TITLE: An African origin for *Mycobacterium bovis*

Chloé Loiseau[1,2], Fabrizio Menardo[12], Abraham Aseffa[3], Elena Hailu[3], Balako Gumi[4], Gobena Ameni[5], Stefan Berg[6], Leen Rigouts[7,8,9], Suelee Robbe-Austerman[10], Jakob Zinsstag[1,2], Sebastien Gagneux[1,2*] and Daniela Brites[1,2*]

[1]Swiss Tropical and Public Health Institute, Basel, Switzerland
[2]University of Basel, Basel, Switzerland
[3]Armauer Hansen Research Centre, Addis Ababa, Ethiopia
[4]Bule Hora University, Department of Animal Science and Range Management, Bule Hora Town, Ethiopia
[5]Addis Ababa University, Aklilu Lemma Institute of Pathobiology, Addis Ababa, Ethiopia
[6]Animal & Plant Health Agency (APHA), Bacteriology Department, Weybridge, Surrey, United Kingdom
[7]Mycobacteriology Unit, Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium
[8]Collection of Mycobacterial Cultures (BCCM/ITM), Institute of Tropical Medicine, Antwerp, Belgium
[9]Department of Biomedical Sciences, Antwerp University, Antwerp, Belgium
[10]National Veterinary Services Laboratories, United States Department of Agriculture, Ames, Iowa, USA

Corresponding authors:
Daniela Brites
Email: d.brites@swisstph.ch
Tel: +41612848185

Sebastien Gagneux
Tel : +41 61 284 8369
Email: sebastien.gagneux@swisstph.ch

Email addresses all other authors:

Abraham Aseffa	aseffaa@gmail.com
Stefan Berg	Stefan.Berg@apha.gov.uk
Leen Rigouts	LRigouts@itg.be
Suelee Robbe-Austerman	Suelee.Robbe-Austerman@aphis.usda.gov
Chloé Loiseau	chloemarie.loiseau@swisstph.ch
Fabrizio Menardo	fabrizio.menardo@swisstph.ch
Jacob Zinsstag	jakob.zinsstag@swisstph.ch
Gobena Ameni	gobena.ameni@aau.edu.et
Balako Gumi	balako.gumi@yahoo.com
Elena Hailu	elenahailu@yahoo.com

Heading Title: Phylogeography of Mycobacterium bovis

Lay Summary: Mycobacterium bovis and M. caprae are both the most important agents of tuberculosis in livestock and zoonotic tuberculosis in humans. Using phylogenetic and molecular clock inferences based on an extensive collection of whole-genome sequences we provide new insights into the global population structure, phylogeography and evolutionary history of these pathogens.

Word count abstract: 180

Word count main text: 4366

Number of Figures: 3

ABSTRACT

Background and objectives

Mycobacterium bovis and *Mycobacterium caprae* are the two most important agents of tuberculosis (TB) in livestock and the most important causes of zoonotic TB in humans. However, little is known about the global population structure, phylogeography and evolutionary history of these pathogens.

Methodology

We compiled a global collection of 3364 whole-genome sequences from *M. bovis* and *M. caprae* originating from 35 countries and inferred their phylogenetic relationships, geographic origins and age.

Results

Our results resolve the phylogenetic relationship among the four previously defined clonal complexes of *M. bovis*, and another eight newly described here. Our inferences indicate that *M. bovis* emerged in East Africa likely between the 4th and 10th century. While some *M. bovis* groups remained restricted to East- and West Africa, others have subsequently dispersed to different parts of the world.

Conclusions and implications

Our results allow a better understanding of the global population structure of *M. bovis* and its evolutionary history. This knowledge can be used to define better molecular markers for epidemiological investigations in settings where whole genome sequencing cannot easily be implemented.

BACKGROUND AND OBJECTIVES

Tuberculosis (TB) remains an important burden for global health and the economy [1]. TB is the number one cause of human death due to infection globally, with an estimated 10.0 million new cases and 1.6 million deaths occurring every year [1]. TB is caused by members of the *Mycobacterium tuberculosis* complex (MTBC), which includes seven human-adapted lineages, and several animal-adapted ecotypes including *M. bovis* and *M. caprae.* Animal TB complicates the control of human TB due to the zoonotic transfer of TB bacilli from infected animals to exposed human populations through e.g. the consumption of unpasteurized milk or handling of contaminated meat [2]. *M. bovis* and *M. caprae* are the most important agents of TB in livestock and the most important agents of zoonotic TB in humans, causing at least 147 000 new human cases and 12 500 deaths yearly [1, 3]. Zoonotic TB caused by *M. bovis* also poses a challenge for patient treatment, due to its natural resistance to pyrazinamide (PZA), one of the four first-line drugs used in the treatment of TB. In addition, TB in livestock accounts for an estimated loss of three billion US dollars per year [4]. In Africa its prevalence is highest in peri-urban dairy belts of larger cities and remains at a low levels of endemic stable transmission in rural areas [5], threatening also wildlife populations [6].

During the last few years, analyses of large globally representative collections of whole genome sequences (WGS) from the human-adapted MTBC lineages have enhanced our understanding of the global population structure, phylogeography and evolutionary history of these pathogens [7]. By contrast, little corresponding data exist for the various animal-adapted ecotypes of the MTBC such as *M. bovis*.

Current knowledge about global *M. bovis* populations stems mostly from spoligotyping [8, 9]. This method has been highly valuable for showing that *M. bovis* populations vary by geography, and defining strain families based on the presence or absence of spacers in the Direct Repeat region of the MTBC genome [8]. However, the discriminatory capacity of spoligotyping is limited since diversity is measured at a single locus prone to convergent evolution and phylogenetic distances cannot be reliably inferred [10].

In addition to spoligotyping, other markers such as genomic deletions [11-14] and single nucleotide polymorphisms (SNPs) [14], have given insights into the biogeography of *M. bovis*. These markers were used to define four major groups of genotypes within *M. bovis*, known as clonal complexes European 1 and 2 (Eu1, Eu2) and African 1 and 2 (Af1 and Af2) [11-14]. Bovine TB in West Africa and East Africa is caused to a large extent by clonal complexes Af1 and Af2, respectively [11, 12]. Bovine TB in Europe and in the Americas is caused by clonal complex Eu1, which affects mostly the British Islands and former trading

105 countries of the UK [13] while Eu2 is prevalent mostly in the Iberian Peninsula and Brazil
106 [14].
107 More recently, studies based on WGS have brought deeper insights into the population
108 dynamics of *M. bovis* and showed that unlike *M. tuberculosis*, wild animals act as *M. bovis*
109 reservoirs in different regions of the world [15-18] . However, most studies using WGS have
110 aimed at investigating local epidemics, and little is known about the global population
111 structure and evolutionary history of *M. bovis*. Recently, we suggested a scenario for the
112 evolution of the animal-adapted MTBC, in which we propose that *M. caprae* and *M. bovis*
113 might have come out of Africa together with humans [19]. Here we gathered 3356 *M. bovis*
114 and *M. caprae* WGS from the public domain, to which we added 8 *M. bovis* sequences from
115 strains isolated in East Africa. Our results provide a phylogenetic basis to better understand
116 the global population structure of *M. bovis*. Moreover, they point to East Africa as the most
117 likely origin of contemporary *M. bovis*.
118
119

METHODS

Data collection

A total of 3929 *M. bovis* genomes were retrieved from EBI: 3834 BioSamples were registered on EBI with the taxon id 1775 (corresponding to "*Mycobacterium tuberculosis* variant bovis") and downloaded on the 11th of March 2019 and 95 *M. bovis* genomes were registered under taxon id 1765 (corresponding to "*Mycobacterium tuberculosis")*.

Of these, 457 were excluded because they were part of pre-publications releases from the Sanger Institute, 130 were excluded because they were registered as BCG – Bacille Calmette Guérin, the vaccine strain derived from *M. bovis*, one genome was excluded because it was wrongly classified as *M. bovis*, and three samples were excluded because they corresponded to RNA-seq libraries.

In addition, we added 81 publically available *M. caprae* genomes and eight previously unpublished sequences from *M. bovis* isolated in Ethiopia (n=7) and Burundi (n=1). The sequencing data has been deposited in the European Nucleotide Archive (EMBL-EBI) under the study ID PRJEB33773.

From this total of 3427 genomes, 63 sequences were excluded because they did not meet our criteria for downstream analyses (average whole-genome coverage below 7, ratio of variable SNP to fixed SNP above 1), yielding a final dataset of 3364 genomes (Fig. S1, Table S1). Geographical origin of the isolates, date of isolation and host metadata was recovered from EBI (Table S1).

Whole genome sequence analysis

All samples were subject to the same whole-genome sequencing analysis pipeline, as described in [20]. In brief, reads were trimmed with Trimmomatic v0.33 [21]. Only reads larger than 20 bp were kept for the downstream analysis. The software SeqPrep (https://github.com/jstjohn/SeqPrep) was used to identify and merge any overlapping paired-end reads. The resulting reads were aligned to the reconstructed ancestral sequence of the MTBC [22] using the MEM algorithm of BWA v0.7.13 [23]. Duplicated reads were marked using the MarkDuplicates module of Picard v2.9.1 (https://github.com/broadinstitute/picard). The RealignerTargetCreator and IndelRealigner modules of GATK v 3.4.0 were used to perform local realignment of reads around InDels [24]. Finally, SNPs were called with Samtools v1.2 mpileup [25] and VarScan v2.4.1 [26] using the following thresholds: minimum mapping quality of 20, minimum base quality at a position of 20, minimum read

depth at a position of 7X, maximum strand bias for a position 90%. SNPs were annotated using snpEff v4.1144 [27], using the *M. tuberculosis* H37Rv reference annotation (NC_000962.3) as the genome of *M. bovis* (AF2122/97) has no genes absent from H37Rv except for TbD1 (contains *mmpS6* and the 5′ region of *mmpL6*) [28].

