from WGS to define groups.
Results: We propose to divide *M. bovis*, *M. caprae* and *M. orygis* in three main phylogenetic lineages, which we named La1, La2 and La3, respectively. Within La1, we identified several monophyletic groups,

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 01 Dec 2021	✓ view	✓ view
version 1 25 Aug 2021	? view	? view

1. **Lorraine Michelet**, Paris-Est University, Marne-la-Vallée, France

2. **Liliana C. M. Salvador**, University of Georgia, Athens, USA

Any reports and responses or comments on the article can be found at the end of the article.

which we propose to classify into eight sublineages (La1.1-La1.8). These sublineages differed in geographic distribution, with some being geographically restricted and others globally widespread, suggesting different expansion abilities. To ease molecular characterization of these MTBC groups by the community, we provide phylogenetically informed, single nucleotide polymorphisms that can be used as barcodes for genotyping. These markers were implemented in KvarQ and TB-Profiler, which are platform-independent, open-source tools.
Conclusions: Our results contribute to an improved classification of the genetic diversity within the livestock-associated MTBC, which will benefit future molecular epidemiological and evolutionary studies.

Keywords
zoonotic tuberculosis, genetic diversity, mycobacterium tuberculosis complex, phylogenetics, whole-genome sequencing

This article is included in the Microbiology gateway.

This article is included in the European Research Council (ERC) gateway.

This article is included in the Evolution and Ecology gateway.

This article is included in the Genetics and Genomics gateway.

This article is included in the Horizon 2020 gateway.

This article is included in the Evolutionary Biology collection.

Corresponding author: Daniela Brites (d.brites@swisstph.ch)

Author roles: Zwyer M: Data Curation, Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Çavusoglu C**: Resources, Writing – Review & Editing; **Ghielmetti G**: Resources, Writing – Review & Editing; **Pacciarini ML**: Resources, Writing – Review & Editing; **Scaltriti E**: Resources, Writing – Review & Editing; **Van Soolingen D**: Resources, Writing – Review & Editing; **Dötsch A**: Methodology, Writing – Review & Editing; **Reinhard M**: Methodology, Writing – Review & Editing; **Gagneux S**: Funding Acquisition, Investigation, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Brites D**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 883582) This work was also supported by the Swiss National Science Foundation (CRSII5_177163 and 310030_188888)
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2021 Zwyer M *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Zwyer M, Çavusoglu C, Ghielmetti G *et al.* **A new nomenclature for the livestock-associated *Mycobacterium tuberculosis* complex based on phylogenomics [version 2; peer review: 2 approved]** Open Research Europe 2021, **1**:100 https://doi.org/10.12688/openreseurope.14029.2

First published: 25 Aug 2021, **1**:100 https://doi.org/10.12688/openreseurope.14029.1

REVISED **Amendments from Version 1**

Definition of livestock associated MTBC members was added to the introduction.

Rational for considering M. orygis as livestock associated MTBC member was added to the introduction.

10 duplicated WGS were removed from the analysis.

Suite of markers to identify lineages and sublineages was added to TBProfiler.

Any further responses from the reviewers can be found at the end of the article

Plain language summary

Tuberculosis affects humans and livestock species. Its etiological agents are different bacteria belonging to the *Mycobacterium tuberculosis* complex (MTBC). In recent years, whole-genome sequencing (WGS) has become essential in both basic and clinical tuberculosis research. Based on WGS, different human-adapted MTBC genotypes have been classified into lineages and sublineages, which have been shown to differ in their geographic distribution and in virulence. Studies based on WGS are starting to emerge also for livestock-associated MTBC pathogens, but an overarching operational nomenclature systematically covering all known genetic diversity is missing. After gathering several thousands of WGS, we propose here a backbone of operational nomenclature to classify the genetic diversity uncovered by genomic studies of livestock-associated MTBC. Furthermore, a set of molecular markers are provided which can be used to identify the newly proposed lineages and sublineages.

Introduction

Tuberculosis is a leading cause of morbidity and mortality in humans[1]. Moreover, bovine TB (bTB) remains a major economic problem and continues to be a zoonotic threat in many places around the world[2,3]. Human TB is caused mostly by members of the *Mycobacterium tuberculosis* complex (MTBC) collectively known as *Mycobacterium tuberculosis,* and *M. africanum*, whereas bTB in livestock is primarily caused by *M. bovis*. These organisms, and the other members of the MTBC[4], share more than 99% identical nucleotide sequences but can vary considerably in gene content[5]. The MTBC comprises several unique phylogenetic lineages that differ mostly by chromosomal deletions and point mutations. No significant homologous recombination between strains or gene insertion via horizontal gene transfer occurs in the MTBC[6–8]. Despite their high genetic similarity and strict clonality, these lineages exhibit striking differences in host tropism, infecting a wide range of mammalian hosts[9]. For the human-adapted MTBC, a good understanding of the population structure has emerged through comparative analyses of whole-genome sequences (WGS) of TB patient isolates from all over the world. The human-adapted MTBC can be classified into nine phylogenetic lineages: Lineage 1 (L1) to L7, and more recently, two new lineages, L8[10] and L9[11], have been described but remain poorly characterized. Lineages 1-4 and L7 correspond collectively to *M.tuberculosis sensu stricto*, whereas L5 and L6 are traditionally known as *M. africanum*. Further subdivisions among the human-adapted MTBC lineages have been proposed by many different studies to highlight existing within-lineage differences in geographic distribution and genetic differentiation. By contrast, the animal-adapted members of the MTBC remain much less well characterized, and are typically named according to the host species from which they were first, or most commonly, isolated. Considering the growing number of WGS available for many of these pathogens, a more comprehensive and systematic nomenclature beyond the species name is necessary for assisting comparative and molecular epidemiology studies. This is of most relevance for those animal-adapted MTBC members which are a significant cause of TB in livestock species and which also have a high zoonotic potential. In this study, we considered as livestock-associated those MTBC lineages whose evolutionary success is linked to their ability to cause infection and transmit within livestock populations in addition to other host species; *M. bovis*, *M. caprae* and *M. orygis*. Occasionally, TB in livestock can be caused by *M. tuberculosis sensu stricto* or *M. microti*, but these members of the complex have not been shown to transmit within livestock. The low virulence of *M. tuberculosis sensu stricto* in cattle compared to *M. bovis* has also been demonstrated in experimental infections of cattle[12].

For *M. bovis*, there are currently thousands of WGS in the public domain. However, until recently, genetic diversity of *M. bovis* populations was described based on four major groups of genotypes defined by genomic deletions and SNPs. These groups were known as clonal complexes European 1 and 2 (Eu1 and Eu2), and African 1 and 2 (Af1 and Af2). The study of these clonal complexes brought major insights into the genetic diversity underlying bTB in Europe, the Americas and New Zealand (Eu1 and Eu2), as well as in West- and East Africa (Af1 and Af2, respectively)[13–16]. More recently, we and others, have gathered several thousands WGS of *M. bovis*, generating initial insights into the worldwide population structure of this pathogen based on complete genomes[17–21]. Through these efforts, several *M. bovis* sub-populations were identified, and while some corresponded to the previously identified clonal complexes[13–16,21], several others remained unclassified[17–21].

Whereas *M. caprae* is a known cause of infection in livestock species, the association of *M. orygis* with livestock infections is less well established. *M. orygis*, initially thought to be a pathogen of antelope species, has in the meantime been isolated from different hosts[22–24]. Importantly, most available strains today were isolated from humans of South Asian origin[25–30]. In South Asia, *M. orygis* has recently been proposed to be the main cause of zoonotic TB[30]. The main reservoirs of *M. orygis* remain poorly understood, yet it has been isolated from cattle in India and Bangladesh[23,25], and also shown to actively transmit within cattle[23]. India is the country with the biggest cattle population of the world, often living in close proximity with humans, favoring the hypothesis that livestock is the most likely source of zoonotic infections caused by *M. orygis*. Due to its high zoonotic prevalence, the number of *M. orygis* WGS available is steadily increasing, which urges for new definitions aiding comparative genomics.

Here, we propose a comprehensive nomenclature, based on phylogenetic principles and genetic diversity patterns, for the main groups found in what is currently known as *M. bovis*,

M. caprae and *M. orygis*. The nomenclature used for the different members of the MTBC has been repeatedly revised over time, with a particular focus on whether the different MTBC members should be considered separated species or the same species given their high genomic similarity[31]. Classifying the different MTBC members into ecotypes has also been proposed, to better accommodate the differences in host range of the different MTBC members[32,33]. The nomenclature we propose here is not intended as a replacement but rather to serve as an operational nomenclature to assist genomic comparative studies. We propose to take the same hierarchical levels of classification as has been adopted for the human-adapted MTBC lineages and sublineages, as it has proven to be robust and flexible enough to capture diversity both at a global and local level, and is also adequate to describe newly discovered diversity (e.g. L9 and L8). Given the difficulties in defining populations in bacteria, we would like to emphasize that the nomenclature proposed here, might, but does not necessarily have to reflect cohesive groups sharing biological properties. It is rather a pragmatic attempt to find a classification that will usefully describe the genetic diversity and the phylogeographic patterns observed in the MTBC affecting livestock.

Methods

Data collection

Representative dataset for livestock-associated MTBC. We searched the US National Center for Biotechnology Information (NCBI) for new publicly available WGS of *M. bovis*, *M. caprae* and *M. orygis*, using names as search terms: for example for *M. bovis,* "*Mycobacterium tuberculosis variant bovis* [organism]" was searched. Our search was restricted to the time period between the 11th of March 2019, when we already had gathered 3,364 WGS[17], until the 4th of November 2020. A total of 5,383 new genomes concordant with our search terms were available. From these genomes, we excluded those that met the following criteria prior to analysis: genomes registered as bacillus Calmette-Guérin (BCG), as laboratory strains, with unknown country of isolation or isolated in countries already over-represented in previous analyses (Mexico, USA, UK, New Zeeland)[17], and genomes corresponding to strains isolated in patients from low endemic countries with unknown country of origin. Genomes that were publicly available but unpublished at the time of WGS retrieval, were also excluded after a preliminary analysis, as they did not provide new main phylogenetic clades once compared to the representative set of genomes of *M. bovis* and *M. caprae* previously published[17]. Finally, WGS that did not meet our criteria for downstream analysis (average whole-genome coverage > 15x and ratio of heterogenous SNPs to fixed SNPs < 1) were excluded. Furthermore, we newly sequenced 19 genomes from Turkey isolated from humans, two genomes from Italy isolated in cattle in Apulia and Sicily[34], and four genomes from Switzerland isolated in cattle[35]. The selected genomes were added to a previous reference set representing the world-wide diversity of *M. bovis* (n=464) and *M. caprae* (n=12) selected after an initial compilation of 3,364 WGS[17]. For *M. orygis*, 14 newly sequenced genomes isolated from patients and from different zoo animals of South Asian origin[22] were obtained and analysed together with 77 publicly available WGS (*Extended data*, Table 1). In total, 829 representative genomes were considered, of which 675 were *M. bovis*, 63 *M. caprae*, and 91 *M. orygis*. With respect to our previous representative dataset[17], 211 new genomes were added to the downstream analysis for *M. bovis* (*Extended data*, Table 1). Most of these were from animal strains isolated in Brazil (n=19), France (n=83), Germany (n=40), Ethiopia (n=37) and Mali (n=3)[18–21,36], while few derived from human isolates from Tanzania (n=1), Indonesia (n=1), Kazakhstan (n=2) and Moldova (n=1)[37–39] (*Extended data,* Table 1). In the case of *M. caprae*, 51 genomes isolated from Spain were added[40]. The 39 newly sequenced genomes were uploaded to EBI under the study accession numbers PRJEB46653 and PRJEB46575 (*Extended data,* Table 1).

Representative dataset for the complete MTBC. In order to obtain a representative set of world-wide sampled MTBC genomes from both animal and human isolates with a discernible tree topology, we randomly selected genomes from a large in-house collection of WGS (approximately 50,000), for the human lineages 1-6 and for *M. bovis*. The genomes were selected according to the following scheme: 50 random genomes per continent (Africa, America, Asia, Europe, and Oceania) for each lineage. For lineage 1-6, genomes isolated in Northern America, Europe (except Eastern Europe), and Oceania were required to have information about the country of birth of the patient to be considered. Furthermore, WGS from the following strains were added: three strains belonging to the proto Beijing sublineage, eight pyrazinamide susceptible *M. bovis* strains[17], five L9 strains, 23 L7 strains, two L8 strains, 57 *M. caprae* strains, 15 *M. microti* strains, 84 *M. orygis* strains, six *M. pinnipedii* strains, two ancient genomes from Peruvian mummies, one each of Chimp and Dassie bacillus, and one each of *M. mungi* and *M. suricattae*. A complete list containing the accession numbers of all genomes included (n=1,221) can be found in the supplementary data (*Extended data*, Table 2).

