Unraveling the epidemiology of Mycobacterium bovis using whole-genome sequencing combined with environmental and demographic data

Title

Unravelling the epidemiology of Mycobacterium bovis using whole genome sequencing combined with environmental and demographic data

Authors

Gianluigi Rossi[1,2+], Barbara Shih[1], Franklyn N. Egbe[3], Paolo Motta[4], Florian Duchatel[1], Robert Kelly[5], Lucy Ndip[6,7], Melissa Sander[8], Vincent Tanya[9], Samantha J. Lycett[1,2], Barend Mark de Clare Bronsvoort[1,2], Adrian Muwonge[1]

Affiliations

[1]The Roslin Institute, R(D)SVS, University of Edinburgh, Easter Bush Campus, Midlothian, EH25 9RG, Scotland

[2]Centre of Expertise on Animal Diseases Outbreaks, EPIC, Scotland

[3]School of Life Sciences, University of Lincoln, Brayford Pool, Lincoln LN6 7TS United Kingdom

[4]The European Commission for the Control of Foot-and-Mouth Disease (EuFMD), Food and Agriculture Organization of the United Nations, Rome, Italy

[5]Royal (Dick) School of Veterinary Studies and The Roslin Institute, University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, United Kingdom

[6]Laboratory for Emerging Infectious Diseases, University of Buea, Buea, Cameroon

[7]Department of Biomedical Sciences, Faculty of Health Sciences, University of Buea, Buea, Cameroon

[8]Tuberculosis Reference Laboratory, Bamenda, P.O. Box 586, Cameroon

[9]Cameroon Academy of Sciences, P.O. Box 1457, Yaoundé, Cameroon

[+]Corresponding author: g.rossi@ed.ac.uk

Author Contributions

BCMB, LN, and VT conceived the original project. BCMB, RK, LN, VT, MS, and NE designed the field study, the databases, and the survey instrument. RK, NE, VT, and BB developed the field SOPs and collected the data. PM collected the cattle movement data. GR, AM, FD and FE cleaned the data. GR, BCMB, SJL and AM conceived the quantitative analysis. GR and SJL run the phylogenetic analysis. BS implemented the bioinformatical pipelines. GR performed

the machine learning analysis. GR was responsible for writing the initial drafts. All authors contributed comments for the final draft. All authors contributed to the article and approved the submitted version.

Acknowledgements

We want to thank Dr. Isaac D. Otchere for providing the isolation year of the *M. bovis* sequences sampled in Ghana, Dr. Stella Mazeri and Dr. William Harvey, Roslin Institute University of Edinburgh, for data collection and analysis support, and Prof. Kim Vanderwaal, University of Minnesota for the useful comments.

Fundings

The primary data used in this work was generated with funding from Wellcome Trust (WT094945), with BCMB as the Principal investigator. BS, SJL, MB, BMCB, and AM are supported by Biotechnology and Biological Sciences Research Council (BBSRC) programme grant to Roslin Institute (Award Numbers BBS/E/D/20002172 and BBS/E/D/200021723), and AM later by his BBSRC Future leader Fellowship and current Chancellor's fellowship. GR, SJL, BCMB additional received support from received support from the Scottish Government Rural and Environment Science and Analytical Services Division as part of Centre of Expertise on Animal Disease Outbreaks (EPIC). We are also grateful to the staff of MENIPIA, especially the veterinarians and delegates who diligently supported the primary fieldwork that generated this data. BS was partially funded by a BBSRC Core Capability Grant BB/CCG1780/1 awarded to The Roslin Institute. *M. bovis* sequencing was carried out by Edinburgh Genomics, The University of Edinburgh, which is partly supported through core grants from NERC (R8/H10/56), MRC (MR/K001744/1) and BBSRC (BB/J004243/1).

Abstract

When studying the dynamics of a pathogen in a host population, one crucial question is whether it transitioned from an epidemic (i.e. the pathogen population and the number of infected hosts are increasing) to an endemic stable state (i.e. the pathogen population reached an equilibrium). For slow-growing and slow-evolving clonal pathogens like *Mycobacterium bovis*, the causative agent of bovine (or animal) and zoonotic tuberculosis, it can be challenging to discriminate between these two states. This is a result of the combination of suboptimal detection tests, so that the actual extent of the pathogen prevalence is often unknown, as well as of the low genetic diversity, which can hide the temporal signal provided by the accumulation of mutations in the bacteria DNA.

In recent years, the increased availability, efficiency and reliability of genomic reading techniques, such as whole-genome sequencing (WGS), has significantly increased the amount of information we can use to study infectious diseases, and therefore it has improved the precision of epidemiological inferences for pathogens like *M. bovis*.
In this study, we use WGS to gain insights into the epidemiology of *M. bovis* in Cameroon, a developing country where the pathogen has been reported for decades. Ninety-one high-quality sequences were obtained from tissue samples collected in four abattoirs, 64 of which with complete metadata. We combined these with environmental, demographic, ecological and cattle movement data to generate inferences using phylodynamic models.

Our findings suggest *M. bovis* in Cameroon is slowly expanding its epidemiological range over time, therefore endemic stability is unlikely. This suggests that animal movement plays an important role in transmission. The simultaneous prevalence of *M. bovis* in co-located cattle and humans highlights the risk of such transmission being zoonotic. Therefore, using genomic tools as part of surveillance would vastly improve our understanding of disease ecology and control strategies.

Keywords

Mycobacterium bovis; whole-genome sequencing; phylodynamic analysis; genomic surveillance; livestock epidemics; zoonotic tuberculosis; One-health

1. Introduction

In the last two decades, the increased availability, efficiency and reliability of genomic reading techniques, such as whole-genome sequencing (WGS) techniques, have ignited a profound transformation in understanding disease ecology and epidemiology. This, coupled with improved statistical methodologies and high-performance computing, has enhanced our understanding of pathogen dynamics and evolution (1).

Techniques such as WGS can identify polymorphisms in the genetic material, which is generated by transcription errors that can occur to the pathogen while replicating within their host (2). As the pathogen is transmitted through the host population, the accumulation of polymorphisms in its DNA/RNA can be used as a "transmission signature". Therefore, by tracking these mutations across bacterial genomes sampled in a host population, we are now able to infer transmission events between individual hosts, sub-populations, geographical areas or species, while at the same time gather insights about the evolutionary trajectory of a pathogen (2). Furthermore, when accurate spatial information on the sampled isolates is available, we can combine it with pathogen genetic data to disentangle the spatio-temporal dynamics of outbreaks, particularly in natural or other scarcely sampled animal populations (3).

Despite these advances, many challenges still exist, including the reconciliation between the temporal signal of outbreaks with pathogen mutations (4). *Mycobacterium tuberculosis* Complex (MTBC) members are clonal species, and therefore recombination has been considered rare (although a recent publication showed otherwise (5)). A few mutations are expected to occur for these species per year, generating little diversity during outbreaks in host populations. Consequently, there is inherent uncertainty in establishing infection patterns within the infected population and their associated infections. Therefore, combining genomic information with metadata is essential for accurate transmission chain estimation (6).

Mycobacterium bovis, a member of the MTBC group, is the aetiological agent of animal or bovine tuberculosis (bTB) in bovids and other mammalians and of zoonotic tuberculosis (TB) in humans (7). Its infections are characterised by chronic disease, with or without a latent period, where infected cattle are hard to identify, making it hard to quantify potential infectious contacts (8). The estimation of *M. bovis* prevalence is often affected by several factors, including the inaccuracy of diagnostic tests (9), and the potential co-infection with

other pathogens (10). Such challenges explain why *M. bovis* has only been successfully eliminated or controlled in a few countries. Yet, it still represents a significant threat to cattle industries and human health in many other countries. For example, zoonotic tuberculosis due to *M. bovis* is a major public health problem in low and medium-income countries (LMICs), where close interaction between people and livestock is common and the limited access to pasteurized milk (7,11). Indeed, the magnitude of this burden is likely underestimated since human-animal transmission is predominantly via ingestion of infected products and presenting with a range of non-specific symptoms (12).

In Cameroon *M. bovis* is circulating in the cattle population, both in the southern areas (13) and, in particular, in the northern regions, where a previous study on cattle sampled at four regional abattoirs showed a sampled population prevalence ranging from 2.75% (31 positive over 1'129 cattle inspected, Northwest) to 21.25% (34 over 160, North (14)). Abattoirs surveillance, where carcases are inspected for TB-like lesions, is the only surveillance strategy regularly implemented in the country; in Bamenda (Northwest region), Awah-Ndukum and colleagues (15) showed that the TB-like lesion in cattle increased in the period from 1994 to 2010.

Commonly to many LMICs, bTB control in Cameroon is also made difficult by the absence of detailed records on cattle population, by local rearing practices such as pastoralism which expose animals to contacts with other herds and potential reservoir wildlife species, and by the transhumance cattle movements westward towards Nigeria, where the demand of meat is driven by a fast human population increase (16).

In a previous study, Egbe et al. (16) employed two molecular typing techniques to understand the relatedness of *M. bovis* strains circulating in the region. These are spoligotyping and MIRU-VNTR typing: the former is based on the presence of multiple spacer oligonucleotides in the genome Direct Repeat region, while the latter is based on 12 loci containing variable numbers of tandem repeats of mycobacterial interspersed repetitive units (17,18). Compared to WGS, these techniques consider a limited genome region and can be more subject to homoplasy (19). The results reported by Egbe (16) showed that most of the isolates belonged to the Af1 clonal complex (*n* = 250/total *n* = 255), while the remaining ones had an unidentified clonal complex. They also highlighted an unexpectedly high genetic diversity, as showed by the 37 sampled spoligotypes, of which 19 newly

observed, and a total of 97 genotypes, obtained by combining spoligotypes with MIRU-VNTR (16).

While those techniques are instrumental to investigating potential infection clusters at a broader level, they can be limited for a more in-depth understanding of the spatio-temporal dynamics of the disease. This study aimed to fill these gaps and enhance our understanding of the *M. bovis* epidemiology and spatial dynamics in Cameroon using WGS. We applied novel phylogenetic techniques to determine whether there was endemic stability across Cameroon's cattle-rearing regions while examining the role of environmental and ecological variables and animal movements in the pathogen spread.

We used 91 high-quality *M. bovis* sequences obtained from cattle's tissues sampled at regional abattoirs as described by Egbe et al. (14). After determining the single nucleotide polymorphisms (SNPs), we built a tree by joining the Cameroonian WGSs with other African sequences obtained from publicly available repositories, in order to understand how the sampled population fit in the continent context. Then, we ran a continuous space phylogeographical analysis with *BEAST* (20) on the Cameroonian sequences while testing different random walk diffusion models (21). This was possible because the origin village of the cattle tested at the abattoir was known for 64 *M. bovis* cattle isolates, allowing us to associate spatial coordinates to these sequences. Further, we tested the association between the spatial pathogen distribution obtained with the georeferenced phylogenetic tree and environmental, anthropic and ecological factors (22), and we finally ran a machine learning analysis to test whether the empirical cattle movement network (23) or other variables could explain the genetic diversity across isolates.