In silico spoligotyping and previously defined clonal complexes

The spoligotype pattern of the 3364 genomes was determined *in silico* using Kvarq [29]. The results were submitted to the *Mycobacterium bovis* spoligotype database https://www.mbovis.org/ [30] and SB numbers obtained.

All 3364 genomes were screened *in silico* for the presence of molecular markers defining the previously described *M. bovis* clonal complexes; i.e. for the presence or absence of the deletions RDAf1, RDAf2, RDEu1 [11-13] and in the case of Eu2 [14] for the presence of SNP 3813236 G to A with respect to the H37Rv (NC_000962.3).

Phylogenetic analyses

All phylogenetic trees were inferred with RAxML (v.8.2.12) using alignments containing only polymorphic sites and excluding the variable positions in drug resistance-related genes. Maximum likelihood phylogenies were computed using the general time-reversible model of sequence evolution (-m GTRCAT -V options), 1,000 rapid bootstrap inferences, followed by a thorough maximum-likelihood search performed through CIPRES [31]. All phylogenies were rooted using a *M. africanum* Lineage 6 genome from Ghana (SAMEA3359865).

Obtaining a representative dataset of _M. bovis_ genomes - Subsampling 1

Our phylogenetic reconstruction indicated that sequences belonging to clonal complex Eu1 and Eu2 were over-represented in the initial 3,364 genomes dataset, particularly from the USA, Mexico, New Zealand and the UK. To obtain a smaller dataset with a more even representation of the different phylogenetic groups, we pruned the 3364 genomes using the following criteria: 1) we removed all genomes with non-available country metadata (n=739), which resulted in 2625 genomes; 2) we used Treemmer v0.2 [20] with the option *-RTL 99* to keep 99% of the original tree length and the option *–lm* to include a list of taxa to protect from pruning. This list included all genomes belonging to clonal complexes Af1 and Af2, as well

as any genome belonging to any unclassified clade; 3) we visually identified monophyletic clades in which all taxa come from the same country and used Treemmer v0.2 [20], using options *-lmc* and *–lm*, to only keep a few representatives of each of these clades. To have representatives of the BCG clade, we kept 11 BCG genomes from [32]. This selection process rendered a dataset of 476 genomes.

Ancestral reconstruction of geographic ranges - Subsampling 2

To infer the geographic origin of the ancestors of the main groups of *M. bovis* and *M. caprae*, we used the 476 genomes dataset (see subsampling 1) and excluded all BCG genomes and all *M. bovis* from human TB cases or from unknown hosts, which could be humans, if the strains were isolated in a low incidence TB country (Europe, North America, Oceania). This is justified by the fact that the majority of such cases correspond to immigrants from high incidence countries that were infected in their country of origin, i.e. country of isolation does not correspond to the native geographic range of the strain and is thus not informative for the geographic reconstruction. *M. bovis* from patients in high incidence countries were kept (Table S1). The resulting dataset was composed of 392 genomes.

For the ancestral reconstruction of geographic ranges, we used the geographic origin of the strains and the phylogenetic relationships of the 392 genomes. Geographic origin was treated as a discrete character to which 13 states, corresponding to UN-defined regions, were assigned. To select the best model of character evolution the function fitMk from the package phytools 0.6.60 in R 3.5.0 [33] was used to obtain the likelihoods of the models ER, SYM and ARD [34]. A Likelihood Ratio Test (LRT) was used to compare the different log-Likelihoods obtained. According to the former, the best fitting model was SYM, a model that allows states to transition at different rates in a non-reversible way, i.e. reverse and forward transitions share the same parameters (Table S2). The function *make.simmap* in phytools package 0.6.60 in R 3.5.0 [33, 35] was used to apply stochastic character mapping as implemented in SIMMAP [36] on the 392 phylogeny inferred from the best-scoring ML tree rooted on L6, using the SYM model with 100 replicates. We summarized the results of the 100 replicates using the function *summary* in phytools package 0.6.60 in R [33].

Molecular Dating of *M. bovis* and *M. caprae* – Subsampling 3

For the molecular clock analyses we considered only genomes for which the date of isolation was known (n=2,058). We used a pipeline similar to that reported in [37]. We built SNPs alignments including variable positions with less than 10% of missing data. We added a Lineage 6 strain as outgroup (SAMEA3359865) and inferred the Maximum Likelihood tree as described above. Since the alignment contained only variable positions, we rescaled the branch lengths of the trees: rescaled_branch_length = ((branch_length * alignment_length) / (alignment_length + invariant_sites)). To evaluate the strength of the temporal signal we performed root-to-tip regression using the R package ape [38]. Additionally, we used the least square method implemented in LSD v0.3-beta [39] to estimate the molecular clock rate in the observed data and performed a date randomization test with 100 randomized datasets. To do this, we used the QPD algorithm and calculated the confidence interval (options -f 100 and -s).

We also estimated the molecular clock rates using a Bayesian analysis. For this, we reduced the dataset to 300 strains with Treemmer v0.2 in the following way: we randomly subsampled strains, maintaining the outgroup and at least one representative of four small basal clades of the tree that would have disappeared with simple random subsampling strategy (Af2 clonal complex: G42133; Af1 clonal complex: G02538; *PZA_ sus_unknown1*: G04143, G04145, G04147; *M. caprae*: G42152, G42153, G37371, G37372, G41838; Table S1).

We used jModelTest 2.1.10 v20160303 [40] to identify the best fitting nucleotide substitution model among 11 possible schemes, including unequal nucleotide frequencies (total models = 22, options -s 11 and -f). We performed Bayesian inference with BEAST2 [41]. We corrected the xml file to specify the number of invariant sites as indicated here: https://groups.google.com/forum/#!topic/beast-users/QfBHMOqImFE, and used the tip sampling year as calibration.

We used a relaxed lognormal clock model [42], the best fitting nucleotide substitution model according to the results of jModelTest (all criteria selected TVM as best model), and two different coalescent priors: constant population size and exponential population growth. We chose a 1/x prior for the population size [0- 10^9], a 1/x prior for the mean of the lognormal distribution of the clock rate [10^{-10} – 10^{-5}], and a normal(0,1) prior for the standard deviation of the lognormal distribution of the clock rate [0 – infinity]. For the exponential growth rate prior, we used the standard Laplace distribution [-infinity – infinity]. For both analyses, we performed two runs andused Tracer 1.7.1 [43] to identify and exclude the burn-in, to evaluate convergence among runs and to calculate the estimated sample size (ESS). We stopped the

255 runs when they reached convergence, and the ESS of the posterior and of all parameters were
256 larger than 200.

257

RESULTS AND DISCUSSION

Phylogenetic inference of *M. bovis* and *M. caprae* populations

The phylogenetic reconstruction of all *M. bovis* and *M. caprae* sequences obtained (n=3364) confirmed that these two ecotypes correspond to two monophyletic groups, despite infecting similar hosts [44, 45] (Fig. S1). The range of host species from which *M. bovis* was isolated is broad, confirming that *M. bovis* can cause infection in many different mammalian species (Table S1). Our compilation of *M. caprae* included genomes from Japan (isolated in Elephants from Borneo) [46], China (isolated in Primates, Table S1) and Peru (host information unavailable but possibly human [47]), suggesting that the host and geographic distribution of this species ranges well beyond Southern- and Central Europe [48, 49].

For all *M. bovis*, we determined *in silico* clonal complexes Eu1, Eu2, Af1 and Af2 and spoligotypes, and mapped them on the phylogenetic tree and onto a world map (Fig. 1, Fig. S2, Table S1). All previously described clonal complexes corresponded to monophyletic groups in our genome-based phylogeny (Fig. S2). The phylogenetic tree also revealed *M. bovis* representatives that did not fall into any of the previously described clonal complexes (n=175, 5.3%, Fig. 1, Fig. S2, Table S1). These belonged to 8 monophyletic clades with unknown classification and to a few singleton branches (Fig. S2). The tree topology showed a strong bias towards closely related strains, in particular among Eu1, which reflects the different sampling and WGS efforts in the different geographic regions (Fig. 1, Fig. S2). Closely related genomes inform the local epidemiology but not the middle/long term evolutionary history of the strains and were thus excluded from further analysis (Subsampling 1, see Methods).

Two deep divergence events in *M. bovis* populations were notorious: one giving rise to an unclassified lineage we named *M. bovis PZA_sus_unknown1* (RD4 deleted as other *M.bovis*), which included 5 samples from Uganda (isolated from *Bos taurus* cattle, Table S1), 3 from Malawi (isolated from Humans, Table S1) and one isolated from an antelope in Germany (Table S1). These *M. bovis* isolates lacked the *PncA* H57D mutation that is responsible for the intrinsic pyrazinamide resistance of canonical *M. bovis* as reported previously [50] (Fig. 2). The second deep branching lineage included all other *M. bovis* strains descendent from an ancestor that acquired the *PncA* H57D mutation and therefore encompasses all previously described clonal complexes [8, 14] as well as the other previously unknown clades we describe here.