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[41]. Sequencing libraries were prepared using NEXTERA XT DNA and the EBNext Ultra II DNA Library Preparation Kits (Illumina, San Diego, USA). Multiplexed libraries were paired-end and single-end sequenced using Illumina HiSeq 2500 (Illumina, San Diego, USA), Illumina NovaSeq 6000 (Illumina, San Diego, USA) and MiSeq (Illumina, San Diego, USA) with 151, 101 and 250 cycles, respectively.

Bioinformatic analysis

Whole-genome sequence analysis. All WGS downloaded, as well as those generated in-house, were analyzed using the WGS analysis pipeline described in 42. Briefly, the retrieved FASTQ files were processed with Trimmomatic v0.3[43] to remove the Illumina adaptors and to trim low quality reads. Only reads of at least 20 bp were kept for further analysis. SeqPrep v 1.2 was then used to merge overlapping paired-end reads (overlap size = 15). We then mapped the resulting reads using BWA v0.7.13[44] (mem algorithm) with respect to the

chromosome of the *M. tuberculosis* H37Rv (NC_000962.3, NCBI). As a reference sequence, we used a reconstructed ancestral sequence of the MTBC[45] where at each position of the chromosome NC_000962.3 the inferred nucleotide of the ancestor of MTBC is the reference. Duplicated reads were marked by the Mark Duplicates module of Picard v 2.9.1 and then excluded. We further performed local realignment of reads around INDELs using the RealignerTargetCreator and IndelRealigner modules of GATK v 3.4.0[46]. Samtools v1.2 mpileup[47] and VarScan v2.4.1[48] were then used for SNP calling with the subsequent thresholds: minimum mapping quality of 20, minimum base quality at a position of 20, minimum read depth at a position of 7x and maximum strand bias of 90%. Only SNPs with a frequency of ≥ 90% within an isolate were considered, and for those with a frequency of ≤ 10% the ancestor state was called. The *M. tuberculosis* H37Rv reference annotation (NC_000962.3, NCBI) was used as the reference genome of *M. bovis* (AF2122/97, NCBI) has no genes absent from H37Rv, except for TbD1[49]. SNPs were annotated with SnpEff v4.11[50]. Positions falling in PE/PPE genes, phages, insertion sequences, and in regions with at least 50 bp identity to other genomic regions were excluded[51].

In silico spoligotyping. All WGS were *in silico* spoligotyped using KvarQ[52]. The respective SB numbers were retrieved by entering the spoligotype patterns into the *Mycobacterium bovis* Spoligotype Database and are reported in *Extended data*, Table 1.

Phylogenetic analyses. The phylogenetic trees were constructed from alignments of variable positions with a percentage of missing data of ≤ 10%. With RAxML v 8.2.11[53] maximum-likelihood phylogenies were constructed by using the general time-reversible model of sequence evolution (-m GTRCAT –V), a rapid bootstrap analysis with 1000 bootstraps and search for the best-scoring maximum-likelihood phylogeny. The MTBC phylogeny was rooted with *M. canetti* (SAMN00102920, NCBI) while all other phylogenies were rooted with a MTBC lineage 6 strain (SAMEA3359865, NCBI). Phylogenetic trees were plotted with ggtree[54] and Figtree.

Population structure and genetic distances. Population structure was evaluated using a Principal Component Analysis (PCA) based on all polymorphic positions obtained from the 1,221 dataset,using the R package adegenet[55] in R 3.5.2. Between and within group genetic distances were measured as raw pair-wise SNP differences for the different groups using the R package ape[56].

Maps of geographic distribution. The geographical origin of the isolates and host-related metadata were recovered from NCBI and used to inform geographic ranges. Strains isolated from zoo animals or isolated from humans living in Europe, Oceania, or North America with unknown place of birth were not taken into account. Since WGS is not performed on a regular basis in all countries, relying only on WGS data would underestimate the geographical distribution of certain clades. To adjust for that, we used the *in silico* SB numbers shown to be phylogenetically informative[17], and searched for publications reporting those SB numbers and their associated geography (*Extended data*, Table 3). The countries of isolation identified in this way were added to those obtained from the WGS and were used to obtain geographic distributions using the rworldmap package[57] in R 3.5.2[58].

Validation of lineage- and sublineage- specific markers. In order to obtain a list of polymorphic positions specific to all members of a defined lineage or sublineage, the variant calls obtained from the 829 La1, La2, and La3 WGS were merged using BCFtools. On the merged dataset, the following filtering steps were applied: First only positions mutated in at least seven genomes were kept using VCFtools (--mac 7)[59], second only positions with a FILTER flag PASS were kept using BCFtools. The first filtering step was included, since we were only interested in SNPs that were common to all members of a sublineage and the lowest number of WGS for a sublineage was seven (unknown6). A genotype matrix was created using the R package VariantAnnotation[60] and by using customized python scripts, those variants mutated in all members of a specific lineage or sublineage, or in a monophyletic group of multiple sublineages (e.g. La1.3 and La1.2) were extracted. This resulted in a list containing 2,203 variants specific to 19 different sublineages and combinations of multiple sublineages. Additionally, we created a list of polymorphic positions using 4,742 WGS representing the genetic diversity of human-adapted lineages L1-L7 and L9[42] and to ensure that our SNPs defining lineages and sublineages among livestock-associated MTBC were specific, we excluded all positions that were polymorphic in the set of 4,742 genomes. This way, a final list of 1,959 SNPs specific to a lineage, sublineage or sublineage combinations within livestock-associated MTBC, and not polymorphic in any of human-adapted MTBC lineages, was generated (*Extended data*, Table 4). Out of the 1863 SNPs, 80 (two to five variants per lineage and sublineage or sublineage combinations) were selected to create a new test suite[61] specific for the livestock-associated MTBC in KvarQ[52]. In order to validate the specificity of the 87 SNPs used in the new KvarQ test suite, we scanned 2,861 livestock-associated WGS from Loiseau *et al*. 2020[17] that were not included in the 829 dataset, and 66 additional WGS randomly chosen from recent publications[62–64] (*Extended data*, Table 5). The 2,927 fastq files were also processed using the workflow described in the WGS sequence analysis section and a phylogenetic tree was inferred as described above. The phylogenetic tree was compared to the lineage and sublineage identity as determined by KvarQ, to assess the accuracy and specificity of the test suite.

Results and discussion

Classification of livestock-associated MTBC into new lineages

After screening an extensive collection of approximately 50,000 WGS, we compiled a comprehensive set of 1,221 WGS representing all MTBC members from all continents in the world (except Antarctica). For the human-adapted lineages (L1 to L6) as well as for *M. bovis*, a large number of WGS is available, and in order to obtain an even representation of these groups with a discernable topology, 50 representatives were randomly selected from each continent and from each lineage. The phylogenetic relationships of these randomly selected 1,221 MTBC strains are represented in Figure 1A. The results

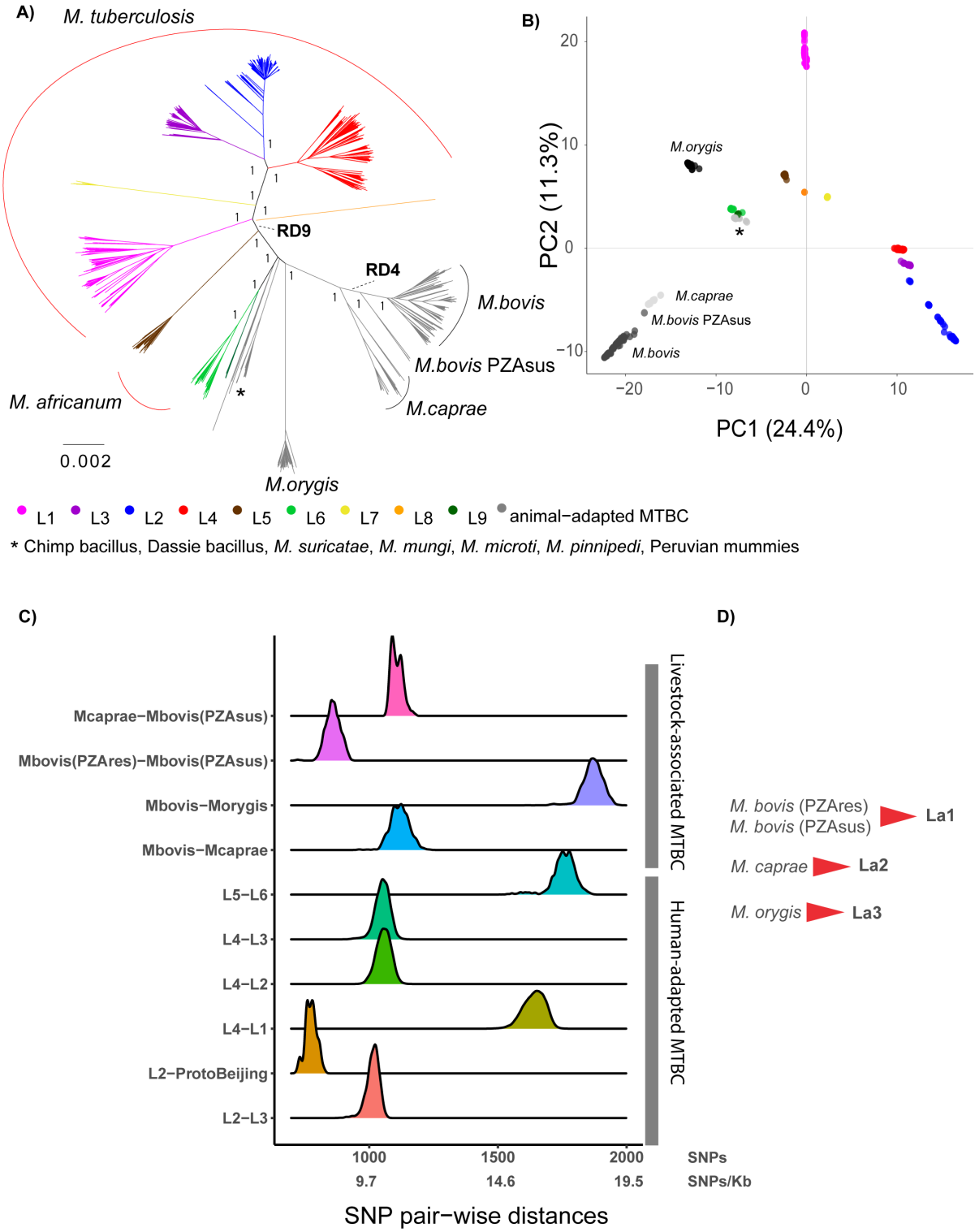

Figure 1. A) Maximum Likelihood topology of 1,221 MTBC representatives, where 50 representatives were randomly selected from each continent and from each lineage (see methods). The tree was inferred from an alignment containing 103,843 polymorphic positions. Branch lengths are proportional to nucleotide substitutions. Support values correspond to bootstrap values. Members of the human-adapted MTBC have tips colored according to their lineage. **B**) Principal Component Analysis (PCA) derived from the same alignment as the phylogeny. The two first principal components are shown. **C**) Distribution of the raw pairwise SNP distances between human adapted MTBC lineages and between different animal adapted MTBC members. **D**) Proposed lineage nomenclature for *M. bovis* susceptible and resistant to pyrazinamide, *M. caprae* and *M. orygis*.

indicated that the human-adapted MTBC members are paraphyletic, given that the group defined by the Region of Difference 9 (RD9)[65] comprises human (L6 and L9) and animal-adapted members (Figure 1A), in line with previous findings[4]. While the distinct clades of the human-adapted members were separated into different lineages and have been named accordingly (Lineage 1-9), the animal-adapted members are still only referred to by their species name. Recent studies, and our searches for WGS from the public domain, indicated that there is a wealth of WGS, in particular for *M. bovis*, representing different geographical areas, hosts, and epidemiological settings of the world[17–21,36,66–68]. *M. orygis*, which has been recently suggested to be the main cause of zoonotic TB in South Asia and possibly a pathogen of cattle in that region[25], also has a growing number of genomes available. There is, however, a lack of consistent nomenclature to assist in the comparative analysis of these genomes. Therefore, we propose to adopt a lineage nomenclature that covers the main groups found in what is currently known as *M. bovis*, *M. caprae* and *M. orygis* based on phylogenetics and genetic diversity patterns. For the remaining animal-adapted MTBC members, *M. mungi, M. suricattae*, the Dassie and Chimpanzee bacillus, as well as *M. pinnipedii and M. microti*, still too few WGS were available to allow for any meaningful within-lineage diversity analysis. In addition, the host range and ecology of these ecotypes remain poorly understood. We reasoned that these cases would require more extensive sampling, and thus focused the remaining of our analyses on *M. bovis*, *M. caprae* and *M. orygis*.