Our findings strengthen the call for an improved *M. bovis* molecular surveillance in underrepresented regions and countries, so to gather insights on potential patterns that can be missed when limiting the studies to areas of low genetic diversity, consequence of strict control practices such as test-and-cull.

2. Materials and methods

2.1. Data collection

Four regional abattoirs were sampled between 2012 and 2013, in the Northwest (Bamenda), Adamawa (Ngaoundere), North (Garoua) and Extreme North (Maroua) regions of Cameroon (Figure S1). As part of the regular operations, cattle carcases were inspected for the presence of TB-like lesions. The tissues, including lymph nodes, of all animals with lesions and of some randomly chosen without lesions were collected to be cultured, and information about the animal (age, breed, village of provenance, among others) were taken. A detailed description of the data collection and bacterial isolation can be found in Egbe et al. (14). The DNA extraction was conducted in BSL 3 facilities (Tuberculosis Reference Laboratories in Bamenda, Cameroon), and the procedure is fully described in Egbe et al. (16). Sequencing was also attempted for *M. bovis* isolates sampled in human hosts at the Bamenda hospital (Northwest region) during a cross-sectional study within the wider project. We reported a summary of the number of sampled animals and the number of *M. bovis* positive ones in Table S1.

2.2. Whole genome sequencing processing

The sequencing was carried out at Edinburgh Genomic facilities (University of Edinburgh). Samples were prepared with 1 TruSeq Nano 550 bp insert, 76 Pippin selected library from the supplied genomic DNA, while MiSeq v2 (Illumina) was used to generate 250 base paired-end sequence from library to yield at least 11M+11M reads (1 run) at 30x coverage. The output was read from a 4 lane Miseq. A total of 124 *M. bovis* WGSs were obtained (two from human hosts), while for nine isolates (one from human) the sequencing failed.

We used adapted *BovTB*-nf pipeline (24) for quality control. Reads were deduplicated using *fastuniq*, trimmed using *Trimmomatics* (25) (-phred33 ILLUMINACLIP:$adapters:2:30:10 SLIDINGWINDOW:10:20 MINLEN:36), and mapped to the reference genome using *bwa-mem2* (26). The mapped reads were filtered (-ShuF 2308 -) and sorted using *Samtools* (27), and then classified using *Kraken2* (28) (--quick) against a prebuilt Kraken 2 database (*Minikraken* v2 (28)). The Kraken2 output was summarized with Braken (29) (-r 150 -l S), and the top 20 list of species from the Bracken output was used to determine if the sample was contaminated with other microorganisms. Variants were called using *bcftools* (30) (--IndelGap 5 -e 'DP<5 && AF<0.8') and strain-specific SNPs were used for classifying whether

213 the samples were *M. bovis* or not (custom script and differentiating SNPs taken from (24)).
214 The percentage of coverage (>60%) on the reference, read depth (>10), and number of
215 reads (> 600,000) were used to identify and remove samples with insufficient data. To
216 curate aligned core-variants for the downstream phylogenetic analysis, variants were called
217 and filtered using *Snippy* v4.6.0 (31) using the default settings (minimum coverage = 10,
218 minimum VCF variant call quality = 100), with the *M. bovis* AF2122/97 genome (GeneBank:
219 LT708304.1) as the reference genome. Variants from repeated regions were removed (mask
220 for repetitive regions taken from (24)). Core-SNPs were determined by snippy-core function
221 within *Snippy*, where a genomic position was considered to be a core-site when present in
222 all samples. We defined as "high-quality" sequences the ones with genome coverage > 90%
223 and reading depth > 10 (32), and we renamed the sequences with a string composed by the
224 following information: host species, location (administrative subdivision, or country, see
225 Section 2.3), sequential number, and date.
226 For each sequence, the spoligotype and the clonal complex were retrieved from Egbe et al.
227 (16). In one case a sequence was missing the spoligotype number, however, it was assessed
228 with the *vSNP* pipeline (33). For all bioinformatics tools we used the default settings, unless
229 stated otherwise.
230 We checked if divergent sequences belonged to other mycobacteria species. We tested the
231 presence of RD (regions of difference) 1, 4, 9 and 12 patterns (34) in the outlier samples,
232 raw reads from each sample were aligned to *M. tuberculosis* (NC_000962.3) with *Burrows-*
233 *Wheeler Aligner* v0.7.17 (35), and sorted and indexed with *SAMtools* v1.10 (36). Primer
234 flanking regions for the RDs on *M. tuberculosis* were determined through querying the
235 sequences using NCBI web nucleotide BLAST with the default parameters (37), while the
236 presence of RDs were manually determined by examining the read alignment in *Integrative*
237 *Genomics Viewer* v2.14.1 (38).
238
239 ***2.3. Cameroonian M. bovis sequences in the African context***
240 We obtained other *M. bovis* genomes from online repositories: first, from the *Patric* (now
241 BV-BRC) dataset (39), and second, selecting the appropriate genomes among the ones listed
242 by Loiseau et al. (40) and obtained from the EBI dataset (for details and references, see
243 Table S2). We selected all the available sequences sampled in Africa, in order to qualitatively
244 detect potential genetic similarities between the sampled Cameroonian *M. bovis* population

and other isolates from the African continent, and thus provide a broader context to our analysis.

When analysing sequences from *Patric*, genomes were shredded into pseudo by *Snippy* followed by the process of alignment and SNP identification described above. The core-SNP alignments were made with and without the other African genomes. We used *iqtree* web server (41,42) to compute a phylogenetic tree (*n* = 212) which included all the Cameroonian high-quality sequences (*n* = 91) and the other African ones plus the 1997 reference from UK (*n* = 121).

2.4. Cameroonian sequences phylogenetic analysis

The quantitative analyses were performed on a subsample of the Cameroonian sequences, obtained after removing the non-cattle ones, the ones missing the geographical coordinates, and potential outliers, i.e., isolates not clustering within the main Cameroonian clade. We initially joined the remaining sequences (*n* = 64) tree using the TN93 genetic distance model and Neighbour-Joining (NJ) algorithm *ape* package (43) in *R* v4.0.5 (44) with the sole purpose of estimating a temporal signal within the sample in *TempEst* v1.5.3 (45). We then used the sequences SNP alignment completed with sampling dates, to infer time-scaled phylogenetic trees using *BEAST* v1.10.4 (20) with the *BEAGLE* library (46), and evaluated the results with *Tracer* v1.7.2 (47). Since the sequences had associated geographical location metadata, we included latitudes and longitudes as an additional continuous space variable for phylogeographic inference.

To select the best model, we ran a series of exploratory models using a HKY (48) substitution model, similar to other studies (49–51), and a strict molecular clock. We sequentially selected the best continuous trait model first, then the best bacterial population size model (tree prior). We tested the Brownian random walk, Cauchy Relaxed Random Walk (RRW), lognormal RRW and Gamma RRW for the former, and constant population, exponential growth and Bayesian *Skygrid* (52,53) for the latter. In the exploratory *BEAST* runs, we chose a truncated (between 0 and 0.1) normally distributed clock rate prior, with mean and standard deviation set as the slope in the root-to-tip obtained in *Tempest*; the chain length was set to 10^8, sampled every 10^4 steps. The models were compared using marginal likelihood estimation (MLE), with path sampling (PS) and stepping-stone sampling (SS), if they reached a satisfactory effective sample size (>200).

Once the model features were selected, we ran a final one setting the chain length to 10^9 steps, sampled every 10^5 steps. In this case, we used the clock rate posterior of the selected exploratory model as a prior for the final model. The maximum clade credibility (MCC) tree was extracted with *TreeAnnotator* v1.10.4 (part of the *BEAST* suite), and clades were visually defined within the MCC tree branches. The MCC tree was plotted against the sequences spoligotype and MIRU-VNTR typing to visually assess the correspondence between molecular typing and clades.

2.5. Spatial statistics and environmental factors analysis

From the final *BEAST* run, we extracted a set of 100 trees from the posterior distribution and further analysed using *seraphim* v1.0 (22,54) to obtain the spatial spread statistics: branch velocity and epidemic wavefront. The former was calculated for each branch dividing the geographical distance from the origin to the destination nodes by the time branch time duration. The epidemic wavefront shows the geographical range of the epidemic over time: at each time it is calculated as the geographical distance between the positions of the tree estimated root and the most distant node (spatial distance wavefront), or accounting for the distance of nodes closer to the root (patristic distance wavefront).

Additionally, *seraphim* allows to statistically test hypothesis on the effect of environmental layers on the epidemic dynamics; the effect can either be of "conductance", when the layer favours the pathogen diffusion, or "resistance", when it hampers it. We tested nine layers: elevation, cattle population density, human population density, two describing the roads infrastructure (number of intersections and total road length), and four land cover types (waterbodies, forest, grassland and grazeland, and other vegetation types: mosaic, shrub, sparse vegetation). The original raster layers were downloaded from online repositories (see Table S3 for the sources) and adapted to a 5km x 5km grid using *QGIS* v3.26.1. For each cell, elevation, cattle and human populations were averaged for the 5x5km grids, while roads intersections were counted, and roads length were measured starting from the same road original raster. For land cover, each value represents the percentage of that cell covered by each land cover type. The original land cover raster included 38 different cover types. To ease computation, we selected the most relevant for the study and merged them in four layers: waterbodies, forest, cropland/grassland, and other vegetation, including mosaic, shrub, and partial cover (Table S4).

First, we ran a preliminary analysis on each variable, to determine if it could have played a role as conductance or resistance in the pathogen spread. For each of the 100 extracted trees, we estimated the correlation between dispersal duration and environmental distance. Results are summarised by two statistics: the number of positive variable's coefficient of determination out of the 100 trees, and the number of positive *Q* statistic, calculated as $Q = R^2_{var} - R^2_{null}$, that is the difference between the correlation R^2 for the variable's raster and for a null raster, again calculated for each tree (54). For the analysis, we used two path models: straight line (where the branch "weight" is calculated as the by summing the cells values through which the straight-line passes), and least cost path (where the branch "weight" is calculated by summing the values between adjacent cells along the least-cost path).

Once we identified the potential resistance or conductance factor, we performed ten tree randomisations and calculated the statistics again. In this case, we used the Bayes Factor (BF_e), calculated as $BF_e = p_e /(1 - p_e)$, were p_e is the probability that $Q_{observed} > Q_{randomised}$. We used two criteria for trees randomisations: 1) randomisations of nodes positions while maintaining the branches lengths, the tree topology and the location of the most ancestral node; and 2) randomisations of nodes positions while maintaining only the branches lengths.