From the *M. bovis* PZA resistant ancestor strains, two main splits occurred; one split led to the ancestor of Af2 and its previously unclassified sister clade which contains the BCG vaccine which we call here *unknown2* (Fig. 2). *M. bovis* strains with spoligotyping patterns similar to BCG have previously been referred to as "BCG-like". However, our genome-based phylogeny shows that BCG-like spoligotyping patterns have little discriminatory power both due to convergent evolution leading to homoplasy in the Direct Repeat locus of the MTBC [10] or due to common ancestry (Table S3). The other split led to the ancestor, from which all remaining *M. bovis* strains evolved, i.e. Af1, Eu2 and Eur1 as well as other groups (Fig. 2). Interestingly, Af1 does not share a MRCA with Af2 but with Eu1 and Eu2 as well as with another unclassified group, which we called *unknown3* (Fig. 2). Clonal complexes Eu1 and Eu2, despite being more closely related to each other than to Af1, are not monophyletic. Eu2 is more closely related to other clades, which we have called here *unknown4 and 5*, than to Eu1 (Fig. 2). Eu1 in turn shares a common ancestor with three other clades *unknown6, 7 and 8* (Fig. 2).

The temporal and geographic origin of *M. bovis*

Our reconstruction of ancestral geographical ranges points to East Africa as the most likely origin for the ancestor of all *M. bovis* (Fig. 2, Fig. S2). This result is probably driven by the fact that the most basal group of contemporary *M. bovis* (*PZA_sus_ unknown1*) has an exclusively East African distribution. For *M. caprae*, the sampling was too small and biased (Table S1) and no conclusions can be confidently drawn. We performed tip-dating calibration using the isolation dates of the strains with both Bayesian methods and LSD (see methods). Both the tip-to-root regression and the randomization tests performed indicated a temporal signal in the data (Fig. S4). The common ancestor of *M. bovis* was estimated to have evolved in the year 710 AD (95% HPD: 398-985) by Bayesian analysis and in the year 388 AD by LSD (Fig. 3, TreeS1.txt, TreeS2.txt). Together, these estimates suggest that *M. bovis* has emerged in East Africa sometime during the period spanning the 4th to the 10th century AD (Fig. 3). There are archaeological findings of approximately 2,000 years old *M. bovis* in South Siberia that are difficult to reconcile with these results [51]. Our findings are also at odds with the archaeological finding of RD9/RD4-deleted MTBC from human bones in Central Germany from the period between 5400 and 4800 BC [52]. The tip-dating calibration provided accurate results for the emergency of BCG strains (TreeS2.txt), indicating that the

method can reliably infer divergence times at least for events occurring in the last 100 years. However, extrapolating the clock rate to date older divergence events should be done with caution [37].

Insights into the detailed population structure of *M. bovis* around the world

Understanding the evolutionary history of the *M. bovis* populations requires understanding their geographic distribution at a continental scale. Our WGS data set has limited geographical resolution due to both oversampling and undersampling of certain regions of the world, partial unavailability of associated metadata and an origin in foreign-born TB patients from Western countries. To get more insights into the geographical ranges of the different *M. bovis* clades, we used the spoligotype patterns inferred from the WGS data and searched for references describing the prevalence of those in different regions of the world (Table S3). Patterns SB0120 and SB0134, known as "BCG-like" and reported to be relatively prevalent [9], as well as SB0944, are phylogenetically uninformative; SB0120 is present in several clades probably by common ancestry given that it contains the spacers deleted in most other spoligotypes and SB0134 and SB0944 have evolved independently in two different *M. bovis* populations (Fig. 2, Table S3).

Our results suggest that the most basal group of all contemporary *M. bovis*, *PZA_sus_unknown1,* is restricted to East Africa. The same holds true for Af2, which is in accordance to previous reports [8, 12]. Our findings further suggest that the geographical distribution of the Af2 sister clade *unknown2* includes East Africa (Eritrea, Ethiopia), but also Southern Europe (Spain and France). Informative spoligotypes of the *unknown2* clade suggest that it also circulates in North-Africa (Fig. 2, Table S3). Of note, the original strain, from which all BCG vaccine strains were derived, was isolated in France [53]. Our inferences suggest that a common ancestor of Af2 and *unknown2* evolved in East Africa, and while Af2 remained geographically restricted, its sister clade *unknown2* has subsequently dispersed (Fig. 2).

All remaining *M. bovis* descended from a common ancestor, for which the geographical origin was impossible to infer reliably with our data. However, the tree topology showed that from this ancestor several clades evolved that are worldwide important causes of bovine TB today (i.e., the clonal complexes Eu1, Eu2, Af1 and Af2; Fig.2). The ancestors of these clades have most likely evolved during a period of around 500 years between the 11th and 16th centuries AD (Fig. 3).

The most basal clade within this group is *unknown3*, which contained 25 genomes mostly from human studies (Table S2). The *in silico* derived spoligotypes suggest that the geographical spread of *unknown3* ranges from Western Asia to Eastern Europe, but also includes East Africa (Fig. 2, Table S3). The next split in our phylogeny corresponds to Af1, which has been characterized extensively using the deletion RDAf1 and spoligotyping, and shown to be most prevalent in countries from West- and Central Africa [11]. Here we could only compile nine Af1 genomes, of which five originated in Ghana [54], and the remaining had either a European or an unknown origin. The small diversity of Af1 spoligotyping patterns found in our WGS dataset [11], indicates strong undersampling (Fig. 2, Table S3). Nevertheless, it was possible to estimate the divergence of the Af1 clade from the remaining *M. bovis* to a period ranging from the year 1003 to 1361 AD (Fig. 3), making it unlikely that Af1 was originally brought to West Africa by Europeans [55].

The next split comprises clades *unknown4*, *unknown5* and Eu2. Clade *unknown4* was composed of 33 genomes with little geographic information and for which the most common spoligotyping pattern was the uninformative SB0120 (n=19). Additional *unknown4* spoligotyping patterns indicate that strains belonging to this clade circulate in Southern Europe, Northern and Eastern Africa (Fig. 2, Table S3), supporting dispersion events between Africa and Southern Europe. Clade *unknown5* comprised only 9 genomes isolated mostly from Zambian cattle. Its corresponding spoligotyping pattern is also SB0120, limiting further geographical inferences.

In contrast to the strains from clades *unknown4* and *unknown5*, among the 323 Eu2 genomes, no genomes of East African origin were found, and Africa was only represented by nine South African genomes [56]. By far, most Eu2 were isolated in the Americas. Previous studies have shown that Eu2 dominates in Southern Europe, particularly in the Iberian Peninsula [14], thus possibly the source of Eu2 in the Americas. There were no representatives of Eu2 from the Iberian Peninsula in our dataset. However, our molecular dating analysis revealed that the common ancestor of Eu2 evolved during the period 1459 to 1662 AD (Fig. 3), which would be compatible with an introduction from Europe into the Americas.

Clonal complex Eu1, *unknown6*, *unknow7* and *unknown8*, form a sister group to the previously described. Eu1 has previously been characterized based on the RDEu1 deletion

and spoligotyping, showing that it is highly prevalent in regions of the world that were former trading partners of the UK [8, 13]. That geographic range is well represented in our dataset including many genomes from the UK (n=215) and Ireland (n=45) (Table S2). The latter were very closely related, suggesting that there was probably fixation of just a few genotypes in this region as previously proposed [8]. In contrast, most branching events within Eu1 correspond to *M. bovis* isolated in North- and Central America as well as New Zealand, resulting from the expansion of clonal families not seen in the British Islands. Consequently, most of the genetic diversity of Eu1 exists outside of its putative region of origin. Our molecular dating is compatible with this view, indicating that the ancestor of Eu1 is likely to have emerged between the year 1294 to 1541 AD (Fig. 3), with several Eu1 sub-clades evolving in the last 200-300 years (TreeS1.txt, TreeS2.txt).

The closest relative to Eu1 is a genome from Ethiopia (*unknown8*) with the spoligotyping pattern SB1476, commonly found in Ethiopia [12]. *Unknown6* comprised seven genomes from North America (Fig. 2, Table S2, Table S3), whereas *unknown7* included eight genomes, four of which were isolated in Western Europe and another four without country of origin available. Spoligotyping patterns indicate that identical strains are common in Southern Europe, Northern and Eastern Africa expanding the geographic range of *unknown7*.

CONCLUSIONS AND IMPLICATIONS

We screened the public repositories and compiled 3,364 genome sequences of *M. bovis* and *M. caprae* from 35 countries. Although this dataset mostly represents local epidemics, it provides novel insights into the phylogeography of *M. bovis* and *M. caprae*. Our whole-genome based phylogeny showed that although certain spoligotypes show specific associations with monophyletic groups, prevalent patterns such as the so-called "BCG-like" should not be used to infer phylogenetic relatedness. Moreover, our data extend the previously known phylogenetic diversity of *M. bovis* by eight previously uncharacterized clades in addition to the four clonal complexes described previously. Among those, *unknown 1* shares a common ancestor with the rest of *M. bovis*, has an exclusively East African distribution and does not share the PncA mutation H57D, conferring intrinsic resistance to PZA. We provide here a global phylogenetic framework that can be further exploited to find better molecular markers for settings where WGS cannot be easily implemented.

Our further inferences suggest that *M. bovis* evolved in East Africa between the 4th and 10th century. While some *M. bovis* groups remained restricted to East Africa, others have dispersed to different parts of the world. The contemporary geographic distribution of *M.*

bovis clades suggest that East- and North Africa, Southern Europe and Western Asia have played an important role in shaping the population structure of these pathogens. However, these regions were underrepresented in our dataset. Thus, more *M. bovis* genomes from these regions are necessary to generate a more complete picture, particularly given the central role of the latter in the history of cattle domestication [57].