A phylogenomics-based nomenclature for *M. bovis*, *M. caprae* and *M. orygis*

These three members of the MTBC evolved from a common ancestor not shared by any other group within the MTBC (Figure 1A). The visual inspection of the phylogeny and the PCA plot suggested that among these three groups, there are four main phylogenetic clades: *M. orygis*, *M. caprae*, the pyrazinamide-susceptible *M. bovis*[17] and the pyrazinamide-resistant *M. bovis* (Figure 1 A&B). The long branches leading to these clades indicate that many genetic changes have occurred in their founding ancestor populations, and this was also reflected in the pair-wise SNP distances between these clades estimated from the 1,221 whole-genomes dataset (Figure 1C). We suggest classifying these four clades into three main lineages within the MTBC analogously to the human lineages, considering *M. bovis* pyrazinamide- resistant and -susceptible as one lineage, and *M. caprae* and *M. orygis* as two other main lineages. We propose adopting the numerical lineage nomenclature used for the human-adapted MTBC members, adding the lower-letter "a" standing for "animal". This nomenclature distinguishes the human-adapted from the remaining members of the complex, which can be of relevance for clinicians; simultaneously, for the non-human adapted MTBC members, it has the advantage of being agnostic with respect to the host species, which can be multiple. In this way, we suggest naming La1, La2 and La3 the groups currently known as pyrazinamide-resistant and -susceptible *M. bovis*, *M. caprae* and *M. orygis*, respectively (Figure 1D).

The pyrazinamide-susceptible *M. bovis* group is composed of pyrazinamide-susceptible strains within *M. bovis*, and is geographically restricted to East Africa (*Extended data*, Table 1, Figure 2)[17]. This group of strains were quite divergent from the pyrazinamide-resistant *M. bovis*, yet closer to the latter than to *M. caprae* (Figure 1C). A similar situation occurred within the human-adapted L2 when comparing the so-called Proto-Beijing group with the remaining strains of L2 (Figure 1C). The available WGS of pyrazinamide-susceptible *M. bovis* came from strains isolated in humans, cattle and a zoo antelope, and no new WGS in our current analysis have been added with respect to previous studies[17,20]. *In silico* determination of spoligotypes (*Extended data*, Table 1) revealed that similar patterns are common in cattle from Tanzania and Uganda[15,69,70], and have also been observed in different wild animal species in Tanzania[70]. In our extensive WGS collections of MTBC isolates from TB patients in Uganda and Tanzania (unpublished), we did not find any representatives of pyrazinamide-susceptible *M. bovis*, suggesting that zoonotic transfers of this group of strains are rare, like for other *M. bovis* strains.

Unlike the human-adapted lineages of the MTBC, La1, La2 and La3 are multi-host pathogens known to infect livestock and other wild mammal species, and occasionally humans[9,25]. The multiple host species from which these isolates were obtained, are in line with that notion (*Extended data*, Table 1). Despite this general broad host range, these lineages differ substantially in their geographic distribution, suggesting local adaptation to different hosts and/or different dispersion abilities of their host populations (Figure 2). The evolutionary success of La1, and its broad distribution around the world, are linked with the ability of La1 to infect different species of livestock, in particular cattle. Additionally, its broad host tropism also contributes to this success, as demonstrated by the difficulties in eradicating bovine TB even in high-resource countries, where La1 can be maintained in different wildlife species that live in close proximity to livestock such as badgers, deer, or wild boar, or possums[71]. Various molecular markers, and more

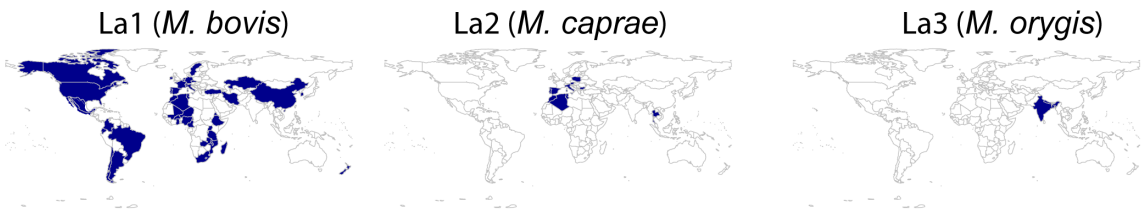

Figure 2. Geographic distribution of La1, La2 and La3 informed by WGS and *in silico* spoligotype patterns.

recently WGS, suggest no preferential association of La1 genotypes with particular host species[68,72,73]. It is thus still unclear whether La1 infections in non-bovid species are the result of spillover events from cattle populations (i.e. La1 is better adapted to cattle than to other animal species), or if La1 has an intrinsically broader host spectrum that can lead to similarly successful infectious cycles in many different animal species. Interestingly, despite its broad host repertoire and the ability to cause zoonotic TB, La1 is not able to sustainably cause infectious cycles in immune-competent humans. Despite being much less studied, a similar rationale might apply to La2 and La3, as we shall discuss next.

La2, or *M. caprae*, is globally associated with a much lower burden of disease compared to La1, and that is presumably also reflected in a much lower number of WGS available. La2 is, however, a significant regional cause of animal TB as it is the main cause of TB in goats in the Iberian Peninsula[74], affects several livestock and wild animal species populations in Central Europe[75,76], and is occasionally a source of zoonotic TB[77]. Indeed, a study in Germany showed that up to one third of zoonotic TB cases in that country were caused by La2[78]. Two of our newly sequenced genomes belonged to La2, with one corresponding to an isolate from cattle in Switzerland[35] and the other from a patient in Turkey (Figure 4). Both were closely related to La2 strains isolated in Spain and in Germany[40,79]. The geographic distribution of La2 obtained from the WGS metadata and from searching the literature using the spoligotypes patterns determined *in silico* (*Extended data,* Table 3) confirms, as previously suggested, that La2 is not restricted to

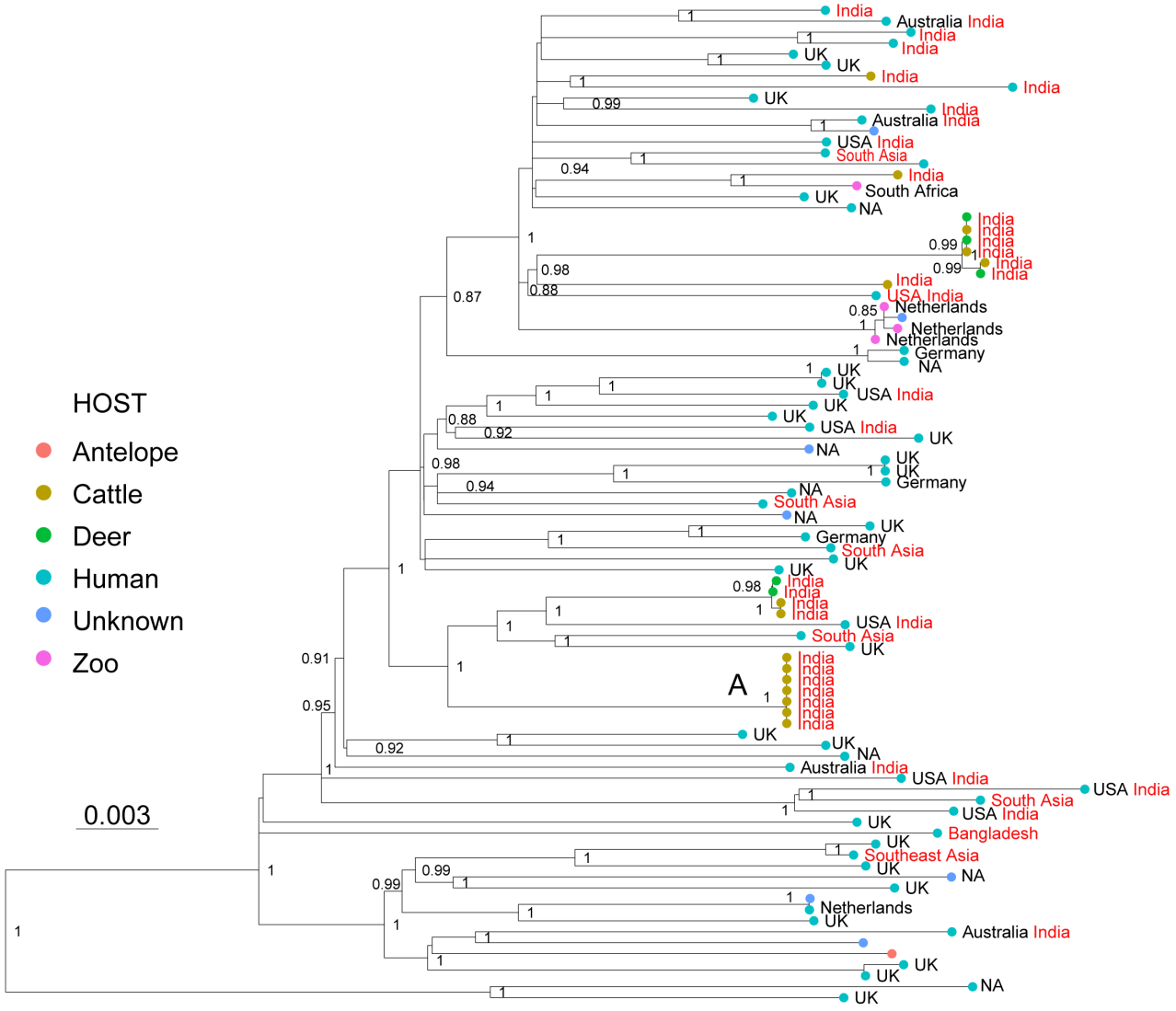

Figure 3. Maximum Likelihood topology based on 2,114 polymorphic positions derived from 91 WGS of La3, after conservatively filtering out several repetitive regions of the genome (see methods). Branch lengths are proportional to nucleotide substitutions and the topology is rooted with one L6 WGS. Support values correspond to bootstrap values. The different colors of the tips correspond to different hosts indicated in the legend. Country of isolation is indicated, followed by the country of birth in the case of human isolates, when known. Isolates with origin in South Asia are indicated in red. A cluster of WGS obtained from cattle isolated is indicated with A.

Europe[17] but also occurs in Africa, South America and East Asia (*Extended data*, Table 1, Figure 2). Our phylogenetic reconstruction also revealed that La2 exhibits strong population divisions, in particular between isolates of Asian and European origin (Figure 1A & Figure 3). However, better sampling, including more isolates from Africa, America and Asia, will be necessary to better understand the biogeography and evolutionary history of La2.