2.6. Genetic distance regression and role of the cattle movements

We finally tested which variables can better explain the genetic distances between the sampled *M. bovis* isolates, so to understand the signatures of temporal, spatial, and demographic factors (56,57). We ran this analysis using a Boosted Regression Trees (BRT) regression model (58) in *R* (packages *dismo*(59) and *gbm*(60)), a very flexible tool which combines decision trees and boosting techniques (61). In this model, the dependent variable was the genetic distance between *M. bovis* strains, expressed as SNPs. We tested a total of 28 relational variables, calculated for each pair of isolates (Table S5). Except for the temporal and spatial distance (which were calculated from the original isolates metadata), and for a binary variable indicating whether two sequences have the same spoligotype, MIRU-VNTR and clade (yes/no), the other variables are associated to the *M. bovis* isolates administrative subdivision.

We built two subdivision-level contact networks. The first one is a spatial network where nodes represent subdivisions and edges between them are positive if they share a border. This network is undirected (edges are not directional) and unweighted (all edges values are set to one). For this network we computed six variables to be associated with each pair of isolates: degree and betweenness centrality (62) of both isolates' subdivisions; shortest path and a binary variable indicating whether the two subdivisions belonged to the same network's community.

The second network represented the cattle movements, and edges correspond to the number of animals moved between subdivisions over a year. We built this network by aggregating the empirical data collected by Motta et al. (23), which originally reported the monthly number of cattle exchanged between markets. For this network we computed eight variables: degree, strength and betweenness centrality of both isolates' subdivisions; shortest path and the same community binary variable. The degree counts the number of each subdivision's connections, while the strength is the sum of the number of cattle moved to and from each subdivision. All networks' metrics were computed using the *R* package *igraph* (63).

Once we computed all the variables (the full list is reported in Table S5), we trained the BRT model using 75% of the observations, while the remaining 25% were used for testing. We evaluated the models based on pseudo-R^2 and Root Mean Squared Error (RMSE) on the test dataset. These were both calculated using the package *caret* (64). For BRT the relative influence of the variables is determined by the times each variable is selected to split the data in a decision tree, which in turn is weighted by the improvement in the model fit that resulted from that variable being used at each split (58). All models were fitted with a 10-fold cross validation. The BRT algorithm has two main parameters: the learning rate, which controls the contribution of each tree to the final model, and the tree complexity, which corresponds to the number of nodes in the tree. We ran some preliminary tests to tune the BRT in order to improve the predictions. Finally, we set the learning rate to 0.05 and the tree complexity to 8.

3. Results

3.1. Cameroonian sequences in the African context

We analysed 124 *M. bovis* sequences (nine of the original 133 failed), with 91 having enough read depth and genome coverage to allow further analyses (see Table S6 for further details). Two of these sequences came from isolates sampled humans, while for a third the sequencing failed. One of the excluded sequences was marked as not-*M. bovis*, and based on the presence of the four RD1, 4, 9 and 12 patterns (34), it was likely *M. tuberculosis*. All the high-quality *M. bovis* sequences were merged in a tree with other 22 obtained from the *Patric* dataset, 99 from EBI, and the 1997 UK Reference to provide a continental context. The qualitative phylogenetic tree in Figure 1 shows that most of the Cameroonian sequences (two of which obtained from human tissue samples) cluster with the Ghanaian human samples, and two Nigerians ones recovered from unreported hosts. All human samples from West Africa cluster with cattle sequences except for the Malian human sequence. Most sequences (*n* = 89) belonged to Af1 clonal complex and except one, the spoligotypes were already known; for the other, we identified a new pattern (hex code: 6F-1F-5F-7F-BF-40). Being characterised by the absence of spacer 30, this spoligotype was considered as Af1 (65). The dominant spoligotype was SB0944 (n = 32/89).

Two outlier sequences did not cluster with the rest of the sampled Cameroonian population. Their average distance from the rest of the Cameroonian population (respectively 235 and 231 SNPs) was slightly higher than the average distance of the 1997 UK reference from the Cameroonian isolates (222 SNPs), and they did not cluster with any other WGS sequence sampled in Africa (Figure 1). For both outlier sequences, the spoligotype was SB2332, found for the first time in Cameroon and submitted for classification at *www.Mbovis.org* by Egbe et al. (16). Following Warren et al. (34), we tested the presence of RD1, 4, 9 and 12 patterns, finding only the first one, confirming that they are likely *M. bovis*. We compared this spoligotype pattern with all the others from the *www.Mbovis.org* database, and we identified four patterns differing by two spacers: SB0858 sampled in Spain (66) (different spacers 20 and 22), SB1102 sampled in Chad (65) and Cameroon (13) (different spacers 33 and 34), SB2333 reported by Egbe et al. (16) (different spacers 22 and 34) and SB2691 sampled in France (not found in publications, different in spacers 20 and 34). We also identified eleven patterns differing by three spacers, sampled in France (67), Portugal (68), and Spain (66).

3.2. *M. bovis* evolutionary time scale in Cameroon

A total of 1'540 SNPs were determined from the *Snippy* core-SNP analysis on Cameroonian *M. bovis* genomes (Figure S2). This reduced to 1'106 SNPs when the dataset was reduced to the 64 samples with complete metadata and excluding the non-cattle ones (two sampled in humans), which were used for the downstream quantitative analysis. The median SNP distance among the remaining high-quality sequences was 70 SNPs (mean 68, range from 0 to 144, 2.5th and 97.5thquantiles 14 and 118). For two cattle (one from Bibemi, the other from Touboro), two *M. bovis* isolates sequenced were available (obtained from different tissues). In both cases, the two strains were identical (Bibemi 3 and 4, Touboro 7 and 8, Figure 2), which suggests a single infection disseminated in different organs, rather than two separate infections.

The analysis in *Tempest* showed a slightly positive temporal signal (coefficient of determination 0.11, and correlation coefficient 0.33) and a slope of 1.267 x 10^{-2} (Figure S3). We used a sequential approach in *BEAST* to select the best spatial model and bacterial population models. Based on the MLE estimation of the exploratory models (Table S7) we determined the best model included a Gamma Relaxed Random Walk (RRW) spatial model (first step of the sequential analysis) and the SkygGrid population model (second step). The final *BEAST* model was run with 10 bins and a cut-off of 400 years. The population trend is shown in Figure S4. The model estimates suggest the mean age of the root was in July 1950 (95th high-posterior density, HPD, April 1938 – August 1961), while the average clock rate was 1.32×10^{-7} substitution/site/year (95th HPD 1.20×10^{-7} – 1.44×10^{-7}). The maximum clade credibility (MCC) tree is reported in Figure 2, which also shows the division in four clades: clade 1 (green, 22 isolates), clade 2 (blue, 17 isolates), clade 3 (purple, 19 isolates) and clade 4 (red, 5 isolates). One sequence was excluded from all clades (Belel 4, Figure 2, reported as "no clade" in the figures). The geographical distribution of the clades is reported in Figure 3, showing the number of *M. bovis* isolates per administrative subdivision, which ranged from 1 to 17 (see Table S8 for the number of isolates per clade by regional abattoir). In Figure 4, we superimposed the MCC tree with spoligotypes; the most prevalent spoligotype, SB0944, occurred 26 times (out of 64 sequences) and was present in three of the four clades. The second most prevalent spoligotypes were SB0953 and SB2312, the first occurring five times in two clades, the latter occurring five times in one clade only (clade 2).

We also superimposed the MIRU-VNTR types as shown in Figure S5. The most prevalent MIRU-VNTR type in the sampled population was V89, which occurred nine times; V82 and V37 respectively occurred six and four times; and V81, V76 and V100 all occurred three times. Seven MIRU-VNTR types occurred twice, while 39 types occurred only once.

3.3. Spatio-temporal pathogen expansion

The estimated mean branch velocity was 53.1 km/year (95^{th} CI 18.4 – 219.0, temporal trend reported in Figure S7). The wavefront statistics in Figure 5 suggests that the pathogen expansion was slow until the mid 1960s, but accelerated thereafter to reach the entire study area, with a slow but constant expansion in the following period. This is reflected in an increase of the branch velocity at the same time (Figure S7), which is approximately the period when the branches formed the observed clades (Figure 2). The timing of the different branches in space is reported in Figure 6 (95^{th} HPD in Figure S8, with nodes coloured by estimated/observed date).

We tested the association between nine geographical variables with the dispersal duration. Table 1 shows the results obtained using the straight line and the least cost path models, the latter run considering the variables as potential conductance or resistance factor. Six variables resulted in a significant association (positive coefficients for all at least 95 out of 100 trees, and above 75% of positive *Q*): mosaic, shrub and other vegetation cover (with both path models, as resistance in the least cost one); forest cover, elevation and waterbodies cover (all as conductance); and cattle density (as resistance). However, when their statistical significance was tested through the randomisation, only forest cover and elevation (both as conductance) showed a Bayes Factor significant (≥ 3 (69)). The result was robust against two different trees randomisation algorithms for the forest layer, while for the elevation this was true only when maintaining only the branches length and excluding the other tree topological characteristics.

3.4. Factors associated with genetic distance

The RMSE of the boosted regression trees BRT model ran using all 28 variables was 20.23, while the R^2 was 0.450. We simplified the model using the *dismo* package, which tests the performance of the model by dropping the less important variables with a procedure similar to backward selection in regression (58). The algorithm brought to eliminating 12 variables

(see Table S5), nonetheless the model run using the remaining 16 variables performed very similarly to the original one (RMSE = 20.22 and R^2 = 0.452). Therefore, we used the latter to calculate the variable importance (Figure 7).

As expected, the most relevant variables were the temporal distance between the samples (1st) and the binary variable indicating whether the two *M. bovis* isolates belonged to the same clade in the MCC tree (2nd). The variables describing the subdivisions' population were also relevant in the model (population.y, 3rd, and population.x, 5th), as well as whether two isolates shared the same MIRU-VNTR (4th). This was more relevant than if two isolates shared the same spoligotypes (11th), suggesting the former as more useful to discriminate closer *M. bovis* strains. The markets movement network strength (i.e. the number of cattle moved from/to a subdivision) was the most important (6th and 9th) among network-related variables, while the betweenness (8th and 10th) was the only spatial network variable retained in the simplified model. Interestingly, when both variables were selected for the same metric, the one related to the youngest isolate (marked by *y*) was always preferred to the one related to the oldest isolate (marked by *x*). The partial dependency plots, showing the relationship between SNP distance and variables, are reported in Figure S9.

4. Discussion

We sought to unravel the characteristics of the spread of a pathogen with zoonotic potential in time and space to improve our understanding and inform control and preparedness strategies. Our basic premise is that the accumulation of mutations in the pathogen's genome can be used as signatures of transmission events from host to host across time and space. Within space, the environment can create barriers which influence the population dynamics of diseases, i.e., altering host-to-host and pathogen-host interactions has direct effects on the genetic structure of the pathogen (70). The availability of high-throughput genomic techniques means we can interrogate the structural changes linked to the environment over time to gain critical insights into how the epidemic has evolved. In this study, we aimed to characterise *M. bovis* sampled from cattle in Cameroon using genetic and demographic data to understand whether the pathogen is in a stable endemic state and the influence on the spread dynamic of environmental and ecological factors and cattle movements.