ACKNOWLEDGEMENTS

Calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing core facility at University of Basel. This work was supported by the Swiss National Science Foundation (grants 310030_166687, IZRJZ3_164171, IZLSZ3_170834 and CRSII5_177163), the European Research Council (309540-EVODRTB) and SystemsX.ch.

REFERENCES

1. WHO; Global tuberculosis report 2018. Geneva: World Health Organization, 2018.

2. Olea-Popelka F, Muwonge A, Perera A, et al.; Zoonotic tuberculosis in human beings caused by Mycobacterium bovis-a call for action. *Lancet Infect Dis* 2017;**17**(1):e21-e25. doi: 10.1016/S1473-3099(16)30139-6.

3. Muller B, Durr S, Alonso S, et al.; Zoonotic Mycobacterium bovis-induced tuberculosis in humans. *Emerg Infect Dis* 2013;**19**(6):899-908. doi: 10.3201/eid1906.120543.

4. Waters WR, Palmer MV, Buddle BM, et al.; Bovine tuberculosis vaccine research: historical perspectives and recent advances. *Vaccine* 2012;**30**(16):2611-22.

5. Tschopp R, Hattendorf J, Roth F, et al.; Cost estimate of bovine tuberculosis to Ethiopia. *Current topics in microbiology and immunology* 2013;**365**:249-68.

6. Michel AL, Muller B, van Helden PD; Mycobacterium bovis at the animal-human interface: a problem, or not? *Veterinary microbiology* 2010;**140**(3-4):371-81.

7. Gagneux S; Ecology and evolution of Mycobacterium tuberculosis. *Nat Rev Microbiol* 2018;**16**(4):202-213. doi: 10.1038/nrmicro.2018.8.

8. Smith NH; The global distribution and phylogeography of Mycobacterium bovis clonal complexes. *Infect Genet Evol* 2012;**12**(4):857-65. doi: 10.1016/j.meegid.2011.09.007.

9. Ghavidel M, Mansury D, Nourian K, et al.; The most common spoligotype of Mycobacterium bovis isolated in the world and the recommended loci for VNTR typing; A systematic review. *Microbial Pathogenesis* 2018;**118**:310-315. doi: 10.1016/j.micpath.2018.03.036.

10. Comas I, Homolka S, Niemann S, et al.; Genotyping of genetically monomorphic bacteria: DNA sequencing in mycobacterium tuberculosis highlights the limitations of current methodologies. *PLoS ONE* 2009;**4**(11):e7815. doi: 10.1371/journal.pone.0007815.

11. Muller 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. doi: 10.1128/JB.01590-08.

12. Berg S, Garcia-Pelayo MC, Muller B, et al.; African 2, a clonal complex of Mycobacterium bovis epidemiologically important in East Africa. *J Bacteriol* 2011;**193**(3):670-8. doi: 10.1128/JB.00750-10.

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):1340-51. doi: 10.1016/j.meegid.2011.04.027.

14. Rodriguez-Campos S, Schurch AC, Dale J, et al.; European 2--a clonal complex of Mycobacterium bovis dominant in the Iberian Peninsula. *Infect Genet Evol* 2012;**12**(4):866-72. doi: 10.1016/j.meegid.2011.09.004.

15. 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.

16. Orloski K, Robbe-Austerman S, Stuber T, et al.; Whole Genome Sequencing of Mycobacterium bovis Isolated From Livestock in the United States, 1989-2018. *Front Vet Sci* 2018;**5**:253. doi: 10.3389/fvets.2018.00253.

17. 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. doi: 10.1111/mec.15061.

18. 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. doi: 10.3389/fvets.2018.00272.

19. Brites D, Loiseau C, Menardo F, et al.; A New Phylogenetic Framework for the Animal-Adapted Mycobacterium tuberculosis Complex. *Frontiers in microbiology* 2018;**9**:2820.

20. 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. doi: 10.1186/s12859-018-2164-8.

21. Bolger AM, Lohse M, Usadel B; Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**(15):2114-20. doi: 10.1093/bioinformatics/btu170.

22. 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. doi: 10.1038/ng.590.

23. Li H, Handsaker B, Wysoker A, et al.; The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**(16):2078-9. doi: btp352 [pii] 10.1093/bioinformatics/btp352.

24. 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):1297-303. doi: 10.1101/gr.107524.110.

25. Li H; A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 2011;**27**(21):2987-93. doi: 10.1093/bioinformatics/btr509.

26. 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):568-76. doi: 10.1101/gr.129684.111.

27. 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 w1118; iso-2; iso-3. *Fly (Austin)* 2012;**6**(2):80-92. doi: 10.4161/fly.19695.

28. 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):7877-82.

29. Steiner A, Stucki D, Coscolla M, et al.; KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. *BMC Genomics* 2014;**15**:881. doi: 10.1186/1471-2164-15-881.

30. Smith NH, Upton P; Naming spoligotype patterns for the RD9-deleted lineage of the Mycobacterium tuberculosis complex; www.Mbovis.org. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* 2012;**12**(4):873-6.

31. Miller MA, Pfeiffer W, Schwartz T; Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *Proceedings of the Gateway Computing Environments Workshop (GCE)*. New Orleans, LA, 2010, 1-8.

531 32. Copin R, Coscolla M, Efstathiadis E, et al.; Impact of in vitro evolution on antigenic
532 diversity of Mycobacterium bovis bacillus Calmette-Guerin (BCG). *Vaccine*
533 2014;**32**(45):5998-6004. doi: 10.1016/j.vaccine.2014.07.113.

534 33. Revell LJ; phytools: an R package for phylogenetic comparative biology (and other
535 things). *Methods in Ecology and Evolution* 2011;**3**(2):217-223.

536 34. Lewis PO; A likelihood approach to estimating phylogeny from discrete
537 morphological character data. *Syst Biol* 2001;**50**(6):913-25. doi:
538 10.1080/106351501753462876.

539 35. Team RC; R: A language and environment for statistical computing. R Foundation for
540 Statistical Computing. Vienna, Austria, 2019.

541 36. Bollback JP; SIMMAP: stochastic character mapping of discrete traits on phylogenies.
542 *BMC Bioinformatics* 2006;**7**:88. doi: 10.1186/1471-2105-7-88.

543 37. Menardo F, Duchêne S, Brites D, et al.; The molecular clock of Mycobacterium
544 tuberculosis. *PloS Pathogens* 2019;***in progress***.

545 38. Paradis E, Schliep K; ape 5.0: an environment for modern phylogenetics and
546 evolutionary analyses in R. *Bioinformatics* 2019;**35**(3):526-528. doi:
547 10.1093/bioinformatics/bty633.

548 39. To TH, Jung M, Lycett S, et al.; Fast Dating Using Least-Squares Criteria and
549 Algorithms. *Systematic Biology* 2016;**65**(1):82-97. doi: 10.1093/sysbio/syv068.

550 40. Darriba D, Taboada GL, Doallo R, et al.; jModelTest 2: more models, new heuristics
551 and parallel computing. *Nature Methods* 2012;**9**(8):772-772. doi: DOI 10.1038/nmeth.2109.

552 41. Bouckaert R, Heled J, Kuhnert D, et al.; BEAST 2: a software platform for Bayesian
553 evolutionary analysis. *PLoS Comput Biol* 2014;**10**(4):e1003537. doi:
554 10.1371/journal.pcbi.1003537.

555 42. Drummond AJ, Ho SYW, Phillips MJ, et al.; Relaxed phylogenetics and dating with
556 confidence. *Plos Biology* 2006;**4**(5):699-710. doi: ARTN e88
557 10.1371/journal.pbio.0040088.

558 43. Rambaut A, Drummond AJ, Xie D, et al.; Posterior Summarization in Bayesian
559 Phylogenetics Using Tracer 1.7. *Systematic Biology* 2018;**67**(5):901-904. doi:
560 10.1093/sysbio/syy032.

561 44. Rodriguez S, Bezos J, Romero B, et al.; Mycobacterium caprae Infection in Livestock
562 and Wildlife, Spain. *Emerging Infectious Diseases* 2011;**17**(3):532-535. doi:
563 10.3201/eid1703.100618.

564 45. Prodinger WM, Brandstatter A, Naumann L, et al.; Characterization of
565 Mycobacterium caprae isolates from Europe by mycobacterial interspersed repetitive unit
566 genotyping. *J Clin Microbiol* 2005;**43**(10):4984-92.

567 46. Yoshida S, Suga S, Ishikawa S, et al.; Mycobacterium caprae Infection in Captive
568 Borneo Elephant, Japan. *Emerg Infect Dis* 2018;**24**(10):1937-1940. doi:
569 10.3201/eid2410.180018.

570 47. Consortium C, the GP, Allix-Beguec C, et al.; Prediction of Susceptibility to First-
571 Line Tuberculosis Drugs by DNA Sequencing. *N Engl J Med* 2018;**379**(15):1403-1415. doi:
572 10.1056/NEJMoa1800474.

573 48. Aranaz A, Cousins D, Mateos A, et al.; Elevation of Mycobacterium tuberculosis
574 subsp. caprae Aranaz et al. 1999 to species rank as Mycobacterium caprae comb. nov., sp.
575 nov. *Int J Syst Evol Microbiol* 2003;**53**(Pt 6):1785-9.

576 49. Broeckl S, Krebs S, Varadharajan A, et al.; Investigation of intra-herd spread of
577 Mycobacterium caprae in cattle by generation and use of a whole-genome sequence. *Vet Res*
578 *Commun* 2017;**41**(2):113-128. doi: 10.1007/s11259-017-9679-8.