The most distantly related group within the livestock-associated lineages is La3, commonly known as *M. orygis*. La3 was originally isolated from a captive *oryx* antelope, and has since then been isolated from many different wild, zoo and domestic animals, and from patients of South Asian origin in low endemic TB countries[22,26–29]. In India, Bangladesh and Nepal, La3 has been isolated from humans, cattle, primates, deer and a wild rhinoceros[23–25]. The native geographic distribution of this pathogen seems to be restricted to South Asia where it is possibly the main cause of zoonotic TB[25]. Here, we compiled 91 WGS of La3 from different sources: 1) isolates from low TB endemic countries from patients of South Asian (n=13) or unknown origin (n=35)[30], 2) isolates from patients in Southern India (n=5)[25] and one patient from Bangladesh[80], 3) isolates from cattle (n=15) and deer (n=5) from different Indian regions[25]. The remaining publicly available genomes were of unknown origin and unknown host species. The 14 newly sequenced La3 isolates were obtained from zoo animals and from patients of South Asian origin in the Netherlands[22] (*Extended data*, Table 1). The genetic relationships among the 91 WGS showed that the isolates from low TB endemic countries, isolates from zoo animals, and isolates obtained in India, both from patients and from veterinarian samples, appeared intermingled in the phylogenetic tree; they were separated by relatively long branches, suggesting a common origin of infection in South Asia (Figure 3). Little is known about the transmission of La3, and the host preferences of this pathogen also remain unclear[25,81]. The phylogenetic relationships presented here are consistent with direct transmission from cattle-to-cattle in India (Figure 3, cluster A), but they remain inconclusive with respect to direct transmission among and between the other host species. Cattle-to-cattle transmission of La3 inferred through mini-satellite markers (MIRU-VNTR) has been reported previously in Bangladesh[23]. In contrast, no evidence of patient-to-patient transmission has been shown yet, although transmission from one TB patient to cattle has been reported[26], suggesting that humans are not necessarily a dead end for La3. The La3 patient samples analyzed here are not well-suited to capture direct transmission given that they mostly represent active TB cases in emigrated patients who most likely have acquired their infection in their country of origin. One exception was the data published by Duffy and colleagues[25], which was the first to report infection in patients by La3 within the endemic geographic range of this pathogen. Their findings suggest that human infections by La3 are relatively rare when compared to *M. tuberculosis*, given that, of the almost 1,000 patient samples collected in a referral hospital in southern India, only 0.7% belonged to La3. In addition, patients reported to be infected with La3 were often associated with non-pulmonary TB[25,27]. This is indirect evidence pointing to La3 not being very successful at maintaining infectious cycles in humans, in a way that is reminiscent of zoonotic infections by *M. bovis*, as already suggested by 25. Future studies are needed to better understand the host preferences of La3 and how this lineage is transmitted between species. However, given that bTB is endemic in India, which also harbors the largest population of cattle in the world[82], a plausible scenario is that cattle may play an important role in the dynamics of La3 infections.

Sublineages within La1

Lineage a1 is the most studied member of the animal-adapted MTBC, since bTB has a major economic impact and it is the most common cause of zoonotic TB[9,83]. In recent years, several studies have compared large collections of WGS of *M. bovis*, bringing new insights into the local transmission dynamics and into to the global population structure, phylogeography and evolutionary history of this pathogen[17–20,66,68]. In a previous study, after an initial compilation of 3,364 genomes representing 35 countries around the world, we defined a reference set of 476 WGS representing the global diversity of *M. bovis* (n=464) and *M. caprae* (n=12)[17]. Our results revealed that a large proportion of these genomes belonged to the clonal complex Eu1[13], reflecting biases in sampling and WGS efforts towards the United Kingdom and its former trading partners. Other regions of the world with high *M. bovis* prevalence remained comparatively under-sampled, and yet, we identified several clades within *M. bovis* that did not belong to either Eu1 or to any of the clonal complexes known at the time[17]. Here, we aimed to improve the WGS representation of these previously unclassified clades and to identify new clades by including WGS from countries that were previously under-sampled. After a new search of 5,383 entries on the public domain, following a set of exclusion criteria (see M), and our own sequencing efforts (19 strains from Turkey, four from Switzerland and two from Italy), we added 221 *M. bovis* and 63 *M. caprae* WGS to the previous identified reference set of 476 WGS[17]. In total, we newly analysed 675 La1 and 63 La2 WGS. The phylogenetic reconstruction of the 738 genomes represented in Figure 4 revealed several clades diverging early from the common ancestor of all La1. All these clades were previously identified, and while some formed monophyletic groups corresponding to clonal complexes already defined (Eu1, Eu2, Af1 and Af2), the remaining clades were named transiently as unknown1 to unknown9[17,20]. In the present analysis, several WGS were added to these unclassified clades, but including more WGS available at NCBI (see methods) did not uncover any new deeply rooted and divergent clades. We therefore considered that the 675 genomes selected here provided a good representation of the global diversity of La1 in its main groups, and could be used to delineate a systematic nomenclature to assist future comparative studies.

Branches that represent deep splits from the most recent common ancestor of La1 leading to monophyletic groups represent evolutionary successful populations deriving from common founder events, and thus from a common genetic pool. The strains belonging to these monophyletic groups might share biological properties, which are more similar within than between groups. We have used this rationale to split La1 into

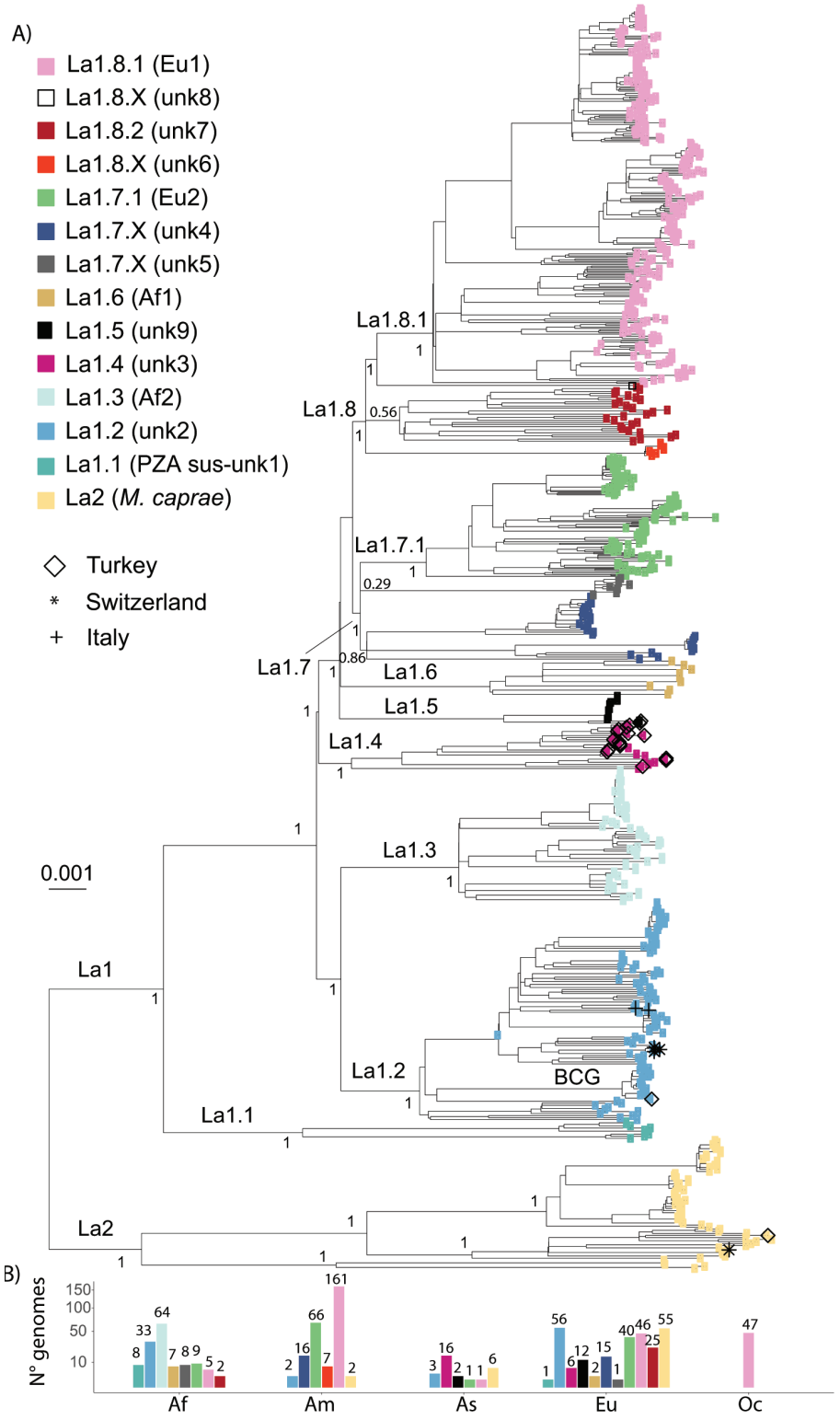

Figure 4. A) Maximum Likelihood topology based on 34,308 polymorphic positions derived from 675 La1 and 63 La2 WGS, after conservatively filtering out several repetitive regions of the genome (see methods). Branch lengths are proportional to nucleotide substitutions and the topology is rooted with one L6 WGS. Support values indicated for the main divisions correspond to bootstrap values. Monophyletic clades corresponding to sublineage divisions are indicated in color as in the legend. Newly sequence La1 in this study are indicated by different symbols as in the legend. **B**) Number of WGS included in the phylogenetic tree per continent (Af = Africa, Am = America, As = Asia, Eu = Europe, Oc = Oceania) and sublineage shown on a square-root scale. The bars are colored according to the sublineage.

several clades, which we hereafter call sublineages in analogy to the human-adapted sublineages of the MTBC. In addition, to increase the operative value of this nomenclature, we have taken into account the geographic distribution of these groups whenever possible, and have attempted to be consistent with the clonal complex nomenclature already in use[17,19]. The correspondence between the nomenclature we propose here and previously defined groups based on WGS[19,21] is given in *Extended data*, Table 1.

Sublineages La1.1 to La1.3. A sublineage classification was attributed to all well-resolved monophyletic clades showing a strong statistical support and separating deeply from the most recent common ancestor of La1 (Figure 4 & Figure 5). This was clearly the case for the pyrazimamide-susceptible *M. bovis*[17], the unknown2 group[17] and clonal complex Af2[15]. We thus propose to classify these groups as sublineages La1.1, La.1.2 and La1.3, respectively (Figure 4 & Figure 5). The shape of the pairwise SNP distances between and among these sublineages also reflects that they have diverged markedly from each other (*Extended data*, Figure 1). Sublineage La1.2 appeared as one of the main genotypes circulating in continental Europe (Figures 4 A & B and Figure 5). This sublineage had been recently called clonal complex European 3 (Eu3)[36,84]. The

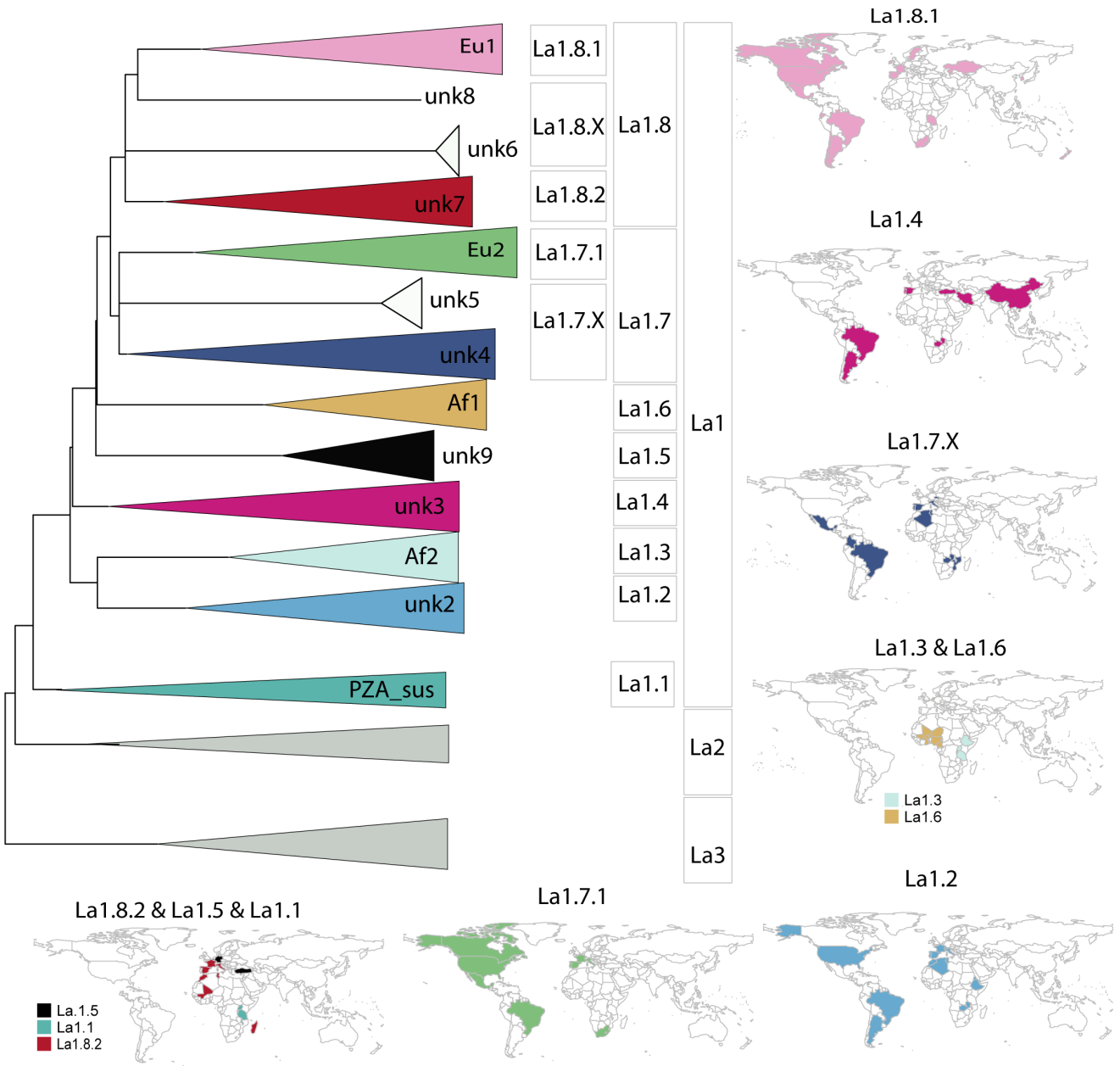

Figure 5. A) Schematic illustration of La1 sublineages. The length of the branch is not proportional to genetic distances. Groups without strong bootstrap support (<80) are shown as polytomies. Color codes are the same as in Figure 4. **B**) Geographic distribution of La1 sublineages informed by WGS and *in silico* spoligotype patterns (Extended data: Table 3).