4.1. Evidence of dynamic endemicity

An important question was whether the *M. bovis* outbreak in North Cameroon was in a steady state, at an endemic equilibrium, or if it was expanding. Determining whether a pathogen is endemic has implications on risk perception and, consequently, on resource allocation. At the same time, the chances of zoonotic transmission are likely to be higher in the case of endemicity. In our analysis, the Bayesian model estimation with SkyGrid as a population model showed an increasing pathogen effective population size, corresponding to a constant increase in the disease velocity after the sudden jump during the mid-to-late 1960s. This suggests that the pathogen is not in a state of endemic stability, instead it has been expanding at various rates over the years. This is in agreement with a previous publication using spoligotypes and MIRU-VNTR (16) and with the work by Awah-Ndukum et al. (15). The expansion of *M. bovis* might represent an issue for livestock and humans, particularly as we showed that the bacterium is circulating in both. At the moment, disease control in the area is absent, while on the other hand, the dairy industry in Africa is generally expanding. A lack of widespread milk pasteurization could lead to an increase in zoonotic TB cases, which already represent a problematic issue in the region (12).

4.2. Genetic diversity of M. bovis in Cameroon

We observed a high diversity of *M. bovis*, confirming earlier observations with molecular typing techniques providing less granular information (16), considering the short time span of the sampling campaign and the small sample size. This contrasts with areas such as Great Britain and other European countries, where strict control measures, such as routine testing and stamping out of positive individuals, have been in place for decades. This can act as a bottleneck with a consequent reduction in the pathogen's genetic variability by reducing the time a pathogen has to develop inside a domestic host, therefore, the likelihood of substitutions in the DNA. As an example, Crispell et al. (57) reported a similar SNP distance range (0 to 150), albeit across a much bigger sample (*n* = 230), with a lower median (20 SNPs) and with isolates dating back two decades, while in a similar size monophyletic outbreak (*n* = 64), Rossi et al. reported a maximum SNP distance of only 6 SNPs (56). In Spain, Pozo et al. (71) found a similar SNP distance average and range (62, and 0 to 150) in a bigger *M. bovis* population, sampled in both cattle and wildlife over 13 years. It is noteworthy that high diversity can be associated with dynamic epidemiology and not with endemic stability.

All 64 core isolates belonged to the clonal complex Af1, which was observed in the region in previous studies (65). The most common spoligotype, SB0944, was found by Müller et al. (65) as the most prevalent in West Africa and considered as the original of the Af1 clonal complex. Our findings also suggest zoonotic transmission in West Africa, as sequences recovered from humans in Cameroon and Ghana clustered with Cameroonian cattle *M. bovis* isolates (72). Because it is known that zoonotic TB represents a minoritarian but still crucial part of all TB cases in Africa, these results strengthen the case for One Health approaches to control, that involve humans, livestock, wildlife and environmental health (12,73). Except for the one sequence in Mali and the two Cameroonian outliers, all the sequences from West Africa clustered together, hinting to a high connectivity likely caused by cattle movements throughout the area, as previously showed by another study (74). Our results showed that the areas with the highest *M. bovis* diversity were in the Adamawa and North regions, both reporting all the clades identified by the maximum clade credibility (MCC) tree. All clades were also sampled in the towns of Touboro and Tchollire, both located in the North region but close to the Adamawa border. Previous studies reported that this area receive cattle from neighbouring country as part of the transhumance

546 migration, suggesting that cattle movements and markets play an important role in defining
547 the dynamics of the pathogen, and therefore influencing its genetic diversity (16,23). The
548 Northwest region was underrepresented in the sample, with only five high-quality
549 sequences on 31 infected cattle detected at the abattoir. This inherently reduces the level of
550 diversity, which is far lower than reported using spoligotypes and MIRU-VNTR (16).
551 Despite covering a smaller portion of the genome and the higher occurrence of homoplasy
552 with respect to WGS, in other contexts spoligotypes have been used as a proxy cluster, or to
553 narrow down potential transmission within the study population (57,75). Our results
554 showed that this cannot be done for areas with high diversity such as the one we
555 considered, as we observed little correspondence between the MCC tree branches and the
556 spoligotypes. Similarly, other studies pointed out the limitations of such typing techniques
557 (19), in case of an expanding infection where transmission is steadily ongoing, compared to
558 point-source ones (76). The high SNP distances among the sampled isolates also precluded
559 the use of methods to infer direct transmission between hosts (8,77).
560 When considering the entire sampled population, therefore including the sequences with
561 incomplete metadata, we found two of the 91 sequences not belonging to the clonal
562 complex Af1. In their spoligotype pattern (SB2332), we noted the absence of spacer 21 (78),
563 and the closest relatives analysed by Loiseau et al. (40) were identified as part of the clonal
564 complex Eu2, including isolates sampled both in South-western Europe (SB0837, SB1090,
565 SB1308) and West Africa (SB1102, isolated in Cameroon as well (13)). We can then
566 speculate that these sequences likely belong to Eu2 as well, although we could not exclude
567 one of the “unknown” clonal complexes identified by other studies (40,79). Further
568 development on this point was beyond the scope of this study, as we focused on the 64 core
569 sequences to gather insights on the pathogen dynamics in the area.
570
571 ***4.3. Tracking the spread of M. bovis in Cameroon***
572 We acknowledge that our estimates for the most recent common ancestor (MRCA) have a
573 wide credible interval around it (23 years). This uncertainty is likely due to the short
574 duration of the samples collection campaign, which also generated a weak temporal signal,
575 although the coefficient of determination was similar to other *M. bovis* studies in highly
576 sampled populations (56,57). Nonetheless, our estimates coalesce around 1950, suggesting
577 that the pathogen has been spreading in the area for at least six decades at the time of

578 sampling. For the same reason, the estimated clock rate was higher than others in the
579 literature but in the same order of magnitude (0.67-1.26 $\times$ 10^{-7}, *n* = 2625 (40)).
580 The estimated MCC tree located the most recent common ancestor (MRCA) in Touboro
581 (North region) and, from there, a rapid expansion of the outbreak reaching most of the
582 study area by the early 1970s. From the estimated origin, the pathogen likely spread first
583 northward to Garoua (in the same region) and westward, to the Nort West region, and later
584 to the Extreme North and Adamawa regions and again to the Northwest.
585 The results of the spatial factors analysis showed that forest cover and elevation were the
586 only significant ones, both acting as "conductance". Forest cover could be a proxy for
587 potential wildlife interactions, as *M. bovis* is known to be quite effective in spreading at the
588 wildlife-livestock (and humans) interface (73,80). The elevation as conductance was
589 counter-intuitive, however, this could be linked to cattle movements in pastoralist
590 communities within the plateau located in the study area. This is important because, if
591 confirmed, altitude could be used as a proxy for the missing pastoralist movements.
592 Our regression model performed reasonably well, although the amount of variability
593 explained was below 50%. However, our objective was to understand which variables could
594 better explain the genetic distance between *M. bovis* isolates, expressed as SNP distance.
595 Except for the between isolates temporal distance and clade, the demographic variables
596 were the most effective in explaining SNP distance, particularly the administrative
597 subdivision human population size. These variables had a negative effect on the SNP
598 distance, meaning that smaller population was associated to a close relatedness of the *M.*
599 *bovis* strains. This could be an effect of the population distribution in the country because
600 the northern regions, where cattle are most concentrated, are less populated compared to
601 the cities in the south. The simplified model performed similarly to the full model,
602 suggesting some variables were not important in explaining the genetic distance. Beyond
603 the human population size, also the other demographic variables (population and cattle
604 density) were all retained. Conversely, only five network related variables were retained,
605 three for the cattle movement network (out of eight) and two for the spatial network (out
606 of six). All network related variables had a positive effect on the SNP distance, with the
607 number of cattle moved in or out a subdivision (i.e., strength) having the higher predictive
608 effect. Interestingly, this result was similar to other studies where cattle movements alone
609 could not fully capture *M. bovis* genetic diversity (56,57).

610

611 **4.4. Limitations**

612 The major limitation of this dataset was the short data collection time window, less than a
613 year and a half, which resulted in uncertainty in the MRCA estimate and a weak temporal
614 signal. While we can speculate the sampled bacterial population already reached the entire
615 study area before the 1970s, a wider sampling time window would likely allow a stronger
616 temporal signal and improve our estimate of the MRCA, which might be prior with respect
617 to the current estimate. In turn, this affected the pathogen's expansion patterns, including
618 the branch velocity and wavefront, which are also limited by the sampled area size. The
619 spatial uncertainty might also be affected by the absence of dense cattle movements
620 records, so the known spatial coordinates associated with each sequence correspond to the
621 last village the animal lived in. The Adamawa and Northwest regions are home to 1.25
622 million and 450.000 cattle respectively (23), and while this abattoir-based study provides a
623 very informative snapshot of the *M. bovis* population in North Cameroon, it adds to the calls
624 to improve cattle records and movements routine data collections in LMICs (81), as well as
625 bTB detection efforts.
626 The low-quality WGSs disproportionally affected the Northwest region, as reported in Table
627 S7. This could have hampered the representativeness of the *M. bovis* diversity in that
628 region, reducing the number of clades observed. The Adamawa region was the most
629 represented, despite most of the sequences excluded from the quantitative analysis
630 because of missing coordinates, came from the Ngaoundere abattoir. The bacterium
631 diversity in the Northwest might also be affected by the demographic of the slaughtered
632 cattle in the region (14): because the region is highly populated by humans and more
633 isolated in the trade network (23), local animals of both sexes and at any age are
634 slaughtered. Conversely, young male calves from the Adamawa, North and Extreme North
635 regions are often sent to richer southern regions to maximise their economic values, leaving
636 the older cows to be slaughtered. By being exposed to the *M. bovis* for longer, the latter
637 have more chances to develop lesions. On the other hand, these trends likely reduce the
638 impact of missing information on the previous location of the animals, because these
639 animals have more chances of being reared locally.
640 In agreement with many studies, and with the *vSNP* analysis result, we used AF2122/97 as
641 reference genome (50,51,56,57,79,82,83). In order to account for genes, absent in *M. bovis*,

Loiseau et al. (40) used *M. tuberculosis* H37Rv, a choice driven by the different purpose of their work compared to ours (define the origin and the global population structure of *M. bovis*). Generally, the pipelines used to call the SNPs differed in many of the aforementioned studies, contributing to the estimates uncertainty and potentially generating biases the analysis results and the clock rate calculations.

5. Conclusion

In conclusion, our study indicates endemic stability of *M. bovis* is unlikely in North Cameroon, but rather the disease is slowly expanding over time. Our findings highlight the importance of collecting data in underrepresented areas to enrich insights in the current body of literature, predominantly from developed countries. Moreover, our results pave the way for future research aimed to understand whether the observed *M. bovis* high genetic diversity affects the spread dynamics.