579 50. Loiseau C, Brites D, Moser I, et al.; Revised Interpretation of the Hain Lifescience
580 GenoType MTBC To Differentiate Mycobacterium canettii and Members of the
581 Mycobacterium tuberculosis Complex. *Antimicrobial Agents and Chemotherapy* 2019;**63**(6).
582 doi: ARTN e00159-19
583 10.1128/AAC.00159-19.

584 51. Taylor GM, Murphy E, Hopkins R, et al.; First report of Mycobacterium bovis DNA
585 in human remains from the Iron Age. *Microbiology* 2007;**153**(Pt 4):1243-9.

586 52. Nicklisch N, Maixner F, Ganslmeier R, et al.; Rib lesions in skeletons from early
587 neolithic sites in Central Germany: on the trail of tuberculosis at the onset of agriculture. *Am J*
588 *Phys Anthropol* 2012;**149**(3):391-404. doi: 10.1002/ajpa.22137.

589 53. Oettinger T, Jorgensen M, Ladefoged A, et al.; Development of the Mycobacterium
590 bovis BCG vaccine: review of the historical and biochemical evidence for a genealogical tree.
591 *Tuber Lung Dis* 1999;**79**(4):243-50. doi: 10.1054/tuld.1999.0206.

592 54. Otchere ID, van Tonder AJ, Asante-Poku A, et al.; Molecular epidemiology and whole
593 genome sequencing analysis of clinical Mycobacterium bovis from Ghana. *PLoS One*
594 2019;**14**(3):e0209395. doi: 10.1371/journal.pone.0209395.

595 55. Muwonge A, Franklyn E, Mark B, et al.; Molecular Epidemiology of Mycobacterium
596 bovis in Africa. In: B. DA, J. KNP, O. TCs (eds). *Tuberculosis in Animals: An African*
597 *Perspective*. Switzerland: Springer, 2019, 127-170.

598 56. Dippenaar A, Parsons SDC, Miller MA, et al.; Progenitor strain introduction of
599 Mycobacterium bovis at the wildlife-livestock interface can lead to clonal expansion of the

disease in a single ecosystem. *Infect Genet Evol* 2017;**51**:235-238. doi: 10.1016/j.meegid.2017.04.012.

57. Decker JE, McKay SD, Rolf MM, et al.; Worldwide Patterns of Ancestry, Divergence, and Admixture in Domesticated Cattle. *Plos Genetics* 2014;**10**(3). doi: ARTN e1004254 10.1371/journal.pgen.1004254.

58. Rambaut A; FigTree. Edinburgh: Institute of Evolutionary Biology, University of Edinburgh, 2010.

FIGURE LEGENDS

Figure 1 – Geographic distribution of the *M. bovis* samples used in this study according to isolation country. The circles correspond to pie charts and are coloured according to clonal complexes.

Figure 2 – Maximum likelihood phylogeny of 476 of the 3,364 genomes included in this study (redundant genomes were removed), and inferred from 22,492 variable positions. The scale bar indicates the number of substitutions per polymorphic site. The phylogeny is rooted on a *M. tuberculosis* Lineage 6 genome from Ghana (not shown) and bootstrap values are shown for the most important splits. The coloured bars on the side of the phylogeny show the different clonal complexes. Other "unknown" monophyletic clades are coloured in black and additionally the branches of the eight clades are coloured to show their phylogenetic position more precisely. The pie charts mapped on the tree represent the summary posterior probabilities (from 100 runs) of the reconstructed ancestral geographic states and are coloured according to geographical UN region. Inferred spoligotype patterns from WGS described in *M. bovis* spoligotype database [30] are indicated for the unknown clades.

Figure 3 – The inferred age of main monophyletic clades according to LSD and Beast dating analyses. The confidence intervals reported correspond to the BEAST analysis. The Beast analysis was based on 300 genomes and the LSD analysis was based on 2058 genomes (see methods section for subsampling strategy). Only one genome from the Af1 clonal complex was included in the dating analyses and therefore the dates reported correspond to the node where Af1 diverged.

Supporting material

Figure S1 – Flow chart showing the selection of genomes.

Figure S2 – Maximum likelihood phylogeny of all 3,364 genomes, based on 45,981 variable positions. The scale bar indicates the number of substitutions per polymorphic site. The phylogeny is rooted on a *M. tuberculosis* Lineage 6 genome from Ghana. The outer ring indicates the geographical region from which the strains were isolated. The four clonal complexes are highlighted on the tree. Branches corresponding to BCG genomes are coloured in grey and the *PncA* mutation H57D is indicated by a yellow star.

Figure S3 – Phylogeographic reconstruction of *M. bovis* and *M. caprae*, inferred from 392 genomes. Thirteen UN-defined geographic regions were assigned to the discrete character geographic origin, and mapped onto the phylogeny. Pie charts at internal nodes represent the summary posterior probabilities (from 100 runs) of the reconstructed ancestral geographic states and are coloured according to geographical UN region

Figure S4 – A) Tip-to-root regression and B) Date randomization tests (DTR). Significance levels the clock rate estimate for the observed data does not overlap; 1) with the range of estimates obtained from the randomized sets (simple test); 2) with the confidence intervals of the estimates obtained from the randomized sets (intermediate test) and 3) the confidence interval of the clock rate estimate for the observed data does not overlap with the confidence intervals of the estimates obtained from the randomized sets (stringent test).

Table S1 - List of genomes included in this study along with metadata used for the analyses.

Table S2 - Comparison of models for discrete character evolution using likelihood ratio tests.

Table S3 - Spoligotype patterns determined *in silico* for different clonal complex groups with reference to other studies.

TreeS1.txt – The maximum clade credibility tree inferred using Bayesian evolutionary analysis in BEAST2 [41] based on 300 genomes. Tips labels are present in Table S1. Ages in years before present and correspondent 95% High Posterior Density (HPD) can be visualized as node_heights and height_95%_HPD respectively, in FigTree [58].