WGS we have obtained from cattle in Switzerland and Italy belonged to La1.2 (Figure 4 A), as did a high proportion of genomes isolated from different host species in France[21]. The BCG group of strains belongs to La1.2, and is closely related to a veterinarian sample from France, in line with the origin of the BCG vaccine strain in that country[85]. Several genomes isolated in Ethiopia also belonged to La1.2[36], reinforcing the notion that this sublineage has a strong presence in both Western Europe and East-Africa[17,36].

Sublineages La1.4 to La1.8. The topology of the phylogenetic tree suggested that the remaining extant groups of La1 were founded in many instances by very closely related ancestral populations, as shown by the very short internal branches connecting them (Figure 4 A). This could be explained by a history of several migrations occurring more or less simultaneously, followed by rapid diversification in different parts of the world resulting in extant groups with a markedly different geographic distribution (e.g. the clonal complex Af1 and unknown9 groups, the unknown4 and unknown5 groups, and the unknown6 and unknown7 groups, Figure 4 A & Figure 5). Yet, the splits leading to the unknown3, unknown9 and clonal complex Af1 groups are well supported statistically (Figure 4), and the distributions of their within- and between- pair-wise SNP distances differ markedly (*Extended data*, Figure 1). Moreover, these groups also occupy different geographic regions (Figure 5). We therefore propose to classify unknown3, unknown9 and Af1 as sublineages La1.4, La1.5 and La1.6, respectively (Figure 4 & Figure 5). Most of our isolates from Turkey belonged to La1.4. The geographical distribution of La1.4 based on the WGS and *in silico* spoligotyping suggests a broad distribution spanning Asia, Europe and South America. As for La1.5, the WGS analysed here were isolated from several captive animal species in Germany[20], originally classified as group 09[20], and from humans in Turkey (this study). All these genomes had the spoligotype pattern SB0989, of which we found reports only in the mentioned geographical regions and in Albania[20,86,87]. As for La1.6, only few WGS were available; however, the work by Muller *et al.*,[16], which provides a very comprehensive description of clonal complex Af1, highlighted the restriction of this group of strains to West-Africa.

The unknown4, clonal complex Eu2 and unknown5 groups form a well-supported monophyletic group. However, the relationships between these groups are unresolved (Figure 4). Clonal complex Eu2 has diverged from the remaining strains, forming a well-supported monophyletic clade (Figure 4). The strains classified as unknown4 form quite a diverse group, as indicated by the relatively long branches coalescing to their common ancestor and by the distribution of their within-pair-wise SNP distances (Figure 4 & *Extended data*, Figure 1). These latter strains were mostly isolated in Brazil[18], France[21] and Germany[20]. The modes of the pair-wise SNP distribution of unknown4 and unknown5 suggest, when compared to densely sampled groups like Eu2, that sampling might be incomplete (*Extended data*, Figure 1). Finally, clonal complex Eu2 and unknown4 occur in Western Europe, America and Southern Africa, overlapping in their geographic distribution and possibly reflecting dispersion events between South America and Western Europe (Figure 5). As for unknown5, only eight closely related genomes from strains isolated in Zambia were available. All eight genomes have the phylogenetically uninformative spoligotype pattern SB0120, limiting further inferences[17]. Based on the above discussed points, we suggest to include clonal complex Eu2, unknown4 and unknown5 in one sublineage, hereafter called La1.7, to further classify clonal complex Eu2 as a subgroup within La1.7 called La1.7.1 and to classify unknown4 and unknown5 as sublineage La1.7.X (Figure 4 & Figure 5). Further studies with better sampling of the unknown4 and 5 groups are necessary to better understand their population structure.

The remaining genomes classified as clonal complex Eu1, unknown7 and unknown6, and a single genome with an origin in Ethiopia (unknown8)[17], also form a well-supported monophyletic clade. But similarly to the example discussed above, the phylogenetic relationships of their ancestors are not well resolved. Clonal complex Eu1 and unknown7 each form well-supported monophyletic groups and have distinct geographic distributions. While clonal complex Eu1 has a broad distribution all around the world, strains classified as unknown7 seem much more geographically restricted (Figure 5). As for unknown6, only seven closely related genomes were available, all corresponding to strains isolated in cervids in the USA[66]. Unknown8 is most closely related to clonal complex Eu1, and as discussed elsewhere[17,36,88], probably belongs to a group of strains circulating in Ethiopia. We suggest classifying these groups into one sublineage, La1.8, and to further subdivide clonal complex Eu1 and the group unknown7 into La1.8.1 and La1.8.2, respectively (Figure 4 & Figure 5). La1.8.1 is among the best characterised groups of strains within *M. bovis*, known to be prevalent in the UK and in regions of the world known to be former UK trading partners. L1.8.2 was mostly composed of WGS from isolates from France[21] and a few from Ethiopia[36]. Similar spoligoytypes have been reported in Western- and Southern Europe, suggesting that this might be another common genotype circulating in continental Europe (*Extended data*, Table 3). In addition, similar spoligotypes have been described in different African countries including Madagascar (Figure 5). As for unknown6 and 8, we suggest a transient classification as La1.8.X which can be revised once more genomes become available. Given the bTB surveillance measurements taking place specially in Western countries, it is also possible that some of these groups are now rare or have even become extinct.

Validation of lineage- and sublineage- specific markers

We identified SNPs that are specific to La1, La2 and La3 lineages and La1 sublineages, and which can be used as genotyping markers (see Methods). To ensure specificity, the resulting list of phylogenetic SNPs obtained from the 829 dataset was compared to a set of polymorphic positions (370,449) obtained from 4,742 WGS representing the genetic diversity of human-adapted lineages L1-L7 and L9[42]. After excluding those SNPs, occurring in at least one out of the 4,742 genomes representing human-adapted MTBC, 1,959 SNPs remained that were specific for La1, La2, La3 and the described La1 sublineages (*Extended data*, Table 4). Thereof 87 were selected as phylogenetic markers to create a test suite for KvarQ[52] (See analysis code, *Extended data*). KvarQ is a user-friendly and platform-independent tool that enables scanning fastq files

for a given list of SNPs, without the need for aligning sequencing reads to a reference genome or *de novo* assemblies[52]. We validated the test suite with 2,774 WGS from Loiseau *et al.*,[17] not included in our initial 831 dataset and 66 WGS randomly chosen from recently published WGS isolated in Brazil and Algeria[62–64]. In parallel, the WGS used to validate KvarQ were aligned with respect to the genome of reference, as indicated in the Methods and used together with the WGS from the 829 dataset to infer a new phylogenetic tree. According to the KvarQ results (*Extended data*, Table 5), all WGS belonged to one of the defined La1 sublineages, or to La2, and the visual inspection of the phylogenetic tree indicated that all lineage/sublineage assignments by KvarQ were correct.

Thus, here we provide a specific set of polymorphic positions that can be used to develop molecular assays to classify strains. These are provided in Table 4 (*Extended data*) both as coordinates with respect to our genome of reference as well as with respect to the first position of genes. In the cases for which WGS exist, sequencing reads can be queried with a new suite of markers (See Zenodo repository)[61], using KvarQ and bypassing the need to run conventional alignment approaches and phylogenetic analysis for strain classification. The same markers have been implemented in TBProfiler[89].

Conclusions

In recent years, several thousands of WGS became available for *M. bovis*, *M. caprae* and *M. orygis*, in particular for the former. Previous phylogenomic studies have unveiled that these pathogens, despite being associated with livestock species, exhibit a broad host species range and marked differences in the geographic distribution of various genotypes. Hypothetically, these genotypes might also differ in pathogenicity as observed in the case of the human-adapted MTBC members. As the number of WGS of livestock-associated MTBC continues to grow in public repositories, there is a need for a practical nomenclature allowing comparative analysis and hypothesis testing. After gathering several thousands of WGS and selecting representatives of different genotypes and geographic regions, we have obtained an exhaustive phylogenetic depiction of *M. bovis*, *M. caprae* and *M. orygis* as well as of the main genetic groups currently known within *M. bovis*. In analogy with the nomenclature in use by the scientific community for the human-adapted members of the MTBC, we proposed here a body of operational nomenclature hierarchically classifying genetic groups within the livestock-associated members in lineages and sublineages. This nomenclature classifies all main genetic groups that are known currently, and is flexible so as to accommodate new genetic diversity uncovered by future studies. We also provided specific marker SNPs that can be used to develop molecular assays to identify each of the lineages and sublineages proposed. Furthermore, we developed a new SNP test suite implemented in KvarQ and TBProfiler, which allows querying WGS without requiring a lot of bioinformatics expertise.

Data availability

Underlying data

European Nucleotide Archive (EBI-EMBL): A new nomenclature for the livestock-associated Mycobacterium tuberculosis complex based on phylogenomics. Accession number: PRJEB46653, https://identifiers.org/ena.embl:PRJEB46653

European Nucleotide Archive (EBI-EMBL): Whole Genome sequencing (WGS) of Mycobacterium bovis spoligotype SB0120 and SB0841 isolates circulating in Italy. Accession number PRJEB46575, https://identifiers.org/ena.embl:PRJEB46575

Zenodo: A new nomenclature for the livestock-associated Mycobacterium tuberculosis complex based on phylogenomics, https://doi.org/10.5281/zenodo.5153095[90]

This project contains the following underlying data:

- Table 1: Accession numbers and metadata associated with the 829 WGS used of La1, La2 and La3.
- Table 2 - Accession numbers and metadata associated with the 1,221 WGS used as representatives of the whole MTBC.

Extended data

Zenodo: A new nomenclature for the livestock-associated Mycobacterium tuberculosis complex based on phylogenomics, https://doi.org/10.5281/zenodo.5730685

This project contains the following extended data:

- Figure 1: Distribution of the raw pairwise SNP distances between and within main La1 groups.
- Table 3: Spoligotypes patterns inferred from the WGS and used to complement the geographic distribution of La1 sublineages.
- Table 4: Single nucleotide polymorphisms (SNPs) specific to livestock-associated MTBC lineages and sublineages. Coordinates based on the *M. tuberculosis* H37Rv annotation (NC_000962.3) are given (Position_ref), and the lineage and or sublineage classification (PhylogeneticSNP). Additionally, the gene-based position is indicated (Position_gene) as well as the kind of mutation based on SnpEff annotation[50]. SNPs used to create the new KvarQ testsuite are indicated.
- Table 5: KvarQ results of lineage and sublineage typing done with the new testsuite implemented.

Analysis code available from :

- https://github.com/dbrites/LivestockAssociatedMTBC
- Archived analysis code at time of publication: DOI: 10.5281/zenodo.5730644
- License:GNU

Test suite and sublineages implementable in KvarQ[52] available from:

https://github.com/dbrites/LivestockAssociatedMTBC/tree/main/KvarQ_testsuite/MTBC_animals

- Archived analysis code at time of publication: DOI: 10.5281/zenodo.5730644
- License: GNU

Acknowledgements

Calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing core facility at University of Basel. Data were obtained from the TB Portals (https://tbportals.niaid.nih.gov), which is an open-access TB data resource supported by the National Institute of Allergy and Infectious Diseases (NIAID) Office of Cyber Infrastructure and Computational Biology(OCICB) in Bethesda, MD. These data were collected and submitted by members of the TB Portals Consortium (https://tbportals.niaid.nih.gov/Partners) and other data contributors that originally submitted the data to the TB Portals did not participate in the design or analysis of this study.