Our findings underscore the need to adopt a one-health surveillance strategy for *M. bovis* control (12). More work on combining tools such as phylogeography, statistical modelling, landscape and ecology will be beneficial to map spread patterns and effectively inform control and preparedness strategies (56).

660 **References**

661 1. Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious
662 disease. *Nat Rev Genet* (2009) 10:540–550. doi: 10.1038/nrg2583
663 2. Kao RR, Haydon DT, Lycett SJ, Murcia PR. Supersize me: How whole-genome
664 sequencing and big data are transforming epidemiology. *Trends Microbiol* (2014)
665 22:282–291. doi: 10.1016/j.tim.2014.02.011
666 3. Pybus OG, Suchard MA, Lemey P, Bernardin FJ, Rambaut A, Crawford FW, Gray RR,
667 Arinaminpathy N, Stramer SL, Busch MP, et al. Unifying the spatial epidemiology and
668 molecular evolution of emerging epidemics. *Proc Natl Acad Sci* (2012) 109:15066–
669 15071. doi: 10.1073/pnas.1206598109
670 4. Campbell F, Strang C, Ferguson N, Cori A, Jombart T. When are pathogen genome
671 sequences informative of transmission events? *PLoS Pathog* (2018) 14:1–21. doi:
672 10.1371/journal.ppat.1006885
673 5. Patané JSL, Martins J, Castelão AB, Nishibe C, Montera L, Bigi F, Zumárraga MJ, Cataldi
674 AA, Junior AF, Roxo E, et al. Patterns and Processes of *Mycobacterium bovis* Evolution
675 Revealed by Phylogenomic Analyses. *Genome Biol Evol* (2017) 9:521–535. doi:
676 10.1093/gbe/evx022
677 6. Campbell F, Strang C, Ferguson N, Cori A, Jombart T. When are pathogen genome
678 sequences informative of transmission events? *PLoS Pathog* (2018) 14:1–21. doi:
679 10.1371/journal.ppat.1006885
680 7. WHO. Zoonotic tuberculosis factsheet. (2017) http://www.who.int/tb/areas-of-
681 work/zoonotic-tb/en [Accessed October 26, 2022]
682 8. Rossi G, Crispell J, Balaz D, Lycett S, Delahay R, Kao R. Identifying likely transmission
683 pairs with pathogen sequence data using Kolmogorov Forward Equations; an
684 application to *M. bovis* in cattle and badgers. *Sci Rep* (2020)1–13. doi:
685 10.1101/2020.06.11.146894
686 9. Bernitz N, Kerr TJ, Goosen WJ, Chileshe J, Higgitt RL, Roos EO, Meiring C, Gumbo R, de
687 Waal C, Clarke C, et al. Review of Diagnostic Tests for Detection of *Mycobacterium*
688 *bovis* Infection in South African Wildlife. *Front Vet Sci* (2021) 8:1–11. doi:
689 10.3389/fvets.2021.588697
690 10. Kelly RF, Gordon LG, Egbe NF, Freeman EJ, Mazeri S, Ngwa VN, Tanya V, Sander M,
691 Ndip L, Muwonge A, et al. Bovine Tuberculosis Antemortem Diagnostic Test

692 Agreement and Disagreement in a Naturally Infected African Cattle Population.
693 (2022) 9: doi: 10.3389/fvets.2022.877534
694 11. Sichewo PR, Vander Kelen C, Thys S, Michel AL. Risk practices for bovine tuberculosis
695 transmission to cattle and livestock farming communities living at wildlife-livestock-
696 human interface in northern Kwazulu Natal, South Sfrica. *PLoS Negl Trop Dis* (2020)
697 14:1–18. doi: 10.1371/journal.pntd.0007618
698 12. Olea-Popelka F, Muwonge A, Perera A, Dean AS, Mumford E, Erlacher-Vindel E,
699 Forcella S, Silk BJ, Ditiu L, El Idrissi A, et al. Zoonotic tuberculosis in human beings
700 caused by *Mycobacterium bovis*—a call for action. *Lancet Infect Dis* (2017) 17:e21–
701 e25. doi: 10.1016/S1473-3099(16)30139-6
702 13. Koro Koro F, Ngatchou AF, Portal JL, Gutierrez C, Etoa FX, Eyangoh SI. The genetic
703 population structure of *Mycobacterium bovis* strains isolated from cattle slaughtered
704 at the Yaoundé and Douala abattoirs in Cameroon. *OIE Rev Sci Tech* (2015) 34:1001–
705 1010. doi: 10.20506/rst.34.3.2390
706 14. Egbe NF, Muwonge A, Ndip L, Kelly RF, Sander M, Tanya V, Ngwa VN, Handel IG,
707 Novak A, Ngandalo R, et al. Abattoir-based estimates of mycobacterial infections in
708 Cameroon. *Sci Rep* (2016) 6:1–14. doi: 10.1038/srep24320
709 15. Awah-Ndukum J, Kudi AC, Bradley G, Ane-Anyangwe I, Titanji VPK, Fon-Tebug S,
710 Tchoumboue J. Prevalence of bovine tuberculosis in cattle in the highlands of
711 Cameroon based on the detection of lesions in slaughtered cattle and tuberculin skin
712 tests of live cattle. *Vet Med (Praha)* (2012) 57:59–76. doi: 10.17221/5252-VETMED
713 16. Egbe NF, Muwonge A, Ndip L, Kelly RF, Sander M, Tanya V, Ngwa VN, Handel IG,
714 Novak A, Ngandalo R, et al. Molecular epidemiology of *Mycobacterium bovis* in
715 Cameroon. *Sci Rep* (2017) 7:1–17. doi: 10.1038/s41598-017-04230-6
716 17. Kamerbeek J, Schouls L, Kolk A, Van Agterveld M, Van Soolingen D, Kuijper S,
717 Bunschoten A, Molhuizen H, Shaw R, Goyal M, et al. Simultaneous detection and
718 strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J*
719 *Clin Microbiol* (1997) 35:907–914. doi: 10.1128/jcm.35.4.907-914.1997
720 18. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, Savine
721 E, De Haas P, Van Deutekom H, Roring S, et al. Proposal for standardization of
722 optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat
723 typing of *Mycobacterium tuberculosis*. *J Clin Microbiol* (2006) 44:4498–4510. doi:

724 10.1128/JCM.01392-06
725 19. Reis AC, Salvador LCM, Robbe-Austerman S, Tenreiro R, Botelho A, Albuquerque T,
726 Cunha M V. Article whole genome sequencing refines knowledge on the population
727 structure of *Mycobacterium bovis* from a multi-host tuberculosis system.
728 *Microorganisms* (2021) 9: doi: 10.3390/microorganisms9081585
729 20. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian
730 phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* (2018)
731 4:1–5. doi: 10.1093/ve/vey016
732 21. Lemey P, Rambaut A, Welch JJ, Suchard MA. Phylogeography takes a relaxed random
733 walk in continuous space and time. *Mol Biol Evol* (2010) 27:1877–1885. doi:
734 10.1093/molbev/msq067
735 22. Dellicour S, Rose R, Faria NR, Lemey P, Pybus OG. SERAPHIM: Studying environmental
736 rasters and phylogenetically informed movements. *Bioinformatics* (2016) 32:3204–
737 3206. doi: 10.1093/bioinformatics/btw384
738 23. Motta P, Porphyre T, Handel I, Hamman SM, Ngu Ngwa V, Tanya V, Morgan K,
739 Christley R, Bronsvoort BMD. Implications of the cattle trade network in Cameroon
740 for regional disease prevention and control. *Sci Rep* (2017) 7:1–13. doi:
741 10.1038/srep43932
742 24. MMM O. BovTB-nf-docker. (2022) https://github.com/oxfordmmm/BovTB-nf-docker
743 [Accessed October 31, 2022]
744 25. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina
745 sequence data. *Bioinformatics* (2014) 30:2114–2120. doi:
746 10.1093/bioinformatics/btu170
747 26. Md V, Misra S, Li H, Aluru S. Efficient architecture-aware acceleration of BWA-MEM
748 for multicore systems. *Proceedings - 2019 IEEE 33rd International Parallel and*
749 *Distributed Processing Symposium, IPDPS 2019*. IEEE (2019). p. 314–324 doi:
750 10.1109/IPDPS.2019.00041
751 27. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
752 Durbin R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* (2009)
753 25:2078–2079. doi: 10.1093/bioinformatics/btp352
754 28. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome*
755 *Biol* (2019) 20:1–13. doi: 10.1186/s13059-019-1891-0

29. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: Estimating species abundance in metagenomics data. *PeerJ Comput Sci* (2017) 2017:1–17. doi: 10.7717/peerj-cs.104
30. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* (2011) 27:2987–2993. doi: 10.1093/bioinformatics/btr509
31. Seemann T. snippy: Rapid haploid variant calling and core genome alignment. (2022) https://github.com/tseemann/snippy [Accessed October 31, 2022]
32. Jiang Y, Jiang Y, Wang S, Zhang Q, Ding X. Optimal sequencing depth design for whole genome re-sequencing in pigs. *BMC Bioinformatics* (2019) 20:1–12. doi: 10.1186/s12859-019-3164-z
33. USDA. vSNP. https://github.com/USDA-VS/vSNP [Accessed January 16, 2023]
34. Warren RM, Pittius NCG Van, Barnard M, Hesseling A, Engelke E, Kock M De, Gutierrez MC, Chege GK, Victor TC, Hoal EG, et al. Differentiation of *Mycobacterium tuberculosis* complex by PCR amplification of genomic regions of difference. *Int J Tuberc Lung Dis* (2006) 10:818–822.
35. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* (2009) 25:1754–1760. doi: 10.1093/bioinformatics/btp324
36. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. Twelve years of SAMtools and BCFtools. *Gigascience* (2021) 10:1–4. doi: 10.1093/gigascience/giab008
37. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* (1990) 215:403–410. doi: 10.1016/S0022-2836(05)80360-2
38. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nat Biotechnol* (2011) 29:24–26. doi: 10.1038/nbt.1754
39. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, Conrad N, Dietrich EM, Disz T, Gabbard JL, et al. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res* (2017) 45:D535–D542. doi: 10.1093/nar/gkw1017
40. Loiseau C, Menardo F, Aseffa A, Hailu E, Gumi B, Ameni G, Berg S, Rigouts L, Robbe-Austerman S, Zinsstag J, et al. An African origin for *Mycobacterium bovis*. *Evol Med Public Heal* (2020) 2020:49–59. doi: 10.1093/EMPH/EOAA005

41. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R, Teeling E. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* (2020) 37:1530–1534. doi: 10.1093/molbev/msaa015
42. Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res* (2016) 44:W232–W235. doi: 10.1093/NAR/GKW256
43. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in {R}. *Bioinformatics* (2018) 35:526–528.
44. R Core Team. R: A Language and Environment for Statistical Computing. (2021) https://www.r-project.org/
45. Rambaut A, Lam TT, Carvalho LM, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* (2016) 2:1–7. doi: 10.1093/ve/vew007
46. Ayres DL, Darling A, Zwickl DJ, Beerli P, Holder MT, Lewis PO, Huelsenbeck JP, Ronquist F, Swofford DL, Cummings MP, et al. BEAGLE: An application programming interface and high-performance computing library for statistical phylogenetics. *Syst Biol* (2012) 61:170–173. doi: 10.1093/sysbio/syr100
47. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst Biol* (2018) 67:901–904. doi: 10.1093/sysbio/syy032
48. Hasegawa M, Kishino H, Yano T aki. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* (1985) 22:160–174. doi: 10.1007/BF02101694
49. Crispell J, Zadoks RN, Harris SR, Paterson B, Collins DM, De-Lisle GW, Livingstone P, Neill MA, Biek R, Lycett SJ, et al. Using whole genome sequencing to investigate transmission in a multi-host system: Bovine tuberculosis in New Zealand. *BMC Genomics* (2017) 18:180. doi: 10.1186/s12864-017-3569-x
50. Duault H, Michelet L, Boschiroli M-L, Durand B, Canini L. A Bayesian evolutionary model towards understanding wildlife contribution to F4-family *Mycobacterium bovis* transmission in the South-West of France. *Vet Res* (2022) 53:1–12. doi: 10.1186/s13567-022-01044-x
51. Salvador LCM, O'Brien DJ, Cosgrove MK, Stuber TP, Schooley A, Crispell J, Church S,

Grohn YT, Robbe-Austerman S, Kao RR. 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)1–14. doi: 10.1111/mec.15061

52. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. Improving bayesian population dynamics inference: A coalescent-based model for multiple loci. *Mol Biol Evol* (2013) 30:713–724. doi: 10.1093/molbev/mss265

53. Hill V, Baele G. Bayesian Estimation of Past Population Dynamics in BEAST 1.10 Using the Skygrid Coalescent Model. *Mol Biol Evol* (2019) 36:2620–2628. doi: 10.1093/molbev/msz172

54. Dellicour S, Rose R, Pybus OG. Explaining the geographic spread of emerging epidemics: A framework for comparing viral phylogenies and environmental landscape data. *BMC Bioinformatics* (2016) 17:1–12. doi: 10.1186/s12859-016-0924-x

55. McRae BH. Isolation By Resistance. *Evolution (N Y)* (2006) 60:1551. doi: 10.1554/05-321.1

56. Rossi G, Crispell J, Brough T, Lycett SJ, White PCL, Allen A, Ellis RJ, Gordon S V., Harwood R, Palkopoulou E, et al. Phylodynamic analysis of an emergent *Mycobacterium bovis* outbreak in an area with no previously known wildlife infections. *J Appl Ecol* (2022) 59:210–222. doi: 10.1111/1365-2664.14046

57. Crispell J, Benton CHCH, Balaz D, De Maio N, Ahkmetova A, Allen A, Biek R, Presho ELEL, Dale J, Hewinson G, et al. Combining genomics and epidemiology to analyse bi-directional transmission of *Mycobacterium bovis* in a multi-host system. *Elife* (2019) 8:1–36. doi: https://doi.org/10.7554/eLife.45833.001

58. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. *J Anim Ecol* (2008) 77:802–813. doi: 10.1111/j.1365-2656.2008.01390.x

59. Hijmans RJ, Phillips S, Leathwick J, Elith J. dismo: Species Distribution Modeling. (2021) https://cran.r-project.org/package=dismo

60. Greenwell B, Boehmke B, Cunningham J, Developers GBM. gbm: Generalized Boosted Regression Models. (2022) https://cran.r-project.org/package=gbm

61. Brock PM, Fornace KM, Grigg MJ, Anstey NM, William T, Cox J, Drakeley CJ, Ferguson HM, Kao RR. Predictive analysis across spatial scales links zoonotic malaria to deforestation. *Proc R Soc B Biol Sci* (2019) 286: doi: 10.1098/rspb.2018.2351

62. Newman M. *Networks: An Introduction*. New York, NY, USA: Oxford University Press,

852 Inc. (2010).
853 63. Csárdi G, Nepusz T. The igraph software package for complex network research. *J*
854 *Comput Appl* (2014) Complex Sy:9. doi: 10.3724/SP.J.1087.2009.02191
855 64. Kuhn M, Weston S, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel B, Team RC,
856 Benesty M, Lescarbeau R, et al. Classification and Regression Training. (2016)198.
857 https://cran.r-project.org/package=caret
858 65. Müller B, Hilty M, Berg S, Garcia-Pelayo MC, Dale J, Boschiroli ML, Cadmus S,
859 Ngandolo BNR, Godreuil S, Diguimbaye-Djaibé C, et al. African 1, an epidemiologically
860 important clonal complex of *Mycobacterium bovis* dominant in Mali, Nigeria,
861 Cameroon, and Chad. *J Bacteriol* (2009) 191:1951–1960. doi: 10.1128/JB.01590-08
862 66. Rodríguez S, Romero B, Bezos J, de Juan L, Álvarez J, Castellanos E, Moya N, Lozano F,
863 González S, Sáez-Llorente JL, et al. High spoligotype diversity within a *Mycobacterium*
864 *bovis* population: Clues to understanding the demography of the pathogen in Europe.
865 *Vet Microbiol* (2010) 141:89–95. doi: 10.1016/j.vetmic.2009.08.007
866 67. Hauer A, De Cruz K, Cochard T, Godreuil S, Karoui C, Henault S, Bulach T, Bañuls AL,
867 Biet F, Boschiroli ML. Genetic evolution of *Mycobacterium bovis* causing tuberculosis
868 in livestock and wildlife in France since 1978. *PLoS One* (2015) 10:1–17. doi:
869 10.1371/journal.pone.0117103
870 68. Matos F, Cunha M V., Canto A, Albuquerque T, Amado A, Botelho A. Snapshot of
871 *Mycobacterium bovis* and *Mycobacterium caprae* infections in livestock in an area
872 with a low incidence of bovine tuberculosis. *J Clin Microbiol* (2010) 48:4337–4339.
873 doi: 10.1128/JCM.01762-10
874 69. Kass RE, Raftery AE. Bayes Factors. *J Am Stat Assoc* (1995) 90:773–795. doi:
875 10.1080/01621459.1995.10476572
876 70. Real LA, Biek R. Spatial dynamics and genetics of infectious diseases on
877 heterogeneous landscapes. *J R Soc Interface* (2007) 4:935–948. doi:
878 10.1098/rsif.2007.1041
879 71. Pozo P, Lorente-Leal V, Robbe-Austerman S, Hicks J, Stuber T, Bezos J, de Juan L, Saez
880 JL, Romero B, Alvarez J. Use of Whole-Genome Sequencing to Unravel the Genetic
881 Diversity of a Prevalent *Mycobacterium bovis* Spoligotype in a Multi-Host Scenario in
882 Spain. *Front Microbiol* (2022) 13: doi: 10.3389/fmicb.2022.915843
883 72. Otchere ID, van Tonder AJ, Asante-Poku A, Sánchez-Busó L, Coscollá M, Osei-Wusu S,

884 Asare P, Aboagye SY, Ekuban SA, Yahayah AI, et al. Molecular epidemiology and
885 whole genome sequencing analysis of clinical *Mycobacterium bovis* from Ghana. *PLoS*
886 *One* (2019) 14:1–13. doi: 10.1371/journal.pone.0209395
887 73. Mohamed A. Bovine tuberculosis at the human–livestock–wildlife interface and its
888 control through one health approach in the Ethiopian Somali Pastoralists: A review.
889 *One Heal* (2020) 9:100113. doi: 10.1016/j.onehlt.2019.100113
890 74. Valerio VC, Walther OJ, Eilittä M, Cissé B, Muneepeerakul R, Kiker GA. Network
891 analysis of regional livestock trade in West Africa. *PLoS One* (2020) 15:1–20. doi:
892 10.1371/journal.pone.0232681
893 75. Trewby H, Wright DM, Skuce RA, McCormick C, Mallon TR, Presho EL, Kao RR, Haydon
894 DT, Biek R. Relative abundance of *Mycobacterium bovis* molecular types in cattle: a
895 simulation study of potential epidemiological drivers. *BMC Vet Res* (2017) 13:268. doi:
896 10.1186/s12917-017-1190-5
897 76. Rodriguez-Campos S, Aranaz A, De Juan L, Sáez-Llorente JL, Romero B, Bezos J,
898 Jiménez A, Mateos A, Domínguez L. Limitations of spoligotyping and variable-number
899 tandem-repeat typing for molecular tracing of *Mycobacterium bovis* in a high-
900 diversity setting. *J Clin Microbiol* (2011) 49:3361–3364. doi: 10.1128/JCM.00301-11
901 77. Campbell F, Cori A, Ferguson N, Jombart T. Bayesian inference of transmission chains
902 using timing of symptoms, pathogen genomes and contact data. *PLoS Comput Biol*
903 (2019) 15:1–20. doi: 10.1371/journal.pcbi.1006930
904 78. Rodriguez-Campos S, Schürch AC, Dale J, Lohan AJ, Cunha M V., Botelho A, Cruz K De,
905 Boschiroli ML, Boniotti MB, Pacciarini M, et al. European 2 - A clonal complex of
906 *Mycobacterium bovis* dominant in the Iberian Peninsula. *Infect Genet Evol* (2012)
907 12:866–872. doi: 10.1016/j.meegid.2011.09.004
908 79. Zimpel CK, Patané JSL, Guedes ACP, de Souza RF, Silva-Pereira TT, Camargo NCS, de
909 Souza Filho AF, Ikuta CY, Neto JSF, Setubal JC, et al. Global Distribution and Evolution
910 of *Mycobacterium bovis* Lineages. *Front Microbiol* (2020) 11:1–19. doi:
911 10.3389/fmicb.2020.00843
912 80. Renwick AR, White PCL, Bengis RG. Bovine tuberculosis in southern African wildlife: A
913 multi-species host-pathogen system. *Epidemiol Infect* (2007) 135:529–540. doi:
914 10.1017/S0950268806007205
915 81. Chaters GL, Johnson PCD, Cleaveland S, Crispell J, De Glanville WA, Doherty T,

Matthews L, Mohr S, Nyasebwa OM, Rossi G, et al. Analysing livestock network data for infectious disease control: An argument for routine data collection in emerging economies. *Philos Trans R Soc B Biol Sci* (2019) 374: doi: 10.1098/rstb.2018.0264

82. Reis AC, Cunha M V. The open pan-genome architecture and virulence landscape of *Mycobacterium bovis*. *Microb Genomics* (2021) 7: doi: 10.1099/MGEN.0.000664

83. Crispell J, Cassidy S, Kenny K, McGrath G, Warde S, Cameron H, Rossi G, Macwhite T, White PCL, Lycett S, et al. *Mycobacterium bovis* genomics reveals transmission of infection between cattle and deer in Ireland. *Microb Genomics* (2020) 6:1–8. doi: 10.1099/mgen.0.000388