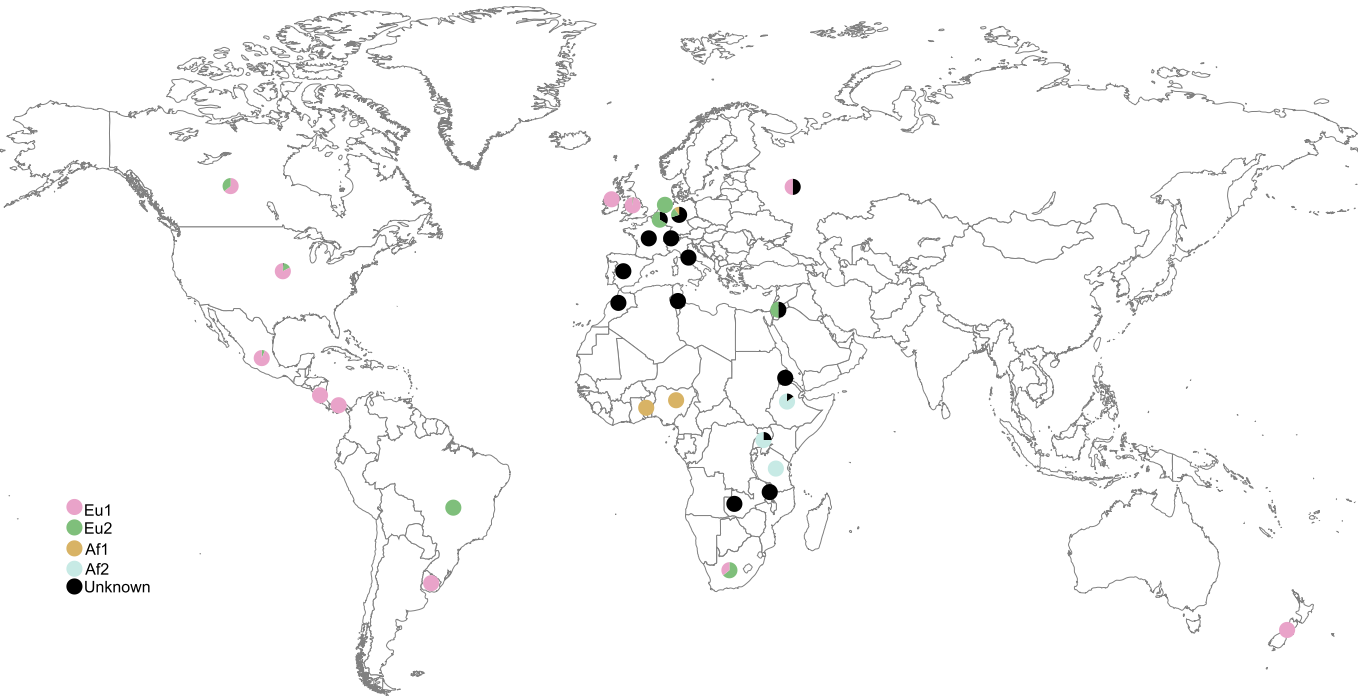
668

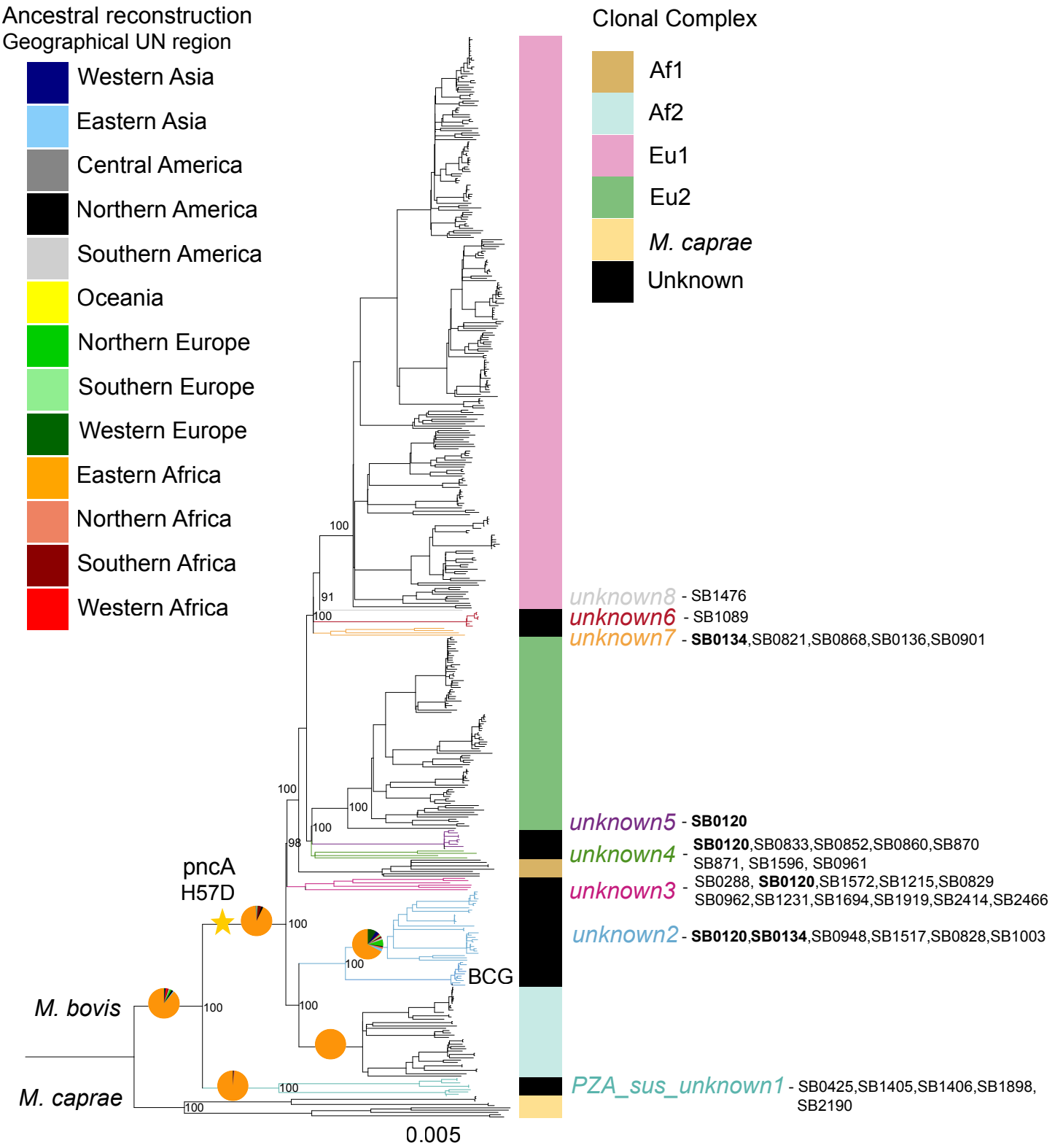
669 **TreeS2.txt** – Least-squares dating using LSD [39] based on 2058 genomes. Tips labels are
670 present in Table S1. Ages in years before present can be visualized as Node heights in
671 FigTree [58].