References

1. WHO: **Global Tuberculosis Report.** WHO, 2020.
Reference Source

2. Ayele WY, Neill SD, Zinsstag J, *et al.*: **Bovine tuberculosis: an old disease but a new threat to Africa.** *Int J Tuberc Lung Dis.* 2004; **8**(8): 924–937.
PubMed Abstract

3. Reviriego Gordejo FJ, Vermeersch JP: **Towards eradication of bovine tuberculosis in the European Union.** *Vet Microbiol.* 2006; **112**(2–4): 101–109.
PubMed Abstract | Publisher Full Text

4. Brites D, Loiseau C, Menardo F, *et al.*: **A New Phylogenetic Framework for the Animal-Adapted *Mycobacterium tuberculosis* Complex.** *Front Microbiol.* Original Research, 2018; **9**: 2820.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Bolotin E, Hershberg R: **Gene Loss Dominates As a Source of Genetic Variation within Clonal Pathogenic Bacterial Species.** *Genome Biol Evol.* 2015; **7**(8): 2173–2187.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Godfroid M, Dagan T, Kupczok A: **Recombination Signal in *Mycobacterium tuberculosis* Stems from Reference-guided Assemblies and Alignment Artefacts.** *Genome Biol Evol.* 2018; **10**(8): 1920–1926.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Namouchi A, Didelot X, Schöck U, *et al.*: **After the bottleneck: Genome-wide diversification of the *Mycobacterium tuberculosis* complex by mutation, recombination, and natural selection.** *Genome Res.* 2012; **22**(4): 721–734.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Boritsch EC, Khanna V, Pawlik A, *et al.*: **Key experimental evidence of chromosomal DNA transfer among selected tuberculosis-causing mycobacteria.** *Proc Natl Acad Sci U S A.* 2016; **113**(35): 9876–9881.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Malone KM, Gordon SV: ***Mycobacterium tuberculosis* Complex Members Adapted to Wild and Domestic Animals.** *Adv Exp Med Biol.* 2017; **1019**: 135–154.
PubMed Abstract | Publisher Full Text

10. Ngabonziza JCS, Loiseau C, Marceau M, *et al.*: **A sister lineage of the *Mycobacterium tuberculosis* complex discovered in the African Great Lakes region.** *Nat Commun.* 2020; **11**(1): 2917.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Coscolla M, Gagneux S, Menardo F, *et al.*: **Phylogenomics of *Mycobacterium africanum* reveals a new lineage and a complex evolutionary history.** *Microb Genom.* 2021; **7**(2): 000477.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Villarreal-Ramos B, Berg S, Whelan A, *et al.*: **Experimental infection of cattle with Mycobacterium tuberculosis isolates shows the attenuation of the human tubercle bacillus for cattle.** *Sci Rep.* 2018; **8**(1): 894.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Smith NH, Berg S, Dale J, *et al.*: **European 1: a globally important clonal complex of *Mycobacterium bovis*.** *Infect Genet Evol.* 2011; **11**(6): 1340–1351.
PubMed Abstract | Publisher Full Text

14. Rodriguez-Campos S, Schürch AC, Dale J, *et al.*: **European 2--a clonal complex of *Mycobacterium bovis* dominant in the Iberian Peninsula.** *Infect Genet Evol.* 2012; **12**(4): 866–872.
PubMed Abstract | Publisher Full Text

15. Berg S, Garcia-Pelayo MC, Müller B, *et al.*: **African 2, a clonal complex of *Mycobacterium bovis* epidemiologically important in East Africa.** *J Bacteriol.* 2011; **193**(3): 670–678.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Müller B, Hilty M, Berg S, *et al.*: **African 1, An Epidemiologically Important Clonal Complex of *Mycobacterium bovis* Dominant in Mali, Nigeria, Cameroon, and Chad.** *J Bacteriol.* 2009; **191**(6): 1951–1960.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Loiseau C, Menardo F, Aseffa A, *et al.*: **An African origin for *Mycobacterium bovis*.** *Evol Med Public Health.* 2020; **2020**(1): 49–59.
PubMed Abstract | Publisher Full Text | Free Full Text

18. da Conceição ML, Conceição EC, Furlaneto IP, *et al.*: **Phylogenomic Perspective on a Unique *Mycobacterium bovis* Clade Dominating Bovine Tuberculosis Infections among Cattle and Buffalos in Northern Brazil.** *Sci Rep.* 2020; **10**(1): 1747.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Zimpel CK, Patané JSL, Guedes ACP, *et al.*: **Global Distribution and Evolution of *Mycobacterium bovis* Lineages.** *Front Microbiol.* 2020; **11**: 843.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Kohl TA, Kranzer K, Andres S, *et al.*: **Population Structure of *Mycobacterium bovis* in Germany: a Long-Term Study Using Whole-Genome Sequencing Combined with Conventional Molecular Typing Methods.** *J Clin Microbiol.* 2020; **58**(11): e01573–20.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Hauer A, Michelet L, Cochard T, *et al.*: **Accurate Phylogenetic Relationships Among *Mycobacterium bovis* Strains Circulating in France Based on Whole Genome Sequencing and Single Nucleotide Polymorphism Analysis.** *Front Microbiol.* 2019; **10**: 955.
PubMed Abstract | Publisher Full Text | Free Full Text

22. van Ingen J, Rahim Z, Mulder A, *et al.*: **Characterization of *Mycobacterium orygis as M. tuberculosis* complex subspecies.** *Emerg Infect Dis.* 2012; **18**(4): 653–655.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Rahim Z, Thapa J, Fukushima Y, *et al.*: **Tuberculosis Caused by *Mycobacterium orygis* in Dairy Cattle and Captured Monkeys in Bangladesh: a New Scenario of Tuberculosis in South Asia.** *Transbound Emerg Dis.* 2017; **64**(6): 1965–1969.
PubMed Abstract | Publisher Full Text

24. Thapa J, Nakajima C, Maharjan B, *et al.*: **Molecular characterization of *Mycobacterium orygis* isolates from wild animals of Nepal.** *Jpn J Vet Res.* 2015; **63**(3): 151–158.
PubMed Abstract | Publisher Full Text

25. Duffy SC, Srinivasan S, Schilling MA, *et al.*: **Reconsidering *Mycobacterium bovis* as a proxy for zoonotic tuberculosis: a molecular epidemiological surveillance study.** *Lancet Microbe.* 2020; **1**(2): e66–e73.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Dawson KL, Bell A, Kawakami RP, *et al.*: **Transmission of *Mycobacterium orygis* (*M. tuberculosis* complex species) from a tuberculosis patient to a dairy cow in New Zealand.** *J Clin Microbiol.* 2012; **50**(9): 3136–3138.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Marcos LA, Spitzer ED, Mahapatra R, *et al.*: ***Mycobacterium orygis* Lymphadenitis in New York, USA.** *Emerg Infect Dis.* 2017; **23**(10): 1749–1751.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Lipworth SIW, Jajou R, de Neeling H, *et al.*: **A novel multi SNP based method for the identification of subspecies and associated lineages and sub-lineages of the Mycobacterium tuberculosis complex by whole genome sequencing.** *bioRxiv.* 2017.
Publisher Full Text

29. Lavender CJ, Globan M, Kelly H, *et al.*: **Epidemiology and control of tuberculosis in Victoria, a low-burden state in south-eastern Australia, 2005-2010.** *Int J Tuberc Lung Dis.* 2013; **17**(6): 752–758.
PubMed Abstract | Publisher Full Text

30. CRyPTIC Consortium and the 100,000 Genomes ProjectAllix-Béguec C, Arandjelovic I, *et al.*: **Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing.** *N Engl J Med.* 2018; **379**(15): 1403–1415.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Riojas MA, McGough KJ, Rider-Riojas CJ, *et al.*: **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 *Mycobacterium tuberculosis*.** *Int J Syst Evol Microbiol.* 2018; **68**(1): 324–332.
PubMed Abstract | Publisher Full Text

32. Cohan FM: **Towards a conceptual and operational union of bacterial systematics, ecology, and evolution.** *Philos Trans R Soc Lond B Biol Sci.* 2006; **361**(1475): 1985–1996.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Smith NH, Kremer K, Inwald J, *et al.*: **Ecotypes of the *Mycobacterium tuberculosis* complex.** *J Theor Biol.* 2006; **239**(2): 220–5.
PubMed Abstract | Publisher Full Text

34. Marianelli C, Amato B, Boniotti MB, *et al.*: **Genotype diversity and distribution of *Mycobacterium bovis* from livestock in a small, high-risk area in northeastern Sicily, Italy.** *PLoS Negl Trop Dis.* 2019; **13**(7): e0007546.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Ghielmetti G, Scherrer S, Friedel U, *et al.*: **Epidemiological tracing of bovine tuberculosis in Switzerland, multilocus variable number of tandem repeat analysis of *Mycobacterium bovis* and *Mycobacterium caprae*.** *PLoS One.* 2017; **12**(2): e0172474.
PubMed Abstract | Publisher Full Text | Free Full Text

36. Almaw G, Mekonnen GA, Mihret A, *et al.*: **Population structure and transmission of *Mycobacterium bovis* in Ethiopia.** *Microb Genom.* 2021; **7**(5): 000539.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Katale BZ, Mbelele PM, Lema NA, *et al.*: **Whole genome sequencing of *Mycobacterium tuberculosis* isolates and clinical outcomes of patients treated for multidrug-resistant tuberculosis in Tanzania.** *BMC Genomics.* 2020; **21**(1): 174.
PubMed Abstract | Publisher Full Text | Free Full Text

38. Rosenthal A, Gabrielian A, Engle E, *et al.*: **The TB Portals: an Open-Access, Web-Based Platform for Global Drug-Resistant-Tuberculosis Data Sharing and Analysis.** *J Clin Microbiol.* 2017; **55**(11): 3267–3282.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Beckert P, Sanchez-Padilla E, Merker M, *et al.*: **MDR *M. tuberculosis* outbreak clone in Eswatini missed by Xpert has elevated bedaquiline resistance dated to the pre-treatment era.** *Genome Med.* 2020; **12**(1): 104.
PubMed Abstract | Publisher Full Text | Free Full Text

40. Ciaravino G, Vidal E, Cortey M, *et al.*: **Phylogenetic relationships investigation of *Mycobacterium caprae* strains from sympatric wild boar and goats based on whole genome sequencing.** *Transbound Emerg Dis.* 2021; **68**(3): 1476–1486.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Belisle JT, Sonnenberg MG: **Isolation of genomic DNA from mycobacteria.** *Methods Mol Biol.* 1998; **101**: 31–44.
PubMed Abstract | Publisher Full Text

42. Menardo F, Loiseau C, Brites D, *et al.*: **Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity.** *BMC Bioinformatics.* 2018; **19**(1): 164.
PubMed Abstract | Publisher Full Text | Free Full Text

43. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics.* 2014; **30**(15): 2114–2120.
PubMed Abstract | Publisher Full Text | Free Full Text

44. Li H, Handsaker B, Wysoker A, *et al.*: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics.* 2009; **25**(16): 2078–2079.
PubMed Abstract | Publisher Full Text | Free Full Text

45. Comas I, Chakravartti J, Small PM, *et al.*: **Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved.** *Nat Genet.* 2010; **42**(6): 498–503.
PubMed Abstract | Publisher Full Text | Free Full Text

46. McKenna A, Hanna M, Banks E, *et al.*: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res.* 2010; **20**(9): 1297–1303.
PubMed Abstract | Publisher Full Text | Free Full Text

47. Li H: **A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.** *Bioinformatics.* 2011; **27**(21): 2987–2993.
PubMed Abstract | Publisher Full Text | Free Full Text

48. Koboldt DC, Zhang Q, Larson DE, *et al.*: **VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.** *Genome Res.* 2012; **22**(3): 568–576.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Garnier T, Eiglmeier K, Camus JC, *et al.*: **The complete genome sequence of *Mycobacterium bovis*.** *Proc Natl Acad Sci U S A.* 2003; **100**(13): 7877–7882.
PubMed Abstract | Publisher Full Text | Free Full Text

50. Cingolani P, Platts A, Wang le L, *et al.*: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: *SNPs in the genome of Drosophila melanogaster strain w^{1118}; iso-2; iso-3*.** *Fly (Austin).* 2012; **6**(2): 80–92.
PubMed Abstract | Publisher Full Text | Free Full Text