Figures

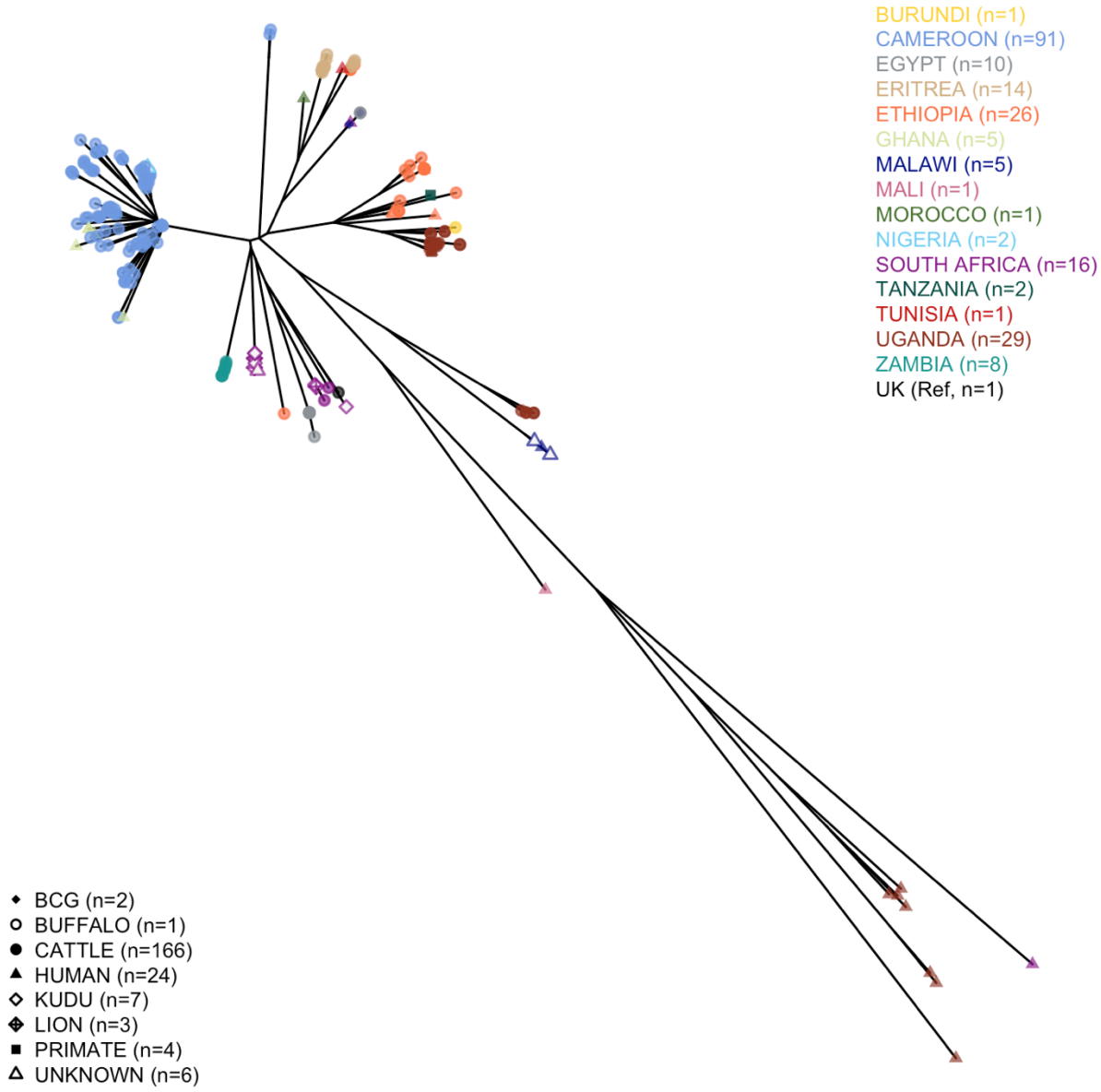


Figure 1: Phylogenetic tree of the African *Mycobacterium bovis* whole-genome sequences considered in the study. The tree includes 91 high-quality Cameroonian sequences, 101 from the EBI dataset, 20 from Patric and the 1997 UK *M. bovis* reference.

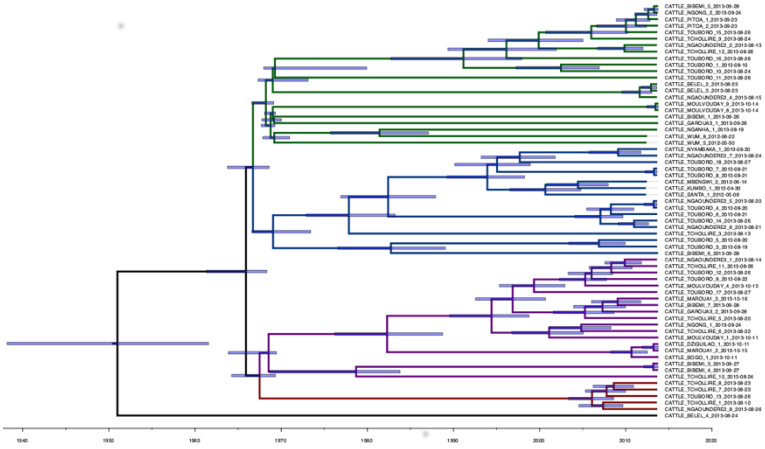


Figure 2: Phylogenetic time scaled MCC tree of the 64 high-quality *M. bovis* whole-genome sequences sampled in Cameroon in 2012 and 2013. The thin lines represent the 95th HPD of the internal node dates, while the branch colours represent different clades: 1 (green), 2 (blue), 3 (purple) and 4 (red). A non-time scaled tree showing the genetic distance between the 64 sequences is reported in Figure S6.

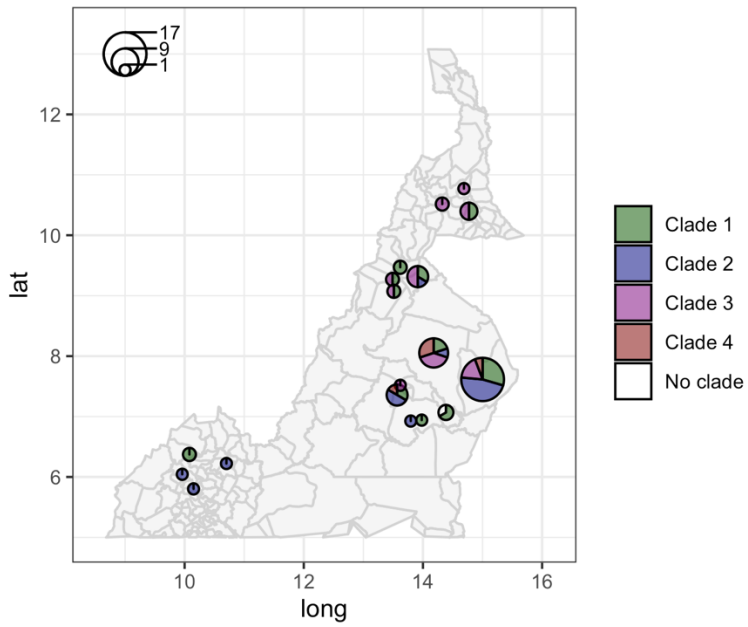


Figure 3: Geographic distribution of the 64 high-quality *M. bovis* whole-genome sequences in Cameroon. Circle sizes correspond to the number of sequences per administrative subdivision, and colours represent different clades (clade 1 green, clade 2 blue, clade 3 purple and clade 4 red).

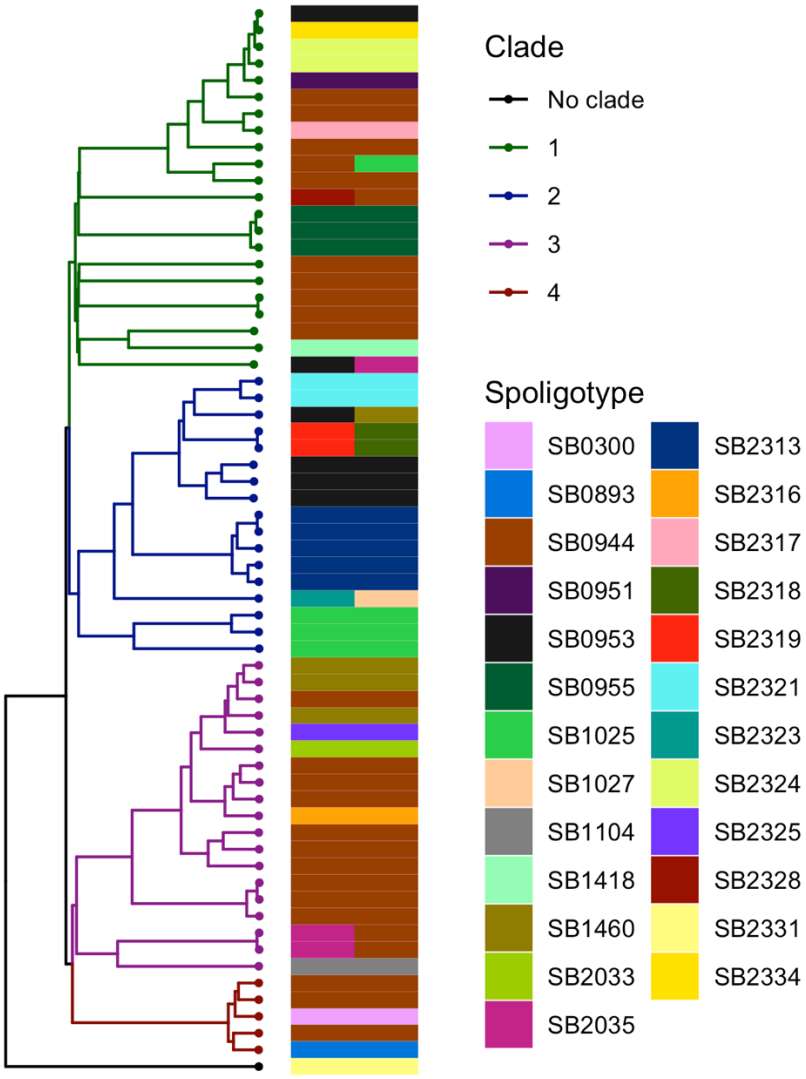


Figure 4: Visual comparison between the *M. bovis* phylogenetic time scaled MCC tree and the spoligotypes obtained by Egbe et al. (16). Ten sequences were associated with two spoligotypes, because multiple samples from the same animal (up to three) were submitted for spoligotyping.