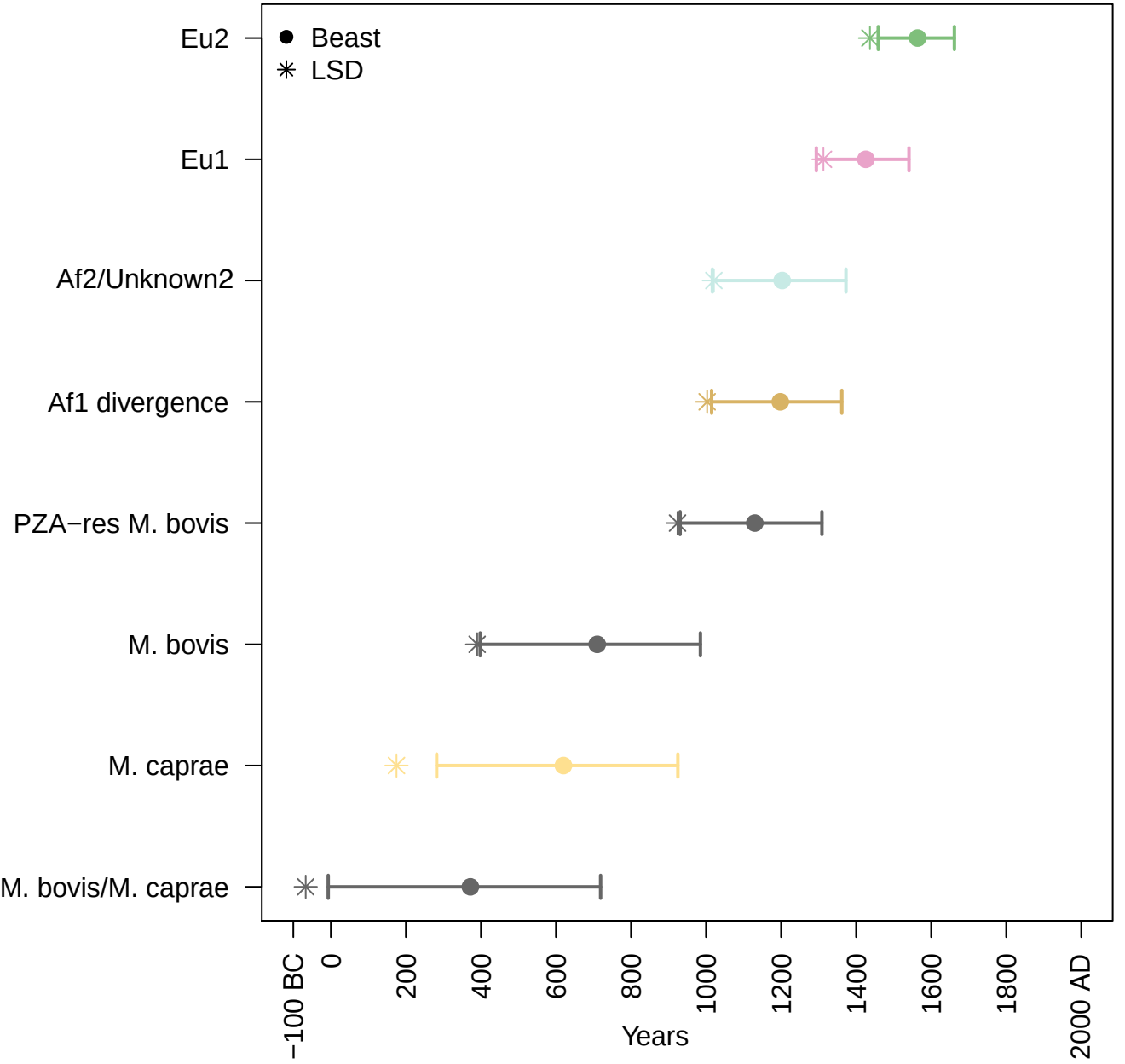
672

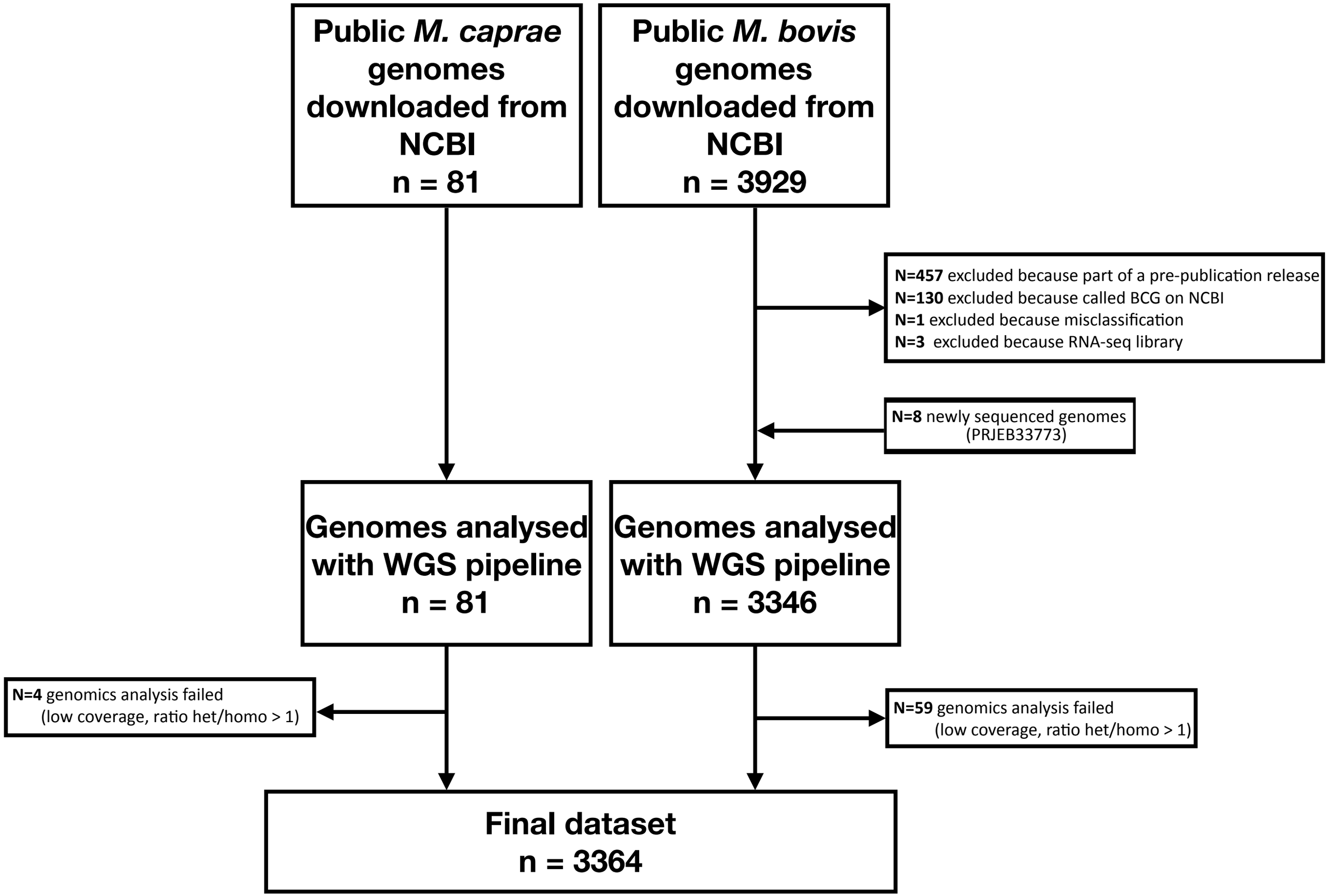
673

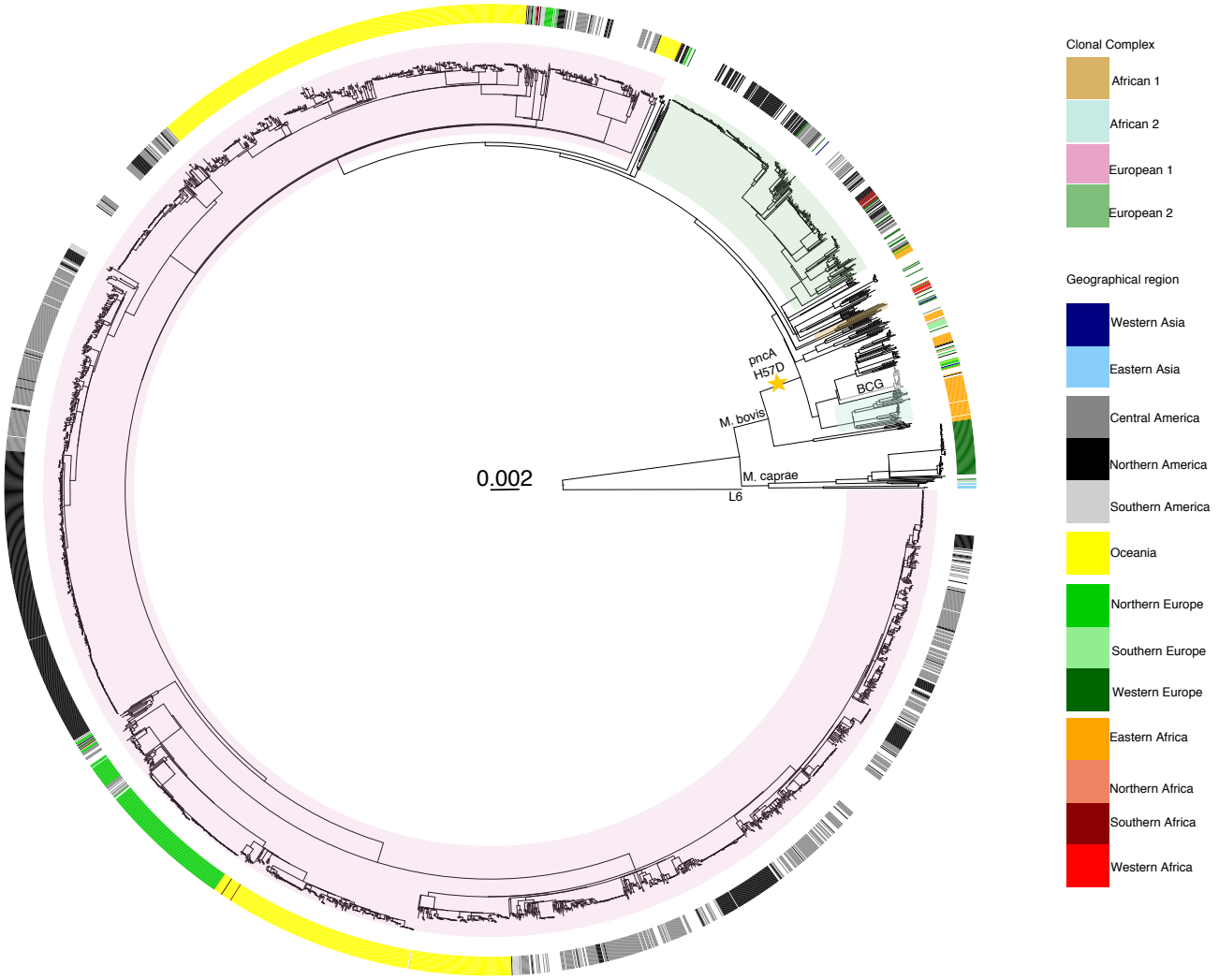
674

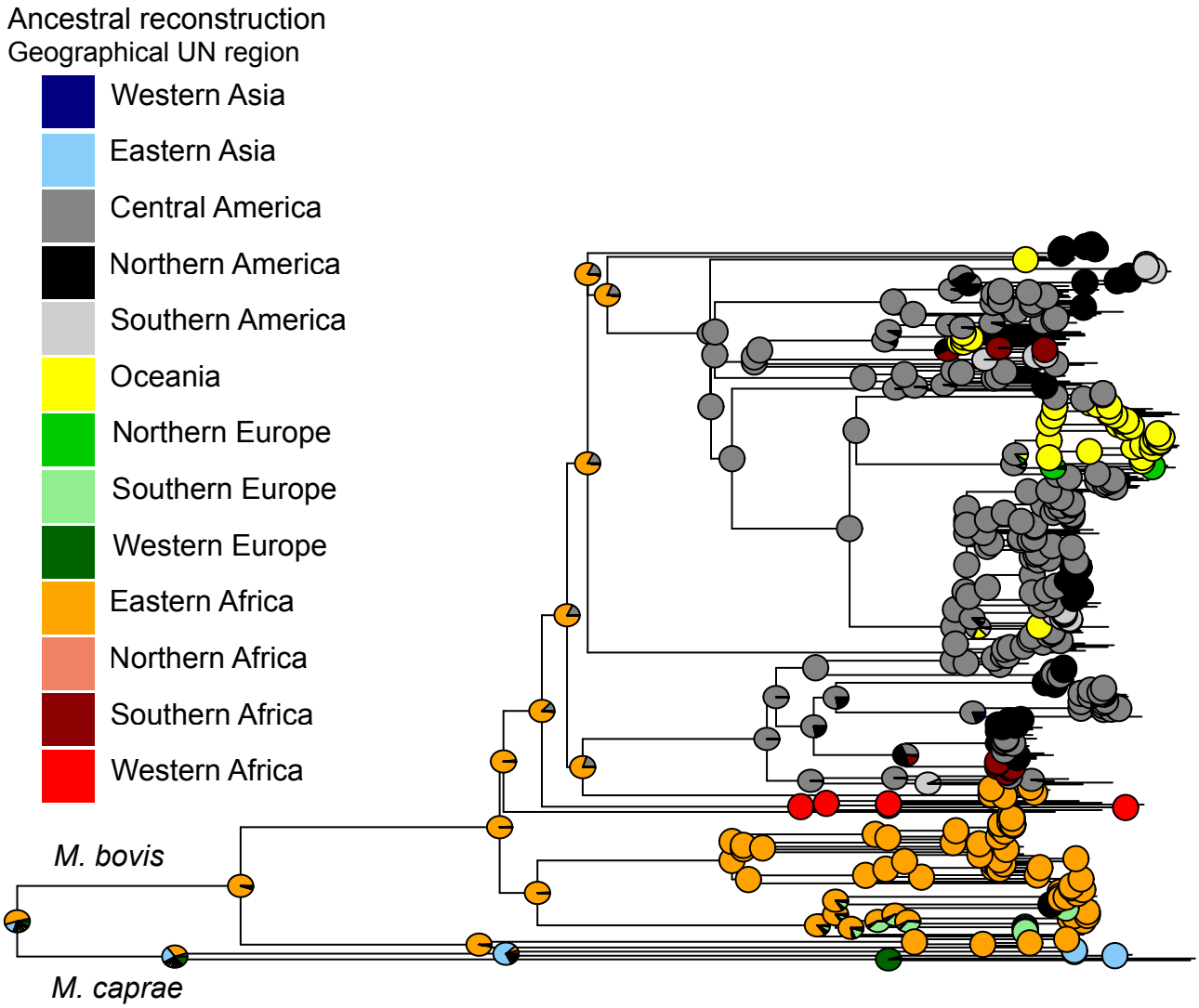


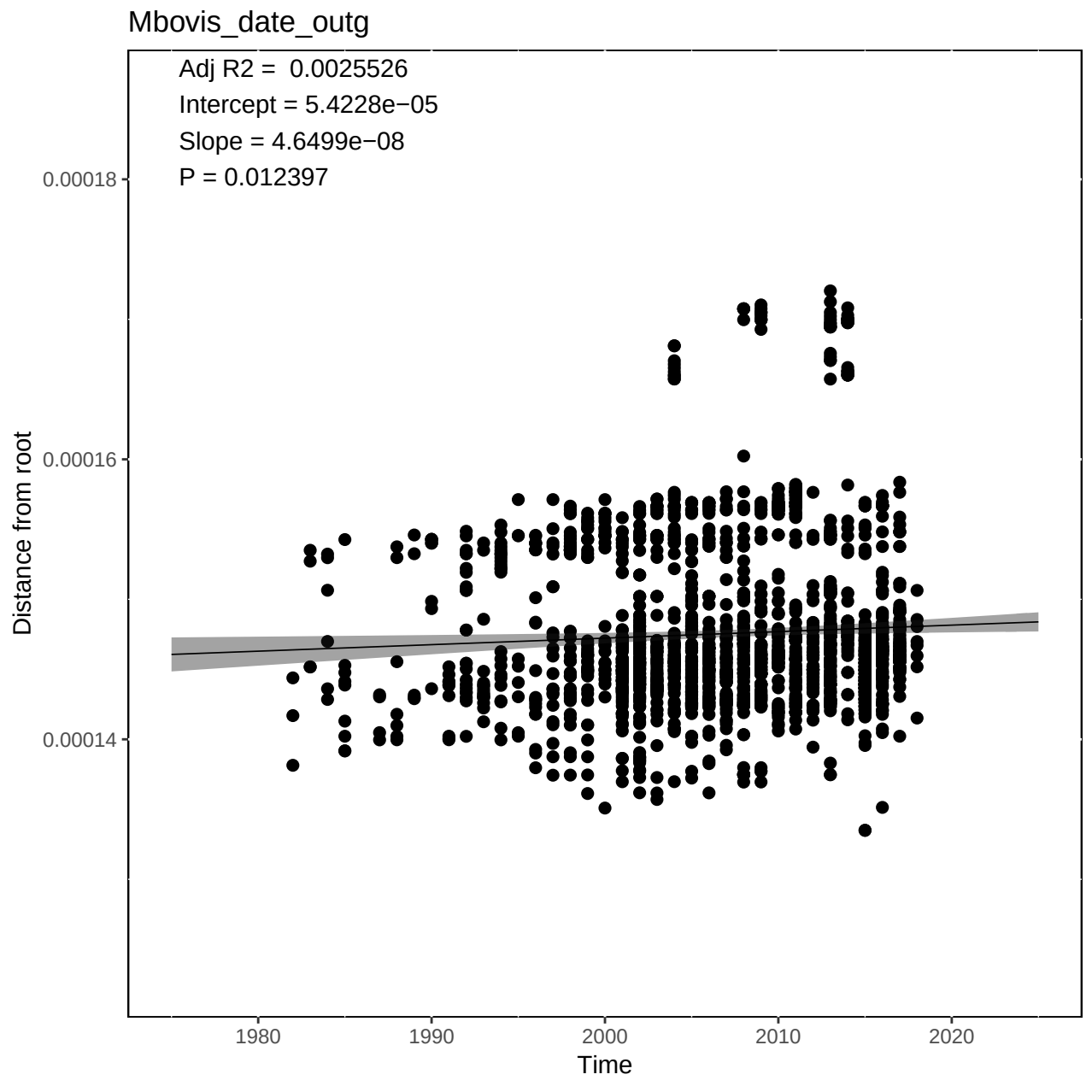

Ancestral reconstruction
Geographical UN region
Western Asia
Eastern Asia
Central America
Northern America
Southern America
Oceania
Northern Europe
Southern Europe
Western Europe
Eastern Africa
Northern Africa
Southern Africa
Western Africa
Clonal Complex
Af1
Af2
Eu1
Eu2
M. caprae
Unknown
unknown8 - SB1476
unknown6 - SB1089
unknown7 - SB0134,SB0821,SB0868,SB0136,SB0901
unknown5 - SB0120
unknown4 - SB0120,SB0833,SB0852,SB0860,SB870 SB871, SB1596, SB0961
unknown3 - SB0288, SB0120,SB1572,SB1215,SB0829 SB0962,SB1231,SB1694,SB1919,SB2414,SB2466
unknown2 - SB0120,SB0134,SB0948,SB1517,SB0828,SB1003
PZA_sus_unknown1 - SB0425,SB1405,SB1406,SB1898, SB2190
pncA H57D
BCG
M. bovis
M. caprae
100
91
98
0.005

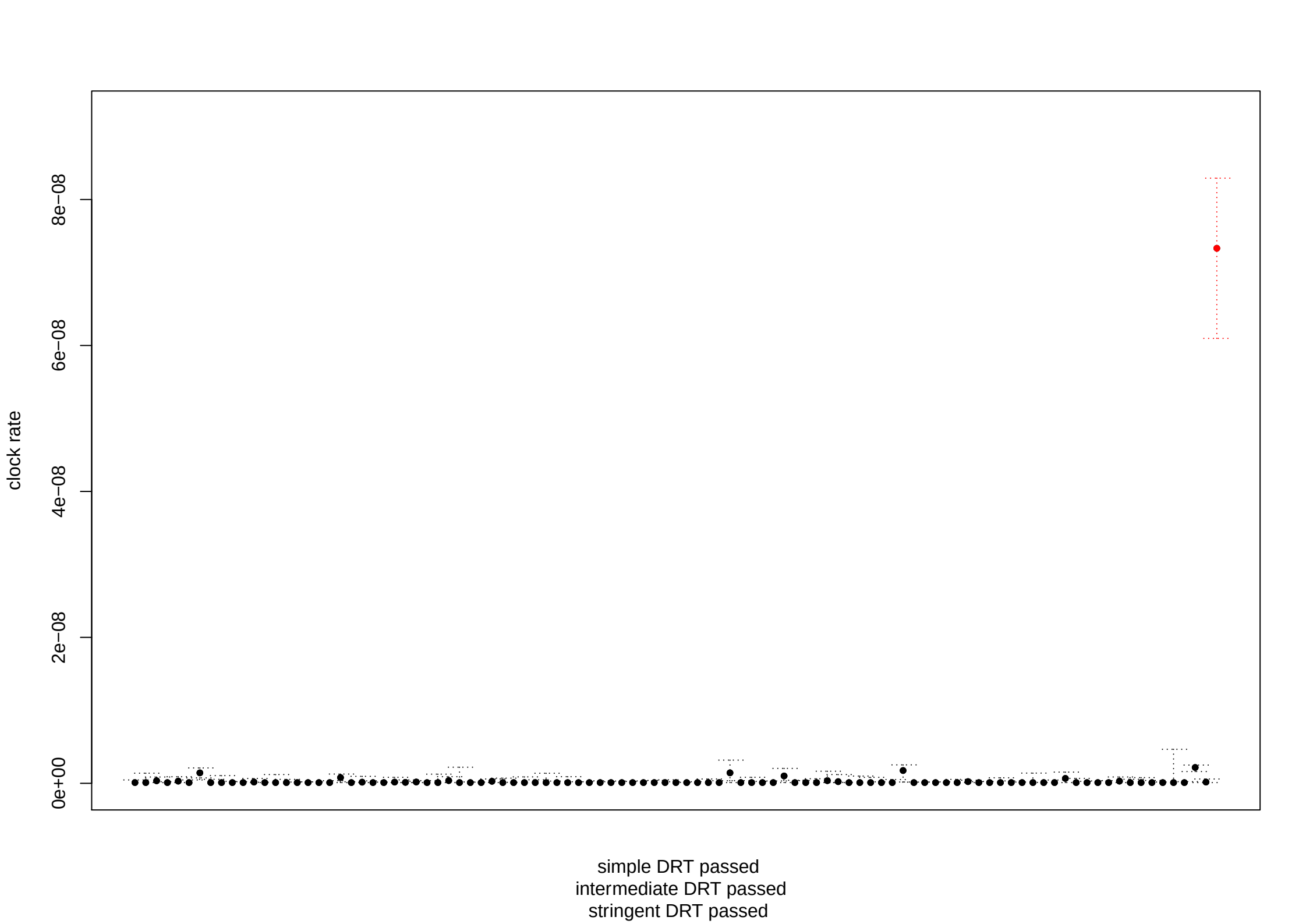

Beast
LSD
Eu2
Eu1
Af2/Unknown2
Af1 divergence
PZA−res M. bovis
M. bovis
M. caprae
M. bovis/M. caprae
−100 BC
0
200
400
600
800
1000
1200
1400
1600
1800
2000 AD
Years