51. Stucki D, Brites D, Jeljeli L, *et al.*: ***Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages.** *Nat Genet.* 2016; **48**(12): 1535–1543.
PubMed Abstract | Publisher Full Text | Free Full Text

52. Steiner A, Stucki D, Coscolla M, *et al.*: **KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes.** *BMC Genomics.* 2014; **15**(1): 881.
PubMed Abstract | Publisher Full Text | Free Full Text

53. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.** *Bioinformatics.* 2014; **30**(9): 1312–1313.
PubMed Abstract | Publisher Full Text | Free Full Text

54. Yu G, Smith DK, Zhu H, *et al.*: **ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data.** *Methods Ecol Evol.* 2017; **8**(1): 28–36.
Publisher Full Text

55. Jombart T: ***adegenet*: a R package for the multivariate analysis of genetic markers.** *Bioinformatics.* 2008; **24**(11): 1403–1405.
PubMed Abstract | Publisher Full Text

56. Paradis E, Claude J, Strimmer K: **APE: Analyses of Phylogenetics and Evolution in R language.** *Bioinformatics.* 2004; **20**(2): 289–290.
PubMed Abstract | Publisher Full Text

57. South A: **rworldmap: A New R package for Mapping Global Data.** *The R Journal.* 2011; **3**(1): 35–43.
Publisher Full Text

58. Team RC: **R: A Language and Environment for Statistical Computing**. Vienna, Austria; 2018.
Reference Source

59. Danecek P, Auton A, Abecasis G, *et al.*: **The variant call format and VCFtools.** *Bioinformatics.* 2011; **27**(15): 2156–2158.
PubMed Abstract | Publisher Full Text | Free Full Text

60. Obenchain V, Lawrence M, Carey V, *et al.*: **VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants.** *Bioinformatics.* 2014; **30**(14): 2076–2078.
PubMed Abstract | Publisher Full Text | Free Full Text

61. gagneux-lab: **dbrites/LivestockAssociatedMTBC: (1.1.0).** *Zenodo.* 2021.
http://www.doi.org/10.5281/zenodo.5730644

62. Carneiro PA, Zimpel CK, Pasquatti TN, *et al.*: **Genetic Diversity and Potential Paths of Transmission of *Mycobacterium bovis* in the Amazon: The Discovery of M. bovis Lineage Lb1 Circulating in South America.** *Front Vet Sci.* 2021; **8**: 630989.
PubMed Abstract | Publisher Full Text | Free Full Text

63. Rodrigues RA, Ribeiro Araujo F, Rivera Davila AM, *et al.*: **Genomic and temporal analyses of *Mycobacterium bovis* in southern Brazil.** *Microb Genom.* 2021; **7**(5): 000569.
PubMed Abstract | Publisher Full Text | Free Full Text

64. Tazerart F, Saad J, Sahraoui N, *et al.*: **Whole Genome Sequence Analysis of *Mycobacterium bovis* Cattle Isolates, Algeria.** *Pathogens.* 2021; **10**(7): 802.
PubMed Abstract | Publisher Full Text | Free Full Text

65. Brosch R, Gordon SV, Marmiesse M, *et al.*: **A new evolutionary scenario for the Mycobacterium tuberculosis complex.** *Proc Natl Acad Sci U S A.* 2002; **99**(6): 3684–3689.
PubMed Abstract | Publisher Full Text | Free Full Text

66. Salvador LCM, O'Brien DJ, Cosgrove MK, *et al.*: **Disease management at the wildlife-livestock interface: Using whole-genome sequencing to study the role of elk in *Mycobacterium bovis* transmission in Michigan, USA.** *Mol Ecol.* 2019; **28**(9): 2192–2205.
PubMed Abstract | Publisher Full Text

67. Crispell J, Zadoks RN, Harris SR, *et al.*: **Using whole genome sequencing to investigate transmission in a multi-host system: bovine tuberculosis in New Zealand.** *BMC Genomics.* 2017; **18**(1): 180.
PubMed Abstract | Publisher Full Text | Free Full Text

68. Crispell J, Benton CH, Balaz D, *et al.*: **Combining genomics and epidemiology to analyse bi-directional transmission of *Mycobacterium bovis* in a multi-host system.** *Elife.* 2019; **8**: e45833.
PubMed Abstract | Publisher Full Text | Free Full Text

69. Oloya J, Kazwala R, Lund A, *et al.*: **Characterisation of mycobacteria isolated from slaughter cattle in pastoral regions of Uganda.** *BMC Microbiol.* 2007; **7**: 95.
PubMed Abstract | Publisher Full Text | Free Full Text

70. Clifford DL, Kazwala RR, Sadiki H, *et al.*: **Tuberculosis infection in wildlife from the Ruaha ecosystem Tanzania: implications for wildlife, domestic animals, and human health.** *Epidemiol Infect.* 2013; **141**(7): 1371–1381.
PubMed Abstract | Publisher Full Text

71. Fitzgerald SD, Kaneene JB: **Wildlife reservoirs of bovine tuberculosis worldwide: hosts, pathology, surveillance, and control.** *Vet Pathol.* 2013; **50**(3): 488–499.
PubMed Abstract | Publisher Full Text

72. Price-Carter M, Brauning R, de Lisle GW, *et al.*: **Whole Genome Sequencing for Determining the Source of *Mycobacterium bovis* Infections in Livestock Herds and Wildlife in New Zealand.** *Front Vet Sci.* 2018; **5**: 272.
PubMed Abstract | Publisher Full Text | Free Full Text

73. Zimpel CK, Brandao PE, de Souza Filho AF, *et al.*: **Complete Genome Sequencing of *Mycobacterium bovis* SP38 and Comparative Genomics of *Mycobacterium bovis* and *M. tuberculosis* Strains.** *Front Microbiol.* 2017; **8**: 2389.
PubMed Abstract | Publisher Full Text | Free Full Text

74. Rodriguez S, Bezos J, Romero B, *et al.*: **Mycobacterium caprae infection in livestock and wildlife, Spain.** *Emerg Infect Dis.* 2011; **17**(3): 532–535.
PubMed Abstract | Publisher Full Text | Free Full Text

75. Orlowska B, Krajewska-Wedzina M, Augustynowicz-Kopec E, *et al.*: **Epidemiological characterization of Mycobacterium caprae strains isolated from wildlife in the Bieszczady Mountains, on the border of Southeast Poland.** *BMC Vet Res.* 2020; **16**(1): 362.
PubMed Abstract | Publisher Full Text | Free Full Text

76. Broeckl S, Krebs S, Varadharajan A, *et al.*: **Investigation of intra-herd spread of Mycobacterium caprae in cattle by generation and use of a**

whole-genome sequence. *Vet Res Commun.* 2017; **41**(2): 113–128.
PubMed Abstract | Publisher Full Text

77. Prodinger WM, Indra A, Koksalan OK, *et al.*: ***Mycobacterium caprae* infection in humans.** *Expert Rev Anti Infect Ther.* 2014; **12**(12): 1501–1513.
PubMed Abstract | Publisher Full Text

78. Kubica T, Rusch-Gerdes S, Niemann S: ***Mycobacterium bovis* subsp. *caprae* caused one-third of human *M. bovis*-associated tuberculosis cases reported in Germany between 1999 and 2001.** *J Clin Microbiol.* 2003; **41**(7): 3070–3077.
PubMed Abstract | Publisher Full Text | Free Full Text

79. Domogalla J, Prodinger WM, Blum H, *et al.*: **Region of difference 4 in alpine *Mycobacterium caprae* isolates indicates three variants.** *J Clin Microbiol.* 2013; **51**(5): 1381–1388.
PubMed Abstract | Publisher Full Text | Free Full Text

80. Zignol M, Cabibbe AM, Dean AS, *et al.*: **Genetic sequencing for surveillance of drug resistance in tuberculosis in highly endemic countries: a multi-country population-based surveillance study.** *Lancet Infect Dis.* 2018; **18**(6): 675–683.
PubMed Abstract | Publisher Full Text | Free Full Text

81. Parsons SDC: ***Mycobacterium orygis*: a zoonosis, zooanthroponosis, or both?** *Lancet Microbe.* 2020; **1**(6): E241.
Publisher Full Text

82. Refaya AK, Bhargavi G, Mathew NC, *et al.*: **A review on bovine tuberculosis in India.** *Tuberculosis (Edinb).* 2020; **122**: 101923.
PubMed Abstract | Publisher Full Text

83. Muller B, Durr S, Alonso S, *et al.*: **Zoonotic *Mycobacterium bovis*-induced tuberculosis in humans.** *Emerg Infect Dis.* 2013; **19**(6): 899–908.
PubMed Abstract | Publisher Full Text | Free Full Text

84. Branger M, Loux V, Cochard T, *et al.*: **The complete genome sequence of *Mycobacterium bovis* Mb3601, a SB0120 spoligotype strain representative of a new clonal group.** *Infect Genet Evol.* 2020; **82**: 104309.
PubMed Abstract | Publisher Full Text

85. Oettinger T, Jørgensen M, Ladefoged A, *et al.*: **Development of the *Mycobacterium bovis* BCG vaccine: review of the historical and biochemical evidence for a genealogical tree.** *Tuber Lung Dis.* 1999; **79**(4): 243–250.
PubMed Abstract | Publisher Full Text

86. Çavuşoğlu C, Yilmaz FF: **Molecular Epidemiology of Human *Mycobacterium bovis* Infection in Aegean Region, Turkey.** *Mikrobiyol Bul.* 2017; **51**(2): 165–170.
PubMed Abstract | Publisher Full Text

87. Koni A, Tafaj S, Loda D, *et al.*: **Genotyping and spoligotyping of *Mycobacterium bovis* isolates from positive tuberculin skin test cattle in Albania.** *Albanian J Agric Sci.* 2018.
Reference Source

88. Berg S, Firdessa R, Habtamu M, *et al.*: **The burden of mycobacterial disease in ethiopian cattle: implications for public health.** *PLoS One.* 2009; **4**(4): e5068.
PubMed Abstract | Publisher Full Text | Free Full Text

89. Phelan JE, O'Sullivan DM, Machado D, *et al.*: **Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs.** *Genome Med.* 2019; **11**(1): 41.
PubMed Abstract | Publisher Full Text | Free Full Text

90. Brites D: **A new nomenclature for the livestock-associated Mycobacterium tuberculosis complex based on phylogenomics**. 2021.
http://www.doi.org/10.5281/zenodo.5153095

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 10 December 2021

https://doi.org/10.21956/openreseurope.15458.r28121

© **2021 Michelet L.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Lorraine Michelet**
Tuberculosis National Reference Laboratory, Laboratory for Animal Health, ANSES, Paris-Est University, Marne-la-Vallée, France

I acknowledge the authors for their responses. The animal lineages definition has been clarified and justified in the introduction.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: microbiology, molecular biology, phylogenetics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 08 December 2021

https://doi.org/10.21956/openreseurope.15458.r28122

© **2021 Salvador L.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Liliana C. M. Salvador**
Department of Infectious Diseases, College of Veterinary Medicine, University of Georgia, Athens, GA, USA

I appreciate that the authors have addressed all my concerns. I have no further comments to make. Thank you.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: phylogenetics, phylodynamics, bioinformatics, M. bovis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 13 October 2021

https://doi.org/10.21956/openreseurope.15117.r27485

© 2021 Salvador L. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Liliana C. M. Salvador
[1] Department of Infectious Diseases, College of Veterinary Medicine, University of Georgia, Athens, GA, USA
[2] Department of Infectious Diseases, College of Veterinary Medicine, University of Georgia, Athens, GA, USA

Summary
Zwyer *et al.* proposed a new nomenclature for three species of the Mycobacterium tuberculosis complex *M. bovis*, *M. orygis* and *M. caprae* by combining phylogenetics with genomics.The goals of their study was to define lineages and sub-lineages based on genetic diversity patterns by analysing a subset of 830 whole genomes extracted from (and representative of) different parts of the world). The authors have found three main lineages La1, La2, and La3, representing respectively, *M. bovis*, *M. caprae* and *M. orygis*. Within La1, the authors have identified eight sublineages (La1.1-La1.8), which presented distinct geographical patterns (some restricted to an area, while others globally widespread). The authors have also found specific markers (SNPs) to molecularly characterize the different MTBC groups.