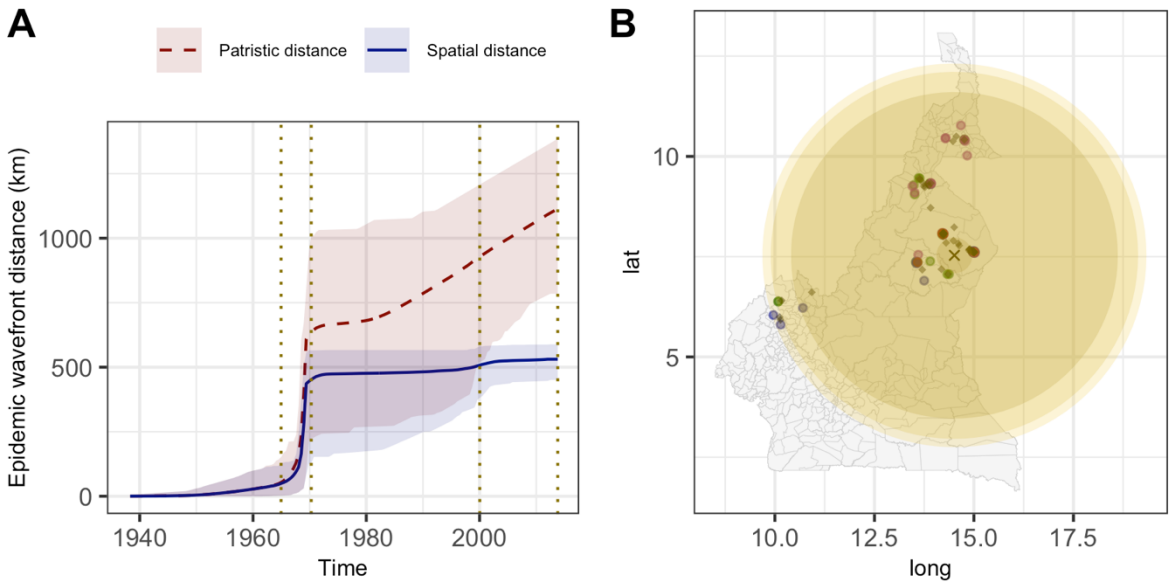
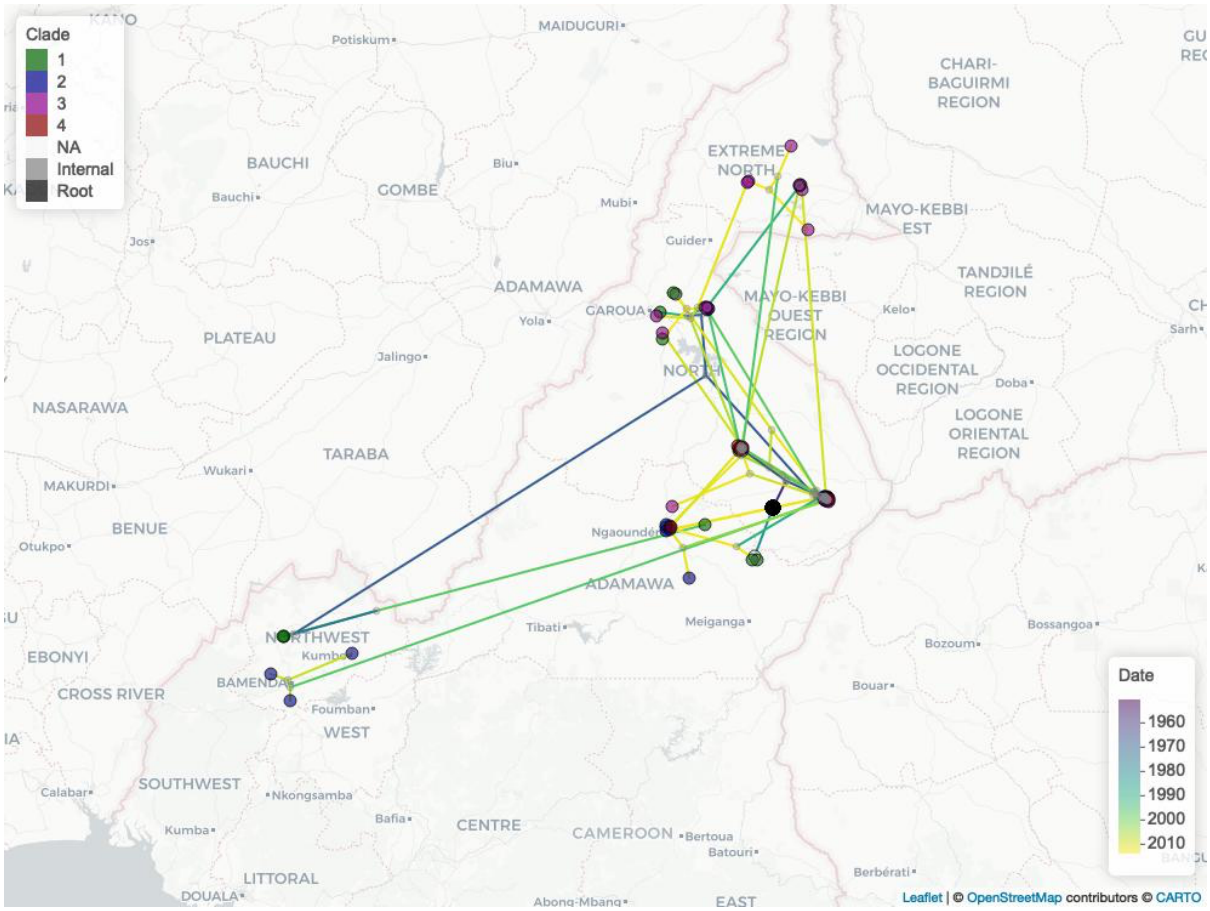


Figure 5: The estimated epidemic wavefront over time (panel A) and the expansion of the epidemic wavefront on the map (panel B). A: mean (lines) and 95th HPD (shades) of the epidemic wavefront spatial distance (blue) and patristic distance (red) over time. B: different yellow shades represent the epidemic wavefront at sequential point in time (marked by vertical dotted lines in panel A), and lighter shades of yellow correspond to more recent expansion; the estimated tree's root location is indicated by the black cross, diamonds represent the internal nodes estimated locations, and circles the sampled isolates (coloured by clade: 1 green, 2 blue, 3 purple and 4 red).



962

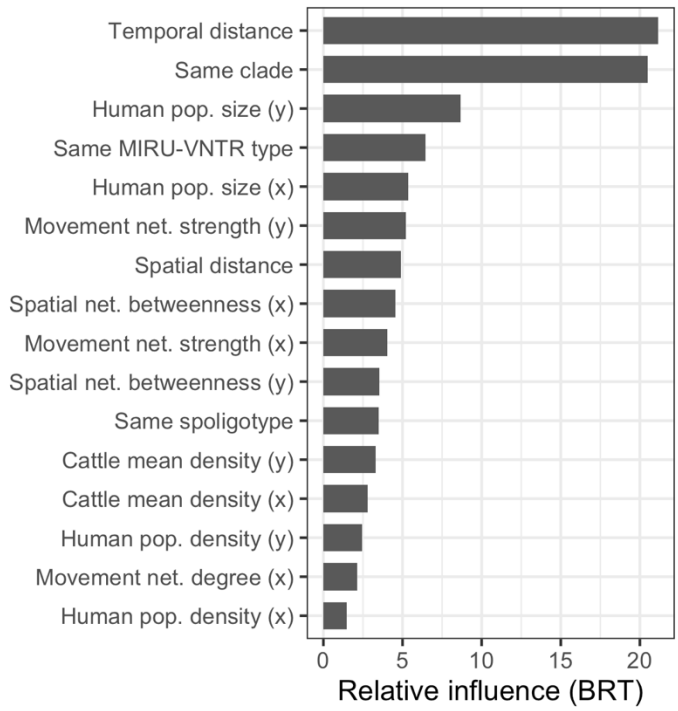
963 **Figure 6**: The Cameroonian *M. bovis* epidemic estimated expansions in space and time.

964 Nodes are coloured by clade (1, green; 2, blue; 3 purple; 4, red; no clade, light grey; internal

965 nodes, dark grey; tree root, black), while the branches are coloured by estimated movement

966 date, from 2007 (purple) to 2013 (yellow).

967



968

969 **Figure 7**: Relative influence of the most relevant variables in the simplified boosted
970 regression tree (BRT) model. The purpose of the BRT model was to explain the SNP distance
971 between the 64 high-quality *M. bovis* isolates. Many variables are calculated between
972 isolates pairs, *x* refers to the oldest isolate's subdivision, and *y* to the youngest one.

973

Variable	Type	Path model	Number of positive coefficients	Number of positive Q statistic	Mean Bayes Factor	
					(Randomisation #1)	(Randomisation #2)
Mosaic_shrub_otherv	Resistance	Least cost	*100*	*99*	1.89	1.83
Forest	**Conductance**	**Least cost**	***100***	***96***	***3.66***	***3.66***
Mosaic_shrub_otherv	NA	Straight line	*100*	*89*	*0.62*	*1.32*
Elevation	**Conductance**	**Least cost**	***100***	***88***	**2.39**	**3.00**
Waterbodies	Conductance	Least cost	100	87	1.10	0.97
Cattle_density	Resistance	Least cost	99	77	2.49	2.88
Cattle_density	NA	Straight line	100	73	Not run	Not run
Cattle_density	Conductance	Least cost	100	70	Not run	Not run
Grassland_cropland	Resistance	Least cost	100	66	Not run	Not run
Grassland_cropland	NA	Straight line	100	56	Not run	Not run
Mosaic_shrub_otherv	Conductance	Least cost	100	44	Not run	Not run
Roads_intersections	Conductance	Least cost	100	42	Not run	Not run
Waterbodies	Resistance	Least cost	100	38	Not run	Not run
Waterbodies	NA	Straight line	100	27	Not run	Not run
Grassland_cropland	Conductance	Least cost	100	15	Not run	Not run
Forest	Resistance	Least cost	100	12	Not run	Not run
Elevation	NA	Straight line	100	7	Not run	Not run
Forest	NA	Straight line	100	3	Not run	Not run
Pop_density	Conductance	Least cost	99	40	Not run	Not run
Elevation	Resistance	Least cost	99	15	Not run	Not run
Roads_length	NA	Straight line	97	0	Not run	Not run
Pop_density	NA	Straight line	96	16	Not run	Not run
Pop_density	Resistance	Least cost	96	7	Not run	Not run
Roads_length	Conductance	Least cost	92	11	Not run	Not run
Roads_length	Resistance	Least cost	59	0	Not run	Not run
Roads_intersections	NA	Straight line	56	1	Not run	Not run
Roads_intersections	Resistance	Least cost	6	1	Not run	Not run

Table 1: Results of the analysis on nine spatial variables, assuming two path models, straight line and least cost, and for the least cost path, whether the variable worked as a conductance or resistance. Results show the number of positive coefficients for the 100 sampled trees, the number of positive *Q* statistics, and the mean Bayes factor calculated over 10 randomisations, testing two algorithms: 1) randomisations of nodes positions while maintaining branches lengths, tree topology and location of the most ancestral node; and 2) randomisations of nodes positions while maintaining only the branches lengths.