Public *M. caprae* genomes downloaded from NCBI n = 81
Public *M. bovis* genomes downloaded from NCBI n = 3929
N=457 excluded because part of a pre-publication release
N=130 excluded because called BCG on NCBI
N=1 excluded because misclassification
N=3 excluded because RNA-seq library
N=8 newly sequenced genomes (PRJEB33773)
Genomes analysed with WGS pipeline n = 81
Genomes analysed with WGS pipeline n = 3346
N=4 genomics analysis failed (low coverage, ratio het/homo > 1)
N=59 genomics analysis failed (low coverage, ratio het/homo > 1)
Final dataset n = 3364


0.002
pncA
H57D
BCG
M. bovis
M. caprae
L6
Clonal Complex
African 1
African 2
European 1
European 2
Geographical region
Western Asia
Eastern Asia
Central America
Northern America
Southern America
Oceania
Northern Europe
Southern Europe
Western Europe
Eastern Africa
Northern Africa
Southern Africa
Western Africa




Mbovis_date_outg
Adj R2 = 0.0025526
Intercept = 5.4228e−05
Slope = 4.6499e−08
P = 0.012397
0.00018
0.00016
0.00014
Distance from root
1980
1990
2000
2010
2020
Time


clock rate
0e+00
2e−08
4e−08
6e−08
8e−08
simple DRT passed
intermediate DRT passed
stringent DRT passed