General comments
This manuscript is highly relevant for the molecular characterization of MTBC and for future molecular epidemiological and evolutionary studies. There is a large need to extend the current classification system (L1-L7 involving *M. tuberculosis* and *M. africanum*) to other species of the complex and the authors did it using a dataset of 853 genomes from across the world. Overall, I find the manuscript well thought out, very relevant, and timely. I only have a few comments below regarding the analysis performed, which could impact the presented results.

General comments/questions
- Selection of isolates: authors mentioned that they chose 50 genomes per continent. What if the same amount of genomes was taken from an endemic place versus a non-endemic place. How does the local prevalence of the disease affect the genetic diversity that we see in the 50 isolates? Are the 50 genomes enough to characterize well the genetic diversity of a

specific place?

- What is MTBC-Livestock associated? How is it defined? Are there any other MTBC species that infect livestock that are not included? Are there any “MTBC-Livestock associated” lineages that infect the same or even more other types of animals like, for example, badgers or white-tailed deer (both reservoirs for *M. bovis*)?

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: phylogenetics, phylodynamics, bioinformatics, M. bovis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 18 Nov 2021

Daniela Brites

We thank the reviewer for her positive feedback.

As for the selection of isolates; the aim of this analysis was to have balanced phylogenetic inference of the total diversity of the MTBC. However, we were not so interested in maximizing genetic diversity within each lineage but rather to have an even representation of endemic and non-endemic places (which we believe to have achieved by sampling randomly each lineage in each continent) at a global scale and show that there is no reason to treat *M. bovis* differently than the human adapted lineages; at least not in what concerns the tree topology. Certainly, as very well pointed out by the reviewer, 50 genomes might not

be enough to represent all diversity of *M. bovis* in Africa for instance, but that should be true also for other groups of the MTBC. As to the definition of livestock associated, a similar question was also raised by the other reviewer, revealing that we failed to pass on more clearly why we consider *M. bovis*, *M. orygis* and *M. caprae* as livestock-associated MTBC members. We have added to the Introduction a few considerations defending this point of view, which we hope addresses the reviewer's concerns.

At the core of considering these species as livestock-associated is the strong evidence that the evolutionary success of *M. bovis*, *M. orygis* and *M. caprae* is associated with the fact that these pathogens are able to cause sustainable infections in livestock species. The fact that these are multi-host pathogens allows them to linger in other animal reservoirs, and in more recent times (post bTB surveillance), this might play a crucial role for these lineages by avoiding local extinctions. However at a global and evolutionary scale, migration and population expansion of *M. bovis* (possibly also *M. caprae* and *M. orygis*, to be confirmed as more data becomes available) are most likely a consequence of cattle husbandry and movements. We think that distinguishes *M. bovis*, *M. orygis* and *M. caprae* from other animal MTBC members, which are also multi-host pathogens (e.g. *M. microti*).

Competing Interests: No competing interests were disclosed.

Reviewer Report 16 September 2021

https://doi.org/10.21956/openreseurope.15117.r27482

© 2021 Michelet L. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Lorraine Michelet
[1] Tuberculosis National Reference Laboratory, Laboratory for Animal Health, ANSES, Paris-Est University, Marne-la-Vallée, France
[2] Tuberculosis National Reference Laboratory, Laboratory for Animal Health, ANSES, Paris-Est University, Marne-la-Vallée, France

This publication propose a new nomenclature for livestock-associated *Mycobacterium tuberculosis* complex members, like what exist for human-adapted MTBC. The authors proposed to call the lineages corresponding to *M. bovis*, *M. caprae* and *M. orygis*, respectively La1 to La3 and subdivided La1 (M. bovis) into 8 sublineages. This work was necessary in order to facilitate phylogenetic analyzes and to be able to compare studies from different countries.

However, I feel uncomfortable with the use of the term livestock-associated. Indeed, *M. bovis* and *M. caprae* are the main agents responsible for bovine (and caprine) tuberculosis but *M. orygis* is not. If we refer to the OIE Terrestrial Animal Health Code, bovine tuberculosis is caused particularly by *M. bovis* but also by *M. caprae* and, to a lesser extent, by *M. tuberculosis*. I do not understand why the authors include *M. orygis* in the livestock-associated MTBC members. In this

case, why do the authors not include *M. microti*, which is known to infect cattle (Jahans et al, 2004) and has recently been identified in several times in goat and cattle.

I can understand that more WGS data are available for *M. orygis* and that allow to better defining this lineage. However, this MTBC member cannot be included as a livestock-associated. Even if few WGS data are available for *M. microti* and *M. pinnipedii*, I would have expected that the authors include these species as animal-adapted and define two other lineages for these species. The authors could provide some specific markers based on the few available genomes.

In the introduction, I consider that the authors could have more described the previous works done on *M. bovis*, especially on the definition of four clonal complexes. These studies are the result of wide collaborations and have provided a much better understanding of the population structure of *M. bovis* before WGS era. Some other recent works have tried to defined lineages or clusters which are confirmed by this work but the parallel between them is not done. Some references are missing in this part of the introduction (Hauer et al, definition of 9 clusters).

One aspect of the evolution of *M. bovis* is not discussed at all, it is the time. The year of isolation of each sample is an important data that is missing here. The authors suggest further studies on unknown 4, 5, 6 and 8 group to better understand their population structure. However, maybe this could not be done if these lineages have disappeared. For example, all unknown6 samples have been isolated in the 90's expect one in 2008. It is well known that selection pressure due to the surveillance and eradication measures put in place in each country has led to a reduction in the genetic diversity of *M. bovis*. This aspect of *M. bovis* evolution could be discussed.

Minor comments

- Page 4 $ Population structure and genetic distance: add a space between "dataset," and "using the R package".
- Page 7 $ La2, or M. caprae: the authors refer twice to Figure 3 but this should be the Figure 4 (Figure 3 refer to La3 M. orygis).
- Page 12 $ The remaining genomes: there is a mistake in Spoligotypes.
- Extended table 1: some samples are in duplicate. For example, SRR7851305 rely to G47578 and G49365. Please check carefully your data.

References

1. Jahans K, Palmer S, Inwald J, Brown J, et al.: Isolation of Mycobacterium microti from a male Charolais-Hereford cross.*Vet Rec*. 2004; **155** (12): 373-4 PubMed Abstract
2. Hauer A, Michelet L, Cochard T, Branger M, et al.: Accurate Phylogenetic Relationships Among Mycobacterium bovis Strains Circulating in France Based on Whole Genome Sequencing and Single Nucleotide Polymorphism Analysis.*Front Microbiol*. 2019; **10**: 955 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: microbiology, molecular biology, phylogenetics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 18 Nov 2021

Daniela Brites

(Reviewer comments in italics)

This publication propose a new nomenclature for livestock-associated Mycobacterium tuberculosis complex members, like what exist for human-adapted MTBC. The authors proposed to call the lineages corresponding to M. bovis, M. caprae and M. orygis, respectively La1 to La3 and subdivided La1 (M. bovis) into 8 sublineages. This work was necessary in order to facilitate phylogenetic analyzes and to be able to compare studies from different countries.

However, I feel uncomfortable with the use of the term livestock-associated. Indeed, M. bovis and M. caprae are the main agents responsible for bovine (and caprine) tuberculosis but M. orygis is not. If we refer to the OIE Terrestrial Animal Health Code, bovine tuberculosis is caused particularly by M. bovis but also by M. caprae and, to a lesser extent, by M. tuberculosis. I do not understand why the authors include M. orygis in the livestock-associated MTBC members. In this case, why do the authors not include M. microti, which is known to infect cattle (Jahans et al, 2004) and has recently been identified in several times in goat and cattle.

Author Response: We understand the reservations of the reviewer with respect to the livestock-associated terminology, as not much is known about the host range of *M. orygis*. Solid molecular evidence for *M. orygis* infections only started to emerge with whole-genome sequencing relatively recently, and we are in a time where new research about *M. orygis* is

defying what we know about bovine TB and zoonotic TB in countries such as India, which has the biggest livestock population in the world and where cattle and humans co-exist very closely. Several recent studies suggest that *M. orygis* is a main cause of zoonotic TB in South Asian countries. That is particularly well demonstrated by the study of Duffy *et al* in India and by a series of other studies reporting infection caused by *M. orygis* in patients of South Asian origin living in low burden TB countries (cited in our study). There is also direct evidence of *M. orygis* infecting livestock species, and very importantly, transmitting within cattle herds. This is not the case for *M. microti* or *M. tuberculosis*, which can cause TB in livestock species, but are not able to maintain cycles of infection. Transmission of *M. orygis* within cattle is also supported by our genomic analysis. We think that these are enough evidence to consider that livestock species might be one of the reasons for the evolutionary success of *M. orygis* like for *M. bovis* but unlike *M. microti* and *M. tuberculosis*. The reviewer is right when referring to the OIE Terrestrial Animal Health Code and bovine TB. On the human TB side, the WHO has also considered until the latest report in 2020 that zoonotic TB was synonymous of infections by *M. bovis*. We think that reflects the fact that the molecular characterization of the agents of bovine TB in Asia is lagging behind. We have added now more explicitly these considerations to the introduction and thank the reviewer for pointing out that this rational was not clear in the first version of our manuscript.

I can understand that more WGS data are available for M. orygis and that allow to better defining this lineage. However, this MTBC member cannot be included as a livestock-associated. Even if few WGS data are available for M. microti and M. pinnipedii, I would have expected that the authors include these species as animal-adapted and define two other lineages for these species. The authors could provide some specific markers based on the few available genomes. Response: We believe that our previous points now demonstrate more clearly why we considered M. orygis as livestock associated. Our interest was to contribute to the molecular characterization of MTBC species that cause a high burden of disease, because they affect some of the most dense animal populations of the world, humans and livestock. We consider this necessary in the light of the growing numbers of WGS available. That is not the case for M. microti and M. pinnipedii. In addition, in this manuscript, we present a rational for defining lineages based in an interval of SNPs, which is consistent with work that has been previously done for the human-adapted lineages. As shown by us in a previous publication (Brites et al, 2018, Figure 5) M. microti and M. pinnipedii are separated by only 400-500 SNPs, and thus could not be classified as different lineages if we take the same thresholds as for the rest of MTBC complex. Perhaps their split as independent phylogenetic lineages has occurred more recently, or they might have different substitution rates, etc. In any case, we consider that having a better sampling of M. microti and M. pinnipedii is necessary to be able to draw more informed scenarios.

In the introduction, I consider that the authors could have more described the previous works done on M. bovis, especially on the definition of four clonal complexes. These studies are the result of wide collaborations and have provided a much better understanding of the population structure of M. bovis before WGS era. Some other recent works have tried to defined lineages or clusters which are confirmed by this work but the parallel between them is not done. Some references are missing in this part of the introduction (Hauer et al, definition of 9 clusters).

Author Response: Thank you for this comment. We have added these considerations into the introduction. As to the recent works that have tried to define lineages or clusters based

on WGS, we think the reviewer refers to Zimpel *et al* 2020 and Hauer *et al* 2019. We have added to Table 1 the correspondence between the groups defined in these two studies and the sub-lineages we are proposing here. We explicitly mention this in the text. "The correspondence between the nomenclature we propose here and previously defined groups based on WGS is given in *Extended data,* Table 1". The study of Hauer *et al* is now also mentioned in the Introduction and not only in the Results and Discussion section.

One aspect of the evolution of M. bovis is not discussed at all, it is the time. The year of isolation of each sample is an important data that is missing here. The authors suggest further studies on unknown 4, 5, 6 and 8 group to better understand their population structure. However, maybe this could not be done if these lineages have disappeared. For example, all unknown6 samples have been isolated in the 90's expect one in 2008. It is well known that selection pressure due to the surveillance and eradication measures put in place in each country has led to a reduction in the genetic diversity of M. bovis. This aspect of M. bovis evolution could be discuss. Response: We agree with the reviewer and have added these points to the end of section Sublineages La1.4 to La1.8.

Minor comments

- *Page 4 $ Population structure and genetic distance: add a space between "dataset," and "using the R package".*
- *Page 7 $ La2, or M. caprae: the authors refer twice to Figure 3 but this should be the Figure 4 (Figure 3 refer to La3 M. orygis).*
- *Page 12 $ The remaining genomes: there is a mistake in Spoligotypes.*
- *Extended table 1: some samples are in duplicate. For example, SRR7851305 rely to G47578 and G49365. Please check carefully*

Author Response: We thank the reviewer for the careful revision. Indeed, we inadvertently downloaded from public databases a few genomes twice. We have removed them from our alignments and redone the analysis and figure 4 and corrected tables 1 and table 2.

Competing Interests: No competing interests were disclosed.